Published in final edited form as:
Biometrics. 2011 March ; 67(1): 194–202. doi:10.1111/j.1541-0420.2010.01446.x.

Statistical Inference for a Two-Stage Outcome-Dependent Sampling Design with a Continuous Outcome

Haibo Zhou,
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420 zhou@bios.unc.edu

Rui Song,
Department of Statistics, Colorado State University, Fort Collins, CO 80523 song@stat.colostate.edu

Yuanshan Wu, and
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420 and School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China yswu@bios.unc.edu

Jing Qin
Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, NIH, 6700B Rockledge Drive MSC 7609 jingqin@niaid.nih.gov

Summary

The two-stage case-control design has been widely used in epidemiology studies for its cost-effectiveness and improvement of the study efficiency (White, 1982; Breslow and Cain, 1988). The evolution of modern biomedical studies has called for cost-effective designs with a continuous outcome and exposure variables. In this paper, we propose a new two-stage outcome-dependent sampling scheme with a continuous outcome variable, where both the first-stage data and the second-stage data are from outcome-dependent sampling schemes. We develop a semiparametric empirical likelihood estimation for inference about the regression parameters in the proposed design. Simulation studies were conducted to investigate the small sample behavior of the proposed estimator. We demonstrate that, for a given statistical power, the proposed design will require a substantially smaller sample size than the alternative designs. The proposed method is illustrated with an environmental health study conducted at National Institute of Health.

Keywords

Biased sampling; Empirical likelihood; Outcome dependent; Sample size; Two-stage design

7. Supplementary Materials

The Web Appendixes A, B, and C are available under the Paper Information link at the Biometrics web site http://www.biometrics.tibs.org.

1. Introduction

Retrospective sampling scheme where one observes the covariates with a probability that depends on the outcome variable has long been used to enhance the study efficiency. The case-control design is the most well-known such design for a binary outcome and a rare disease situation (Cornfield, 1951; Prentice and Pyke, 1979). Using the double sampling strategy, the two-stage case-control design has been shown to further improve efficiency and reduce study costs in epidemiology studies. In a typical two-stage design, disease variable Y is usually observed in the first-stage of sampling, while the covariate is observed in the second-stage of the sampling, and the sampling probability of the second-stage data is dependent on Y. White (1982) proposed a stratified two-stage case-control design for a rare disease and exposure scenario, where a large preliminary random sample is drawn in the first stage, from which strata are identified based on the disease status and the exposure. Further subsamples are then drawn in stage two from the strata identified in stage one and potential confounding variables are then assembled only for those subjects in the stage two subsamples. Greater efficiency may be achieved through the double sampling for stratification, which balances the number of exposed and non-exposed individuals within the case and control samples for whom covariate information is ascertained. The nice features of the two-stage sampling design have generated a great deal of interest in the statistical literature (Zhao and Lipsitz, 1992; Breslow et al., 2003; Wang and Zhou, 2006, 2010, etc). Variations of two-stage sampling based on exposure-and-outcome dependent sampling design have been proposed. Breslow and Cain (1988) extended the design by considering the preliminary sample itself to be separate samples from subpopulations of diseased and non-diseased subjects. They demonstrated that large efficiency can be gained when both the disease and exposure are rare. Similar ideas were also seen in nested case-control samples using the counter-matching method (Langholz et al., 1995) and the “partial-questionnaire-design” (Wacholder and Weinberg, 1994) where investigators try to reduce the burden to study subjects (and consequent reduction in data quality) by ascertaining complete- or partial-questionnaires on different subjects.

Statistical estimation procedures that efficiently combine the information in the first- and second-stage are generally challenging. It should be noted that most current designs have been developed for binary outcomes with a logistic regression. As the scope of biomedical studies inquiry grows, so does the need for efficient study designs and inference procedures to study the determinants of a continuous outcome's level. This need is especially clear when the measurement of primary exposure (X) is expensive, as evidenced with more and more biomedical and genetics studies measuring expensive biomarkers. For example, Gray et al. (2005) studied background-level in utero exposure to the neurodevelopmental toxicant polychlorinated biphenyls (PCBs) in relation to performance on the Bayley Scale of Infant Development (BSID). Maternal pregnancy serum was available from a previously completed cohort study in which BSID had been measured, and PCB concentration in the maternal serum could provide a good surrogate measure of in utero exposure. Biased sampling problems with a continuous outcome variable has been studied recently (Zhou et al., 2002, 2007; Lawless et al., 1999; Chatterjee et al., 2003; Weaver and Zhou, 2005; Song et al., 2009, etc). The design with a continuous outcome variable is called an outcome-

dependent sampling (ODS) design. The principal idea of an ODS design is to concentrate resources on where there is the greatest amount of information.

In this paper, we discuss a new and general two-stage ODS design with a continuous outcome. We allow both prospective data and outcome-dependent data in both stages of the sampling process. We assume that there exists an auxiliary covariate W for primary exposure variable X which is sampled in the first-stage. We handle the marginal distribution of W using the empirical likelihood method. Our proposed method is semiparametric in the sense that the marginal distribution of the covariate W is left unspecified. The proposed method is a likelihood-based approach that profiles out the nuisance distribution via maximization on a restricted likelihood function.

The rest of the article is organized as follows. We describe the two-stage ODS design for a continuous outcome and derive the corresponding likelihood in Section 2. In Section 3, we propose a semiparametric empirical likelihood estimator and establish its asymptotic properties. Extension of the proposed method to allow the strata to depend on both response and covariate is outlined as well. We conduct a simulation study to evaluate the finite-sample behavior and the robustness to model misspecification of the proposed method and to compare its efficiency with some alternative methods in Section 4. We also use the simulation study to show that the proposed design only requires a fraction of the cost to conduct compared with the alternative designs. In Section 5, we illustrate the proposed method with a real data set, followed by a brief remark in Section 6.

