Published in final edited form as:
Biometrics. 2011 September ; 67(3): 876–885. doi:10.1111/j.1541-0420.2010.01500.x.

A Partial Linear Model in the Outcome Dependent Sampling Setting to Evaluate the Effect of Prenatal PCB Exposure on Cognitive Function in Children

Zhou Haibo[1,*], **Qin Guoyou**[1,2,3], and **P. Longnecker Matthew**[4]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7420, U.S.A.

[2]Department of Biostatistics, School of Public Health, Fudan University, Shanghai, 200032, China

[3]Key Laboratory of Public Health Safety, Ministry of Education of China (Fudan University)

[4]Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, 27709, U.S.A.

Summary

Outcome-dependent sampling (ODS) has been widely used in biomedical studies because it is a cost effective way to improve study efficiency. However, in the setting of a continuous outcome, the representation of the exposure variable has been limited to the framework of linear models, due to the challenge in terms of both theory and computation. Partial linear models (PLM) are a powerful inference tool to nonparametrically model the relation between an outcome and the exposure variable. In this article, we consider a case study of a partial linear model for data from an ODS design. We propose a semiparametric maximum likelihood method to make inferences with a PLM. We develop the asymptotic properties and conduct simulation studies to show that the proposed ODS estimator can produce a more efficient estimate than that from a traditional simple random sampling design with the same sample size. Using this newly developed method, we were able to explore an open question in epidemiology: whether *in utero* exposure to background levels of PCBs is associated with children's intellectual impairment. Our model provides further insights into the relation between low-level PCB exposure and children's cognitive function. The results shed new light on a body of inconsistent epidemiologic findings.

Keywords

Cost-effective designs; Empirical likelihood; Outcome dependent sampling; Partial linear model; Polychlorinated biphenyls; P-spline

1. Introduction

Polychlorinated biphenyls (PCBs) are ubiquitous, persistent environmental pollutants. PCBs were once made in large quantities for use in capacitors, transformers, carbonless copy paper, and other applications (Dobson and van Esch, 1993). Although production has been banned worldwide due to persistence and toxicity, low-level human exposure to these ubiquitous environment contaminants continues via diet (Syracuse Research, 2000). Early

*zhou@bios.unc.edu .

SUPPLEMENTARY MATERIALS Web Appendices A and B referenced in Section 3, are available under the Paper Information link at the Biometrics website http://www.tibs.org/biometrics.

life exposure has been associated with cognitive deficits among children (Schantz et al., 2003). Exposure to PCBs in utero has been implicated as neurotoxic more often than has subsequent exposure (Patandin et al., 1999). *In vitro* and *in vivo* models demonstrate that PCBs cause altered synaptic transmission in the central nervous system, reduction in the striatum dopamine concentration, and decreased thyroxine levels (e.g., Gilbert and Liang, 1998). Although it has been demonstrated in rodent and primate studies that perinatal PCB exposure is usually associated with impairment of learning and cognition, human observational studies of perinatal exposure to background level PCBs in relation to cognitive functioning in children, on the other hand, have given inconsistent results. For example, a Michigan study of children followed from birth found that prenatal PCB exposure was associated with lower performance on the verbal and memory scales for children at 4 years of age and with lower full-scale and verbal intelligence quotient (IQ) scores at age 11 years (Jacobson and Jacobson, 1996). In a North Carolina birth cohort (Rogan and Gladen, 1991), however, performance on similar tests at ages 3, 4, and 5 years was unrelated to PCB exposure. These and other studies (Patandin et al., 1999; Schantz et al., 2003) suggest that in order to improve assessment of risks associated with PCB exposure, a better understanding of the relation between low-level PCB exposure and performance on examinations of cognitive function is needed. A more flexible and powerful statistical approach could play a critical role in understanding the risks associated with PCB exposure.

Our research was motivated by addressing the above scientific question through taking advantage of a large study conducted by the NIH and reanalyzing the data of Gray et al. (2005), where the maternal third trimester serum PCB levels were obtained in an outcome-dependent way for a subgroup of children born in the Collaborative Perinatal Project (CPP). The CPP study was a prospective study designed to identify determinants of neurodevelopmental deficits in children. Pregnant women were recruited from 1959 to 1965 from 12 US study centers resulting about 55,908 pregnancies recruited into the study (Longnecker et al., 2001). Gray et al. (2005) employed the outcome-dependent sample design (ODS) design developed by Zhou et al. (2002) based on eligible children who met the following criteria: 1) they were liveborn, 2) they were singletons, and 3) a 3-ml third trimester maternal serum specimen was available. From the 43,628 eligible children, 1,256 subjects were selected at random. An additional 207 children were randomly selected from eligible children whose 7-year full scale IQ score was either one or more standard deviations below the mean or one or more standard deviations above the mean. The mother's third trimester's blood specimen for the subgroup of subjects selected in the above ODS fashion were then thawed and the PCB levels were assessed at the US Centers for Disease Control and Prevention laboratory in Atlanta.

In the remainder of this section we give brief overviews of the two statistical areas that play a key role in this paper: the ODS design and the partial linear model (PLM).

To understand the determinants of diseases in humans, epidemiologic observational studies are conducted to characterize the relation between individuals' exposure and a disease outcome. As large cohort studies are usually very costly to conduct, small studies using the case-control design are often preferred, especially for rare diseases, because they can include equal number of individuals with a rare disease in a smaller overall study. The case-control design (Cornfield, 1951) is a binary special case of a general ODS design where investigators can observe the exposure or covariates with a probability that depends on an outcome variable where the outcome can be either continuous or discrete. The principle idea of ODS is to concentrate resources on the parts of the studied population having the greatest amount of information. By allowing selection probability of each individual in the ODS samples to depend on the outcome, the investigators can improve the efficiency and reduce the cost of the studies. Historically, the ODS design has been implemented in the binary case

(White, 1982; Rathouz et al., 2002; Wang and Zhou, 2006; Haneuse and Wakefield, 2007; Schildcrout and Heagerty, 2008), while development of the ODS design with a continuous outcome has lagged behind until recently. For example, Zhou et al. (2002) considered a semiparametric empirical likelihood method for a continuous ODS design. Chatterjee et al. (2003) considered a pseudoscore estimation method and Weaver and Zhou (2005) proposed an estimated likelihood method for inference based on data from the two-stage ODS design. The gains in statistical power and the reduction in sample size for a given budget relative to a traditional simple random sample is shown in Zhou et al. (2007).

