Biostatistics (2011), **12**, 3, *pp.* 506–520
doi:10.1093/biostatistics/kxq070
Advance Access publication on December 14, 2010

Partial linear inference for a 2-stage outcome-dependent sampling design with a continuous outcome

GUOYOU QIN

Department of Biostatistics, School of Public Health and Key Laboratory of Public Health Safety, Ministry of Education of China, Fudan University, Shanghai 200032, People's Republic of China
Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599, USA

HAIBO ZHOU*

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
zhou@bios.unc.edu

SUMMARY

The outcome-dependent sampling (ODS) design, which allows observation of exposure variable to depend on the outcome, has been shown to be cost efficient. In this article, we propose a new statistical inference method, an estimated penalized likelihood method, for a partial linear model in the setting of a 2-stage ODS with a continuous outcome. We develop the asymptotic properties and conduct simulation studies to demonstrate the performance of the proposed estimator. A real environmental study data set is used to illustrate the proposed method.

Keywords: Biased sampling; Partial linear model; P-spline; Validation sample; 2-stage.

1. INTRODUCTION

Two-stage design has been widely recognized as an efficient study design for epidemiological studies. For the discrete outcome, White (1982) proposed a 2-stage case–control design for a rare disease and exposure scenario, where a large simple random sample (SRS) sample is drawn in the first stage, and further subsamples with additional potential confounding variables are drawn in the second stage from the strata identified based on the disease status and the exposure from the first-stage sample. Greater efficiency can be obtained through the 2-stage sampling design (e.g., Breslow and Cain, 1988; Zhao and Lipsitz, 1992; Langholz and Borgan, 1995; Breslow *and others*, 2003).

For the continuous case, Weaver and Zhou (2005) considered a 2-stage outcome-dependent sampling (ODS) design, where in the second stage, in addition to an SRS sample, some supplemental samples were collected in the second stage via an ODS way, that is, the subsamples were drawn from the strata identified by the outcome from the first stage. The ODS design has attracted much attention in the last several decades since it is a cost-effective way to improve study efficiency. For example, the case–control

*To whom correspondence should be addressed.

study is a well-known ODS design with a binary outcome variable. One can ignore the sampling scheme when the underlying model is logistic. However, this is not true for other link function or the nonbinary outcomes. For inference of the data from ODS design with a continuous response, the usual methods will be not appropriate since the ODS is a biased sampling scheme. There are several methods that are developed to inference the data from ODS design, for example, the weighted estimating equation methods by Horvitz and Thompson (1952), the semiparametric empirical likelihood method proposed by Zhou *and others* (2002), the pseudoscore estimation methods by Chatterjee *and others* (2003), and the estimated likelihood method proposed by Weaver and Zhou (2005).

A recent epidemiological study (Gray *and others*, 2005) employed the 2-stage ODS design of Weaver and Zhou. In this study, the investigators were interested in how the children's intelligence quotient (IQ) at 7 years of age is related to an environmental pollutant, polychlorinated biphenyls (PCBs). The study subjects are children who were born into the Collaborative Perinatal Project (CPP) which is a prospective cohort designed to provide precise data for studies of a wide variety of neuropsychological outcomes and birth defects (Niswander and Gordon, 1972). Since it is expensive to obtain the PCB measurement for all the first-stage sample, the investigators decided to obtain the PCB measurement for a sample that was sampled in an ODS way from the first-stage sample based on the observed IQ scores (Gray *and others*, 2005). Then the ODS sample with measured PCBs consists of the second-stage sample. The literature on the ODS (Zhou *and others*, 2002; Weaver and Zhou, 2005) generally assumes that the effect on the outcome of the exposure is linear, which is chosen mainly for its mathematically convenience. However, in practice, the true relation between them is usually unknown. For this CPP data, significant relation between the children's IQ and PCB was not found in the framework of linear model although the efficient ODS design was adopted (e.g., Zhou *and others*, 2002; Weaver, 2001).

The partial linear model (PLM) for continuous outcomes (Zeger and Diggle, 1994; He *and others*, 2002), where the outcome is assumed to depend on some covariate W 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 study of Gray *and others* (2005) to have a more flexible PLM approach to investigate the relation of low-level PCB exposure and childrens cognitive development. Motivated by the CPP study, in this article, we studied the inference of a PLM in a 2-stage ODS design with a continuous outcome, where the first-stage data are an SRS, but the second-stage data are collected via an ODS design. The softwares to fit the generalized additive model (GAM) (Ruppert *and others*, 2003) were generally developed based on methods that assume SRS design, and consequently the likelihood composition would not handle the complex data structures of ODS design. Hence, they cannot be directly applied to the proposed 2-stage ODS design nor a minor modification to softwares for GAM can accomplished the task.

In our simulation study, we compared the proposed estimator, denoted by the P estimator, with 3 competing methods: the semiparametric empirical likelihood estimator incorporating the P-spline method based on the second-stage sample denoted by the SEPLE-ODS estimator, which is an extension of Zhou *and others* (2002), the penalized inverse-probability weighted estimator denoted by the IPW estimator, which is the extension of Horvitz and Thompson (1952) and defined as the maximizer of the following penalized weighted likelihood function:

$$\sum_{k=1}^{K} p_k \sum_{i \in V_k} f(Y_i | X_i; \theta) - \frac{1}{2} N \lambda \theta^T \Psi \theta, \tag{1.1}$$

where $p_k = \frac{1}{n_0/N + n_k/N_k}$ and the penalized maximum likelihood method based on the SRS sample with the sample size equal to the second-stage sample, denoted by the PMLE-SRS estimator. Compared with the SEPLE-ODS method, the proposed method incorporates the outcome information of the first-stage sample. Compared with the PMLE-SRS method, a more efficient ODS design, than the SRS design, is

used to sample second-stage data. Further more outcome information of the first-stage sample are also used for inference. The proposed method is preferred to the IPW method because we utilize the actual observation of the outcome variable, while the IPW method only uses the strata proportion.

The rest of this article is organized as follows. In Section 2, we describe the 2-stage ODS design and the PLM. The penalized likelihood is proposed. The inference method and main asymptotic results of the proposed estimator will be given in Section 3. In Section 4, we present simulation studies to investigate the performance of the proposed method. In Section 5, we illustrate our proposed method with the analysis of the CPP data set. We conclude with a brief discussion in Section 6.