2. Data Structure and the Likelihood

2.1 Two-Stage ODS Design with a Continuous Outcome

To fix notation, let Y be a continuous outcome, X be the exposure variable of interest, and W be an auxiliary covariate for X. X and W can be either continuous or discrete variables. By auxiliary, we mean that for a given X, W does not provide any additional information on the relationship between X and Y , i.e., $f(Y|X, W) = f(Y|X)$. We assume that the relationship between X and Y follows a parametric model $f(Y|X; \beta)$, where β is the regression parameter of interest, and that X is linked with W through a model $f(X|W) = f(X|W; \xi)$. We assume that W has a probability distribution function $f_W(\cdot)$ and a cumulative distribution function $F_W(\cdot)$, where both f_W and F_W are unspecified.

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 satisfying $-\infty = c_0 < c_1 < c_2 < \cdots < c_k = \infty$. Thus, these constants partition the study population into K strata. We assume that the first-stage sample consists of a total of n individuals, and the sampling of these individuals follows a two-component ODS scheme of Zhou et al. (2002). Specifically, we assume that, from the underlying population of interests, n_0 individuals are obtained in a simple random sample (SRS) and n_k are from the kth stratum in an outcome-dependent sampling scheme, where $k = 1, \ldots, K$, respectively. That is, the total first-stage sample size is $n = \sum_{k=0}^{K} n_k$. It is assumed that Y and W are observed for the n individuals in the first-stage. In the second-stage, X is observed on a subsample of the first-stage that consists of m individuals. Among

the m individuals, m_0 are obtained from the SRS sample in the first stage and m_k are from the kth stratum, where $k = 1, \ldots, K$, respectively. Thus, $m=\sum_{k=0}^{K} m_k$.

The generality of the proposed two-stage ODS scheme can be seen from the following special situations where it encompasses several commonly encountered designs. For example, when $n_0 \neq 0$, $m_0 \neq 0$ but $n_k = m_k = 0$, $k = 1, \ldots, K$, then the proposed design reduces to the commonly used validation study design. When Y is the discrete and $n_0 = m_0 = 0$, then the design reduces to the setting of Breslow and Cain (1988). Similarly, for a discrete Y, if $n_0 \neq 0$, $m_0 = 0$ and $n_k \neq 0$, $k = 1, 2$, then it reduces to the two-stage design of White (1982).

Without loss of generality, we reorder the sample such that the X values are obtained for the first m_k out of n_k samples, where $k = 0, \ldots, K$. Hence, the observed data structure for our two-stage ODS design can be summarized as:

$$\begin{array}{l}\text{The first}-\text{stage: } \{y_i, w_i\} \quad \text{for} \quad i=1,\ldots,n_0, n_0+1,\ldots,n_0+n_1,\ldots,n. \\ \text{The second}-\text{stage: } \{x_i | y_i \in C_k\} \quad \text{for} \quad i=1,\ldots,m_0, n_0+1,\ldots,n_0+m_1,\ldots,m.\end{array}$$

A general setting where the second stage sample depends on both outcome Y and other covariate but not X, which is denoted by Z, is given in Section 3.3.

2.2 The Likelihood Function

Let $F(u|x; \beta) \equiv P(Y \leq u|x; \beta)$. Define $\psi(y, w; \beta, \xi) \equiv \int f(y|x; \beta) f(x|w; \xi) dx$ and $\varphi_k(w; \beta, \xi) \equiv \int (F(c_k|x; \beta) - F(c_{k-1}|x; \beta)) f(x|w; \xi) dx$, for $k = 1, \ldots, K$. We denote the observed data as

$$\begin{array}{c}(y_{ik}, x_{ik}, w_{ik}), i=1,2,\ldots,m_k, k=1,\ldots,K. \\ (y_{ik}, w_{ik}), i=m_k+1,\ldots,n_k, k=1,\ldots,K.\end{array}$$

The likelihood for the SRS component can be written as

$$\left\{\prod_{i=1}^{m_0} f(y_{i0}|x_{i0}, \beta) f(x_{i0}|w_{i0}, \xi) f(w_{i0})\right\} \times \left\{\prod_{i=m_0+1}^{n_0} \psi(y_{i0}, w_{i0}; \beta, \xi) f(w_{i0})\right\}. \quad (1)$$

For the ODS component, using the Bayes formula, the conditional probability distribution function for those with X observed is

$$f(y, x, w|y \in C_k) = \frac{f(y|x; \beta) f(x|w; \xi) f(w) I(y \in C_k)}{\int \phi_k(w; \beta, \xi) dF(w)}, \quad (2)$$

for $k = 1, \ldots, K$, where $I(\cdot)$ is the indicator function. Similarly, it can be shown that the conditional probability distribution function for those with X unobserved is,

$$f(y, w|y \in C_k) = \frac{\psi(y, w; \beta, \xi) f(w) I(y \in C_k)}{\int \phi_k(w; \beta, \xi) dF(w)}, \quad (3)$$

for $k = 1, \ldots, K$. It follows from (1)–(3) that the likelihood from the our proposed two-stage ODS design is

$$L_n = \left\{ \prod_{i=1}^{m_0} f\left(y_{i0}|x_{i0};\beta\right) f\left(x_{i0}|w_{i0};\xi\right) f\left(w_{i0}\right) \right\} \times \left\{ \prod_{i=m_0+1}^{n_0} \psi\left(y_{i0}, w_{i0};\beta,\xi\right) f\left(w_{i0}\right) \right\} \times \prod_{k=1}^{K} \left\{ \prod_{i=1}^{m_k} \frac{f\left(y_{ik}|x_{ik};\beta\right) f\left(x_{ik}|w_{ik};\xi\right) f\left(w_{ik}\right)}{\int \phi_k\left(w;\beta,\xi\right) dF\left(w\right)} \times \prod_{i=m_k+1}^{n_k} \frac{\psi\left(y_{ik}, w_{ik};\beta,\xi\right) f\left(w_{ik}\right)}{\int \phi_k\left(w;\beta,\xi\right) dF\left(w\right)} \right\}.$$

