Published in final edited form as:
J R Stat Soc Series B Stat Methodol. 2014 January 1; 76(1): 197–215. doi:10.1111/rssb.12029.

Semiparametric Inference for Data with a Continuous Outcome from a Two-Phase Probability Dependent Sampling Scheme

Haibo Zhou[1], **Wangli Xu**[1,2], **Donglin Zeng**[1], and **Jianwen Cai**[1]

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

[2]Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, 100872, China

Abstract

Multi-phased designs and biased sampling designs are two of the well recognized approaches to enhance study efficiency. In this paper, we propose a new and cost-effective sampling design, the two-phase probability dependent sampling design (PDS), for studies with a continuous outcome. This design will enable investigators to make efficient use of resources by targeting more informative subjects for sampling. We develop a new semiparametric empirical likelihood inference method to take advantage of data obtained through a PDS design. Simulation study results indicate that the proposed sampling scheme, coupled with the proposed estimator, is more efficient and more powerful than the existing outcome dependent sampling design and the simple random sampling design with the same sample size. We illustrate the proposed method with a real data set from an environmental epidemiologic study.

Keywords

Empirical likelihood; Missing data; Semiparametric; Probability sample

1 Introduction

Observational studies in epidemiology that relate disease outcome to individual exposures and other characteristics play a key role in understanding the determinants of diseases in humans. As all studies are conducted with a limited budget, the maximum study sizes are often restricted by the cost of the exposure assessments. Some large cohort studies, e.g., the Women's Health Initiative and the National Children's Study, could cost hundreds of millions of dollars to conduct. Cost-effective study designs for biomedical studies have always been an important research area. Among them, the biased sampling design, represented by the case-control design, has played a significant role in the development of biostatistics methodological research during the last half of the 20th century. It is often the preferred choice of study design for epidemiologic studies because of its efficiency and cost-effectiveness feature compared to cohort studies (e.g., Cornfield, 1951; Anderson, 1972; Prentice and Pyke, 1979).

The fundamental idea of case-control design is to over-sample observations (e.g., cases) that are believed to be more informative regarding the exposure-response relationship. This basic idea motivated the development of research in the area of general Outcome Dependent Sampling (ODS) for a continuous outcome in recent years (e.g., Zhou et al., 2002; Weaver and Zhou, 2005; Song, Zhou and Kosorok, 2009). The general ODS design allows investigators to selectively sample observations based on the observed values of a continuous outcome to achieve improved efficiency for a fixed sample size. The ODS design (Zhou et al., 2002) assumes that the values of the response, denoted by Y, are known for all subjects, but the exposure variable, denoted by X, may be expensive or difficult to assess. This is reasonable in many studies where responses like Intelligence Quotient (IQ) or disease status are easily obtainable, but exposure assessment needs expensive assay or follow up. Assume that the domain of the Y is partitioned into three mutually exclusive intervals: $(-\infty, y_L]\bigcup(y_L, y_U]\bigcup(y_U, \infty)$. The ODS sample proposed by Zhou et al. (2002) has X values ascertained on the following three samples: an overall simple random sample, a supplemental sample conditional on $Y < y_L$, and a supplemental sample conditional on $Y > y_U$. Other recent progresses in the ODS design includes (e.g., Kang and Cai, 2009; Lu and Tsiatis, 2006; Zhou et al. 2011; Qin and Zhou, 2011; Chatterjee, Chen and Breslow, 2003; Manatunga et al, 2008; Schildcrout and Rathouz, 2010; Wang and Zhou, 2006; Zhou, Song, et al, 2011; Zhou, Wu, et al, 2011). Part of the explanation that the ODS design is more efficient than the simple random sampling is because through sampling the response Y at its two distributional tails, the observed exposure values X were also more likely to occur at its distributional tails. Linear model theory shows that the variance of $\widehat{\beta}$, the estimate of the regression coefficient corresponding to X, is inversely proportional to the summed squares of observed X's values. Hence, when the goal is to evaluate the relationship between an exposure X and a response Y, having a sample of subjects whose X values are at its two distributional tails would be more informative than having a sample of subjects whose X values concentrated around its mean.

Assume that the domain of the exposure X is partitioned into three mutually exclusive intervals: $(-\infty, x_L]\bigcup(x_L, x_U]\bigcup(x_U, \infty)$. If an investigator knows which interval each individual's X value falls into, the investigator can draw a supplemental sample from those whose X values are in the upper or lower tail intervals, respectively. Such a strategy, however, is not feasible in practice as investigators do not have knowledge of X in advance. In this paper, we propose a new two-phase design where we select the second phase supplemental sample with a probability-dependent-sampling scheme (PDS) that will allow us to oversample X from its two distributional tails. The proposed two-phase PDS is outlined as follows. Let Y denote the response variable, X the primary exposure variable, and Z the collection of all other covariates. In the first phase of the proposed design, a simple random sample is drawn and the values of (X, Y, Z) are observed. We fit a model for $E(X|Y, Z)$ using the phase one SRS sample. Based on this model, the chances of a new subject's X, conditional on $Y = y$, $Z = z$, will be in $(-\infty, X_L]$ and (X_U, ∞) are predicted by $\widehat{\phi}_1(y,z)=\widehat{Pr}(X<x_L|Y,Z)$ and $\widehat{\phi}_3(y,z)=\widehat{Pr}(X>x_U|Y,Z)$, respectively. We then draw the supplemental samples in the second phase by obtaining a simple random sample from those who are likely to have high or low X values. For example, random samples can be drawn from those with $\widehat{\phi}_1(y,z)=>80\%$ and with $\widehat{\phi}_3(y,z)=>80\%$, respectively. As a result, the final observed data is over-represented by individuals who are more likely to be on the distributional tails of X.

The roots of the proposed two-phase PDS design can also be traced back to Neyman (1938), who introduced the two-phase stratified design to enhance study efficiency. At the first phase of a typical two-phase design, a relatively large random sample is drawn and only Y and Z are measured in the first phase cohort. The ascertainment of X is made at the second phase of the design, where a subsample is drawn randomly, without replacement, from the first phase cohort. Greater efficiency can be obtained through the two-phase sampling design (e.g. Breslow and Cain, 1988; Breslow et al., 2003; Song et al, 2009; and Wang and Zhou, 2010).

The key differences among the traditional two-phase design, the recent work on the two-phase ODS design, and the proposed two-phase PDS design are that: (i) the second phase of the traditional two-phase design is either independent of *Y* and *Z* or is only dependent on binary *Y*, e.g., case-control second phase; (ii) the two-phase ODS allows for continuous *Y* but not *Z* in the 2nd phase drawing; (iii) the two-phase PDS, not only allows for a continuous *Y*, but also allows for any dimension of *Z* in the decision making of 2nd phase drawing. By estimating the chance of the unknown *X*'s range, this approach avoided the impracticability of high dimension stratification of vector *Z*.

For data obtained via complex sampling designs like the PDS designs described above, estimators ignoring the design will be biased unless they properly account for the biased sampling scheme. In practice, some ad hoc or simplification of the data is often made prior to analysis. A commonly used approach in epidemiologic studies is to dichotomize a continuous outcome *Y* and then use available methods for binary outcome for inference (e.g., White, 1982; Amemiya, 1985; Prentice, 1986; Breslow and Cain, 1988; Weinberg and Wacholder, 1993; Langholz and Borgan, 1995; Breslow and Holubkov, 1997; Schildcrout and Heagerty, 2008). In this paper, we propose a semiparametric empirical likelihood method for estimating the regression parameters. The proposed methods are semiparametric in the sense that the marginal distribution of the exposure variable *X* is left unspecified.

The remainder of this paper is organized as follows. In Section 2, we introduce the data structure for the two-phase PDS design. We outline the estimation algorithm for the proposed semiparametric empirical likelihood estimator and establish its asymptotic properties. In Section 3, we present simulation study results comparing the proposed method with some competing designs and estimators. We illustrate the proposed method with a data set from the Collaborative Perinatal Project (CPP) data. Final remarks are given in Section 4.

2 Design and Inference for a Two-phase PDS Study

2.1 Design and Data Structure

Let Y denote a continuous outcome variable, (X, Z) denote the vector of covariates with X being the expensive scalar exposure variable and Z being the easily obtainable covariates. Assume that the regression model of Y given (X, Z) is

$$Y=\beta_0+\beta_1 X+\beta_2 Z+\varepsilon,$$

where $(\beta_0, \beta_1, \beta_2)$ denote the unknown regression parameters and $\varepsilon \sim N(0, \sigma^2)$ is the random error. Let $\beta = (\beta_0, \beta_1, \beta_2, \sigma^2{}_1)$ and x_L and x_U $(x_L < x_U)$ be known constants that partition the domain of X into three mutually exclusive intervals: $A_1 \bigcup A_2 \bigcup A_3 = (-\infty, x_L] \bigcup (x_L, x_U] \bigcup (x_U, \infty)$.

The proposed two-phase PDS scheme is as follows: in the first phase, we observe (Y, X, Z) in a simple random sample (SRS) of size n_0 from the underlying study population. A model of $E(X|Y, Z)$ is then fitted based on this SRS sample and $\varphi_1(Y, Z) = Pr(X \in A_1|Y, Z)$ and $\varphi_3(Y, Z) = Pr(X \in A_3|Y, Z)$ are estimated. In the second phase of the PDS design, we draw a supplemental random sample from those in the study population whose predicted probability $\widehat{\phi}_1=\widehat{Pr}\left(X \in A_1|Y,Z\right)$ satisfies $\widehat{\phi}_1 \geq 80\%$. Likewise, a supplemental sample is drawn from those whose X values are more likely in the upper tail, i.e, from those with $\widehat{\phi}_3=\widehat{Pr}\left(X \in A_3|Y,Z\right) \geq 80\%$. Note that the 80% value here is chosen for the simplicity of illustration. We will use constants c_1 and c_3, where $0 < c_1, c_3 < 1$, in the formulation of the likelihood. The data structure for the proposed two-phase PDS is

$$\begin{array}{rl}\text{The SRS sample:} & \left\{Y_{0i},X_{0i},Z_{0i}\right\},\ i{=}1,\ldots,n_0;\\ \text{The supplemental sample:} & \left\{(Y_{1i},X_{1i},Z_{1i}):Pr\left(X_{1i}\in A_1|Y_{1i},Z_{1i}\right)\ge c_1\right\},\ i{=}1,\ldots,n_1;\\ & \left\{(Y_{3i},X_{3i},Z_{3i}):Pr\left(X_{3i}\in A_3|Y_{3i},Z_{3i}\right)\ge c_3\right\},\ i{=}1,\ldots,n_3.\end{array}\quad (2.1)$$

.