2. DATA STRUCTURE, MODEL, AND THE PENALIZED LOG-LIKELIHOOD

2.1 *Two-stage ODS design, data structure, and model*

Let Y denote a continuous outcome variable. Assume that the domain of Y is a union of K mutually exclusive intervals: $C_k = (c_{k-1}, c_k], k = 1, \ldots, K$, with c_k being some known constants such that $-\infty = c_0 < c_1 < c_2 < \cdots < c_k = \infty$. Thus, the collection of intervals $\{C_k, k = 1, \ldots, K\}$ partition the study population into K strata. We consider a 2-stage ODS design as follows. In the first stage, an SRS sample is drawn from the underlying population of interests with sample size N. It is assumed that the outcome variable for the N individuals is observed. In the second stage, an ODS sample is drawn from the N individuals, which consists of an overall SRS of size n_0 and stratified random samples from these K strata (Supplemental samples) with size n_k for the supplemental sample from the kth stratum. Both the outcome and the covariates will be observed for the individuals who are selected in the second stage, that is, the ODS sample, whereas the covariates will not be observed for those not selected in the second stage.

To fix notations, let $n_V = \sum_{k=0}^{K} n_k$ be the total size of the second-stage sample for which we observe both the outcome and covariates, and let $n_{\bar{V}} = N - n_V$ be the number of individuals in the population for whom only the outcome variable is observed but the covariates are not observed. We refer to the n_V complete observations as the second-stage sample and the $n_{\bar{V}}$ incomplete observations as the incomplete first-stage sample. Let V denote the index set of all observations in the second-stage sample and $\bar{V}$ denote the index set of all observations in the incomplete first-stage sample. Furthermore, let the V_k and $\bar{V}_k$ denote the index sets for the observations in the second-stage sample and incomplete first-stage sample from the kth stratum.

We denote X as the covariates, then the data structure can be summarized as

Stage 1: $\{Y_i\}$, for $i = 1, \ldots, N$;

Stage 2: SRS sample: $\{Y_i, X_i\}$, for $i = 1, \ldots, n_0$;

Supplemental sample: $\{Y_i, X_i | Y_i \in C_k, k = 1, \ldots, K\}$, for $i = n_0 + 1, \ldots, n_V$.

We assume that the conditional mean of the outcome is related to covariates $X = (W^T, Z^T)^T$ as

$$E(Y|W, Z) = g(W) + Z^T\gamma, \tag{2.1}$$

where $g(\cdot)$ is an unknown nonparametric function of the exposure variable W and γ is a vector of p-dimensional regression coefficients. Our goal in this article is to inference $g(\cdot)$ and γ in model (2.1).

To further introduce additional notations: let F_X denote the distribution function of X, $f(Y|X;\theta)$ and $f_Y(Y|\theta)$ denote the conditional and marginal density functions of Y.

2.2 *Penalized log-likelihood function*

Several nonparametric smoothing methods can be adopted to estimate the nonparametric function $g(\cdot)$. An incomplete list of publications include: Lin and Carroll (2001), Yu and Ruppert (2002), Huang *and others* (2007), and Zhu *and others* (2008). As in Yu and Ruppert (2002) who studied the estimation of partially

linear single-index model, we adopt penalized splines (P-spline) to estimate the nonparametric function $g(\cdot)$. P-spline is an extension of smoothing spline allowing more flexible choice of knots and penalty. The penalty is used to achieve a smooth fit of the nonparametric function. Following Yu and Ruppert (2002), under the working assumption that $g(\cdot)$ is a rth degree spline function with T fixed knots $t_1, \ldots, t_T$, we then have $g(w) = \pi^T(w)\alpha$, where $\pi(w) = (1, w, w^2, \ldots, w^r, (w-t_1)_+^r, \ldots, (w-t_T)_+^r)^T$ is a r-degree truncated power spline basis with knots $\{t_i\}_{i=1}^T$, $(w)_+^r = w^r 1_{w \geqslant 0}$ and α is a $r+T+1$-dimensional vector. Thus, we have

$$g(W_i) + Z_i^T \gamma = \pi^T(W_i)\alpha + Z_i^T \gamma = D_i^T \theta, \tag{2.2}$$

where $D_i = (\pi^T(W_i), Z_i^T)^T$ and $\theta = (\alpha^T, \gamma^T)^T$.

For the second-stage sample, the likelihood is

$$L(\theta, F_X) = [\Pi_{i \in V} f(Y_i|X_i;\theta)][\Pi_{i \in V} \mathrm{d}F_X(X_i)][\Pi_{k=1}^K \pi_k(\theta, F_X)^{-n_k}], \tag{2.3}$$

where $\pi_k(\theta, F_X) = \int \psi_k(x;\theta)\mathrm{d}F_X(x)$, $\psi_k(x;\theta) = \int_{C_k} f(y|x;\theta)\mathrm{d}y$.

The likelihood for the incomplete first-stage sample is

$$\Pi_{k=1}^K \Pi_{j \in \bar{V}_k} \frac{f_Y(Y_j;\theta)}{\pi_k(\theta, F_X)}. \tag{2.4}$$

Finally, note that the stratum size for the incomplete first-stage sample $n_{\bar{V},k} = N_k - n_{0,k} - n_k, k = 1, \ldots, K$ follows a multinomial law such that

$$\Pr(\{n_{\bar{V},k}\}) = \frac{(N-n_0)!}{\Pi_{k=1}^K (N_k - n_{0,k})!} \Pi_{k=1}^K [\pi_k(\theta, F_X)]^{N_k - n_{0,k}}, \tag{2.5}$$

where $n_{0,k}$ is a random variable representing the number of observation in the SRS that belong to the kth stratum.

As noted by Weaver and Zhou (2005), by combining (2.3–2.5), the full information likelihood of both the first- and the second-stage samples is shown proportional to

$$L(\theta, F_X) = \left[\prod_{i \in V} f(Y_i|X_i;\theta)\right]\left[\prod_{j \in \bar{V}} f_Y(Y_j;\theta)\right]. \tag{2.6}$$

To achieve a smooth fit, we introduce a penalized term into the log-likelihood, and our penalized log-likelihood is expressed as follows:

$$\mathrm{pl}(\theta, F_X; \lambda) = \sum_{i \in V} \log[f(Y_i|X_i;\theta)] + \sum_{j \in \bar{V}} \log f_Y(Y_j;\theta) - \frac{1}{2} N \lambda \theta^T \Psi \theta, \tag{2.7}$$

where $\Psi = \mathrm{diag}\{(\mathbf{0}_{(r+1)\times 1}^T, \mathbf{1}_{T\times 1}^T, \mathbf{0}_{p\times 1}^T)^T\}$, $\theta^T \Psi \theta$ is a common quadratic penalty function and λ is the smoothing parameter. Some discussion on the selection of knots and smoothing parameter are provided in Section 2 of the supplementary material available at *Biostatistics* online.