Clearly, inference of β based the above likelihood function requires some methods of handling f or F . A simple way is to assume a parametric distribution for f, but this could lead to biased conclusions if the underlying model is misspecified. In the next section, we propose a semiparametric empirical likelihood approach to maximize the likelihood without specifying the distribution function of W . Note that a reduced form of likelihood function L_n can be derived if one ignores the information W (see the online Web Appendix A). The estimation algorithm we propose below will lead to a reduced estimator that is not dependent on parametrization of $X|W$.

3. An Empirical Likelihood Approach

3.1 The Inference Algorithm

Denote $\pi_k = \int \varphi_k(w; \beta, \xi) dF(w)$, $k = 1, \ldots, K$. $p_{ik} = f(w_{ik})$, $i = 1, \ldots, n_k$, $k = 0, \ldots, K$. The loglikelihood can be expressed as

$$l_n\left(\beta,\xi,\{p_{ik}\}\right) = l_{1n}\left(\beta,\xi\right) + \sum_{k=0}^{K}\sum_{i=1}^{n_k} log\, p_{ik} - \sum_{k=1}^{K} n_k\, log\, \pi_k, \quad (4)$$

where

$$l_{1n}\left(\beta,\xi\right) = \sum_{k=0}^{K} \left[\sum_{i=1}^{m_k} \left\{ log\, f\left(y_{ik}|x_{ik},w_{ik};\beta\right) + log\, f\left(x_{ik}|w_{ik};\xi\right) \right\} + \sum_{i=m_k+1} log\, \psi\left(y_{ik},w_{ik};\beta,\xi\right) \right]$$

is a function only involving β and ξ. To estimate β, we first profile the log likelihood (4) over p_{ik}, that is, all distributions whose support contains the observed W values. The corresponding profile likelihood is

$$pl_n\left(\beta,\xi\right) \equiv \sup_{\{p_{ik}\}} l_n\left(\beta,\xi,\{p_{ik}\}\right). \quad (5)$$

The estimator for β can thus be obtained by maximizing (5) over β and ξ.

To get (5), it suffices to maximize

$$\ell_{2n}(p_{ik}) \equiv \sum_{k=0}^{K}\sum_{i=1}^{n_k} log\, p_{ik} - \sum_{k=1}^{K} n_k\, log\, \pi_k, \quad (6)$$

for fixed (β, ξ), subject to

$$\sum_{k=0}^{K}\sum_{i=1}^{n_k} p_{ik}\{\phi_j(w_{ik};\beta,\xi) - \pi_j\} = 0, \quad \text{for} \quad j = 1, \ldots, K,$$

and

$$\sum_{k=0}^{K}\sum_{i=1}^{n_k} p_{ik} = 1.$$

Using ideas similar to Qin and Lawless (1994), for a fixed β and ξ, we can show that a unique maximum $\hat{p}_{ik}$ in (6) which satisfies the above constraints exists if 0 is inside the convex hull formed by the points $\{\varphi_j(w; \beta, \xi) - \pi_j\}$ for $j = 1, \ldots, K$. An explicit expression can be derived by the Lagrange multiplier argument:

$$H(\beta,\xi,\{p_{ik}\}) = l_n(\beta,\xi,\{p_{ik}\}) + \rho\left(1 - \sum_{i,k} p_{ik}\right) - n\sum_{j=1}^{K}\lambda_j \sum_{i,k} p_{ik}\{\phi_j(w_{ik};\beta,\xi) - \pi_j\}, \quad (7)$$

where ρ and λ's are Lagrange multipliers. Taking derivatives of H with respect to p_{ik} and solving the score equations, we obtain that $\rho = n$ and

$$\hat{p}_{ik}(\beta,\xi) = \frac{1}{n} \cdot \frac{1}{1 + \sum_{j=1}^{K}\lambda_j\{\phi_j(w_{ik};\beta,\xi) - \pi_j\}}. \quad (8)$$

We plug $\hat{p}_{ik}(\beta,\xi)$ back into $l_n(\beta, \xi, \{p_{ik}\})$ and obtain the estimator for (β, ξ) by maximizing the resultant profile likelihood function. The above procedure enables us to change an infinite dimension problem to a finite dimension problem at the expense of introducing a 2k-dimensional parameters. The Newton-Raphson procedure can be invoked to get $\left(\hat{\beta}, \hat{\xi}\right)$.

3.2 Asymptotic Properties

We reparametrize $\gamma_k = \lambda_k - \frac{n_k}{n\pi_k}$, $k = 1, \ldots, K$. Denote $\zeta = (\beta, \xi)$, $\eta = (\pi, \gamma)$, and $\theta = (\zeta, \eta)$.

The likelihood function now is

$$l_n(\theta) = l_{1n}(\zeta) - \sum_{i,k} log\left[1 + \sum_{j=1}^{K}\lambda_j\{\phi_j(w_{ik};\zeta) - \pi_j\}\right] - \sum_{k=1}^{K} n_k\, log\, \pi_k.$$

Define $h_k(w_{ik}) = \frac{\phi_k(w_{ik};\zeta) - \pi_k}{Q(w_{ik})}$, where $Q(w_{ik}) = \sum_{k=0}^{K} \frac{n_k}{n\pi_k} \phi_k(w_{ik};\zeta)$. The following theorem summarizes the asymptotic properties of the proposed estimator.