The supplemental samples can be generated with different, perhaps unknown, selection probabilities, e.g., one can choose to select a fixed proportion of the sets {(Y_{ki}, Z_{ki}): φ_k(Y_{ki}, Z_{ki}) ≥ c_k} (k = 1, 3) from a underlying cohort of subjects whose (Y, Z) are known, or, one can select a predetermined number of subjects from the underlying population, in which case, the proportion of the selected set relative to the underlying population is unknown. The total sample size in the two-phase PDS design is $n = n_0 + n_1 + n_3$.

If X is a continuous variable and can be viewed as normally distributed after proper transformation, then a linear model can be used for $\varphi_k(Y, Z) = Pr(X \in A_k|Y, Z)$, k = 1, 3. More specially, we estimate $\varphi_1(Y, Z)$ by $\widehat{P}r\left(X\in A_1|Y,Z\right)=\Phi\left(\left(x_L-\left(\widehat{\gamma}_0+\widehat{\gamma}_1Y+\widehat{\gamma}_2Z\right)\right)/\widehat{\sigma}_1\right)$ and $\varphi_3(Y, Z)$ by $\widehat{P}r\left(X\in A_3|Y,Z\right)=1-\Phi\left(\left(x_U-\left(\widehat{\gamma}_0+\widehat{\gamma}_1Y+\widehat{\gamma}_2Z\right)\right)/\widehat{\sigma}_1\right)$, where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution and $\widehat{\gamma}_i, i{=}0,1,2$ and $\widehat{\sigma}_1$ are estimates using the first phase data based on the following regression model:

$$X{=}\gamma_0{+}\gamma_1Y{+}\gamma_2Z{+}e,\tilde{e}N\left(0,\sigma_1^2\right).\quad (2.2)$$

Another natural estimator for φ_k results from the use of logistic regression model. Denote $\delta_k = I(X \in A_k)$, k = 1, 3. We estimate $\varphi(Y, Z)$ by $\widehat{\phi}_k\left(Y,Z\right)=\left(1{+}exp\left(-\left(\widehat{\alpha}_{0k}{+}\widehat{\alpha}_{1k}Y{+}\widehat{\alpha}_{2k}Z\right)\right)\right)^{-1}$, where ($\widehat{\alpha}_{0k},\widehat{\alpha}_{1k},\widehat{\alpha}_{2k}$) are obtained from fitting

$$Pr\left(\delta_k{=}1|Y,Z\right)=\left(1{+}exp\left(-\left(\alpha_{0k}{+}\alpha_{1k}Y{+}\alpha_{2k}Z\right)\right)\right)^{-1}\quad (2.3)$$

to the first phase SRS data. Alternatively, one can also derive a nonparametric estimator for φ_k, k = 1, 3, by using the kernel method. Note that

$$Pr\left(X\in A_k|Y,Z\right)=\frac{\int I_{x\in A_k}f\left(Y|X,Z\right)dG\left(X|Z\right)}{\int f\left(Y|X,Z\right)dG\left(X|Z\right)},\quad (2.4)$$

where $G(X|Z)$ is the conditional c.d.f. for $X|Z$ that can be estimated by

$$\widehat{G}\left(X|Z\right)=\frac{\sum_{j=1}^{n_0}I\left(X_{0j}\le X\right)\phi_h\left(Z_{0j}-Z\right)}{\sum_{j=1}^{n_0}\phi_h\left(Z_{0j}-Z\right)},$$

where $\varphi_h(\cdot) = \varphi(\cdot/h)$ is a kernel function with a bandwidth h. One can then estimate $\varphi_k(Y, Z)$ by

$$\widehat{\phi}_{Ek}(Y,Z)=\frac{\sum_{j=1}^{n_0} I_{X_{0j}\in A_k} f(Y|X_{0j},Z)\,\phi_h(Z_{0j}-Z)}{\sum_{j=1}^{n_0} f(Y|X_{0j},Z)\,\phi_h(Z_{0j}-Z)}. \tag{2.5}$$

2.2 A Semiparametric Empirical Likelihood Inference

Let $G(X, Z)$ and $g(X, Z)$ denote the joint c.d.f. and p.d.f. of (X, Z), respectively. If $\varphi_k(Y, Z)$, $k = 1, 3$ were known, the likelihood function for the data in (2.1) would be

$$L(\beta,G)=\left\{\prod_{i=1}^{n_0} f_\beta(Y_{0i}|X_{0i},Z_{0i})\,g(X_{0i},Z_{0i})\right\}\left\{\prod_{k=1,3}\prod_{j=1}^{n_k} f_\beta(Y_{kj},X_{kj},Z_{kj}|\phi_k(Y_{kj},Z_{kj})\geq c_k)\right\}. \tag{2.6}$$

Due to the biased sampling of the proposed design, maximizing the likelihood function over β involves addressing $G(X, Z)$. Hence we include G in the above likelihood function. For $k = 1, 3$, define

$$\pi_k=Pr(\phi_k(Y,Z)\geq c_k)=\int\int\int f_\beta(Y|X,Z)\,g(X,Z)\,I_{\{(Y,Z):\phi_k(Y,Z)\geq c_k\}}dY\,dX\,dZ$$

Using the Bayes formula, $L(\beta, G)$ can be expressed as

$$L(\beta,G)=\left\{\prod_{i=1}^{n_0} f_\beta(Y_{0i}|X_{0i},Z_{0i})\,g(X_{0i},Z_{0i})\right\}\left\{\prod_{k=1,3}\prod_{j=1}^{n_k} f_\beta(Y_{kj}|X_{kj},Z_{kj})\,g(X_{kj},Z_{kj})\right\}\left\{\prod_{k=1,3}\pi_k^{-n_k}\right\}. \tag{2.7}$$

We propose a semiparametric likelihood method to maximize the likelihood function without specifying the underlying distribution of $G(X, Z)$. We first profile the likelihood function $L(\beta, G)$ by fixing β and obtaining the empirical likelihood function of $G(X, Z)$ over all distributions whose support contains the observed (X, Z) values. For a fixed β, this is a biased sampling likelihood (Vardi 1982, 1985; Qin 1993). We then maximize the resulting profile likelihood function with respect to β. For simplicity of notation, let $(X_1, \ldots, X_n) = (X_{01}, \ldots, X_{0n_0}, X_{11}, \ldots, X_{1n_1}, X_{31}, \ldots, X_{3n_3})$, $(Z_1, \ldots, Z_n) = (Z_{01}, \ldots, Z_{0n_0}, Z_{11}, \ldots, Z_{1n_1}, Z_{31}, \ldots, Z_{3n_3})$ and $(Y_1, \ldots, Y_n) = (Y_{01}, \ldots, Y_{0n_0}, Y_{11}, \ldots, Y_{1n_1}, Y_{31}, \ldots, Y_{3n_3})$. Then the log-likelihood function can be written as

$$\begin{aligned} l_f(\beta,\{p_i\},\{\pi_k\}) &= \sum_{i=1}^{n} log\, f_\beta(Y_i|X_i,Z_i)+\left\{\sum_{i=1}^{n} log\, p_i-\sum_{k=1,3} n_k\, log(\pi_k)\right\} \\ &=: l_1(\beta)+l_2(\{p_i\},\{\pi_k\}), \end{aligned} \tag{2.8}$$

where $p_i=g(X_i,Z_i)$, $l_1(\beta)=\sum_{i=1}^{n} log f_\beta(Y_i|X_i,Z_i)$ is a function only involving β, and $l_2(\{p_i\},\{\pi_k\})=\sum_{i=1}^{n} log\, p_i-\sum_{k=1,3} n_k log(\pi_k)$.

The first step in deriving the proposed estimator for β is to profile (2.8) over $\{p_i\}$, by fixing (β, π_1, π_3), and obtain the empirical likelihood function of $\{p_i\}$ over all distributions whose support contains the observed values of X and Z. To this end, we need only consider discrete distributions with jumps at each of the observed points (Owen, 1988, 1990). That is, for fixed (β, π_1, π_3), we search for $\{\widehat{p}_i\}$ that mamximize $l_2(\{p_i\}, \{\pi_k\})$ in (2.8) under the following four constraints:

$$\left\{p_i \ge 0; \sum_{i=1}^{n} p_i = 1 \sum_{i=1}^{n} p_i \left(\int f_\beta \left(Y|X_i, Z_i\right) I_{\{(Y,Z_i):\phi_1(Y,Z_i)\ge c_1\}} dY - \pi_1\right) = 0; \right.$$
$$\left. \sum_{i=1}^{n} p_i \left(\int f_\beta \left(Y|X_i, Z_i\right) I_{\{(Y,Z_i):\phi_3(Y,Z_i)\ge c_3\}} dY - \pi_3\right) = 0. \right\} \qquad (2.9)$$

These constraints reflect the properties of $g(X, Z)$ being a discrete distribution function with support points at the observed (X, Z) values, i.e., $\{p_i\}$ are nonnegative probabilities that sum up to unity.