3. Maximum estimated penalized likelihood estimation

The penalized log-likelihood (2.7) involves an unspecified marginal distribution function of x, $F_X(x)$. Inference of θ will depend on how one handles $F_X(x)$. A commonly used idea is to replace the $F_X(x)$ in the likelihood with a consistent estimator and this will lead to an estimated likelihood. Such idea has been

widely used in the statistical literature, for example, Pepe and Fleming (1991), Reilly and Pepe (1995, 1997), Lawless *and others* (1999), Zhou and Pepe (1995); Zhou and Wang (2000), and Weaver and Zhou (2005).

We adopt the following estimator, which is proposed by Weaver and Zhou (2005) to specially accommodate the ODS sampling nature in the second stage, as our estimator of $F_X(x)$:

$$\widehat{F}_X(x) = \sum_{k=1}^{K} \frac{N_k}{N} \hat{F}_k(x), \tag{3.1}$$

where

$$\widehat{F}_k(x) = \sum_{i \in V_k} \frac{I\{X_i \leqslant x\}}{n_k + n_{0,k}} \tag{3.2}$$

is the empirical cumulative distribution function for the covariates based on all sampled observations from the kth stratum. Note that the X_i is a vector representing the covariates $\{W_i, Z_i\}$. The value of $I\{X_i \leqslant x\}$ in (3.2) is equal to 1 if each component of X_i is less than or equal to the corresponding component of x, otherwise $I\{X_i \leqslant x\} = 0$.

Using (3.1), we can obtain a consistent estimate of the marginal distribution of Y as

$$\begin{aligned} \widehat{f}_Y(Y_j;\theta) &= \int f_Y(Y_j|X;\theta)\mathrm{d}\hat{F}_X(X) \\ &= \sum_{k=1}^{K} \frac{N_k}{N(n_k + n_{0,k})} \sum_{i \in V_k} f(Y_j|X_i;\theta), \end{aligned} \tag{3.3}$$

assuming that $n_k + n_{0,k} > 0$ for all $k = 1, \ldots, K$.

Substituting (3.3) into (2.7), we obtain the following estimated penalized log-likelihood function for θ:

$$\begin{aligned} \widehat{\mathrm{pl}}(\theta;\lambda) = &\sum_{i \in V} \log f(Y_i|X_i;\theta) + \sum_{j \in \bar{V}} \log \left\{ \sum_{k=1}^{K} \frac{N_k}{N(n_k + n_{0,k})} \sum_{i \in V_k} f(Y_j|X_i;\theta) \right\} \\ &- \frac{1}{2} N \lambda \theta^T \Psi \theta. \end{aligned} \tag{3.4}$$

We define the maximum estimated penalized likelihood estimator $\hat{\theta}$ to be the maximizer for the estimated penalized likelihood functions, and it can be obtained by invoking the Newton–Raphson procedure.

3.1 *Asymptotic properties of* $\hat{\theta}$

Under some regularity conditions, the asymptotic property of $\hat{\theta}$ is summarized in the following theorem.

THEOREM 3.1 (i) If the smoothing parameter $\lambda = \mathrm{o}(1)$, $\hat{\theta}$ converges to θ_0 with probability one. (ii) If the smoothing parameter $\lambda = \mathrm{o}(1/\sqrt{N})$, the maximizer $\hat{\theta}$ is asymptotically distributed as a normal distribution $\sqrt{N}(\hat{\theta} - \theta_0) \to N(0, \Sigma)$, where

$$\Sigma(\theta) = I^{-1}(\theta) + \sum_{l=1}^{K} \frac{\pi_k^2}{\rho_k \rho_V + \pi_k \rho_0 \rho_V} I^{-1}(\theta) \Sigma_k(\theta) I^{-1}(\theta),$$

with $\pi_k = \int \psi_k(x;\theta)\mathrm{d}F_X(x)$ and $\psi_k(x;\theta) = \int_{C_k} f(y|x;\theta)\mathrm{d}y$,

$$I(\theta) = -\rho_0\rho_V E\left[\frac{\partial^2 \log f(Y|X;\theta)}{\partial\theta\partial\theta^T}\right] - \sum_{k=1}^{K}\left\{\rho_k\rho_V E_k\left[\frac{\partial^2 log f(Y|X;\theta)}{\partial\theta\partial\theta^T}\right]\right.$$
$$\left. +[\pi_k(1-\rho_0\rho_V) - \rho_k\rho_V]E_k\left[\frac{\partial^2 log f_Y(Y;\theta)}{\partial\theta\partial\theta^T}\right]\right\}$$

with E_k denote expectation conditional on $Y \in C_k$, and

$$\Sigma_k(\theta) = \mathrm{var}_{X|Y\in C_k}\left\{\sum_{l=1}^{K}[\pi_l(1-\rho_0\rho_V) - \rho_k\rho_V]E_{Y|Y\in C_l}[M_X(Y;\theta)]\right\}$$

with

$$M_X(Y;\theta) = \frac{f(Y|X;\theta)}{f_Y(Y;\theta)} - \frac{\partial f_Y(Y;\theta)}{[f_Y(Y;\theta)]^2} f(Y|X;\theta).$$

All the quantities involved are evaluated at the true parameter value θ_0.