All models considered in the literature are limited to the framework of a linear model. The PLM for continuous outcomes (He et al., 2002; Wang et al., 2005), where the outcome is assumed to depend on some covariate X nonparametrically and some other covariates Z parametrically, is an important inference tool and has been widely applied in many fields. It would be a particular advantage in the Gray et al. (2005) study to have a more flexible PLM approach to investigate the relation of low level PCB exposure and children's cognitive development. In this article, we consider the PLM for data from an ODS design. The functions offered by standard commercial software for the PLM were generally developed under the simple random sample setting, therefore can not be directly applied to our case. We propose a semiparametric maximum likelihood estimation method to achieve the inference of the PLM using penalized splines (P-spline).

The rest of this article is organized as follows. In section 2, we propose the PLM and derive the penalized likelihood for the data from an ODS design. In section 3, we propose a semiparametric empirical likelihood estimator and present its asymptotics. Some simulation studies are conducted in section 4 to investigate the performance of the proposed method, and in section 5 we show that with our new method, we identified some new findings in the data from the Gray et al. study that will shed some light on the ongoing debate on the relationship of PCB and children's cognitive development.

2. Models and Likelihood Function for PLM with ODS Data

2.1 *PLM Model and ODS Data Structure*

Let Y denote a continuous outcome, X an exposure variable and Z a p-dimensional covariate vector. The notation A^T denotes the transpose of the matrix or vector A. The conditional mean of Y given the X and Z is assumed to be

$$E(Y|X,Z) = g(X) + Z^T\gamma, \tag{1}$$

where $g(\cdot)$ is an unknown smooth function of X and γ is a p-dimensional vector of unknown regression coefficients. Model (1) is the PLM considered in our ODS design. The introduction of the $g(\cdot)$ is a key difference with the previous literature on the ODS design.

Various nonparametric smoothing methods can be applied to estimate the nonparametric function $g(.)$ such as kernel and spline methods, e.g., Lin and Carroll (2001); Yu and Ruppert (2002); Huang et al. (2007). Particularly, Yu and Ruppert (2002) studied estimation of a partially linear single-index model by penalized splines (P-spline) to estimate the nonparametric function, which is also adopted in this article. P-spline is an extension of smoothing splines, allowing a more flexible choice of knots and penalty. Following Yu and Ruppert (2002), under the working assumption that $g(.)$ is a r-degree spline function with T fixed knots $t_1, \ldots, t_T$, we then have $g(x) = M^T(x)\alpha$ where

$M(x) = \left\{1, x, x^2, \ldots, x^r, (x-t_1)_+^r, \ldots, (x-t_T)_+^r\right\}^T$ is a r-degree truncated power spline basis

with knots $\{t_i\}_{i=1}^{T}$, $(x)_+^r = x^r 1_{x\geq 0}$ and α is a $r + T + 1$-dimensional vector. A discussion on how to select the knots can be found in the Supplementary Material (Web Appendix A). Then we have $g(X)+Z^T\gamma = M^T(X)\alpha+Z^T\gamma = D^T\theta$, where $D = \{M^T(X), Z^T\}^T$ and $\theta = (\alpha^T, \gamma^T)^T$.

We consider the Zhou et al. sampling scheme with two components (Zhou et al., 2002). First, an overall random sample of the population is taken. Second, one or more supplemental random samples are taken, in which the probability of selection depends on the level of the outcome variables. Assume that the domain of Y is the union of K non-overlapping intervals: $C_k = (c_{k-1}, c_k]$ with c_k being some known constants satisfying $c_0 < c_1 < c_2 <, \ldots, < c_k = \infty$. Thus, $\{C_k, k = 1, \ldots, K\}$ partition the study populations into K strata. Then the data structure of the ODS sample consists of a simple random sample (the SRS sample) and a simple random sample from each of the K intervals of Y (the supplement samples). The supplement samples can be generated with different, perhaps unknown selection probabilities. Moreover, we assume that, from the underlying population of interest, n_0 individuals are obtained using a simple random sampling (SRS) and n_k individuals are sampled from the kth stratum. Then the total ODS sample size is $n=\sum_{k=0}^{K} n_k$. In summary, the observed data structure for the ODS design is $\{y_{kj}, x_{kj}, z_{kj}\}$ for $k = 0, \ldots, K$, $j = 1, \ldots, n_k$, where $k = 0$ represents the SRS sample.

2.2 A Penalized Log-Likelihood Function

Denote $F(y|x, z; \theta)$ and $f(y|x, z; \theta)$ as the conditional cumulative distribution function and density function for Y given X and Z. Define $\psi_k(x, z; \theta) = F(c_k|x, z; \theta) - F(c_{k-1}|x, z; \theta)$ for $k = 1, \ldots, K$. Moreover, we denote the observed data from the kth stratum as (y_{kj}, x_{kj}, z_{kj}), $j = 1, 2, \ldots, n_k$. Let $F_{X,Z}(x, z)$ and $F_{X,Z}(x, z)$ denote the joint cumulative distribution function and probability density function of X and Z. The likelihood for the observed data from ODS can be written as

$$\left\{\prod_{j=1}^{n_0} f\left(y_{0j}|x_{0j}, z_{0j};\theta\right) f_{XZ}\left(x_{0j}, z_{0j}\right)\right\}\left\{\prod_{k=1}^{K}\prod_{j=1}^{n_k} f\left(y_{kj}, x_{kj}, z_{kj}|y_{kj} \in C_k;\theta\right)\right\}, \tag{2}$$

where $f(y_{kj}, x_{kj}, z_{kj}|y_{kj} \in C_k, \theta)$ denotes the joint density function of $\{y_{kj}, x_{kj}, z_{kj}\}$ conditional on y_{kj} being in the interval C_k. Using Bayes formula, (2) can be written as

$$L_n(\theta) = \left\{\prod_{j=1}^{n_0} f\left(y_{0j}|x_{0j}, z_{0j};\theta\right) f_{XZ}\left(x_{0j}, z_{0j}\right)\right\}\times\prod_{k=1}^{K}\left\{\prod_{j=1}^{n_k}\frac{f\left(y_{kj}|x_{kj}, z_{kj};\theta\right) f_{XZ}\left(x_{kj}, z_{kj}\right)}{\int\psi_k(x, z, \theta)\, dF_{XZ}(x, z)}\right\}.$$

Denote $\pi_k = \int \psi_k(x, z; \theta)$, $k = 1, \ldots, K - 1$, $\pi_K = 1 - \sum_{k=1}^{K-1}\pi_k$ and $p_{kj} = f_{X,Z}(x_{kj}, z_{kj})$, $k = 0, \ldots, K$, $j = 1, \ldots, n_k$. Then the log-likelihood can be written as

$$l_n\left[\theta, \{p_{kj}\}, \{\pi_k\}\right] = l_{1n}(\theta) + \sum_{k=0}^{K}\sum_{j=1}^{n_k} \log p_{kj} - \sum_{k=1}^{K} n_k \log\pi_k, \tag{3}$$

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

As a P-spline is used to estimate the nonparametric function in model (1), incorporating the penalty into (3), we obtain the following penalized log-likelihood function,

$$pl_n\left[\theta,\left\{p_{kj}\right\},\{\pi_k\};q_s\right]=l_{1n}(\theta)+\sum_{k=0}^{K}\sum_{j=1}^{n_k}logp_{kj}-\sum_{k=1}^{K}n_k log\pi_k-\frac{1}{2}nq_s\theta^T\Psi\theta, \tag{4}$$

where $\Psi=\text{diag}\left\{\left(0^T_{(r+1)\times 1},1^T_{T\times 1},0^T_{p\times 1}\right)^T\right\}$ and $\theta^T\Psi\theta$ is a common quadratic penalty function. Here q_s is the smoothing parameter.