For a fixed β, using a similar idea to Qin and Lawless (1994), a unique maximum for $\{p_i\}$ in $l_2(\{p_i\}, \{\pi_k\})$ with constraints (2.9) exists if 0 is inside the convex hull of points $\int f_\beta \left(Y|X_i, Z_i\right) I_{\{(Y,Z_i):\phi_k(Y,Z_i)\ge c_k\}} dY - \pi_k$ for $i = 1, \ldots, n$ and $k = 1, 3$. The Lagrange multiplier argument can be invoked to derive the maximum over $\{p_i\}$. Specifically, write

$$H\left(\beta, \{p_i\}, \{\pi_k\}\right) = l_2\left(\{p_i\}, \{\pi_k\}\right) + \rho\left(1 - \sum_{i=1}^{n} p_i\right) + n \sum_{k=1,3} \lambda_k \sum_{i=1}^{n} p_i \left\{\int f_\beta \left(Y|X_i, Z_i\right) I_{\{(Y,Z_i):\phi_k(Y,Z_i)\ge c_k\}} dY - \pi_k\right\},$$

where ρ and λ_k are Lagrange multipliers. Taking derivatives of $H(\beta, \{p_i\}, \{\pi_k\})$ with respect to $\{p_i\}$ and solving the score equations together with the constraints in (2.9), we can obtain that $\rho = n$ and

$$\widehat{p}_i = n^{-1}\left\{1 + \sum_{k=1,3} \lambda_k \left(\int f\left(Y|X_i, Z_i\right) I_{\{(Y,Z_i):\phi_k(Y,Z_i)\ge c_k\}} dY - \pi_k\right)\right\}^{-1}.$$

Replacing p_i with $\widehat{p}_i$ in (2.8), we have a profile log-likelihood function $l_f\left(\beta, \{\widehat{p}_i\}, \{\pi_k\}\right)$ that is a function of $(\beta, \pi_1, \pi_3, \lambda_1, \lambda_3)$ only. Typically, the true value of the Lagrange multipliers are zero in unbiased sampling problem. However, due to the biased nature of the PDS sampling design, λ_1 and λ_3 are not centered around zero. To unify the notation, we center them by reparameterizing $v_k = \lambda_k - n_k/(n\pi_k)$, $k = 1, 3$. We define $\xi = (\beta, \pi_1, \pi_3, v_1, v_3)$. The resulting profile log-likelihood function l(ξ) can be expressed as

$$l(\xi) = l_1(\beta) - \sum_{i=1}^{n} log\left(1 + v^\tau h\left(X_{i,Z_i}\right)\right) - \sum_{i=1}^{n} log\left(\Delta\left(X_i, Z_i\right)\right) - \sum_{k=1,3} n_k \, log \, \pi_k$$

where $h(X_i, Z_i) = (h_1(X_i, Z_i), h_3(X_i, Z_i))^\tau$ with $h_k(X_i, Z_i) = F_k(X_i, Z_i) - \pi_k / \Delta(X_i, Z_i)$, $F_k(X_i, Z_i) = \int f_\beta(Y|X_i, Z_i) I_{\{(Y,Z_i):\varphi_k(Y,Z_i)\ge c_k\}} dY$, and $\Delta(X_i, Z_i) = q_0 + \Sigma_{k=1,3}\, q_k \pi_k^{-1} F_k(X_i, Z_i)$ with $q_k = n_k/n$ for $k = 0, 1, 3$, respectively.

Finally, replacing $\varphi_k(Y, Z)$ by $\widehat{\phi}_k\left(Y, Z\right)$ in $l(\xi)$, we have the following estimated profile log-likelihood function:

$$\tilde{l}(\xi) = l_1(\beta) - \sum_{i=1}^{n} log\left(1 + v^\tau \widehat{h}\left(X_{i,Z_i}\right)\right) - \sum_{i=1}^{n} log\left(\widehat{\Delta}\left(X_i, Z_i\right)\right) - \sum_{k=1,3} n_k \, log \, \pi_k, \qquad (2.10)$$

where $\widehat{h}\left(X,Z\right)$ and $\widehat{\Delta}\left(X,Z\right)$ are obtained by replacing $\varphi_k(Y, Z)$ by $\widehat{\phi}_k\left(Y,Z\right)$ in $h(X, Z)$ and $\Delta(X, Z)$, respectively. We call $\widehat{\xi}$ the maximum semiparametric empirical likelihood estimator (MSELE) where $\widehat{\xi}$ is the maximizer for $\tilde{l}\left(\xi\right)$. The MSELE for β is $\widehat{\beta}$ is the corresponding portion of $\widehat{\xi}$. The Newton-Raphson iterative procedure can be used to obtain $\widehat{\xi}$. The following theorem summarizes the asymptotic properties for the proposed estimators.

THEOREM 1 (asymptotic properties): Under the regularity conditions outlined in the Appendix, $\widehat{\xi}$ converges in probability to the true value $\xi = (\beta, \pi_1, \pi_3, 0, 0)$, and $n^{1/2}\left(\widehat{\xi}-\xi\right)$ converges in distribution to $N(0, \Sigma)$, where $\Sigma = V^{-1}(\xi)U(\xi)\{V^{-1}(\xi)\}^T$ is given in the Appendix.

Details of the proof are given in the Appendix. It will be shown that the asymptotic variance-covariance of $\sqrt{n}\left(\widehat{\xi}-\xi\right)$ takes a sandwich form $V^{-1}(\xi)U(\xi)\{V^{-1}(\xi)\}^T$. In addition, a consistent estimator of the variance-covariance matrix is given by $\widehat{V}^{-1}\left(\widehat{\xi}\right)\widehat{U}\left(\widehat{\xi}\right)\left\{\widehat{V}^{-1}\left(\widehat{\xi}\right)\right\}^T$, where $\widehat{U}$ and $\widehat{V}$ are obtained by replacing the large-sample quantities in U and V with their corresponding small-sample quantities.

Remark 1 The proposed estimation algorithm enables us to change an infinite dimension problem, with regard to nonparametric G, into a finite dimension problem at the expense of introducing 4 parameters π_1, π_3, λ_1, λ_3.

Remark 2 When $\widehat{\phi}_k\left(Y,Z_i\right)$ is from the logistic regression model, $\widehat{\phi}_k\left(Y,Z_i\right)\geq c_k$ is equal to

$$\begin{cases} y\geq-\frac{\widehat{\alpha}_{0k}\widehat{\alpha}_{2k}Z_i+log\frac{1-c_k}{c_k}}{\widehat{\alpha}_{1k}}, & \text{if }\widehat{\alpha}_{1k}>0;\\ y\leq-\frac{\widehat{\alpha}_{0k}\widehat{\alpha}_{2k}Z_i+log\frac{1-c_k}{c_k}}{\widehat{\alpha}_{1k}}, & \text{if }\widehat{\alpha}_{1k}<0;\end{cases}$$

and $F_k(X_i, Z_i)$ can be simply expressed as

$$F\left(-\left(\widehat{\alpha}_{0k}+\widehat{\alpha}_{2k}Z_i+log\frac{1-c_k}{c_k}\right)/\widehat{\alpha}_{1k}|X_i,Z_i\right)I_{\{\widehat{\alpha}_{1k}<0\}}+\bar{F}\left(-\left(\widehat{\alpha}_{0k}+\widehat{\alpha}_{2k}Z_i+log\frac{1-c_k}{c_k}\right)/\widehat{\alpha}_{1k}|X_i,Z_i\right)I_{\{\widehat{\alpha}_{1k}>0}$$

where $F(u|X_i, Z_i) = \Pr(Y \leq u|X_i, Z_i)$ and $\bar{F}=1-F$.

3 Numerical Analysis

3.1 Simulation Studies

We evaluate the small sample behavior of the proposed estimator using Monte Carlo studies. We assume that the domains of both Y and X are partitioned into three mutually exclusive intervals: $\gamma = B_1 \bigcup B_2 \bigcup B_3$ and $\chi = A_1 \bigcup A_2 \bigcup A_3$, where $B_1 = (-\infty, \mu_Y - a * \sigma_Y]$, $B_2 = (\mu_Y - a * \sigma_Y, \mu_Y + a * \sigma_Y]$, $B_3 = (\mu_Y + a * \sigma_Y, \infty)$, $A_1 = (-\infty, \mu_X - a * \sigma_X]$, $A_2 = (\mu_X - a*\sigma_X, \mu_X + a*\sigma_X]$ and $A_3 = (\mu_X + a*\sigma_X, \infty)$. We assume $n_1 = n_3$, $a = 1, 1.5$, and $c_1 = c_3 = 85\%, 95\%$.

The proposed estimator, denoted by $\widehat{\beta}_{PDS_1}$ for c=95% and $\widehat{\beta}_{PDS_2}$ for c=85%, is compared with five other estimators: (i) The first estimator, denoted by $\widehat{\beta}_X$, is an estimator based on a hypothetical situation where one assumes all X values are available in the study. The supplemental samples are drawn from individuals whose X values are in the two tails of X, defined by $\mu_X \pm a * \sigma_X$. We emphasize that this estimator *is not* available in practice since X is unknown, we include it for comparison purpose only. We use the least square method for estimation in this case. (ii) The second estimator, denoted by $\hat{\beta}_{ODS}$, is the ODS estimator (Zhou *et al*, 2002). The supplemental samples are drawn from individuals whose Y values are in the two tails of the distribution of Y, defined by $\mu_Y \pm a * \sigma_Y$; (iii) The third method, denoted by $\hat{\beta}_{IPW}$, is the inverse probability weighted (IPW) method (Horvitz and Thompson, 1952). The data structure for this estimator is the same as that for estimator $\hat{\beta}_{ODS}$ and we use the weights given by Weaver and Zhou (2005); (iv) The fourth case is the ordinary linear regression estimator, denoted by $\hat{\beta}_{SRS}$, from a simple random sample with the same sample size as the total sample size in the *PDS* design. (V) $\hat{\beta}_N$ is the estimator ignoring the sampling structure and treats the data as if an independent sample. All methods compared are under the same sample size scenarios. The IPW also assumes a known sampling fraction. We first generate a large underlying study cohort (4000) and then subsample from it to compare different designs and methods.