Theorem 1—Suppose $|h_k|$ and $|\partial h_k/\partial\theta|$, as functions of θ, are bounded by some integrable function in a neighborhood of the true value $\theta^0 = (\zeta^0, \pi^0, 0)$. *Then* $\sqrt{n}\left(\hat{\theta}_n - \theta^0\right)$ *converges weakly to* $N(0, \Sigma(\theta_0))$ *in a neighborhood of* θ_0, *where* $\Sigma(\theta^0) = V^{-1}(\theta^0)U(\theta^0)V^{-1}(\theta^0)$ *with V and U as defined in the Appendix.*

An outline of the proof for Theorem 1 is given in the Appendix. A consistent estimator of the covariance matrix Σ is $\hat{V^{-1}}\hat{U}\hat{V^{-1}}$, where $\hat{U}$ and $\hat{V}$ are obtained by replacing the large quantities with their corresponding finite sample quantities.

3.3 Extension to Allow the Second Stage Sample to Depend on {Y, Z}

In this subsection, we show that by simply redefining the corresponding components, we can extend the results in Theorem 1 to allow for the selection of the second stage sample to depend on the first stage covariate Z as well as the outcome Y. We assume that the domain of Z is a union of J mutually exclusive intervals $\{B_j\}_{j=1}^{J}$, where $B_j = (b_{j-1}, b_j]$ with b_j's being some prespecified constants such that $-\infty = b_0 < b_1 < \cdots < b_{J-1} < b_J = \infty$. Thus, Y and Z partition the study population into $K \times J$ strata. For notational simplicity, we rewrite these rectangles as Δ_l for $l = 1, \ldots, L$. Hence, $\{C_k \times B_j : k = 1, \ldots, K \text{ and } j = 1, \ldots, J\} = \{\Delta_l : l = 1, \ldots, L\}$ and $\mathscr{Y} \times \mathscr{Z} = \bigcup_{k=1}^{K}\bigcup_{j=1}^{J} C_k \times B_j = \bigcup_{l=1}^{L} \Delta_l$. The first stage sample with $\{Y, Z, W\}$ observed consists of the simple random sample of size n_0 and the outcome Y and covariate Z dependent sample of size n_l conditioning on $\{(Y, Z) \in \Delta_l\}$ for $l = 1, \ldots, L$. Then the second stage sample with X observed is a subsample of first stage sample that consists of m_0 subsamples from n_0 and m_l subsamples from n_l for $l = 1, \ldots, L$. Then the data structure for this two-stage can be summarized as:

$$\begin{aligned}&\text{The first}-\text{stage}: \{y_i, z_i, w_i\} \quad \text{for} \quad i = 1, \ldots, n_0, n_0+1, \ldots, n_0+n_1, \ldots, n.\\ &\text{The second}-\text{stage}: \{x_i | (y_i, z_i) \in \Delta_l\} \quad \text{for} \quad \text{for} \quad i = 1, \ldots, m_0, n_0+1, \ldots, n_0+m_1, \ldots, m.\end{aligned}$$

Redefine $\pi_l = P\{(Y, Z) \in \Delta_l\}$ and define $Q_l(x, z;\beta) \equiv P((Y, z) \in \Delta_l | X = x, Z = z)\, 1_{\Delta_l^*}(z)$ for $l = 1, \ldots, L$, where $\Delta_l^* = \{z \in \mathscr{Z} : \text{for some } (y, z) \in \Delta_l\}$. Thus,

$$Q_l(x, z;\beta) = \int_{\{y:(y,z)\in\Delta_l\}} f(y|x, z;\beta)\, dy.$$

Redefine $\phi_l(z, w;\beta, \xi) = \int_{\mathscr{X}} Q_l(x, z;\beta)\, f(x|z, w;\xi)\, dx$, then

$$\pi_l = \int_{\mathscr{W}} \int_{\mathscr{Z}} \int_{\mathscr{X}} Q_l(x, z;\beta)\, f(x|z, w;\xi)\, dx f(z, w)\, dz dw.$$

Furthermore, redefine $\psi(y,z,w;\beta,\xi)=\int_{\mathscr{X}} f(y|x,z;\beta)\, f(x|z,w;\xi)\, dx$ and $p_{il} = f(z_{il}, w_{il})$ for $i = 1, \ldots, n_l$, $l = 0, \ldots, L$. Then Theorem 1 still holds by replacing the counterparts correspondingly with these redefinitions.

4. Simulation Studies

We conduct simulation studies to assess the small sample performance of the proposed estimator. For all simulation studies, we generated 2000 simulated datasets, each with 500 independent subjects. The data were generated according to the following model,