3. A Semiparametric Empirical Likelihood Approach

To estimate θ through (4), methods for handling $f_{X,Z}(x, z)$ and $f_{X,Z}(x, z)$ are required. We propose a semiparametric likelihood method to maximize the penalized log-likelihood without specification of the underlying distribution of X and Z. We first profile the penalized log-likelihood function (4) over p_{kj} by fixing θ, and obtain the empirical likelihood function of p_{kj} over all distributions whose support contains the observed values of X and Z. We then maximize the following resulting profile likelihood with respect to θ.

$$ppl_n(\theta;q_s)=\sup_{p_{kj}}pl_n\left[\theta,\left\{p_{kj}\right\},\{\pi_k\};q_s\right]. \tag{5}$$

To get (5), it suffices to maximize

$$l_{np}\left[\left\{p_{kj}\right\}\right]=\sum_{k=0}^{K}\sum_{j=1}^{n_k}\log p_{kj}-\sum_{k=1}^{K}n_k\log\pi_k, \tag{6}$$

for fixed $[\theta, \{\pi_k\}]$ subject to $\sum_{k=0}^{K}\sum_{j=1}^{n_k}p_{kj}\left\{\psi_i\left(x_{kj},z_{kj};\theta\right)-\pi_i\right\}=0$, $i = 1, \ldots, K-1$ and $\sum_{k=0}^{K}\sum_{j=1}^{n_k}p_{kj}=1$. Using a similar idea to Qin and Lawless (1994), for a fixed θ, we can show a unique maximum $\widehat{p}_{kj}$ in (6) which satisfies the above constraints exists if zero is inside the convex hull formed by the points $\{\psi_i(x_{kj}, z_{kj}; \theta) - \pi_i\}$ for $i = 1, \ldots, K-1$, $k = 1, \ldots, K$, $j = 1, \ldots, n_k$. An explicit expression can be derived by Lagrange multiplier argument:

$$H\left[\theta,\left\{p_{kj}\right\},\{\pi_i\}\right]=l_{np}\left[\left\{p_{kj}\right\}\right]+\rho\left(1-\sum_{k,j}p_{kj}\right)+n\sum_{i=1}^{K-1}\lambda_i\sum_{k,j}p_{kj}\left\{\psi_i\left(x_{kj},z_{kj};\theta\right)-\pi_i\right\},$$

where ρ and $\lambda's$ are Lagrange multipliers. Taking derivatives of H with respect to $\{p_{kj}\}$ and solving the score equations, we can obtain that $\rho = n$ and $\widehat{p}_{kj}=\frac{1}{n}\frac{1}{1+\sum_{i=1}^{K-1}\lambda_i\left\{\psi_i\left(x_{kj},z_{kj};\theta\right)-\pi_i\right\}}$. We plug $\widehat{p}_{kj}$ into (4) and get the estimator of θ from maximizing the resulting profile penalized log-likelihood function. The above procedure enables us to change an infinite dimension problem into a finite dimension problem at the expense of introducing $2(K-1)$-dimensional parameters $\{\pi_i, i = 1, \ldots, K-1\}$ and $\{\lambda_i, i = 1, \ldots, K-1\}$.

Typically, the true value of the Lagrange multiplier is zero in unbiased sampling problems. However, due to the nature of the ODS sampling design, the Lagrange multipliers, $\lambda's$ are not centered around zero. To unify the notation, we center them by reparameterizing as

$v_k = \lambda_k - \frac{n_k}{n\pi_k}$, $k = 1, \ldots, K-1$. Denote $\vartheta = (\pi^T, v^T)^T$, where $\pi = (\pi_1, \ldots, \pi_{K-1})^T$, $v = (v_1, \ldots, v_{K-1})^T$ and $\eta = (\theta^T, \vartheta^T)^T$. Then, the total number of the parameters to be estimated is $d = T + r + 1 + p + 2(K-1)$. The profile penalized log-likelihood function (5) can be written in an explicit expression as:

$$\begin{aligned} ppl_n(\eta; q_s) &= l_{1n}(\theta) - \sum_{k=0}^{K}\sum_{j=1}^{n_k} \log\left\{1 + v^T h\left(x_{kj}, z_{kj}\right)\right\} - \sum_{k=1}^{K}\sum_{j=1}^{n_k} \log\left\{Q\left(x_{kj}, z_{kj}\right)\right\} \\ &\quad - \sum_{k=1}^{K} n_k \log\pi_k - \tfrac{1}{2} n q_s \theta^T \Psi \theta, \end{aligned} \tag{7}$$

where $h(x_{kj}, z_{kj}) = \{h_1(x_{kj}, z_{kj}), \ldots, h_{K-1}(x_{kj}, z_{kj})\}^T$, $h_i\left(x_{kj}, z_{kj}\right) = \frac{\psi_i\left(x_{kj}, z_{kj}; \theta\right) - \pi_i}{Q\left(x_{kj}, z_{kj}\right)}$, $i = 1, \ldots, K-1$ and $Q\left(x_{kj}, z_{kj}\right) = \sum_{i=1}^{K-1} \frac{n_i}{n\pi_i} \psi_i\left(x_{kj}, z_{kj}; \theta\right) + \frac{n_0}{n}$. Denote $\widehat{\eta}$ the proposed estimator for η, where $\widehat{\eta}$ is the maximizer for the profile penalized log-likelihood function in (7). A Newton-Raphson algorithm can be used to obtain $\widehat{\eta}$ with fixed smoothing parameter q_s. The estimation procedure is as follows:

(1) Choose initial values $\eta^{(0)} = (\theta^{(0)T}, \pi^{(0)T}, v^{(0)T})^T$. The initial value $\theta^{(0)}$ can be selected as the penalized maximum likelihood estimate based on the SRS sample from the ODS design, $\pi^{(0)}$ can be chosen as the estimates of probability $P(y \in C_k)$, $k = 1, \ldots, K-1$ based on the population of response y, and $v^{(0)}$ is selected as $\mathbf{0}_{(K-1)\times 1}$ where $\mathbf{0}_{(K-1)\times 1}$ denote a (K-1)-dimensional vector with all the components equal to 0. Set $j = 0$.