We generate data from the following regression model:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon, \quad (3.1)$$

where $Z = I_{(\log(|X|)+e)>1}$ describes a dependent but weak relationship between X and Z with $\in$, e and X generated independently from $N(0, 1)$. Tables 1 and 2 summarize the simulation results. Results are based on 1000 independent simulation runs.

We note the following observations from Table 1: (i) except $\hat{\beta}_N$, all estimators for (β_1, β_2) are unbiased. Clearly, $\hat{\beta}_N$ shows that ignoring the sampling scheme will lead to biased estimate for $\beta_1 \neq 0$; (ii) The average of the proposed variance estimator is very close to the empirical variance based on the 1000 simulations; (iii) The nominal 95% confidence interval coverage rates are close to 95%, indicating that the large sample normal approximation works well in these situations. As β_1 is of primary interest, we will concentrate on the efficiency comparison of various estimators for β_1 and note the following observations: (iv) When $\beta_1 \neq 0$, the proposed estimator $\hat{\beta}_{PDS_1}$ is the most efficient among all practically available estimators; (v) When $\beta_1 \neq 0$, as a changes from 1 to 1.5, i.e., when we move the partition of X further towards the tails, $\hat{\beta}_X$, $\hat{\beta}_{PDS_1}$, $\hat{\beta}_{PDS_2}$ and $\hat{\beta}_{ODS}$ all become more efficient, while $\hat{\beta}_{IPW}$ becomes less efficient but $\hat{\beta}_{SRS}$ is not affected; (vi) For a fixed overall sample size $n = n_0 + n_1 + n_3$, as we allocate more samples to the tails, e.g., when (n_0, n_1, n_3) changes from (300, 50, 50) to (200, 100, 100), the efficiency of $\hat{\beta}_{PDS1}$, $\hat{\beta}_{PDS2}$, $\hat{\beta}_{ODS}$ improves while the efficiency of $\hat{\beta}_{IPW}$ decreases; (vii) As overall sample sizes increase from 200 to 400, all estimators' efficiency improved. (viii) in general, $\hat{\beta}_{PDS1}$, which corresponds to $c = 0.95$, is more efficient than $\hat{\beta}_{PDS2}$, which corresponds to $c = 0.85$.

Table 2 lists the power for testing $\beta_1 = 0$, and relative efficiency (RE) for $a = 1.0$, $\sigma^2 = 4$, and (n_0, n_1, n_3) = (100, 50, 50) and (150, 25, 25). RE is defined as the ratio of the standard error for the estimator of interest to that of $\hat{\beta}_X$. At $\beta_1 = 0$, we see that all estimators, except $\hat{\beta}_{IPW}$, have type I error rates close to the nominal level. $\hat{\beta}_{IPW}$ has slightly inflated type I error rate (0.07). As β_1 increases, the proposed estimator $\hat{\beta}_{PDS}$ has almost the same power as $\hat{\beta}_X$ and is more powerful than the other competing estimators. $\hat{\beta}_{SRS}$, the estimator from a simple random sampling, has the least power among all. The observation regarding the relative efficiency is similar to that from the power.

We further conducted additional simulation studies to check on the robustness of the estimation. We considered four different combinations for the covariates *X* and *Z* in model (3.1): (i) *X* is a standard normal distribution, while *Z* is a binary variable with parameter p = 0.45; (ii) both *X* and *Z* are standard normal distributions; (iii) *X* is a exponential distribution with parameter being 1, while *Z* is a binary variable with parameter p = 0.45; (iv) *X* is a log normal distribution with parameters $(\mu, \sigma) = (0.0, 0.6)$, while *Z* is from standard normal distribution. Denote the estimators from the true model For $\beta_0 = 1.0$, $\beta_1 = 0.5$ and $\beta_2 = -0.5$, the simulation results are summarized in Table 3 (Part A). The results show that the proposed methods are consistent under the above mentioned scenarios.

Part B of Table 3 illustrated a situation where overwhelming number of sample are allocated to the tails, in this case, $(n_0, n_1, n_3) = (100, 150, 150)$. Results show that the unbiasedness property of $\widehat{\beta}_{PDS}$ still hold, with the efficiency further improved as more sample are allocated in the tails. However, at some point, the loss of precision in $\widehat{\phi}_k$ as SRS sample getting smaller will impact the efficiency.

3.2 Analysis of the Collaborative Perinatal Project Data

We illustrate our method using data from the Collaborative Perinatal Project (CPP) (Niswander and Gordon, 1972). This study evaluates the effect of mother's maternal pregnancy serum level of polychlorinated biphenyls (PCB) on her child's IQ test performance at age 7. Pregnant mothers were enrolled through university-affiliated medical clinics, and data were collected from mother at each prenatal visit. The study children were also followed for various neurodevelopmental outcomes for up to 8 years. One of the hypotheses is that the PCB levels are related to the performance on the Weschler Intelligence Scale for children at 7 years of age (Longnecker et al., 1997). To investigate the *in utero* exposure of PCB in relation to neurodevelopmental abnormality, the PCB levels were measured by analyzing the third trimester blood serum specimens that had been preserved from mothers in the CPP study. PCB levels are available for a simple random sample of 849 subjects from the underlying population. In addition to the PCB level as the exposure variable of interest, other variables available for all subjects under study include socioeconomic status of the child's family (SES), the gender (SEX, 1=female) and race (RACE, 1=black) of the child, and the mother's education (EDU) and age (AGE).

To illustrate our methods, we select a simple random sample with size $n_0 = 100$ from the cohort of 849 subjects. We then select two supplemental samples with size $n_1 = 50$ and $n_3 = 50$ randomly from the set $\left\{(Y, Z) : \widehat{Pr}\,(X \in A_1 | Y, Z) \geq 85\%\right\}$ and $\left\{(Y, Z) : \widehat{Pr}\,(X \in A_3 | Y, Z) \geq 85\%\right\}$, respectively. Note that the estimator $\widehat{Pr}\,(X \in A_1 | Y, Z)$ and $\widehat{Pr}\,(X \in A_3 | Y, Z)$ are estimated from the logistics model, and the domain of PCB is partitioned into 3 intervals with $a = 1$ as the cutpoint, i.e., $A_1 = (-\infty, \mu_{PCB} - \sigma_{PCB}] = (-\infty, 1.210]$ and $A_3 = (\mu_{PCB} + \sigma_{PCB}, \infty) = (5.037, +\infty)$. The ODS design also partitions the domain of *Y* into three intervals. The supplemental sample with size $n_1 = 50$ and $n_3 = 50$ are from the strata $B_1 = (-\infty, \mu_{IQ} - \sigma_{IQ}] = (-\infty, 81.441]$ and $B_3 = (\mu_{IQ} + \sigma_{IQ}, \infty) = (109.469, +\infty)$, respectively. The variables *EDU* and *AGE* are standardized, and we denote them as *EDU* and *AGE* without loss of generality. We tested the proper fitting of the covariates in the SRS sample and found that the p-values from partial F-test for testing a cubic model for *AGE* and *EDU* versus a quadratic model was 0.7, and a quadratic model versus a linear model is 0.008. Hence, we used the following quadratic model for all estimators compared.

$$IQ = \beta_0 + \beta_1 PCB + \beta_2 EDU + \beta_3 SES + \beta_4 AGE + \beta_5 RACE + \beta_6 SEX + \beta_7 EDU^2 + \beta_8 AGE^2 + \varepsilon, \quad (3.2)$$

The results for the CPP data analysis are summarized in Table 4. $\widehat{\beta}_{Full}$ denotes the full data analysis, which is included for the purpose of comparison.

Results in Table 4 reveal that none of the estimators demonstrated a significant PCB effect on the IQ scores for children at 7 years of age. Nevertheless, the effect of two-phase PDS design can be seen from the fact that the estimator $\widehat{\beta}_{PDS}$ for PCB under the two-phase PDS design has smaller standard error than the estimators $\widehat{\beta}_{ODS}$, $\widehat{\beta}_{IPW}$ and $\widehat{\beta}_{SRS}$. As the result, the 95% confidence interval for $\widehat{\beta}_{PDS}$ is narrower than those from $\widehat{\beta}_{ODS}$, , $\widehat{\beta}_{IPW}$ and $\widehat{\beta}_{SRS}$. It is not surprising that the standard error estimator $\widehat{\beta}_{Full}$ based on all data with a size of 849 for the PCB is the smallest, and consequently, has the narrowest confidence interval (−0.225, 0.665) for the effect of PCB.

4 Concluding Remarks

We proposed an innovative and cost-effective sampling design, the two-phase PDS design, that will enable the investigators to collect more informative samples at a fixed budget. The proposed design is multi-phase based and uses a biased sampling scheme where one observes the main exposure variable with a probability that depends on the outcome variable and other covariates. This research is developed in response to the need for designing more powerful study to effectively utilize the available financial resources in the current ongoing study, the Gulf Long-term Follow-up Study conducted at US National Institute of Environmental Health Sciences (NIEHS) (Sandler et al. 2011). The GuLF Study is a health study specifically for workers and volunteers who helped clean up the 2010 Deep water Horizon oil spill. About 56,000 subjects will be recruited. It is the largest study ever conducted on possible short-term and long-term health effects of oil spills. The budget for assessing benzene level in individuals will only be about 900 individuals. Collaborating with NIEHS scientists, we are in the process of designing a sampling strategy using the proposed two-phase PDS scheme to target for sampling more informative subjects.