$$Y=\beta_0+\beta_1 X+\beta_2 Z+\varepsilon_1,$$

where X denotes a continuous exposure variable of interest. We generate $Z\sim\chi^2(1)$. We assume that $X = \xi_0+\xi_1 W+\varepsilon_0$ and $W=\sqrt{Z}+e$, where $\varepsilon_0\sim N\left(0,\sigma_0^2\right)$ and $e \sim N(0, \nu^2)$. That is, $\xi\equiv\left(\xi_0,\xi_1,\sigma_0^2\right)$ and $f(x|w,\xi)\sim N\left(\xi_0+\xi_1 w,\sigma_0^2\right)$. Through some calculations, $f(y|w;\beta,\xi)\sim N\left(\beta_0+\beta_1\xi_0+\beta_1\xi_1 w+\beta_2 w^2,\sigma_1^2+\beta_1^2\sigma_0^2\right)$. We fix $\beta_0 = 1$, $\beta_2 = 1$, $\sigma_1^2=1$, $\xi_0 = 1$, $\xi_1 = 0.5$, $\nu = 0$, and $\sigma_0^2=1$. Our ODS design consists of a SRS sample of 300, a supplement sample of 100 from individuals with Y values in the lower and upper tails of the marginal distribution of Y respectively (i.e., $n_1 = n_3 = 100$, $n_2 = 0$), defined by cutpoints $\mu_Y \pm a\sigma_Y$, where μ_Y and σ_Y represent the mean and standard deviation of Y, respectively. We choose the second-stage data to be a proportion of the first-stage data ($\rho = 0.2, 0.5$), however, it should be noted that one can choose different proportions of the second stage for different strata. We investigate the effect between Y and X by allowing β_1 to take values 0 and 0.5. In addition to various configurations for the parameter values, we investigate the impact of different second-stage sample sizes on parameter estimates. The cutpoints for the ODS design were $\mu_Y \pm a\sigma_Y$, where a is taken to be 0.6, 1, and 1.2, respectively. To ensure convergence we use adjusted empirical likelihood approach proposed by Chen et al. (2008).

For each setting, we compare the proposed estimator $\left(\hat{\beta}_P\right)$ with seven competing estimators: (i) a modified estimated maximum likelihood estimator (EMLE of Weaver and Zhou, 2005) $\left(\hat{\beta}_Y\right)$ that uses the first stage covariates information but does not parametrize $f(X|W)$, (ii) the inverse probability weighted estimator $\left(\hat{\beta}_W\right)$ using only the second-stage data and strata sampling probabilities but no parametrizing $f(X|W)$, (iii) a modified semiparametric empirical likelihood estimator (SPELE of Zhou et al., 2002) $\left(\hat{\beta}_R\right)$ using only the second-stage data with parametrizing $f(X|W)$, (iv) the naive estimator $\left(\hat{\beta}_V\right)$ of using only the simple random sample in the proposed design, (v) the reduced proposed estimator $\left(\hat{\beta}_{P:no\quad W}\right)$ that maximizes the reduced likelihood function L_n(no W) (for details see the online Web Appendix A). This estimator uses the same inference algorithm but with W stripped out. (vi) The proposed estimator under a moderately misspecified model $f(X|W)$ $\left(\hat{\beta}_{P:\text{model}\quad(1)}\right)$,

and (vii) the proposed estimator under a severely misspecified model $f(X|W)$ $\left(\hat{\beta}_{P:\text{model}\quad(2)}\right)$. Table 1 summarizes the similarity and difference among these estimators with special comments on each estimator.

Simulations results are given in Tables 2-4 with Table 3 on the relative efficiency and Table 4 on the sample sizes required for testing $\beta_1 = 0$ at power levels of 80% and 85%, respectively.

4.1 Robustness of $\hat{\beta}_P$

As the proposed method requires a parametric model for $X|W$, it is of practical importance to see how sensitive the resulting inference on the parameters of interest is regards to Y given X to a misspecified nuisance model of $X|W$. We explore this issue with some modifications of the simulation models, where we generate data from model (1): $X = W^{1/3} + \in_0$ and model (2): $X = \log(W^2+1)+\in_0$, respectively. The working model remains to be: $X = \xi_0+\xi_1 W+\in_0$. Other simulation settings remain unchanged. Model (1) characterizes the relations of X with a monotone transformation of W. Hence, model (1) is closer to the working model compared with model (2).

The simulation results reported in Table 2 suggest that the deviation of the working model from the true model will affect the inference on the parameter of interest β_1: the larger the deviation, the larger the bias. For example, when $\beta_1 = 0$, $a = 0.6$ and the second-stage proportion is 0.2, the estimate of β_1 with true model (1) is −0.005 with empirical 95% coverage probability 94.0%. The estimate of β_1 with true model (2) is 0.057 with empirical 95% coverage probability 88.3%.

These observations clearly indicate that careful attention is needed in making and checking for parametrization of $X|W$. Proper transformations may be carried out to ensure the parametrization of $X|W$ is valid. The SRS sample could be used to validate such assumption. Table 2 also demonstrated that the robust estimator $\beta_{P:\text{ no } W}$ is clearly valid and not impacted by the model misspecification of $X|W$. However, it is not as efficient as β_P when the $X|W$ is correctly specified.

Next, we evaluate the performance of the proposed estimator when $X|W$ is correctly specified. We compare it against four remaining estimators β_V, β_W, β_Y, and β_R in the next subsection.

4.2 Efficiency of $\hat{\beta}_P$

For all the cases considered in Table 2, estimators $\hat{\beta}_V$, $\hat{\beta}_W$, $\hat{\beta}_Y$, $\hat{\beta}_R$, $\hat{\beta}_{P:no\quad W}$, and $\hat{\beta}_P$ are all unbiased, the means of the standard error estimates agree well with the sample standard errors and the confidence intervals attain coverage rates close to the nominal 95% level. As evident in Table 2, the proposed estimator $\hat{\beta}_P$ is most efficient one among all estimators compared. As expected, the estimator $\hat{\beta}_V$ is the least efficient one. Both the estimators $\hat{\beta}_Y$ and $\hat{\beta}_R$ are more efficient than the estimator $\hat{\beta}_W$. Further more, efficiency gains associated

with the different second-stage proportions and positions of the cutpoints of estimators $\hat{\beta}_V$, $\hat{\beta}_W$, $\hat{\beta}_Y$, and $\hat{\beta}_R$, relative to $\hat{\beta}_P$ are represented in Table 3.