The regularity conditions and a brief proof for Theorem 3.1 is provided in Section 1 of the supplementary material available at *Biostatistics* online. A consistent estimator $\hat{\Sigma}$ of the covariance matrix Σ is

$$\hat{\Sigma} = \hat{I}^{-1}(\hat{\theta};\lambda)\hat{I}(\theta)\hat{I}^{-1}(\hat{\theta};\lambda) + \sum_{k=1}^{K}\frac{(N_k/N)^2}{(n_k+n_{0,k})/N}\hat{I}^{-1}(\hat{\theta};\lambda)\hat{\Sigma}_k(\hat{\theta})\hat{I}^{-1}(\hat{\theta};\lambda), \tag{3.5}$$

where $\hat{I}(\theta;\lambda) = -\frac{1}{N}\frac{\partial^2\widehat{\mathrm{pl}}(\theta;\lambda)}{\partial\theta\partial\theta^T}$, $\hat{I}(\theta) = -\frac{1}{N}\frac{\partial^2\hat{l}(\theta)}{\partial\theta\partial\theta^T}$, $\hat{l}(\theta) = \widehat{\mathrm{pl}}(\theta;\lambda) + \frac{1}{2}N\lambda\theta^T\Psi\theta$ and $\hat{\Sigma}_k(\theta) = \widehat{\mathrm{var}}_{X_i;i\in V_k}\left\{\sum_{l=1}^{K}\frac{n_{\bar{V},l}}{N}\hat{\bar{M}}_{X_i,l}(\theta)\right\}$, with $\hat{\bar{M}}_{X_i,l}(\theta) = \sum_{j\in\bar{V}_l}\left\{\frac{\partial f(Y_j|X_i;\theta)/\partial\theta}{\hat{f}_Y(Y_j;\theta)} - \frac{f(Y_j|X_i;\theta)\partial\hat{f}_Y(Y_j;\theta)/\partial\theta}{[\hat{f}_Y(Y_j;\theta)]^2}\right\}/n_{\bar{V},l}$.

REMARK 3.2 One can make inference on the nonparametric function and the regression coefficients using Theorem 3.1. In particular, one can construct joint confidence region and test hypotheses. For example, if we want to test the null hypothesis H_0: $B\theta_0 - s_0 = 0$ where B is a $d_1 \times d (d = T + r + 1 + p)$ matrix with full rank $d_1 \leqslant d$, then the test can be implemented by using the Wald test, with the test statistic $U = (B\hat{\theta} - s_0)^T(B\hat{\Sigma}B^T)^{-1}(B\hat{\theta} - s_0)$, that has an asymptotic chi-square distribution with d_1 degree of freedom.

REMARK 3.3 Particular interest of the Wald test is that one can use it to test if the nonparametric function describing the relation between an exposure variable W and a response Y is linear, that is, whether a linear model is enough to model the relation between Y and W. To do so, we repress the $T + r + 1$-dimensional vector α as (α_1^T, α_2^T), where $\alpha_1 = (\alpha_{11}, \alpha_{12})^T$ is a 2D vector and α_2 is a $T + r - 1$-dimensional vector. We are interested in testing the following null hypothesis: $\alpha_2 = \alpha_2^0 = (0, \ldots, 0)^T$. Under this null hypothesis, we have $g(W) = \alpha_{11} + \alpha_{12}W$, that is, the exposure variable W is related to response Y in a linear fashion.

4. SIMULATION STUDIES

In this section, we use simulated data to evaluate the performance of our proposed estimator. We compare the proposed estimator, denoted by the P estimator, with 3 competing methods: the semiparametric empirical likelihood estimator incorporating the P-spline method based on the second-stage sample denoted

by the SEPLE-ODS estimator, the penalized inverse-probability weighted estimator denoted by the IPW estimator,which is the extension of the Horvitz and Thompson (1952) and defined as the maximizer of the following penalized weighted likelihood function:

$$\sum_{k=1}^{K} p_k \sum_{i \in V_k} f(Y_i | X_i; \theta) - \frac{1}{2} N \lambda \theta^T \Psi \theta,$$

where $p_k = \frac{1}{n_0/N + n_k/N_k}$ and the penalized maximum likelihood method based on the SRS sample with the sample size equal to the second-stage sample, denoted by the PMLE-SRS estimator. For all simulations, we generate 1000 simulated data sets. We consider 2 PLMs with different nonparametric functions. One is a monotonic function, while the other is a unimode function in threshold regions. We adopt a 3-degree truncated power spline basis and choose 10 fixed knots selected as the sample quantiles.

Study 1. The data were generated according to the following partial linear mixed model,

$$Y = g_1(W) + Z\gamma + e,$$

where $g_1(W) = 3\Phi(3.2W)$ is a standard normal distribution function, W denotes a continuous exposure variable of interest and $e \sim N(0, \sigma_0^2)$. We assumed that $W \sim N(0, 0.25^2)$ and $Z \sim N(0, 0.3^2)$. We fixed $\gamma = 1$ and $\sigma_0^2 = 0.4$. In all the simulation presented here, we allowed Y to be observed for the entire study population but assumed that both W and Z were observed only for the second-stage sample. The size of the first-stage sample was set to be $N = 2000$, and the second-stage sample fraction was set to $\rho = 0.20$, that is, the size of the second-stage sample is $400 = 2000 \times 0.20$.

The second-stage sample consists of an SRS sample (n_0) supplemented with additional samples from individuals with Y values in the upper and lower tails of the marginal distribution (i.e., $n_1 = n_3$ and $n_2 = 0$) with cut-points $\mu_Y \pm a\sigma_Y$, where μ_Y and σ_Y represent the mean and standard deviation of Y and a is taken to be 0.7 and 1.0, respectively. We considered 2 cases of the allocation of the sample sizes as $n_0 = 300$, $n_1 = n_3 = 50$, and $n_0 = 200$, $n_1 = n_3 = 100$.

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 401 equal spaced grid points in $[-0.75, 0.75]$ (the mean of X minus and plus 3 times the standard deviation of X) over 1000 replications. The relative efficiency (REF) of the P estimator of nonparametric function over the the other 2 estimators is also calculated. The REF is defined by $\text{AMSE}(\hat{g}_M)/\text{AMSE}(\hat{g}_P)$ where M denote P, SEPLE-ODS, IPW, and PMLE-SRS estimators. Moreover, we calculated mean, Monte Carlo standard error, estimated standard error using large sample properties and coverage probability of 95% nominal confidence interval for the estimator of regression coefficient γ.

The results are summarized in Table 1. From Table 1, we can find that the proposed estimator of nonparametric function is most efficient with the smallest average mean square error (AMSE over 401 grid points) among all the estimators compared. For the estimation of the regression coefficient γ, the proposed estimators are generally more efficient than the other 2 estimators with smaller variance. 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 SEPLE-ODS estimator, which indicates that efficiency gain can be achieved by incorporating actual observation of the outcome from the first stage. Moreover, the SEPLE-ODS estimator is more efficient than the PMLE-SRS one, which means that the ODS design can achieve smaller standard errors of the interested estimates using the same sample size as the SRS design. Therefore, the ODS design is a cost-effective way to improve the study efficiency.