The main advantages of the proposed design is that it allows for a continuous Y and a vector of available covariate Z to be used in selecting a more informative second phase data set. The proposed design avoids the impractical high dimension stratification issue when multiple covariate are included in Z. The proposed semiparametric empirical likelihood method is an efficient and robust way to analyze data from the proposed design. The primary competitors of the PDS design in practice are the ODS design with continuous response variable, the simple random sampling design and the inverse probability weighted method for two phase design, though the IPW method will also require the sample probability to be known. Our simulation results suggest that for the same sample size, the proposed PDS design, coupled with the proposed estimator, is more efficient and more powerful than these competing estimators. Our robustness simulation results also suggest that even though the logistic model estimates of $Pr(X \in A_k|Y, Z)$ is quite robust with respect to misspecification of the true underlying models.

There are a few recommendations for using the two-phase PDS design in practice. One needs to consider how to choose a, c, and how to distribute the supplemental samples. We suggest that a three-category design, $(-\infty, \mu_X - a * \sigma_X]$, $A_2 = (\mu_X - a * \sigma_X, \mu_X + a * \sigma_X]$ and $(\mu_X + a * \sigma_X, \infty)$, with a cut point reasonably away from the mean of the exposure be sufficient. The simulation results and subject matter considerations might support large values of a, e.g., greater than 1, so that it corresponds to a clinically abnormal value. However one has to be cautious selecting observations too far out in the distribution as the reward from choosing a relatively large value of a depends on assumption that f_β $(Y|X)$ is true across the entire range. This assumption may be violated if a is too large and stability could be an issue if observations are sampled from the very extreme tails. We recommend *a* to be between 1 and 1.5. We also recommend the value of *c* to be between 75% and 95% and with an even split for the supplement samples in the two outside tails (i.e, $n_1 = n_3$).

Some interesting future works remain. As we pointed out in earlier, implementing the proposal design can be done in two ways: (i) with underly cohort population and sampling proportion unknown, and stopping recruitment after the pre-specified number of subjects in the tail supplement samples are reached; or, (ii) with all (Y, Z) in the underlying cohort known. In the latter case, the augmented IPW estimator can be explored to get more efficient than the IPW estimator. It would be interesting to explore if the combination of the PDS design with ODS design would results in more efficient designs. Such combination could decompose efficiency gains into those gained with increase variation in Y (from ODS) and those gained with increased variation in X beyond the increased variation in Y. On the theory front, it would be interesting to explore the existence of a semiparametric efficient estimator for the proposed PDS designs. Finally, it would be interesting to explore the possible bias-variance tradeoff with different approaches for estimating φ_1 and φ_3.

Acknowledgments

The authors thank the editor, the associate editor and the referees for their helpful comments. The authors also wish to thank Dr. Jane Monaco for the careful proofreading of the manuscript. This work was partly supported by United States National Institutes of Health grants R01-CA79949, R01-ES021900, RR025747, and P01-CA142538.

Appendix: proof of Theorem 1

Recall that the approximated profile log-likelihood function is

$$\begin{aligned}\tilde{l}\left(\beta,\pi,\nu\right)= & \sum_{i=1}^{n} log\, f_\beta\left(Y_i|X_i,Z_i\right)\\ & -\sum_{i=1}^{n} log\left\{q_0+\sum_{k\in S} q_k\pi_k^{-1}\widehat{F}_k\left(X_iZ_i\right)+\nu^T\left(\widehat{F}_1\left(X_i,Z_i\right)-\pi_1,\widehat{F}_3\left(X_i,Z_i\right)-\pi_3\right)\right\}\\ & -\sum_{k\in S} n_k\, log\,\pi_k,\end{aligned}$$

where $\widehat{F}_k\left(X_i,Z_i\right)=\int f_\beta\left(Y|X_i,Z_i\right) I\left(\widehat{\phi}_k\left(Y,Z_i\right)\ge c_k\right) dY$ and $\mathscr{S}=\{1,3\}$ We define $l^*(\beta, \pi, \nu)$ the same as $\tilde{l}\left(\beta,\pi,\nu\right)$ except that $\widehat{\phi}_k$ is replaced by φ_k. Let $\xi = (\beta, \pi, \nu)$, then we can abbreviate $\tilde{l}\left(\beta,\pi,\nu\right)$ and $l^*(\beta, \pi, \nu)$ as $\tilde{l}\left(\xi\right)$ and $l^*(\xi)$, respectively.

We impose the following assumptions.

(C.1) The log-density log f_β (Y |X, Z) is twice-continuously differentiable with respect β.

(C.2) The proportion n_j/n is a fixed constant $q_j \in (0, 1)$.

(C.3) The class of functions

$$\mathscr{F}\equiv\left\{f_\beta\left(Y|X,Z\right),\frac{\partial^s}{\partial\beta^s}log\, f_\beta\left(Y|X,Z\right),\int\tfrac{\partial^s}{\partial\beta^s}log\, f_\beta\left(Y|X,Z\right)f_\beta\left(Y|X,Z\right)I\left(\phi_k\left(Y,Z\right)\ge c_k\right)dY\!:\!s=0,1,2\text{ and the function ar}\right.$$

is P-Donsker and have an envelope function with finite second moment.

(C.4) The hessian matrix of $E[n^{-1}l^*(\xi)]$ is continuous in a neighborhood of the true ξ (β, π, 0, 0) and is non-singular at ξ.

(C.5) The estimator $\int \frac{\partial^s}{\partial\beta^s} log f_\beta(Y|X,Z)\, I\left(\widehat{\phi}_k(Y,Z)\geq c_k\right) dY$, s = 0, 1, 2, belongs to $\mathscr{F}$ and $Pr\left((Y,Z):I\left(\widehat{\phi}_k(Y,Z)\geq c_k\right)\rightarrow I\left(\phi_k(Y,Z)\geq c_k\right)\right)=1$.

(C.6) It holds

$$\begin{aligned}&E\left[\partial log f_\beta(Y|X,Z)\right.\\&\qquad\left./\partial\beta I\left(\widehat{\phi}_k(Y,Z)\geq c_k\right)\right]-E\left[\partial log f_\beta(Y|X,Z)/\partial\beta I\left(\phi_k(Y,Z)\geq c_k\right)\right]\\&\qquad=n^{-1}\sum_{i=1}^{n}Q_{1k}(Y_i,X_i,Z_i)+o_p(1)\end{aligned}$$

and

$$E\left[I\left(\widehat{\phi}_k(Y,Z)\geq c_k\right)\right]-E\left[I\left(\phi_k(Y,Z)\geq c_k\right)\right]=n^{-1}\sum_{i=1}^{n}Q_{2k}(Y_i,X_i,Z_i)+o_p(1)$$

for k = 1, 3, where $Q_{1k}(Y, X, Z)$ and $Q_{2k}(Y, X, Z)$ are mean 0 random vectors with finite second moments.

Conditions (C.1)–(C.4) are all regular conditions for $f_\beta(Y|X, Z)$ and $\varphi_k(Y, Z)$, which hold for usual regression models and the choices of φ_k. Conditions (C.5) and (C.6) regard the properties of the estimator $\widehat{\phi}_k$. These conditions can be easily verified if $\widehat{\phi}_k$ takes parametric structure such as (2.2) or (2.3). For the kernel estimator (2.4), verifying these two conditions needs some additional work but can be shown to hold if the bandwidth is chosen small enough.

(i) Proof of Consistency At the true value for ξ = (β, π, 0, 0), we calculate the first derivative of $n^{-1}\tilde{l}(\xi)$ so obtain

$$\frac{\partial}{\partial\beta}n^{-1}\tilde{l}(\xi)=n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}log f_\beta(Y_i|X_i,Z_i)$$

$$-n^{-1}\sum_{i=1}^{n}\sum_{k\in\mathscr{S}}q_k\frac{\int\partial log f_\beta(Y|X_i,Z_i)/\partial\beta f_\beta(Y|X_i,Z_i)\, I\left(\widehat{\phi}_k(Y,Z_i)\geq c_k\right)dY}{\pi_k},\quad\text{(A.1)}$$

and for k = 1, 3,

$$\frac{\partial}{\partial\pi_k}n^{-1}\tilde{l}(\xi)=n^{-1}\sum_{i=1}^{n}\frac{q_k\int f_\beta(Y|X_i,Z_i)\, I\left(\widehat{\phi}_k(Y,Z_i)\geq c_k\right)dY}{\pi_k^2}-\frac{q_k}{\pi_k},\quad\text{(A.2)}$$

and

$$\frac{\partial}{\partial \nu_k} n^{-1}\tilde{l}(\xi) = -\, n^{-1}\sum_{i=1}^{n}\left(\int f_\beta (Y|X_i, Z_i)\, I\left(\widehat{\phi}_k (Y, Z_i) \geq c_k\right) dY - \pi_k\right). \tag{A.3}$$

By the Donsker property in (C.3) and (C.5), we apply the Glivenko-Cantelli theorem and obtain

$$|\frac{\partial}{\partial \xi} n^{-1}\tilde{l}(\xi) - \frac{\partial}{\partial \xi} E\left[n^{-1}\tilde{l}(\xi)\right]| \rightarrow_{a.s.} 0.$$

Since $E\left[n^{-1}\tilde{l}(\xi)\right] \rightarrow E\left[n^{-1}l^*(\xi)\right]$, we have

$$\frac{\partial}{\partial \xi} n^{-1}\tilde{l}(\xi) \rightarrow_{a.s.} E\left[n^{-1}l^*(\xi)\right].$$