Note that all entries in Table 3 are less than 1, suggesting that $\hat{\beta}_P$ is the most efficient estimator overall. Another interesting fact is that as *a*, the cutpoint of the ODS design, increases, the relative efficiency of $\hat{\beta}_P$ v.s. $\hat{\beta}_W$ also increases. As the same time, the relative efficiencies of $\hat{\beta}_P$ v.s. $\hat{\beta}_Y$ and $\hat{\beta}_R$ are also increasing but in a much less noticeable manner.

Table 4 shows the sample sizes required to achieve a given power for two local values of β_1 (0.05, 0.15) using two methods β_Y and β_P under the previous simulation settings. The sample sizes are calculated based on the asymptotic normal properties of the corresponding estimators. See the online Web Appendix B for calculation details. Note that although we did not include $\hat{\beta}_W$ in Table 4, the sample sizes for method β_W will be consistently larger than that of β_Y , as β_W has been shown to be consistently less efficient than β_Y . Using our proposed estimator under outcome-dependent sampling requires a smaller sample size. When the cutpoints are $\mu_Y \pm 0.6\sigma_Y$, the proposed method needs about 80% of the subjects who would be needed if the study were conducted with a simple random sampling scheme at the first stage. As the cutpoints are further out, less subjects are needed to achieve a certain power. Furthermore, for a given power, as the true value of β_1 is farther away from 0, relatively fewer subjects are needed to achieve the same power with β_P as compared with β_P , therefore, efficiency gains increase as β_1 is farther away from 0.

5. Analysis of CPP Data

We illustrate the proposed method by analyzing a dataset from the Collaborative Perinatal Project (CPP), a study designed to identify determinants of neurodevelopmental deficits in children (Niswander and Gordon, 1972). Pregnant women were enrolled and data were collected on the mothers at each prenatal visit. The children born into the study were also followed for various outcomes. The investigators are interested in the relationship between *in utero* exposure to polychlorinated biphenyls (PCBs), measured as the third trimester maternal serum PCB level, and cognitive test scores (IQ) at 7 years of age for children (Longnecker et al., 1997). We are mainly interested in the effect of PCB on IQ measurement.

Because of the cost associated with the blood serum assay, the PCB level is measured on a subsample with an ODS scheme from the CPP population. In addition to a random sample of 849 subjects, there are two supplemental subgroups which are defined by children's IQ scores that are one standard deviation above and below the mean of the population IQ scores, with 81 subjects in the low IQ group and 108 subjects in the high IQ group.

The two-stage ODS setting in the CPP study were created as follows. The first-stage data consists of 1038 subjects. In the second-stage, 534 measurements of PCB out of 849, 51 out of 81, and 72 out of 108 in SRS and two tails are randomly taken, respectively. The socioeconomic status (SES) is a continuous variable distributed from 0 to 9.5. We defined a surrogate of SES by discretizing SES into three levels: low (0-3), medium (4-6) and high

(6-9.5). Additional covariates considered to be potential confounders include the highest education level attained by the mother at the time of the child's birth (EDU), the mother's age in years at the time of the child's birth (AGE), the race of the child (WHITE and BLACK) and the gender of the child (SEX). The covariate SEX was coded 0 for males and 1 for females. The model we fit is

$$IQ=\beta_0+\beta_1 PCB+\beta_2 SES+\beta_3 EDU+\beta_4 AGE+\beta_5 WHITE+\beta_6 BLACK+\beta_7 SEX+\varepsilon_1,$$

where ε_1 is assumed to be a normal error with zero mean. The relationship between PCB and the discretized SES is set to be $PCB = \xi_0 + \xi_1 W + \varepsilon_0$, and the error term ε_0 is normally distributed.

We first explored the relationship between the discretized SES and PCB based on the SRS data. The resulting lowess curve indicates a linear association between discretized SES and PCB, which is further verified by a linear model fit with the estimate of slope as 0.154 ($p <$ 0.0001). A scatter plot of PCB vs. SES was provided in the online Web Appendix C.

Table 5 summarizes the results from the same three methods we evaluated in the simulation study. The β_W is not calculated as we do not have the strata proportion here. First, note that the three estimators provided similar point estimates for the regression parameters. The differences in these estimators are in the precisions associated with these estimates. Using the covariate AGE as an example, the β_V is the least efficient method with an estimated standard error for AGE at 0.120. The estimated standard error for AGE is 0.089 for β_R and is 0.055 for β_P . The proposed method, β_P , which takes advantage of both two-stage and ODS design, is the most efficient estimator. Again, using AGE as an example, its relative efficiencies relative to β_V and β_R are 4.76 and 2.62, respectively. These observations are consistent with the results from the simulation study. Overall, the results from the three methods agree well. We observe that SES, EDU, and WHITE were all significantly associated (at the 0.05 level) with IQ, whereas AGE, BLACK, and SEX were not. Although the effect of PCB on IQ is also not significant, the proposed method does provide a tighter 95% nominal confidence interval for the effect of PCB.

6. Remarks

In this article, we proposed a two-stage ODS design for a continuous outcome and exposure variables and developed a semiparametric empirical likelihood-based method to analyze data from a such two-stage ODS design. This proposed method is robust to the misspecification of the probability distribution of the auxiliary covariate W . The proposed estimator has the usual asymptotic normality property. Simulation studies show that the proposed estimators are more efficient than existing methods.