(2) Compute $\eta^{(j+1)} = \eta^{(j)} - \left\{\frac{\partial^2}{\partial\eta\partial n^T} ppl_n\left(\eta^{(j)}; q_s\right)\right\}^{-1} \frac{\partial}{\partial\eta} ppl_n\left(\eta^{(j)}; q_s\right)$. Set $j = j + 1$.

(3) Continue step 2 until convergence is achieved. Choose $\eta^{(j+1)}$ as the final estimate.

The following theorem summarizes the asymptotic properties for the proposed estimators of the PLM from an ODS design.

THEOREM 1: Under conditions A1-A4, (i) If the smoothing parameter $q_s = o(1)$, $\widehat{\eta}$ converges to η_0 with probability one. (ii) If the smoothing parameter $q_s = o\left(1/\sqrt{n}\right)$. The maximizer $\widehat{\eta} = \left(\widehat{\theta}^T, \widehat{\pi}^T, \widehat{v}^T\right)^T$ is asymptotically distributed as a normal distribution as $\sqrt{n}\left(\widehat{\eta} - \eta_0\right) \to N\left(0_{d\times 1}, \Sigma\right)$.

The conditions A1-A4 and a brief proof for Theorem 1 are given in the Supplementary Material (Web Appendix B). A consistent estimator $\widehat{\Sigma}$ of the covariance matrix Σ is $\widehat{V}^{-1}\widehat{U}\widehat{V}^{-1}$, where $\widehat{U}$ and $\widehat{V}$ are also defined in the Web Appendix. The conditions on smoothing parameter q_s in Theorem 1 are commonly assumed (Yu and Ruppert, 2002; Qu and Li, 2006). The generalized cross-validation (GCV) method, given in the Supplementary Material (Web Appendix A), is a very popular method to select a smoothing parameter and performs quite well in practice. However, it is a challenge to obtain the rate of the smoothing parameter chosen by the GCV method, which is an important theoretical problem and deserves further study. We note that the asymptotic properties developed in this paper are for fixed knots P-spline, which means that the number of the knots does not grow as sample size grows to infinity.

From Theorem 1, we can construct a joint confidence region and test hypotheses, e.g., with a Wald test. For example, if we want to test the null hypothesis $H_0 : B\theta_0 - s_0 = 0$ where B is a $d_1 \times \{d - 2(K - 1)\}$ matrix with full rank $d_1 \leq \{d - 2(K - 1)\}$, then the test can be based on the Wald test statistic $W=\left(B\widehat{\theta} - s_0\right)^T \left(B\widehat{\Sigma}_\theta B^T\right)^{-1} \left(B\widehat{\theta} - s_0\right)$ with $\widehat{\Sigma}_\theta$ being the consistent estimate of the covariance matrix of $\widehat{\theta}$, which has an asymptotic chi-square distribution with d_1 degrees of freedom.

For this PLM, it is of particular interest to test whether the nonparametric function describing the relation between exposure variable X and response Y is linear. We can use the previously mentioned Wald test to deal with this problem. To do so, we re-express the $T+r+1$-dimensional vector α as $\left(\alpha_1^T, \alpha_2^T\right)$, where $\alpha_1 = (\alpha_{11}, \alpha_{12})^T$ is a two-dimensional vector and α_2 is a $T + r - 1$ dimensional vector. We are interested in testing the null hypothesis: $\alpha_2=\alpha_2^0=(0, \ldots, 0)^T$. Under this null hypothesis, we have $g(X) = \alpha_{11} + \alpha_{12}X$, i.e. the exposure variable X is related to response Y in a linear fashion.

4. Simulation Studies

We conducted simulations to investigate the performance of our proposed estimator. For all simulations, we generated 1000 simulated datasets, each with 500 independent subjects. We consider two models with different nonparametric functions. The first one is a monotonic function while the second is a unimode function in threshold regions. In the simulation, we adopt a three-degree truncated power spline basis and choose ten fixed knots selected as the equally spaced sample quantiles. The computational aspects of the simulation studies are given in last paragraph of this section.

Study 1. The data were generated according to the following PLM,

$$Y=3\Phi(3.2X)+Z_\gamma+e_0,$$

where $\Phi(.)$ is a standard normal distribution function, X denotes a continuous exposure variable of interest, $e_0 \sim N\left(0, \sigma_0^2\right)$. We assume that $X \sim N(0, 0.25^2)$ and $Z \sim N(0, 0.5^2)$. Then $f(y|x, z;\theta) \sim N\left(g_1(x)+z\gamma, \sigma_0^2\right)$, where $g_1(x) = 3\Phi(3.2x)$. We take $\gamma = 1$, $\sigma_0^2=0.2$.

In this simulation, we investigated the effect of different cut points for creating the supplemental samples and the impact on the estimation efficiency through varying the allocation of the sample size between the SRS and supplemental samples. Similar to Zhou et al. (2002), our ODS design consists of a SRS sample of n_0, a supplemental sample of n_1+n_3 from individuals with Y values in the two tails of the marginal distribution of Y ($n_1 = n_3$, $n_2 = 0$), defined by cut points $\mu_Y \pm a\sigma_Y$ where μ_Y and σ_Y are the mean and standard deviation of Y respectively.

We compared the proposed estimator based on the ODS sample including both the SRS sample and supplemental samples (P estimator) with two additional estimators: (i) a penalized maximum likelihood estimator (PMLE) that is based on the SRS samples only from the ODS sample (V estimator); and (ii) a PMLE (S estimator) that is based on a SRS sample but the same sample size as the ODS sample. The comparison against the V estimator will demonstrate the efficiency gain of the P estimator using more information, and the comparison against the S estimator will demonstrate the efficiency gain of the ODS design over the SRS design with the same sample size. We computed the averages of the mean square error(MSE), the absolute value of the bias and the Monte Carlo variance of the

estimated nonparametric function $g_1(x)$ over 501 equal spaced grid points in [−0.75, 0.75] (the mean of X minus and plus three times standard deviation of X) over 1000 replications. Moreover, we calculated mean, Monte Carlo standard error, and estimated standard error using large sample properties and coverage probability of the 95% nominal confidence interval for the estimator of regression coefficient γ.