Figure 1 presents curves of the true function and the average P-spline estimates of g_1 by P, SEPLE-ODS, IPW, and PMLE-SRS estimators over 1000 simulation and gives the confidence bands obtained by

Table 1. *Simulation results over* 1000 *replications in Study* 1

a		$\hat{g}_1$				$\hat{\gamma}$			
	Methods	AMSE	ABIAS	AVAR	EF	MEAN	SE	$\widehat{\text{SE}}$	CI
				$n_0 = 200, n_1 = n_3 = 100, n_2 = 0$					
0.7	P	0.015	0.033	0.013	1.000	0.994	0.095	0.099	0.969
	ODS	0.019	0.032	0.017	1.267	0.998	0.098	0.101	0.965
	IPW	0.021	0.033	0.020	1.400	1.005	0.109	0.111	0.964
	SRS	0.022	0.032	0.020	1.467	1.007	0.107	0.105	0.957
1.0	P	0.016	0.033	0.014	1.000	0.994	0.096	0.099	0.960
	ODS	0.019	0.032	0.017	1.188	0.993	0.099	0.100	0.952
	IPW	0.022	0.034	0.020	1.375	1.000	0.111	0.116	0.959
	SRS	0.022	0.032	0.020	1.375	1.009	0.108	0.105	0.947
				$n_0 = 300, n_1 = n_3 = 50, n_2 = 0$					
0.7	P	0.017	0.033	0.014	1.000	0.996	0.099	0.099	0.956
	ODS	0.021	0.032	0.019	1.235	1.003	0.104	0.102	0.951
	IPW	0.021	0.033	0.019	1.235	1.006	0.106	0.104	0.948
	SRS	0.021	0.032	0.019	1.235	1.003	0.106	0.105	0.958
1.0	P	0.015	0.033	0.013	1.000	0.998	0.099	0.098	0.950
	ODS	0.020	0.031	0.018	1.333	1.000	0.104	0.100	0.937
	IPW	0.020	0.032	0.018	1.333	1.003	0.107	0.105	0.946
	SRS	0.023	0.033	0.021	1.533	1.003	0.102	0.105	0.969

Note: P: proposed estimator; ODS: the semiparametric empirical penalized likelihood estimator based on the second-stage sample (SEPLE-ODS estimator); IPW: inverse-probability weighted method; *SRS*: penalized maximum likelihood estimator based on the SRS sample with the sample size equal to the second stage sample; AMSE: average of the mean square error (MSE) of the estimator $\hat{g}$ over 451 grid points; ABIAS: average of the absolute bias of $\hat{g}$ over 451 grid points; AVAR: average of variance of $\hat{g}$ over 451 grid points ;EF: ratio of AMSEs over those of P estimators; MEAN: mean of $\hat{\gamma}$; SE: standard error of $\hat{\gamma}$; $\widehat{\text{SE}}$: estimated standard error of $\hat{\gamma}$; CI: coverage probability of 95% nominal confidence interval.

the P, SEPLE-ODS, IPW, and PMLE-SRS estimators for comparison. The confidence bands are based on a normal approximation using the Monte Carlo standard error and the cut-point equal to 0.7 and 1.0. In Figure 1, we can find that the confidence bands by the SEPLE-ODS, IPW, and PMLE-SRS estimators are obviously wider than those by the P estimator, specially obvious in the tail of the distribution of X. The reason is that the actual observation of the outcome from the first stage can provide more useful information in efficiency improvement.

In addition, it appears that there exists some finite-sample bias (see Figure 1). We conducted additional simulation with larger sample size and found that the bias will be reduced when the sample size is increased. The additional simulation suggests that while the method delivers consistent estimates (as n approaches infinity), there is a potential for appreciable finite-sample bias. The additional simulation results are presented in Section 4 of the supplementary material available at *Biostatistics* online.

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

$$Y = g_2(W) + Z\gamma + e,$$

where $g_2(W) = 1.5\sin(1.5W)$, W denotes a continuous exposure variable of interest. $e \sim N(0, \sigma_0^2)$. We assumed that $W \sim N(1, 0.5^2)$ and $Z \sim N(0, 0.3^2)$. Then we fixed γ and σ_0^2 the same value as those in Study 1.

The averages of the MSE, the absolute value of the bias and the Monte Carlo variance of the estimated nonparametric function $g_2(x)$ were also calculated over 401 equal spaced grid points in $[-0.5, 2.5]$ (the