In many practical settings, investigators choose a two-stage design because of budget limitations. Suppose the total budget available for the study is B, denoting the cost of each first-stage observation C_1 and the additional cost of ascertaining second-stage data for a subject C_2. It can be seen that $B = nC_1 + mC_2$. With the budget fixed at B, the optimal design is the study size n and the second-stage sampling fractions $\{r_k\}$, $k = 1, \ldots, K$, which

minimize the variance of $\hat{\beta}_j$, where β_j is the jth entry of β which is of primary interest. It would be worthwhile in the future to derive the seconde-stage sampling fractions under such optimal design with an outcome-dependent sampling scheme in this context.

We would like to stress the importance of careful model checking and model building for $f(X|W)$ when using the proposed method, as failure to do so may lead to biased parameter estimates. In this regard, design with SRS sample in the first stage (i.e., $n_0 > 0$) is useful in helping correctly identifying a parametric model for $f(X|W)$. Future research for a fully nonparametric treatment of $f(X|W)$ is certainly warranted.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank the editor, the associate editor, and the referees for their thoughtful comments and constructive suggestions. This research is supported by the National Institute of Health grant R01 CA 079949 (H.Z., R.S., and Y.W.).

Appendix

Proof of Theorem 1.

Let $l_n(\theta) = l_{n1}(\theta) + l_{n2}(\theta)$, where

$$l_{n1}(\theta) = \sum_{k=0}^{K} \left[\sum_{i=1}^{m_k} \{ log\, f(y_{ik}|x_{ik}, w_{ik}; \beta) + log\, f(x_{ik}|w_{ik}; \xi) \} + \sum_{i=m_k+1}^{n_k} \psi(y_{ik}, w_{ik}; \beta, \xi) \right] - \sum_{k=1}^{K} n_k\, log \quad \pi_k$$

and $l_{n2}(\theta) = -\sum_{k=1}^{K} \sum_i log\,(1+\gamma_k(\theta)\, h_k(z_{ik})) - \sum_{k=1}^{K} log\, Q(z_{ik})$.

The consistency proof is similar to the proof of Theorem 4.1 in Lehmann (1983). We sketch the proof in the following. We will show (a) $l_{n1}(\theta^0) > l_{n1}(\theta)$ a.s. and (b) $l_{n2}(\theta) > l_{n2}(\theta)$ a.s. for $\theta = \theta^0 + un^{-1/3}$ to get the conclusion that $l_n(\theta)$ has a local maximum inside the ball $\| \theta - \theta^0 \| \leq n^{-1/3}$. Part (a) can be proved through a Taylor expansion around θ^0 in a $n^{1/3}$ neighborhood of θ^0. We note that the essential part is to notice that $E[\partial^2 l_{n1}(\theta^0)/\partial\theta\partial\theta']$ is negative definite. For part (b), to show $l_{n2}(\theta^0) > l_{n2}(\theta)$ a.s., we follow closely the proof of Lemma 1 in Qin and Lawless (1994). Based on (a) and (b), we conclude $l_n(\theta^0) > l_n(\theta)$ a.s. for $\theta = \theta^0 n^{-1/3}$. Since $l_n(\theta)$ is a continuous function in θ as θ belongs to the ball $\|\theta - \theta^0\| \leq n^{-1/3}$, $l_n(\theta)$ has a local maximum in a small neighborhood of θ^0. The consistency is achieved by the smoothness of the likelihood function.

Now we prove the asymptotic normality. The first derivative of $l_n(\theta)$ with respect to θ is

$$S_n(\theta)=\frac{\partial l_n(\theta)}{\partial\theta}=\sum_{k=0}^{K}\left\{\sum_{i=1}^{m_k}u_{ik}(y_{ik},x_{ik},w_{ik};\theta)+\sum_{i=m_k+1}^{n_k}v_{ik}(y_{ik},w_{ik};\theta)\right\},$$

where

$$u_{ik}(y_{ik},x_{ik},w_{ik};\theta)=\begin{pmatrix}\frac{\partial log\, f(y_{ik}|x_{ik},w_{ik};\beta)}{\partial\zeta}+\frac{\partial log\, f(x_{ik}|w_{ik};\xi)}{\partial\zeta}-\frac{\partial Q(w_{ik},\zeta)/\partial\zeta}{Q(w_{ik};\zeta)}-\frac{\gamma_k\partial h_k(w_{ik};\theta)/\partial\zeta}{1+\gamma_k h_k(w_{ik};\theta)}\\ -1\{k>0\}\frac{n_k}{\pi_k}-\frac{\partial Q(w_{ik};\theta)/\partial\pi_k}{Q(w_{ik};\theta)}-\frac{\gamma_k\partial h_k(w_{ik};\theta)/\partial\pi_k}{1+\gamma_k h_k(w_{ik};\theta)}\\ -\frac{h_k(w_{ik};\theta)}{1+\gamma_k h_k(w_{ik};\theta)}\end{pmatrix}$$

and By the law of large numbers, one can show

$$\frac{S_n(\theta)}{n}=\sum_{k=0}^{K}\frac{n_k}{n}\left(\frac{m_k}{n_k}\sum_{i=1}^{m_k}\frac{u_{ik}(\theta)}{m_k}+\frac{n_k-m_k}{n_k}\sum_{i=m_k+1}^{n_k}\frac{v_{ik}(\theta)}{n_k-m_k}\right)\to_p\sum_{k=0}^{K}\rho_k\{r_kE(u_{1k})+(1-r_k)E(v_{1k})\}\equiv s(\theta),$$

where $\rho_k = \lim_{n\to\infty} n_k/n$ is the limit of the first-stage proportion, and $r_k = \lim_{n\to\infty 1} m_k/n_k$ is the limit of the second-stage proportion. When $r_k = 1$, $k = 0, 1, \ldots, K$, our likelihood has the same form as that in Zhou et al. (2002). When evaluated at θ^0, we can show $s(\theta^0) = 0$.