Here $n^{-1}l^*(\xi)$ takes the same expression as (A.1)–(A.3) except that $\widehat{\phi}_k$ is replaced by φ_k. On the other hand, using the ODS design fact that

$$E\left[n^{-1}\sum_{i=1}^{n} g_1 (Y_i, X_i, Z_i)\right] = q_0 E\left[g_1 (Y, X, Z)\right] + \sum_{k\in\mathscr{S}} q_k E\left[g_1 (Y, X, Z)\,|\phi_k (Y, Z) \geq c_k\right],$$

$$E\left[n^{-1}\sum_{i=1}^{n} g_2 (X_i, Z_i)\right] = E\left[g_2 (X, Z)\right],$$

and the fact that $\pi_k = E[I(\varphi_k(Y, Z)\ c_k)]$, we can easily calculate $E[n^{-1}l^*(\xi)] = 0$. Thus, $n^{-1}\partial\tilde{l}(\xi)/\partial\xi \rightarrow_{a.s.} 0$; that is, 0 belongs to the image of $n^{-1}\partial\tilde{l}(\xi)$ in any given neighborhood of the true ξ when n is large enough. Similarly, we can show $n^{-1}\partial^2\tilde{l}(\xi)/\partial\xi^2 \rightarrow_{a.s.} E\left[n^{-1}\partial^2 l^*(\xi)/\partial\xi^2\right]$ for ξ in a neighborhood of the true value. Thus, from condition (C.4), $n^{-1}\partial^2\tilde{l}(\xi)/\partial\xi^2$ is invertible in this neighborhood when n is large enough. From the inverse mapping theorem, $n^{-1}\partial\tilde{l}(\xi)/\partial\xi$ is invertible in any small neighborhood of the true ξ. Consequently, we conclude that there exists a solution $\widehat{\xi}$ to $\partial\tilde{l}(\xi)/\partial\xi = 0$ and $\widehat{\xi}$ converges almost surely to the true ξ.

(ii) Proof of Asymptotic Normality From equation

$$n^{-1}\frac{\partial}{\partial \xi}\tilde{l}\left(\widehat{\xi}\right) = 0,$$

we obtain

$$n^{-1}\frac{\partial}{\partial \xi}\tilde{l}\left(\widehat{\xi}\right) - E\left[n^{-1}\frac{\partial}{\partial \xi}\tilde{l}\left(\widehat{\xi}\right)\right] = -E\left[n^{-1}\frac{\partial}{\partial \xi}\tilde{l}\left(\widehat{\xi}\right)\right] + E\left[n^{-1}\frac{\partial}{\partial \xi}\tilde{l}(\xi)\right] - E\left[n^{-1}\frac{\partial}{\partial \xi}\tilde{l}(\xi)\right].$$

We apply the Taylor expansion to the first term on the right-hand side and obtain

$$n^{-1}\frac{\partial}{\partial\xi}\tilde{l}\left(\hat{\xi}\right)-E\left[n^{-1}\frac{\partial}{\partial\xi}\tilde{l}\left(\hat{\xi}\right)\right]=-E\left[n^{-1}\frac{\partial^2}{\partial\xi^2}\tilde{l}\left(\tilde{\xi}\right)\right]\left(\hat{\xi}-\xi\right)-E\left[n^{-1}\frac{\partial}{\partial\xi}\tilde{l}\left(\xi\right)\right], \qquad \text{(A.4)}$$

where $\tilde{\xi}$ is between$\hat{\xi}$ and ξ.

In equation (A.4), the left-hand side can be expressed as an empirical process indexed by functions

$$\left\{\frac{\partial}{\partial\xi}logf_\beta\left(Y|X,Z\right)-\frac{\partial}{\partial\xi}log\left(q_0+\sum_{k\in\mathscr{S}}q_k\pi_k^{-1}\hat{F}_k\left(X_i,Z_i\right)+\nu^T\left(\hat{F}_1\left(X_i,Z_i\right)-\pi_1,\hat{F}_3\left(X_i,Z_i\right)-\pi_3\right)\right)+\sum_{k\in\mathscr{S}}q_k\frac{\partial}{\partial\xi}log\pi_k\text{:}\right.$$
$$\left.\xi\,\text{is in a neighborhood of the true value}\right\}.$$

By conditions (C.3) and (C.5), it is asymptotically equivalent to $n^{-1/2}\sum_{i=1}^{n}U\left(Y_i,X_i,Z_i\right)$, where

$$U\left(Y_i,X_i,Z_i\right)=\begin{pmatrix}\frac{\partial}{\partial\beta}logf_\beta\left(Y_i|X_i,Z_i\right)-\sum\limits_{k\in\mathscr{S}}q_k\pi_k^{-1}\int\frac{\partial}{\partial\beta}logf_\beta\left(Y|X_i,Z_i\right)f_\beta\left(Y|X_i,Z_i\right)I\left(\phi_k\left(Y,Z_i\right)\ge c_k\right)dY\\ q_k\pi_k^{-2}\int f_\beta\left(Y|X_i,Z_i\right)I\left(\phi_k\left(Y,Z_i\right)\ge c_k\right)dY-q_k\pi_k^{-1},\quad k=1,3\\ \int f_\beta\left(Y|X_i,Z_i\right)I\left(\phi_k\left(Y,Z_i\right)\ge c_k\right)dY-\pi_k,\quad k=1,3\end{pmatrix}.$$

According to (C.5), the matrix in the first term of the right-hand side of (A.4) satisfies

$$E\left[n^{-1}\frac{\partial^2}{\partial\xi^2}\tilde{l}\left(\tilde{\xi}\right)\right]\to E\left[n^{-1}\frac{\partial^2}{\partial\xi^2}l^*\left(\xi\right)\right]=V\left(\xi\right).$$

For the second term on the right-hand side of (A.4), we note

$$E\left[n^{-1}\frac{\partial}{\partial\xi}\tilde{l}\left(\xi\right)\right]=E\left[n^{-1}\frac{\partial}{\partial\xi}\tilde{l}\left(\xi\right)\right]-E\left[n^{-1}\frac{\partial}{\partial\xi}l^*\left(\xi\right)\right],$$

which is further simplified as

$$\begin{pmatrix}-\sum\limits_{k\in\mathscr{S}}q_k\pi_k^{-1}\left\{E\left[\frac{\partial}{\partial\beta}logf_\beta\left(Y|X,Z\right)I\left(\hat{\phi}_k\left(Y,Z\right)\ge c_k\right)\right]-E\left[\frac{\partial}{\partial\beta}logf_\beta\left(Y|X,Z\right)I\left(\phi_k\left(Y,Z\right)\ge c_k\right)\right]\right\}\\ q_k\pi_k^{-2}\left\{E\left[I\left(\hat{\phi}_k\left(Y,Z_i\right)\ge c_k\right)\right]-E\left[I\left(\phi_k\left(Y,Z\right)\ge c_k\right)\right]\right\},\quad k=1,3\\ -\left\{E\left[I\left(\hat{\phi}_k\left(Y,Z\right)\ge c_k\right)\right]-E\left[I\left(\phi_k\left(Y,Z\right)\ge c_k\right)\right]\right\},\quad k=1,3\end{pmatrix}.$$

Combining all these results and using condition (C.6), we obtain

$$-\left(V\left(\xi\right)+o\left(1\right)\right)\left(\hat{\xi}-\xi\right)=n^{-1/2}\sum_{i=1}^{n}\left[U\left(Y_i,X_i,Z_i\right)+\begin{pmatrix}-\sum\limits_{k\in\mathscr{S}}q_k\pi_k^{-1}Q_{1k}\left(Y_i,X_i,Z_i\right)\\ q_k\pi_k^{-2}Q_{2k}\Big)\left(Y_i,X_i,Z_i\right),\quad k=1,3\\ -Q_{2k}\left(Y_i,X_i,Z_i\right),\quad k=1,3\end{pmatrix}\right]. \qquad \text{(A.5)}$$

The asymptotic normality of $\widehat{\xi}$ thus follows.

(iii) Consistent estimator of variance From the above derivation, the asymptotic covariance of $\widehat{\xi}$ takes form $V(\xi)^{-1}U(\xi)\{V(\xi)^{-1}\}^{\mathrm{T}}$, where $U(\xi)$ is the variance of each summand on the right-hand side of (A.5). Thus, a consistent estimator of the asymptotic variance for $\sqrt{n}\left(\widehat{\xi}-\xi\right)$ is given by $\widehat{V}\left(\widehat{\xi}\right)^{-1}\widehat{U}\left(\widehat{\xi}\right)\left\{\widehat{V}\left(\widehat{\xi}\right)^{-1}\right\}^{T}$, where $\widehat{V}\left(\widehat{\xi}\right)=n^{-1}\frac{\partial^2}{\partial\xi^2}\tilde{l}\left(\widehat{\xi}\right)$ and $\widehat{U}\left(\widehat{\xi}\right)$ is the sample variance of the sample version of

$$U\left(Y_i,X_i,Z_i\right)+\left(\begin{array}{c} -\sum_{k\in\mathscr{S}}q_k\pi_k^{-1}Q_{1k}\left(Y_i,X_i,Z_i\right) \\ \left.q_k\pi_k^{-2}Q_{2k}\right)\left(Y_i,X_i,Z_i\right), \quad k=1,3 \\ -Q_{2k}\left(Y_i,X_i,Z_i\right), \quad k=1,3 \end{array}\right).$$

.

References

Amemiya, T. Advanced Econometrics. Harvard University Press; Cambridge, Massachusetts: 1985.

Anderson JA. Separate sample logistic discrimination. Biometrika. 1972; 59:19–35.

Breslow NE, Cain KC. Logistic regression for two-stage case-control data. Biometrika. 1988; 75:11–20.

Breslow NE, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. Journal of the Royal Statistical Society, Series B. 1997; 59:447–461.

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

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; ll:1269–1275. [PubMed: 14861651]

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

Kang S, Cai J. Marginal hazards regression for retrospective studies within cohort with possibly correlated failure time data. Biometrics. 2009; 65:405–414. [PubMed: 18565164]

Langholz B, Borgan O. Counter-matching: A stratified nested case-control sampling method. Biometrika. 1995; 82:69–79.