For the sample size $n_0 = 300$, $n_1 = n_3 = 100$, we summarize the results in Table 1. From Table 1, we find that the proposed estimator of the nonparametric function is most efficient with the smallest AMSE (average MSE over 501 grid points) among all the estimators compared. For the estimation of the regression coefficient γ, the proposed estimator is generally more efficient than the other two estimators with smaller variance when a is 1.0 and 1.3. It was also found that the nominal 95% confidence intervals based on the proposed standard errors for the regression coefficient γ provide good coverage. The proposed estimator is more efficient than the S estimator, which supports the notion that ODS design can be more efficient than the SRS design.

Figure 1 presents curves of the true function and the average P-spine estimates of g_1 by the P, V and S estimators over 1000 simulations. The difference between the true function g_1 and its estimates is barely visible, which shows good fit of the P-spline estimates. Figure 1 also gives the confidence bands obtained by the P estimator, V estimator and the S estimator for comparison. The confidence bands are based on a normal approximation using the Monte Carlo standard error with the sample size $n_0 = 300$, $n_1 = n_3 = 100$ and $n_0 = 200$, $n_1 = n_3 = 150$, and the cut point equal to 0.7 and 1.3. In Figure 1, we see that in the tail of the distribution of X, the confidence bands by the S estimator are wider than those by the P estimator, and this is more apparent when the cut point is further out or the proportion of the SRS samples in the ODS is decreased. This is because, compared with SRS design, more individuals in the tails of the distribution of X can be sampled in the ODS design, which leads to improvement of the estimation efficiency of the P estimator. In addition, the confidence bands by the V estimator are the widest ones and the bands become wider when the proportion of the SRS sample is decreased.

Under the same simulation model, we study the relative efficiency of the P estimator of nonparametric function over the S estimator for the proportion of the SRS sample in the ODS ($\rho = n_0/n$). Table 2 shows the relative efficiency $(AMSE(\widehat{g_S})/AMSE(\widehat{g_P}))$ comparison across $\rho = n_0/n$ under $a = 0.7$, 1 and 1.3, where $\widehat{g_S}$ and $\widehat{g_P}$ denote the P and S estimators of g respectively. The total sample size is n and the sample size of n_1 and n_3 are determined by $(1 - \rho)n/2$. We can find that the proposed estimator is more efficient than the S estimator and the efficiency can generally be further improved when the proportion decreases and the cut point is selected to be further out.

Study 2. The data were generated according to the following PLM,

$$Y = \sin(1.5X) + Z\gamma + e_0,$$

where X denote a continuous exposure variable of interest, $e_0 \sim N\left(0, \sigma_0^2\right)$. We assume that $X \sim N(1, 0.5^2)$ and $Z \sim N(0, 0.5^2)$. Then $f(y|x, z;\theta) \sim N\left(g_2(x) + z\gamma, \sigma_0^2\right)$, where $g_2(x) = sin(1.5x)$. We take $\gamma = 1$, $\sigma_0^2 = 0.2$. The results are calculated over 501 equal spaced grid points in [−0.5, 2.5] ($X \pm 3SD_X$) over 1000 replications. The corresponding results are similar to Study 1, and similar conclusions to Study 1 can be drawn.

In the setting of our simulation, convergence is relatively fast and the number of iterations required for convergence is generally less than 25. Non-convergence may arise occasionally, e.g., when the sample size is small and the cut points are selected further out (a=1.3 in our case), but this problem can be avoided by increasing the sample size. At a sample size of 1000, we do not have any problem with the CPP data analysis. The codes for simulation and real data analysis were written in Matlab and can be downloaded from http://www.bios.unc.edu/~zhou/software/Zhou-Qin-Longnecker-Biometrics/.

5. Analysis of the Collaborative Perinatal Project Data Set

As introduced in the beginning of the paper, we hope a more flexible partial linear approach in the ODS setting could provide additional insights into the relationship between the low dose maternal pregnancy serum level of polychlorinated biphenyls (PCBs) and children's subsequent IQ test performance based on the CPP data set. Among the 1463 subjects selected by the ODS design (Gray et al., 2005), only 1038 subjects were observed during pregnancy and at age 7 years and had complete data on all covariates used in the analysis. The details of the data structure available to us are such that, in the data set, there is a simple random sample (SRS) of 849 subjects and two supplemental subsamples. The first supplemental sample is a simple random sample of the children in the CPP population whose IQ scores were at least one standard deviation (14) above the mean (96). The second subsample is a simple random sample from those whose IQ scores were at least one standard below the mean. In all, we have 81 subjects in the low IQ subsample and 108 subjects in the high IQ subsample.

The Weschler Intelligence Scale for children at 7 years of age (IQ) was the outcome variable and the prenatal PCB level, measured in μg/liter, was the exposure variable. We built a partial linear model of these data using our proposed method. In addition to PCBs, other covariates include socioeconomic status of the child's family (SES), the gender (SEX) and race (RACE) of the child and the mother's eduction (EDU). The covariate EDU is the highest education level obtained by the mother at the child's birth divided into 18 levels from the lowest level 1 (EDU=1) to the highest level (EDU=18). Here we introduced two dummy variables EDU1 and EDU2 to replace the original covariate EDU in our analysis to achieve a reasonable fit because the relationship between IQ and EDU is nonlinear. These two dummy variables are coded as: (EDU1,EDU2)=(0,0) corresponds to (EDU≤8) representing low education level; (EDU1,EDU2)=(1,0) corresponds to (9 ≤EDU ≤12) representing middle education level and (EDU1,EDU2)=(0,1) corresponds to (EDU≥13) representing high education level. The covariate SEX was coded 0 for males and 1 for females and the covariate RACE was coded 1 for Black and 0 for others. Table 3 describes the basic characteristics of the data in these three samples.

Zhou et al. (2002) reported no association between PCB exposure and the children's IQ score. Gray et al. (2005), using a probability weighted approach, did not find either a linear or quadratic association. Here, we consider the following PLM to describe the relation between PCB and IQ:

$$IQ = g(PCB) + \beta_1 EDU1 + \beta_2 EDU2 + \beta_3 SES + \beta_4 \mathrm{RACE} + \beta_5 \mathrm{SEX} + e,$$

where e is a normal error with zero mean. To estimate the nonparametric function $g(\cdot)$, we adopted a two-degree truncated power function basis

$M(x) = \left(1, x, x^2, (x - T_1)^2_+, \ldots, (x - T_{10})^2_+\right)^T$ with ten fixed knots $T_1, \ldots, T_{10}$ selected as the equally spaced sample quantiles of PCB level, i.e. 1.16, 1.70, 2.17, 2.61, 3.05, 3.52, 4.06, 4.63, 5.54, 6.88. Under these specifications, the above model can be rewritten as IQ =

$M^T(PCB)\alpha + \beta_1 EDU1 + \beta_2 EDU2 + \beta_3 SES + \beta_4 RACE + \beta_5 SEX + e$, where $\alpha = (\alpha_1, \ldots, \alpha_{13})^T$ is the parameter vector associated with the nonlinear function $g(\cdot)$. To adopt our proposed method to fit this model, we chose the smoothing parameter to be 10^7 by the GCV method. The selection of such large smoothing parameter may be because the GCV score is not sensitive to the smoothing parameter q_s when $q_s \geq 41.75$. It is found that the GCV score values over the grid points which are greater than or equal to 41.75 are almost the same. This also indicates that the fitting with $q_s \geq 41.75$ are very close, which is demonstrated by our reanalyses. The selection of smoothing parameter for this data set has very small influence on the conclusion. The fitted estimates of the regression coefficients are given in Table 4 and the resulting estimate of the $g(\cdot)$ is presented in Figure 2.