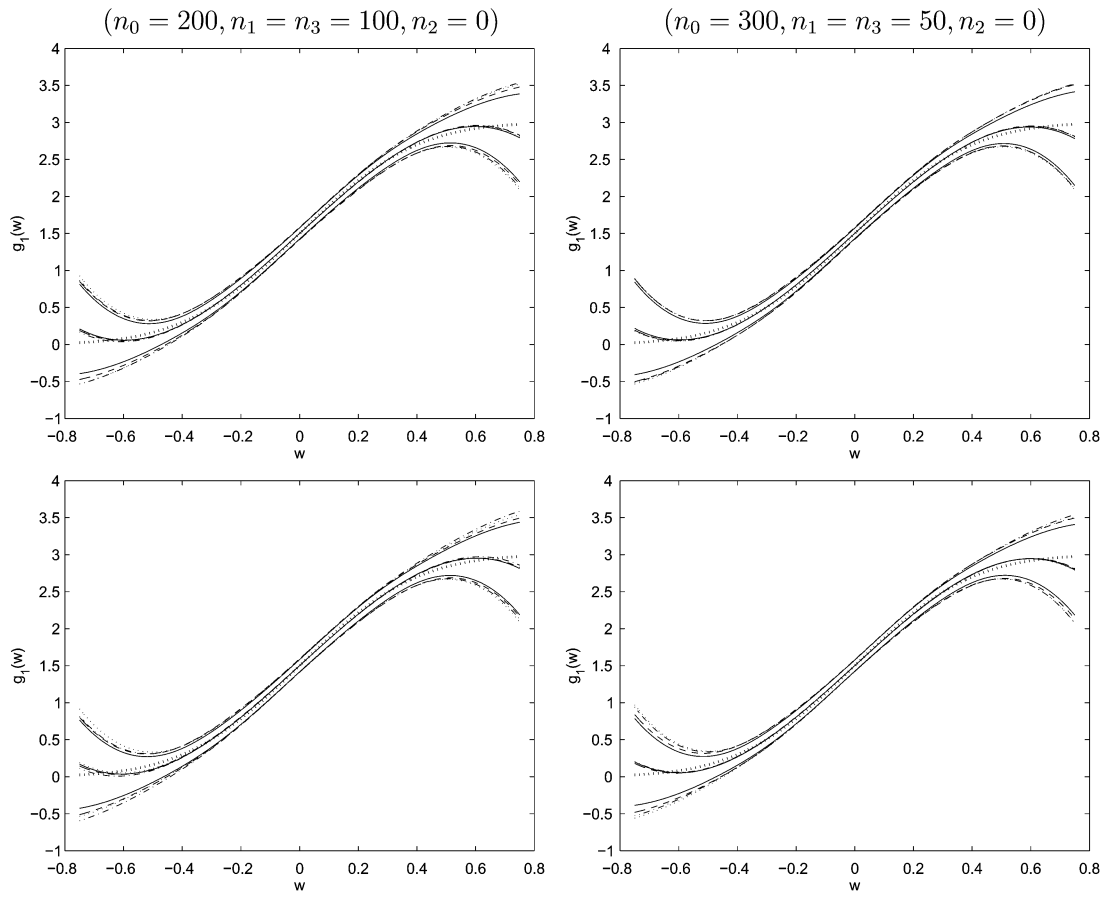

Fig. 1. Confidence band comparison in Study 1. The plots from the first and second lines correspond to cut point $a = 0.7$ and 1.0, respectively. In each plot, the thicker dotted curve in the middle is the true function. The solid, dashed, dot-dashed, and dotted curves in the middle are the average P-spline fits over 1000 simulation, respectively, by the proposed estimator (P), the semiparametric empirical penalized likelihood estimator based on the second-stage sample (SEPLE-ODS), the inverse-probability weighted estimator (IPW), and the penalized maximum likelihood estimator based on the SRS sample with sample size equal to the second-stage sample (PMLE-SRS). For the confidence bands, the solid, dashed, dot-dashed, and dotted curves are the confidence bands obtained by P, SEPLE-ODS, IPW, and PMLE-SRS estimators, respectively.

mean of X minus and plus 3 times the standard deviation of X) over 1000 replications. We also plotted the average P-spline estimate of g_2 by P, SEPLE-ODS, IPW, and PMLE-SRS estimators over 1000 replicates, and the confidence bands obtained by these estimators. The corresponding results including the table and figure are presented in Section 3 of the supplementary material available at *Biostatistics* online, and the conclusions are similar to Study 1.

5. ANALYSIS OF THE COLLABORATIVE PERINATAL PROJECT DATA SET

The proposed method is applied to CPP data set to assess the relationship between maternal pregnancy serum level of PCBs and children's subsequent IQ test performance. We use the Weschler Intelligence Scales for children at 7 years of age (IQ) as outcome variable and the PCBs level as the exposure variable.

We are mainly interested in the effect of PCB on IQ measurement. In addition to PCB, other covariates include socioeconomic status of the child's family (SES), the gender (SEX) and race (RACE) of the child, and the parents eduction (EDU).

There are 38 709 subjects consisting of the first-stage sample with completely observed variables IQ, SES, SEX, RACE, and EDU. In the second stage, an ODS design was conducted based on the first-stage sample with the sample size of 1038. The samples obtained by the ODS design are the completely observed samples with additional measured exposure variable of interest PCB. The second-stage sample consists of an SRS of 849 subjects and 2 supplemental subgroups which are defined by children's IQ scores that are one standard deviation (14) above and below the mean (96) of the population IQ scores, with 81 subjects in the low-IQ group and 108 subjects in the high-IQ group. Let Y denote IQ, W denote PCB, and Z denote (SES, SEX, RACE, and EDU). Then the CPP data structure can be summarized as

Stage 1: $\{Y_i, Z_i\}$, $N = 38\,709$;

Stage 2: SRS sample: $\{Y_i, X_i, Z_i\}$, $n_0 = 849$;

Supplemental sample:

1) $\{Y_i, X_i, Z_i | Y_i < 96.06 - 14.29\}$, $n_1 = 81$;

2) $n_2 = 0$;

3) $\{Y_i, X_i, Z_i | Y_i > 96.06 + 14.29\}$, $n_3 = 108$.

A statistical description for the study variables is presented in Table 2.

Zhou *and others* (2002) analyzed the validation sample of the CPP data in the framework of linear model. We consider a PLM using a nonparametric function to describe the relation between PCB and IQ as

$$\text{IQ} = g(\text{PCB}) + \beta_1 \text{EDU} + \beta_2 \text{SES} + \beta_3 \text{RACE} + \beta_4 \text{SEX} + e,$$

where e is a normal error with zero mean. To estimate the nonparametric function $g(.)$, we here adopt a 2-degree truncated power function basis with 10 fixed knots. Then the above model can be rewritten as $\text{IQ} = \pi^T(\text{PCB})\alpha + \beta_1 \text{EDU} + \beta_2 \text{SES} + \beta_3 \text{RACE} + \beta_4 \text{SEX} + e$, where $\alpha = (\alpha_1, \ldots, \alpha_{13})^T$. We first made the lack of fit test for the considered PLM using the SRS sample and found that the P value of the test is 0.29 which indicates that the PLM is suitable. Then we adopt our proposed method to fit this model. The estimates of the regression coefficient is given in Table 3, and the estimate of nonparametric function is presented in Figure 2.

Table 2. *Description of the variables in CPP data*

	MEAN	STD	25%	75%	MIN	MAX
Population (N = 38 709)						
IQ	96.06	14.29	86.00	106.00	56.00	153.00
EDU	10.67	2.45	9.00	12.00	0.00	18.00
SES	4.67	2.16	3.00	6.30	0.00	9.50
RACE	0.50	0.50	0.00	1.00	0.00	1.00
SEX	0.50	0.50	0.00	1.00	0.00	1.00
ODS ($n_V = 1038$)						
IQ	96.23	16.09	84.00	108.00	56.00	145.00
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
PCB	3.16	1.93	1.88	3.86	0.25	17.61

Note: MEAN = mean of the variable; STD = standard deviation of the variable; 25% = 25% percentile of the variable; 75% = 75% percentile of the variable; MIN = minimum value of the variable; MAX = maximum value of the variable.