Lu W, Tsiatis AA. Semiparametric transformation models for the case-cohort sturdy. Biometrika. 2006; 93:207–214.

Longnecker, M.; Klebanoff, M.; Zhou, H.; Wilcox, A.; Berendes, H.; Hoffman, H. Proposal to study in utero exposure to DDE and PCBs in relation to m,ale hirth defects and neurodevelopmental outcomes in the Collaborative Perinatal Project. Study Proposal, National Institute of Environmental Health Sciences; Washington, D.C.: 1997.

Manatunga A, Chen H, Terrell M, Lyles R, Marcus M. A longitudinal model or repeated highly skewed outcome data. Journal of Applied Statistics. 2008; 9:39–51.

Neyman J. Contribution to the theory of sampling from human populations. Journal of the American Statistical Association. 1938; 33:101–116.

Niswander, KR.; Gordon, M. US. Department of Health, Education, and Welfare Publication (NIH) 73–379. U.S. Government Printing Office; Washington, D.C.: 1972. The women and their pregnancies.

Owen AB. Empirical likelihood ratio confidence intervals for a single functional. Biometrika. 1988; 75:237–249.

Owen AB. Empirical likelihood for confidence regions. The Annals of Statistics. 1990; 18:90–120.

Prentice RL. A case-cohort design for epidemiologic studies and disease prevention trials. Biometrika. 1986; 73:1–11.

Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika. 1979; 66:403–412.

Qin G, Zhou H. Partial linear inference for a 2-stage outcome-dependent sampling design with a continuous outcome. Biostatistics. 2011; 12:506–520. [PubMed: 21156990]

Qin J. Empirical likelihood in biased sample problems. The Annals of Statistics. 1993; 21:1182–1196.

Qin J, Lawless JF. Empirical likelihood and general estimating equations. The Annals of Statistics. 1994; 22:300–325.

Sandler, D., et al. National Institute of Environmental Health Sciences (NIEHS) GuLF Worker Study Draft to IOM-v2. 2010. GuLF Worker Study: Gulf Long-Term Follow-Up Study for Oil Spill Clean-Up Workers and Volunteers.

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]

Schildcrout JS, Rathouz PJ. Longitudinal Studies of Binary Response Data Following Case-Control and Stratified Case-Control Sampling: Design and Analysis. biometrics. 2010; 66:365–373. [PubMed: 19673861]

Song R, Zhou H, Kosorok MR. On semiparametric efficient inference for two-stage outcome dependent sampling with a continuous outcome. Biometrika. 2009; 96:221–228. [PubMed: 20107493]

Vardi Y. Nonparametric estimation in presence of length bias. The Annals of Statistics. 1982; 10:616–620.

Vardi Y. Empirical distribution in selection bias models. The Annals of Statistics. 1985; 13:178–203.

Wang X, Zhou H. A semiparametric empirical likelihood method for biased sampling schemes with auxiliary covariates. Biometrics. 2006; 62:1149–1160. [PubMed: 17156290]

Wang X, Zhou H. Design and inference for cancer biomarker study with an outcome and auxiliary-dependent subsampling. Biometrics. 2010; 66:502–511. [PubMed: 19508239]

Weaver MA, Zhou H. An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. J. Am. Statist. Assoc. 2005; 100:459–469.

Weinberg CR, Wacholder S. Prospective analysis of case-control data under general multiplicative intercept risk models. Biometrika. 1993; 80:461–465.

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

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, Wu Y, Liu Y, Cai J. Semiparametric inference for 2-stage outcome-auxiliary-dependent sampling design with continuous outcome. Biostatistics. 2011; 12:521–534. [PubMed: 21252082]

Zhou H, You J, Qin G, Longnecker MP. A partially linear regression model for data from an outcome-dependent samplign design. Journal of the Royal Statistical Society: Series C. 2011 DOI: 10.1111/j.1467-9876.2010.00756.x.

Zhou H, Song R, Wu Y, Qin J. Statistical inference for a two-stage outcome dependent sampling design with a continuous outcome. Biometrics. 2011; 67:194–202. [PubMed: 20560938]

Table 1

Simulation results PDS design.[†]

			β_1				β_2			
β_1	*a*	**Method**	**Mean**	**SE**	$\widehat{SE}$	**CI**	**Mean**	**SE**	$\widehat{SE}$	**CI**
			$(n_0, n_1, n_3) = (200, 100, 100)$							
0.0	1.0	$\widehat{\beta}_X$	−0.002	0.038	0.038	0.948	−0.502	0.127	0.127	0.951
		$\widehat{\beta}_{PDS1}$	0.000	0.043	0.041	0.930	−0.496	0.150	0.127	0.946
		$\widehat{\beta}_{PDS2}$	0.000	0.037	0.037	0.950	−0.500	0.095	0.096	0.962
		$\widehat{\beta}_{ODS}$	0.002	0.038	0.039	0.942	−0.505	0.120	0.123	0.958
		$\widehat{\beta}_{IPW}$	0.000	0.046	0.046	0.952	−0.510	0.146	0.140	0.931
		$\widehat{\beta}_{SRS}$	0.001	0.050	0.050	0.957	−0.502	0.150	0.154	0.955
		$\widehat{\beta}_N$	0.001	0.047	0.048	0.956	−0.493	0.310	0.119	0.575
0.5	1.0	$\widehat{\beta}_X$	0.499	0.038	0.038	0.957	−0.498	0.126	0.126	0.946
		$\widehat{\beta}_{PDS1}$	0.501	0.039	0.041	0.954	−0.494	0.097	0.106	0.968
		$\widehat{\beta}_{PDS2}$	0.500	0.043	0.042	0.936	−0.495	0.104	0.106	0.950
		$\widehat{\beta}_{ODS}$	0.503	0.044	0.045	0.953	−0.505	0.128	0.131	0.954
		$\widehat{\beta}_{IPW}$	0.502	0.047	0.046	0.941	−0.508	0.155	0.151	0.938
		$\widehat{\beta}_{SRS}$	0.500	0.049	0.050	0.955	−0.496	0.156	0.154	0.944

			β_1				β_2			
β_1	***a***	**Method**	**Mean**	**SE**	$\widehat{SE}$	**CI**	**Mean**	**SE**	$\widehat{SE}$	**CI**
		$\widehat{\beta}_N$	0.633	0.060	0.052	0.302	−0.495	0.122	0.134	0.972
0.5	1.5	$\widehat{\beta}_X$	0.502	0.032	0.032	0.951	−0.500	0.120	0.117	0.940
		$\widehat{\beta}_{PDS1}$	0.499	0.037	0.037	0.938	−0.503	0.100	0.097	0.946
		$\widehat{\beta}_{PDS2}$	0.499	0.038	0.038	0.942	−0.500	0.103	0.099	0.940
		$\widehat{\beta}_{ODS}$	0.503	0.042	0.043	0.956	−0.495	0.118	0.120	0.963
		$\widehat{\beta}_{IPW}$	0.504	0.055	0.052	0.934	−0.508	0.167	0.166	0.937
		$\widehat{\beta}_{SRS}$	0.500	0.049	0.050	0.955	−0.496	0.156	0.154	0.944
		$\widehat{\beta}_N$	0.694	0.108	0.049	0.178	−0.507	0.297	0.128	0.603
			$(n_0, n_1, n_3) = (300, 50, 50)$							
0.5	1.5	$\widehat{\beta}_X$	0.498	0.038	0.038	0.951	−0.497	0.130	0.130	0.949
		$\widehat{\beta}_{PDS1}$	0.501	0.042	0.042	0.952	−0.497	0.114	0.113	0.932
		$\widehat{\beta}_{PDS2}$	0.499	0.043	0.044	0.966	−0.498	0.120	0.121	0.950
		$\widehat{\beta}_{ODS}$	0.503	0.044	0.045	0.955	−0.495	0.126	0.129	0.960
		$\widehat{\beta}_{IPW}$	0.502	0.046	0.045	0.942	−0.498	0.147	0.144	0.940
		$\widehat{\beta}_{SRS}$	0.500	0.049	0.050	0.955	−0.496	0.156	0.154	0.944
		$\widehat{\beta}_N$	0.623	0.072	0.050	0.367	−0.487	0.149	0.130	0.923

β 1	*a*	Method	β 1				β 2			
			Mean	SE	$\widehat{SE}$	CI	Mean	SE	$\widehat{SE}$	CI
			$(n_0, n_1, n_3) = (100, 50, 50)$							
0.5	1.0	$\widehat{\beta}_X$	0.499	0.051	0.053	0.956	−0.501	0.178	0.179	0.951
		$\widehat{\beta}_{PDS1}$	0.494	0.058	0.056	0.934	−0.498	0.150	0.140	0.942
		$\widehat{\beta}_{PDS2}$	0.502	0.059	0.058	0.948	−0.505	0.146	0.150	0.968
		$\widehat{\beta}_{ODS}$	0.502	0.062	0.062	0.955	−0.505	0.188	0.187	0.937
		$\widehat{\beta}_{IPW}$	0.508	0.066	0.064	0.929	−0.507	0.218	0.210	0.922
		$\widehat{\beta}_{SRS}$	0.495	0.072	0.071	0.944	−0.499	0.219	0.219	0.955
		$\widehat{\beta}_N$	0.636	0.084	0.074	0.544	−0.492	0.180	0.192	0.967

[†] Results are based on the model $Y = \beta_0 + \beta_1 X + \beta_2 I_{\log(|x|)+e>1} + \in$, where $e \sim N(0, 1)$, $\in \sim N(0, 1)$, and $X \sim N(0, 1)$; the true parameter values are $\beta_0 = 1.0$ and ($\beta_2 = -0.5$. $\widehat{\beta}_X$, $\widehat{\beta}_{PDS1}$, $\widehat{\beta}_{PDS2}$, $\widehat{\beta}_{ODS}$, $\widehat{\beta}_{SRS}$, $\widehat{\beta}_{IPW}$, and $\widehat{\beta}_N$ are defined as in Section 3.1.