From Figure 2, it can be seen that the IQ is related to PCB nonlinearly. The detection of this nonlinear relationship is due to the use of P-splines and has nothing to do with the sampling design. To formally testing this nonlinear effect, we conduct the proposed Wald test to test the hypothesis: $H_0 : \alpha_3 = \alpha_4 = \ldots = \alpha_{13} = 0$ which is equivalent to testing $H_0 : g(PCB) = \alpha_1 + \alpha_2 PCB$. The H_0 is rejected at the 5% significance level as the Wald statistic $W = 26.37 > \chi^2_{0.05}(11) = 19.68$. The corresponding p-value is 0.006.

Next, we test whether the $g(\cdot)$ is a simple quadratic function. The corresponding null hypothesis is $H_0 : \alpha_4 = \alpha_5 = \ldots = \alpha_{13} = 0$. This null is rejected at 5% level with a p-value at 0.029 based on the χ^2 distribution with 10 degrees of freedom. This result indicates that although the relationship in Figure 2 is unimodal, a simple quadratic function is not enough to describe the relation between the IQ score and the PCB level. Apart from the difference in the fitted PCB effect, the PLM model (β_P) and the linear model (β_L) yields very similar estimates for other confounding variables in the model. This is because these two estimators are using the same data and essentially the same method except that the β_L models PCB as linear while the β_P models PCB as nonlinear. In particular, mother's education level and social economic status were positively related to children's IQ performance. Blacks tended to have lower score while gender did not have any impact on the IQ performance. For both linear and nonlinear components of the model, we can see from Table 4 and Figure 2 that the proposed method β_P yields a smaller variance estimate and corresponding narrower 95% confidence bounds than the β_V estimator.

The fitted relationship in Figure 2 suggests that the level of PCB exposure was positively related to the children's IQ performance within the lower range of exposure, which is possibly due to residual confounding, by either socioeconomic status Gray et al. (2005) or other factors not captured by the existing covariates. This relation began to diminish as the level of PCB exposure moved closer to the mean PCB level and then it became detrimental in the higher range of PCB exposure. The measured PCB level in most of the studies, the CPP study included, was considered background level exposure, which is low compared with the PCB exposure levels in the Yucheng poisoning incident, in which the lower IQ among children exposed at higher levels in utero was clear Rogan et al. (1988). With a large data set and a more powerful and flexible modeling approach, we were able to obtain further insight on the relationship of children's IQ and prenatal PCB exposure.

6. Discussion

The contribution of this paper is twofold. First we introduced a PLM for ODS data and developed a semiparametric inference procedure that is based on a penalized likelihood function. We established the asymptotic properties of the proposed method and most importantly through simulation, we showed that the proposed ODS design with PLM will yield more efficient estimates than the design using a SRS with the same sample size. The advantages of the ODS design is clearly demonstrated in this paper and in practice; this

advantage will translate into more cost-effective studies, or, for a fixed budget, more powerful studies.

Second, the application of the proposed method to the analysis of *in utero* exposure to background levels of polychlorinated biphenyls (PCBs) with children's intellectual development measured by IQ at 7 years of age showed some interesting results. Epidemiologic studies on *in utero* exposure to PCBs have been controversial and inconclusive with some studies suggesting that PCBs cause intellectual impairment among children and others either not supporting this or suggesting opposite effects. We thought the reasons for this phenomenon were due to the following: a) the sample sizes in these studies were usually small (200 to 400); b) the PCB exposure level in Western countries are all background level and thus primarily within the lower range; and c) the statistical approaches were relatively simple and may not be the most powerful.

We had an unique opportunity in that 1) we had a large sample size compared with those published, 2) the study was designed using an ODS strategy such that it is more informative than a usual study with the same sample size; 3) we have developed a flexible yet efficient analytical tool (PLM) to take full advantage of the data from the ODS structure. Indeed, our analysis of the CPP data reveals something that has not been reported before. The nonlinear curvature of the fitted model helps to explain why in the previous smaller studies some find an effect while others do not.

Our simulation results demonstrate the efficiency improvement of the ODS design over some alternative methods that can be used in these situations such as the traditional SRS design with the same sample size. Generally, 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. In practice, the choice of ρ and cut points may be affected by other considerations such as study purpose. Some detailed discussion about this is in Zhou et al. (2002).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Editor, the Associate Editor and two referees for their constructive suggestions that largely improved the presentation of this paper. This work is supported by National Institute of Health grants CA 79949 (Zhou and Qin), the intramural research program of the NIH, National Institute of Environmental Health Sciences (Longnecker) and the National Nature Science Fund of China grants of China 10801039 (Qin).

References

Chatterjee N, Chen YH, Breslow NE. A pseudoscore estimator for regression problems with two-phase sampling. Journal of the American Statistical Association. 2003; 98:158–168.

Cornfield J. A method of estimating comparative rates from clinical data. Applications to cancer of lung, breast and cervix. Journal of the National Cancer Institute. 1951; 11:1269–1275. [PubMed: 14861651]

Dobson, S.; van Esch, GJ. Polychlorinated Biphenyls and Terphenyls. 2nd ed. World Health Organization; Geneva: 1993. p. 79-84.

Gilbert ME, Liang D. Alterations in synaptic transmission and plasticity in hippocampus by a complex PCB mixture, Aroclor 1254. Neurotoxicology and Teratology. 1998; 20:383–389. [PubMed: 9697964]

Gladen BC, Rogan WJ. Effects of perinatal polychlorinated biphenyls and dichlorodiphenyl dichloroethene on later development. The Journal of Pediatrics. 1991; 119:58–63. [PubMed: 1906100]

Gray KA, Klebanoff MA, Brock JW, Zhou H, Darden R, Needham L, Longnecker MP. In utreo exposure to background levels of polychlorinated biphenyls and cognitive functioning among school-age children. American Journal of Epidemiology. 2005; 162:17–26. [PubMed: 15961582]

Haneuse S, Wakefield J. Hierarchical models for combining ecological and case-control data. Biometrics. 2007; 63:128–136. [PubMed: 17447937]

He X, Zhu ZY, Fung WK. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. Biometrika. 2002; 89:579–590.