Table 3. *Analysis results for CPP data*

	$\hat{\beta}_P$	PLM								Linear model	
		SE($\hat{\beta}_P$)	95% CI	$\hat{\beta}_{ODS}$	SE($\hat{\beta}_{ODS}$)	95% CI	$\hat{\beta}_{IPW}$	SE($\hat{\beta}_{IPW}$)	95% CI	$\hat{\beta}_{SRS}$	SE($\hat{\beta}_{SRS}$)
PCB	See Figure 3			See Figure 3			See Figure 3			0.27	0.20
EDU	1.45	0.16	(1.15, 1.76)	1.40	0.19	(1.02 1.77)	1.29	0.20	(0.91 1.67)	1.52	0.22
SES	1.45	0.20	(1.06, 1.85)	1.15	0.22	(0.71 1.57)	1.19	0.22	(0.75 1.62)	1.03	0.24
RACE	−8.02	0.63	(−9.26, −6.78)	−8.50	0.75	(−9.97 −7.04)	−8.39	0.77	(−9.89 −6.89)	−10.24	0.86
SEX	−0.51	0.69	(−1.87, 0.85)	−0.72	0.69	(−2.06 0.63)	−0.88	0.70	(−2.26 0.49)	−0.73	0.77

Note: The result of linear model is from the proposed MSELE estimator in Zhou et al. (2002); $\hat{\beta}_P$, $\hat{\beta}_{ODS}$, and $\hat{\beta}_{IPW}$ correspond to the estimates obtained by the Proposed, SEPLE-ODS, and IPW methods, respectively; SE($\hat{\beta}_P$), SE($\hat{\beta}_{ODS}$), and SE($\hat{\beta}_{IPW}$) are the estimated standard errors of the corresponding estimators.

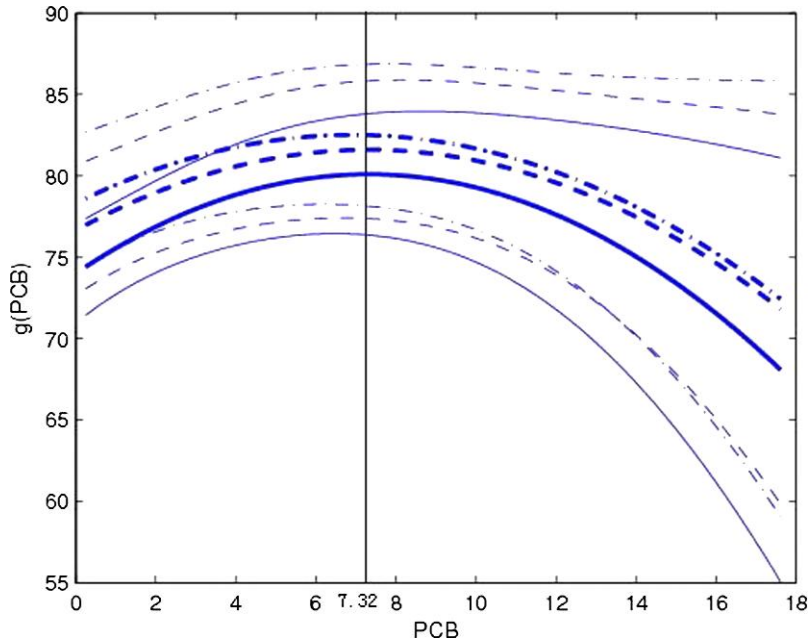

Fig. 2. The estimated function g on PCB for CPP data. Thicker solid, dashed, and dot-dashed curves correspond to estimates obtained by the Proposed, SEPLE-ODS, and IPW methods, respectively. And the solid dashed and dot-dashed curves correspond to estimated confidence bands obtained by the Proposed, SEPLE-ODS, and IPW methods, respectively.

We further conduct the proposed the Wald test to test whether the nonparametric function is a linear function or a quadratic function as follows:

Test 1: $H_0 : \alpha_3 = \alpha_4 = \cdots = \alpha_{13} = 0$ vs H_1: at least an $\alpha_i \neq 0$ for some $i \geqslant 3$.

The H_0 can be also written as $g(\text{PCB}) = \alpha_1 + \alpha_2\text{PCB}$. The Wald test statistic $W = 30.85 > \chi^2_{0.05}(11) = 19.68$, P-value $= 0.0012$. Therefore, the linear relation (H_0) is rejected at the 5% level of significance.

Test 2: $H_0 : \alpha_4 = \alpha_5 = \cdots = \alpha_{13} = 0$ vs H_1: at least one $\alpha_i \neq 0$ for some $i \geqslant 4$.

The H_0 can be also written as $g(\text{PCB}) = \alpha_1 + \alpha_2\text{PCB} + \alpha_3\text{PCB}^2$. The Wald test statistic $W = 21.83 > \chi^2_{0.05}(10) = 18.31$, P-value $= 0.0160$. Therefore, the quadratic relation (H_0) is rejected at the 5% level of significance, and we can conclude that a simple quadratic function is not enough to describe the relation between the IQ and the PCB.

The estimated curve in Figure 2 detects the nonlinear relation between the IQ score and the PCB level. At first, the IQ score increases with the increased PCB level but when the PCB level is higher than 7.32 (according to the nonparametric estimate by the proposed estimator), the IQ score begins to decrease. In addition, it can be found that the proposed estimator (P estimator) provides more precise confidence bands than those obtained by the SEPLE-ODS and IPW estimators.

While the results in Figure 3 may suggest a positive relation of IQ and PCB intake in lower PCB level. We caution not to overinterpret this result as this is likely due to that the effect of low level of PCBs may reflect beneficial aspects of lifestyle not captured by other covariates. For example, fish intake during pregnancy is related to both higher IQ in offspring (Daniels *and others*, 2004) and to higher levels

of serum PCBs (Halldorsson *and others*, 2008). Thus, finding the positive slope at the lower range of exposure alerts us to the possibility of confounding of the PCB coefficient.

From Table 3, the P, SEPLE-ODS, and IPW estimates are similar and the variance estimates for the proposed P estimators are the smallest. All the estimates confirm that the socioeconomic status and education of the parents have a positive impact on the IQ of the children, while there is no evidence that the gender has any effect on the IQ.

6. DISCUSSION

In this article, we considered a PLM in the setting of a 2-stage ODS design with a continuous outcome. Compared with Zhou *and others* (2002) in which only the ODS sample (the second-stage sample in this article) is used to make inference, the first-stage sample is incorporated into the proposed method for efficiency improvement. The inference of the PLM is different from the usual method because the ODS is a biased sampling scheme. We proposed an estimated penalized likelihood method to achieve the inference of PLM. Our simulation results demonstrate the efficiency improvement of the proposed method over some alternative methods can be used in these situations, such as the SEPLE-DOS method and the traditional PMLE-SRS method.