Table 2

Simulations results for the power and relative efficiency.[†]

		β_1		$\beta_2 = -0.5$		β_1		$\beta_2 = -0.5$	
β_1	Method	RE	Size/Power	RE	Power	RE	Size/Power	RE	Power
		$(n_0, n_1, n_3) = (100, 50, 50), \sigma^2 = 4$				$(n_0, n_1, n_3) = (150, 25, 25), \sigma^2 = 4$			
0.0	$\widehat{\beta}_X$	1.000	0.055	1.000	0.284	1.000	0.051	1.000	0.245
	$\widehat{\beta}_{PDS_1}$	0.983	0.056	1.011	0.312	1.001	0.048	1.016	0.296
	$\widehat{\beta}_{PDS_2}$	1.002	0.058	0.780	0.436	1.013	0.044	0.806	0.354
	$\widehat{\beta}_{ODS}$	1.011	0.059	0.934	0.331	1.019	0.044	1.036	0.308
	$\widehat{\beta}_{IPW}$	1.188	0.071	1.085	0.278	1.106	0.055	1.042	0.267
	$\widehat{\beta}_{SRS}$	1.332	0.058	1.227	0.229	1.238	0.058	1.187	0.229
0.1	$\widehat{\beta}_X$	1.000	0.167	1.000	0.286	1.000	0.134	1.000	0.240
	$\widehat{\beta}_{PDS_1}$	1.005	0.172	0.980	0.322	1.000	0.128	0.981	0.252
	$\widehat{\beta}_{PDS_2}$	1.010	0.176	0.795	0.446	1.005	0.148	0.823	0.380
	$\widehat{\beta}_{ODS}$	1.016	0.159	0.940	0.350	1.042	0.131	1.015	0.287
	$\widehat{\beta}_{IPW}$	1.189	0.144	1.099	0.264	1.093	0.129	1.053	0.293
	$\widehat{\beta}_{SRS}$	1.336	0.110	1.234	0.236	1.225	0.110	1.178	0.236
0.5	$\widehat{\beta}_X$	1.000	0.997	1.000	0.291	1.000	0.990	1.000	0.259
	$\widehat{\beta}_{PDS_1}$	1.031	0.999	0.942	0.315	1.000	0.976	0.837	0.316
	$\widehat{\beta}_{PDS_2}$	1.042	0.997	0.812	0.443	1.001	0.980	0.829	0.334
	$\widehat{\beta}_{ODS}$	1.068	0.996	0.962	0.324	1.027	0.985	0.991	0.282
	$\widehat{\beta}_{IPW}$	1.191	0.981	1.117	0.274	1.068	0.978	1.036	0.254
	$\widehat{\beta}_{SRS}$	1.338	0.930	1.237	0.207	1.202	0.930	1.114	0.207

[†] Results are based on the model $Y = \beta_0 + \beta_1 X + \beta_2 I_{\log(|x|)+e>1} + \epsilon$, where $e \sim N(0, 1)$, $\epsilon \sim N(0, \sigma^2)$, and $X \sim N(0, 1)$.

Table 3

Robust property of the PDS estimator.[†]

			β_1				β_2			
β_1	*a*	**Method**	**Mean**	**SE**	$\widehat{SE}$	**CI**	**Mean**	**SE**	$\widehat{SE}$	**CI**
					Part A					
					$(n_0, n_1, n_3) = (200, 100, 100)$					
					$X \sim N(0, 1)$, $Z \sim$ binary(0.45)					
0.5	1.0	$\widehat{\beta}_{PDS_1}$	0.498	0.041	0.042	0.950	−0.500	0.076	0.075	0.944
		$\widehat{\beta}_{PDS_2}$	0.500	0.043	0.044	0.957	−0.495	0.082	0.083	0.950
					$X \sim N(0, 1)$, $Z \sim N(0, 1)$					
		$\widehat{\beta}_{PDS_1}$	0.501	0.043	0.042	0.949	−0.501	0.037	0.037	0.948
		$\widehat{\beta}_{PDS_2}$	0.502	0.045	0.044	0.941	−0.499	0.040	0.041	0.953
					$X \sim$ exp(1), $Z \sim$ binary(0.45)					
		$\widehat{\beta}_{PDS_1}$	0.498	0.039	0.037	0.940	−0.503	0.107	0.097	0.938
		$\widehat{\beta}_{PDS_2}$	0.497	0.044	0.042	0.942	−0.503	0.108	0.098	0.944
					$X \sim$ log-normal(0, 0.6), $Z \sim N(0, 1)$					
		$\widehat{\beta}_{PDS_1}$	0.500	0.047	0.047	0.954	−0.503	0.039	0.038	0.945
		$\widehat{\beta}_{PDS_2}$	0.501	0.056	0.053	0.942	−0.499	0.043	0.042	0.950
					Part B					
					$(n_0, n_1, n_3) = (100, 150, 150)$					

			β_1				β_2			
β_1	***a***	**Method**	**Mean**	**SE**	$\widehat{SE}$	**CI**	**Mean**	**SE**	$\widehat{SE}$	**CI**
0.5	1.0	$\widehat{\beta}_X$	0.500	0.035	0.034	0.941	−0.501	0.114	0.117	0.960
		$\widehat{\beta}_{PDS_1}$	0.500	0.041	0.037	0.940	−0.494	0.111	0.097	0.946
		$\widehat{\beta}_{PDS_2}$	0.501	0.042	0.039	0.941	−0.501	0.100	0.097	0.942
		$\widehat{\beta}_{ODS}$	0.501	0.046	0.044	0.944	−0.497	0.127	0.125	0.946
		$\widehat{\beta}_{IPW}$	0.511	0.058	0.057	0.940	−0.505	0.191	0.185	0.937
		$\widehat{\beta}_{SRS}$	0.500	0.049	0.050	0.955	−0.496	0.156	0.154	0.944

[†] Estimators are defined the same as in Table 1.

Table 4

Analysis results for the CPP data set.[††]

Covariate	Int	PCB	EDU	SES	AGE	RACE	SEX	EDU^2	AGE^2
$\widehat{\beta}_{full}$	93.160	0.219	3.407	0.998	−0.566	−7.712	−0.768	0.579	0.684
$\widehat{SE}\left(\widehat{\beta}_{full}\right)$	1.688	0.227	0.535	0.269	0.525	0.925	0.839	0.225	0.370
upperC.I.	89.851	−0.225	2.357	0.471	−1.595	−9.526	−2.414	0.136	−0.042
lowerC.I.	96.469	0.665	4.457	1.526	0.462	−5.897	0.877	1.022	1.410
$\widehat{\beta}_{PDS}$	92.818	0.153	3.154	1.072	−0.279	−7.453	−1.053	0.665	0.500
$\widehat{SE}\left(\widehat{\beta}_{PDS}\right)$	3.588	0.465	1.102	0.554	1.044	1.907	1.770	0.460	0.735
upperC.I.	85.784	−0.758	0.992	−0.015	−2.326	−11.191	−4.523	−0.238	−0.941
lowerC.I.	99.851	1.064	5.315	2.159	1.767	−3.715	2.416	1.568	1.941
$\widehat{\beta}_{ODS}$	92.228	0.641	5.073	1.273	−0.722	−12.394	−0.070	0.457	0.527
$\widehat{SE}\left(\widehat{\beta}_{ODS}\right)$	4.341	0.516	1.267	0.686	1.330	2.417	2.084	0.459	0.886
upperC.I.	83.718	−0.371	2.588	−0.072	−3.329	−17.132	−4.157	−0.443	−1.210
lowerC.I.	100.738	1.654	7.557	2.619	1.885	−7.656	4.016	1.357	2.265
$\widehat{\beta}_{IPW}$	93.101	0.271	3.589	0.995	−0.602	−7.944	−0.674	0.556	0.678
$\widehat{SE}\left(\widehat{\beta}_{IPW}\right)$	11.757	1.587	3.867	1.851	3.612	6.309	5.882	1.644	2.493
upperC.I.	70.057	−2.839	−3.991	−2.632	−7.682	−20.309	−12.204	−2.667	−4.208
lowerC.I.	116.146	3.382	11.169	4.623	6.478	4.421	10.856	3.780	5.565
$\widehat{\beta}_{SRS}$	91.296	0.579	3.535	0.906	−0.439	−6.143	0.183	0.981	0.832
$\widehat{SE}\left(\widehat{\beta}_{SRS}\right)$	3.527	0.554	1.017	0.547	1.158	1.847	1.685	0.467	0.705
upperC.I.	84.383	−0.508	1.542	−0.166	−2.710	−9.765	−3.119	0.066	−0.550
lowerC.I.	98.209	1.666	5.529	1.979	1.831	−2.521	3.486	1.896	2.215

The outcome is the Weschler Intelligence Scale for children at 7 years of age (IQ). PCB is the level measured from the third-trimester blood serum specimens that have been preserved from mothers in the CPP study; EDU is the standardized mother's education level; SES is the socioeconomic status of the child's family; AGE is standardized mother's age; RACE and SEX are the race and gender of the child. The fitted model is $IQ = \beta_0 + \beta_1 PCB + \beta_2 EDU + \beta_3 SES + \beta_4 AGE + \beta_5 RACE + \beta_6 SEX + \beta_7 EDU^2 + \beta_8 AGE^2 + \varepsilon$, where ε is zero mean normal variable with unknown variance. $\widehat{\beta}_{full}$, $\widehat{\beta}_{PDS}$, $\widehat{\beta}_{ODS}$, $\widehat{\beta}_{IPW}$ and $\widehat{\beta}_{SRS}$ are defined in 3.2.

[††] $a = 1$ and the allocation pattern is $(n_0, n_1, n_3) = (100, 50, 50)$.