Holt D, Smith TMF, Winter PD. Regression analysis of data from complex survey. Journal of Royal Statistical Society, Ser.A. 1980; 143:474–487.

Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association. 1952; 47:663–685.

11. Huang JZ, Zhang L, Zhou L. Efficient estimation in marginal partially linear models for longitudinal/clustered data using spline. Scandinavian Journal of Statistics. 2007; 34:451–477.

Jacobson JL, Jacobson SW. Intellectual impairment in children exposed to polychlorinated biphenyls in utero. New England Journal of Medicine. 1996; 335:783–789. [PubMed: 8703183]

Lehmann, E. Theory of point estimation. Wiley; New York: 1983.

Lin X, Carroll RJ. Semiparametric regression for clustered data using generalized estimating equations. Journal of the American Statistical Association. 2001; 96:1045–1056.

Longnecker MP, Klebanoff MA, Zhou H, Brock JW. Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth. Lancet. 2001; 358:110–114. [PubMed: 11463412]

Patandin S, Lanting CI, Mulder PG, et al. Effects of environmental exposure to polychlorinated biphenyls and dioxins on cognitive abilities in Dutch children at 42 months of age. Journal of Pediatrics. 1999; 134:33–41. [PubMed: 9880446]

Qin J, Lawless JF. Empirical Likelihood and General Estimating Equations. The Annals of Statistics. 1994; 22:300–325.

Qu A, Li R. Quadratic inference functions for varying coefficient models with longitudinal data. Biometrics. 2006; 62:379–391. [PubMed: 16918902]

Rathouz PJ, Satten GA, Carroll RJ. Semiparametric inference in matched case-control studies with missing covariate data. Biometrika. 2002; 89:905–916.

Rogan WJ, Gladen BC, Hung KL, Koong SL, Shih LY, Taylor JS, Wu YC, Yang D, Ragan NB, Hsu CC. Congenital poisoning by polychlorinated biphenyls and their contaminants in Taiwan. Science. 1988; 241:334–336. [PubMed: 3133768]

Rogan WJ, Gladen BC. PCBs, DDE, and child development at 18 and 24 months. Annals of Epidemiology. 1991; 1:407–413. [PubMed: 1669521]

Ruppert D. Selecting the number of knots for penalized splines. Journal of Computational and Graphical Statistics. 2002; 11:735–757.

Ruppert, D.; Carroll, R. Cornell University, School of Operations Research and Industrial Engineering; 1997. Penalized Regression Splines. working paper(Available at www.orie.cornell.edu/david/papers)

Schantz SL, Widholm JJ, Rice DC. Effects of PCB exposure on neuropsychological function in children. Environmental Health Perspectives. 2003; 111:357–376. [PubMed: 12611666]

Schildcrout JS, Heagerty PJ. On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. Biostatistics. 2008; 9:735–749. [PubMed: 18372397]

Syracuse Research. Toxicological Profile for Polychlorinated Biphenyls (PCBs). 2000. p. 556-576.prepared for U.S. Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry, U.S. Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry, Atlanta, GA

Wang N, Carroll RJ, Lin X. Efficient semiparametric marginal estimation for longitudinal/clustered data. Journal of the American Statistical Association. 2005; 100:147–157.

Wang X, Zhou H. A Semiparametric Empirical Likelihood Method For Biased Sampling Schemes In Epidemiologic Studies With Auxiliary Covariates. Biometrics. 2006; 62:1149–1160. [PubMed: 17156290]

Weaver MA, Zhou H. An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. Journal of the American Statistical Association. 2005; 100:459–469.

White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. American Journal of Epidemiology. 1982; 115:119–128. [PubMed: 7055123]

Yu Y, Ruppert D. Penalized Spline Estimation for Partially Linear Single Index Models. Journal of the American Statistical Association. 2002; 97:1042–1054.

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]

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. Epidemology. 2007; 18:461–468.

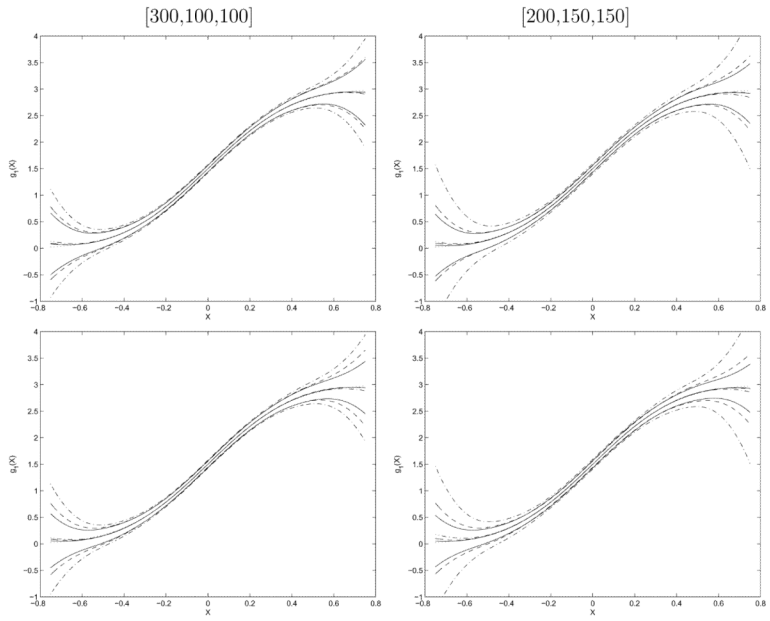

Figure 1.

Simulation results in Study 1. Confidence band comparison with sample size allocation of [300, 100, 100] and [200, 150, 150]. The plots from the first and second lines correspond to cut point $a = 0.7$ and 1.3 respectively. In each plot, the dotted curve in the middle is the true function. The solid, dashed and dot-dashed curves in the middle are the average P-spline fits over 1000 simulation respectively by the proposed method (P), the method based on a SRS sample with the same sample size as the ODS sample (S) and the method based on the SRS samples from the ODS design (V). For the confidence bands, the solid, dashed and dot-dashed curves are the confidence bands obtained by the P, S and V methods respectively.