The IPW method is built on the completely observed data (those with (Y_i, X_i) observed, see formula (1.1)). For those observation with missing X, they are reflected in the IPW method through a weight that is proportional to the missing proportion. The real value of Y for those with missing X is not used, while our proposed method utilizes the actual observation of Y whatever its corresponding X is observed or not. Therefore, the proposed method performs better than the IPW method as more information of the data is used. Our method do rely on the specification of $f(Y|X)$ to be known. Choosing the error distribution as normal is a common practice in linear model analysis. Generally speaking, IPW method and empirical likelihood (EL) method do not require $f(Y|X)$ to be known but only need to specify the conditional expectation $E(Y|X)$. However, the IPW estimator compared in our paper relies on the specification of $f(Y|X)$ because it is really a weighted likelihood estimator and not the usual semiparametric IPW estimator. Likewise, the semiparametric EL estimator compared in our paper requires $f(Y|X)$ to be known because it is also based on the likelihood and the semiparametric EL inference procedure is used to deal with the unknown distribution of the covariates involved in the likelihood, which is treated as a nuisance parameter. The efficiency gain of the proposed method over SEPLE-ODS is due to the inclusion of individuals in the nonvalidation set (those with missing X) in the inference.

All the covariates were completely unobserved for the individuals of the incomplete first-stage sample through the article. However, in practice, some important covariates are easy and cheap to measure, so they can be observed for every individual of the population. How to incorporate the information into the inference is interesting. More details about this issue are referred to Weaver and Zhou (2005).

Some issues on the design of ODS such as the optimal size for the second-stage ODS sample, the optimal allocation of the cut points, and the optimal allocation of the second-stage ODS sample across different strata are deserved further investigation in the future study. Moreover, there are several interesting topics on the PLM for the ODS design that deserve further study in the future. For example, how to apply the doubly robust estimation method to our case and how to conduct the lack of fit test for the PLM in the ODS design.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

We thank the editor and the associate editor for their constructive suggestions that largely improved the presentation of this paper. *Conflict of Interest:* None declared.

FUNDING

National Institute of Health (CA79949 to H.Z., G.Q.); National Natural Science Foundation of China (10801039 to G.Q.).

REFERENCES

BRESLOW, N., MCNENEY, B. AND WELLNER, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *The Annals of Statistics* **31**, 1110–1139.

BRESLOW, N. E. AND CAIN, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika* **75**, 11–20.

CARROLL, R. J. AND WAND, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the American Statistical Association* **98**, 158–168.

CHATTERJEE, N., CHEN, Y. H. AND BRESLOW, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of Royal Statistics Society, Series B* **53**, 573–585.

DANIELS, J. L., LONGNECKER, M. P., ROWLAND, A. S., GOLDING, J. AND ALSPAC STUDY TEAM. UNIVERSITY OF BRISTOL INSTITUTE OF CHILD HEALTH. (2004). Fish intake during pregnancy and early cognitive development of offspring. *Epidemiology*, **15**, 394–402.

GRAY, K. A., KLEBANOFF, M. A., BROCK, J. W., ZHOU, H., DARDEN, R., NEEDHAM, L. AND LONGNECKER, M. P. (2005). In utreo exposure to background levels of polychlorinated biphenyls and cognitive functioning among school-age children. *American Journal of Epidemiology*, **162**, 17–26.

HALLDORSSON, T. I., THORSDOTTIR, I., MELTZER, H. M., NIELSEN, F., OLSEN, S. F. (2008). Linking exposure to polychlorinated biphenyls with fatty fish consumption and reduced fetal growth among Danish pregnant women: a cause for concern? *American Journal of Epidemiology*, **168**, 958–965.

HE, X., ZHU, Z. Y. AND FUNG, W. K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* **89**, 579–590.

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.

HUANG, J. Z., ZHANG, L. AND ZHOU, L. (2007). Efficient estimation in marginal partially linear models for longitudinal/clustered data using spline. *Scandinavian Journal of Statistics* **34**, 451–477.

LANGHOLZ, B. AND BORGAN, Ø. (1995). Counter-matching: a stratified nested case-control sampling method. *Biometrika* **82**, 69–79.

LAWLESS, J. F., KALBFLEISCH, J. D. AND WILD, C. J. (1999). Semiparametric methods for response-selective and missing-data problems in regression. *Journal of Royal Statistics Society, Series B* **61**, 413–438.

LIN, X. AND CARROLL, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association* **96**, 1045–1056.

NISWANDER, K. R. AND GORDON, M. (1972). *The Women and Their Pregnancies. U.S. Department of Health, Education, and Welfare Publication (NIH)*. Washington, DC: Government Printing Office, pp. 73–379.

PEPE, M. S. AND FLEMING, T. R. (1991). A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association* **86**, 108–113.

REILLY, M. AND PEPE, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82**, 299–314.

REILLY, M. AND PEPE, M. S. (1997). The relationship between hot-deck multiple imputation and weighted likelihood. *Statistics in Medicine* **16**, 5–19.

RUPPERT, D., WAND, M. P. AND CARROLL, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

WEAVER, M. A. (2001). Semiparametric Methods for Continuous Outcome Regression Models with Covariate Data from an Outcome-Dependent Subsample, [Doctoral Dissertation]. Chapel Hill, NC: University of North Carolina.

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

WHITE, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology* **115**, 119–128.

YU, Y. AND RUPPERT, D. (2002). Penalized spline estimation for partially linear single index models. *Journal of the American Statistical Association* **97**, 1042–1054.

ZEGER, S. L. AND DIGGLE, P. J. (1994). Semiparametric model for longitudinal data with application to CD4 cell numbers in HIV seroconvertiers. *Biometrics* **50**, 689–699.

ZHAO, L. P. AND LIPSITZ, S. (1992). Designs and analysis of two-stage studies. *Statistics in Medicine* **11**, 769–782.

ZHOU, H. AND PEPE, M. S. (1995). Auxiliary covariate data in failure time regression. *Biometrika* **82**, 139–149.

ZHOU, H. AND WANG, C. Y. (2000). Failure time regression analysis with measurement error in covariates. *Journal of Royal Statistics Society, Series B* **62**, 657–665.

ZHOU, H., 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.

ZHU, Z. Y., FUNG, W. K. AND HE, X. (2008). On the asymptotics of marginal regression splines with longitudinal data. *Biometrika* **95**, 907–917.

[*Received July 26, 2010; revised October 25, 2010; accepted for publication October 26, 2010*]