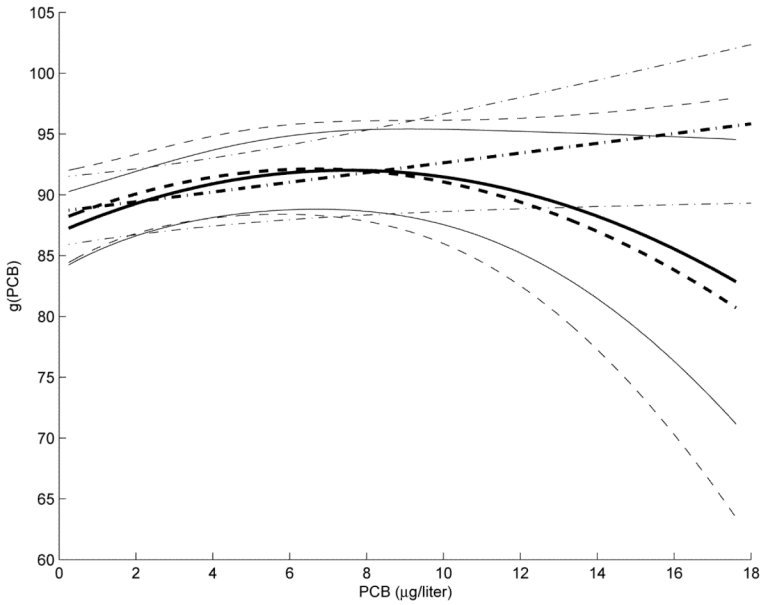

Figure 2.

Analysis results of CPP data. The estimated function *g* on PCB. Thicker solid and dashed curves correspond to estimates obtained by the P and V methods, and the thicker dot-dashed line corresponds to the linear fit using Zhou et al. (2002) method. And the solid, dashed curves and dot-dashed lines correspond to estimated confidence bands obtained by the corresponding methods respectively.

Table 1

Simulation results over 1000 replications for sample size allocation of [300,100,100] in Study 1

		$\hat{g}$			$\hat{\gamma}$			
a	**Methods**	**AMSE**	**ABIAS**	**AVAR**	**MEAN**	**SE**	$\widehat{SE}$	**CI**
0.7	P	0.0104	0.0067	0.0103	1.0019	0.0399	0.0388	0.9360
	V	0.0280	0.0115	0.0276	1.0018	0.0516	0.0514	0.9460
	S	0.0130	0.0084	0.0128	0.9984	0.0397	0.0399	0.9400
1.0	P	0.0100	0.0069	0.0099	1.0013	0.0389	0.0383	0.9510
	V	0.0298	0.0134	0.0291	1.0012	0.0512	0.0515	0.9620
	S	0.0134	0.0089	0.0132	0.9998	0.0402	0.0399	0.9470
1.3	P	0.0078	0.0061	0.0077	1.0001	0.0371	0.0377	0.9490
	V	0.0281	0.0118	0.0277	0.9976	0.0524	0.0516	0.9590
	S	0.0133	0.0089	0.0132	0.9980	0.0396	0.0397	0.9590

NOTE 1: P= the proposed method; V= the method based on the SRS sample from ODS design; S = the method based on a SRS sample with the same sample size as ODS design; AMSE= average of the mean square error (MSE) of the estimator $\widehat{g}$ over 501 grid points; ABIAS= average of the absolute value of the bias of $\widehat{g}$ over 501 grid points; AV AR= average of variance of $\widehat{g}$ over 501 grid points; MEAN= mean of $\widehat{\gamma}$; SE= standard error of $\widehat{\gamma}$; $\widehat{SE}$= estimated standard error of $\widehat{\gamma}$ using large sample approximation; CI= coverage probability of 95% nominal confidence interval; a: the cut-points for the ODS design are $\mu_Y \pm a\sigma_Y$.

Table 2

Relative efficiency (RE) comparison over 1000 *replications*

	a	$\rho = 0.8$	$\rho = 0.6$	$\rho = 0.4$
Study 1	0.7	1.2143	1.2500	1.3980
	1.0	1.3725	1.3400	1.5976
	1.3	1.4947	1.7051	1.8143
Study 2	0.7	1.1219	1.3077	1.3158
	1.0	1.2410	1.4247	1.5303
	1.3	1.3378	1.6875	1.9636

NOTE 2: ρ = the proportion of the SRS sample in the ODS design (ρ = n0/n); $AMSE(\widehat{g_S})/AMSE(\widehat{g_P})$; $\widehat{g_S}$ and $\widehat{g_P}$ denote the P and S estimators of g respectively; a: the cut-points for the ODS design are $\mu_Y \pm a\sigma_Y$.

Table 3

Description of the sample of the variables in CPP data

	MEAN	STD	25% percentile	75% percentile	MIN	MAX
IQ	96.23	16.09	84.00	108.00	56.00	145.00
PCB	3.16	1.93	1.88	3.86	0.25	17.61
EDU	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
RACE	0.49	0.50	0.00	1.00	0.00	1.00
SEX	0.50	0.50	0.00	1.00	0.00	1.00

NOTE 3: MEAN = sample mean of the variable; STD= sample standard deviation of the variable; MIN = minimum value of the sample of the variable; MAX = maximum value of the sample of the variable.

Table 4

Analysis results for CPP data

		Partial linear model						Linear Model		
a	**Methods**	**AMSE**	**ABIAS**	**AVAR**	**MEAN**	**SE**	$\widehat{SE}$	**CI**		
	$\hat{\beta}_P$	$SE(\hat{\beta}_P)$	**95% CI**	$\hat{\beta}_V$	$SE(\hat{\beta}_V)$	**95% CI**	$\hat{\beta}_L$	$SE(\hat{\beta}_L)$	**95% CI**
PCB		See Figure 2			See Figure 2			See Figure 2	
EDU1	2.79	1.04	(0.75, 4.83)	2.73	1.24	(0.30, 5.16)	2.72	1.04	(0.68, 4.76)
EDU2	10.44	1.66	(7.19, 13.69)	9.39	2.02	(5.43, 13.35)	10.47	1.66	(7.22, 13.72)
SES	1.39	0.20	(1.00, 1.78)	1.31	0.24	(0.84, 1.78)	1.40	0.20	(1.01, 1.79)
RACE	−7.97	0.75	(−9.44, −6.50)	−7.73	0.89	(−9.47, −5.99)	−7.82	0.75	(−9.29, −6.35)
SEX	−0.81	0.69	(−2.16, 0.54)	−0.75	0.84	(−2.40, 0.90)	−0.79	0.69	(−2.14, 0.56)

NOTE 4: The result of linear model is obtained using the Zhou et al. (2002) method, and in the case of linear model, g(PCB) = $\alpha_1 + \alpha_2$PCB; $\widehat{\beta}_P$ and $\widehat{\beta}_V$ correspond to the estimates obtained by the P and V methods respectively; $SE\left(\widehat{\beta}_P\right)$ and $SE\left(\widehat{\beta}_V\right)$ are the estimated standard errors of corresponding estimators.