Similarly, we have $V_n(\theta)=\frac{1}{n}\frac{\partial^2 l_n(\theta)}{\partial\theta\partial\theta'}$ converges to $V(\theta)$ in probability, where

$$V(\theta)=\sum_{k=0}^{K}\rho_k\left\{r_kE\left(\frac{\partial u_{ik}(\theta)}{\partial\theta}\right)+(1-r_k)E\left(\frac{\partial v_{ik}(\theta)}{\partial\theta}\right)\right\}.$$

It follows from central limit theorem that $\sqrt{n}S_n\left(\hat{\theta}_n\right)$ converges in distribution to $N(0, U(\theta))$, where

$$U(\theta)=\sum_{k=0}^{K}\rho_k(r_kcov(u_{ik})+(1-r_k)cov(v_{ik})).$$

Expanding $S_n(\theta)$ at the true value θ^0, we have $\sqrt{n}\left(\hat{\theta}_n-\theta^0\right)=V_n(\theta)^{-1}\frac{1}{\sqrt{n}}S_n(\theta)+o_p(1)$. Thus Theorem 1 holds by Slutsky's theorem. A consistent estimator of Σ is $\hat{\Sigma}=\hat{V}^{-1}\hat{U}\hat{V}^{-1}$, where $\hat{V}=V_n\left(\hat{\theta}_n\right)$, and $\hat{U}=\frac{1}{n}\sum_{k,i}S_n\left(\hat{\theta}_n\right)$, by consistency results and the continuous mapping

References

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.

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

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

Chen J, Mulayath A, Abraham B. Adjusted empirical likelihood and its properties. Journal of Computational and Graphical Statistics. 2008; 17:426–443.

Cornfield J. A method of estimating comparatice rates from clinical data. Journal of the National Cancer Institute. 1951; 11:1269–1275. [PubMed: 14861651]

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

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

Lawless JF, Kalbfleisch JD, Wild CJ. Semiparametric methods for response-selective and missing data problems in regression. Journal of the Royal Statistical Society, Series B. 1999; 61:413–438.

Lehmann, EL. Theory of Point Estimation. Springer-Verlag; New York: 1983.

Longnecker, M.; Klebanoff, M.; Zhou, H.; Wilcox, A.; Berendes, H.; Hoffman, H. Study Proposal. National Institiute of Environmental Health Sciences; Washington, D.C.: 1997. Proposal to study in utero exposure to dde and pcbs in relation to male birth defects and neurodevelopmental outcomes in the collaborative perinatal project..

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

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

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

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

Wacholder S, Weinberg CR. Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling. Biometrics. 1994; 50:350–357. [PubMed: 8068835]

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, in press. 2010

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

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

Zhao LP, Lipsitz S. Designs and analysis of two-stage studies. Statistics in Medicine. 1992; 11:769–782. [PubMed: 1594816]

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

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

Table 1

Summary for different methods compared in stimulation study.

Method	Design 1st/2nd	Stage of data used in inference	Parametrization $f(x\|w)$	Comment
β_P	ODS/ODS	1st and 2nd	Yes	proposed method
β_R	ODS/ODS	2nd only	Yes	modified SPELE
β_W	SRS/ODS	2nd only	No	sampling probability needed
β_Y	SRS/ODS	1st and 2nd	No	modified EMLE
$\beta_{P:\text{ no } W}$	ODS/ODS	1st and 2nd	No	reduced $\hat{\beta}_P$, no W used
$\beta_{P:\text{ model (1)}}$	ODS/ODS	1st and 2nd	Yes	moderately misspecified $X\|W$
$\beta_{P:\text{ model (2)}}$	ODS/ODS	1st and 2nd	Yes	severely misspecified $X\|W$

Table 2

Simulation results. Results are based on 2000 simulations with $n_0 = 300$, $n_1 = n_3 = 100$.

			$\hat{\beta}_1$				$\hat{\beta}_2$			
a	β_1	Methods	Mean	SE	$\widehat{\text{SE}}$	C.I.	Mean	SE	$\widehat{\text{SE}}$	C.I.
0.6	0.5	β_V	0.500	0.132	0.131	0.947	1.006	0.445	0.444	0.950
		β_W	0.503	0.102	0.100	0.939	1.000	0.067	0.066	0.939
		β_Y	0.502	0.080	0.081	0.952	1.000	0.049	0.049	0.949
		β_R	0.497	0.084	0.085	0.940	1.005	0.062	0.063	0.941
		β_P	0.501	0.066	0.067	0.945	1.003	0.030	0.030	0.946
		$\beta_{P:\text{ no }W}$	0.499	0.094	0.090	0.945	1.005	0.067	0.064	0.944
		$\beta_{P:\text{ model (1)}}$	0.515	0.054	0.052	0.955	1.002	0.030	0.030	0.959
		$\beta_{P:\text{ model (2)}}$	0.451	0.067	0.071	0.903	1.123	0.033	0.031	0.930
	0	β_V	0.000	0.133	0.132	0.944	1.005	0.448	0.440	0.953
		β_W	0.001	0.099	0.099	0.940	1.003	0.065	0.066	0.947
		β_Y	0.002	0.091	0.091	0.940	1.000	0.044	0.045	0.952
		β_R	–0.003	0.082	0.084	0.943	1.003	0.062	0.063	0.940
		β_P	–0.001	0.064	0.064	0.939	1.002	0.027	0.027	0.951
		$\beta_{P:\text{ no }W}$	0.000	0.077	0.083	0.960	1.002	0.065	0.064	0.947
		$\beta_{P:\text{ model (1)}}$	–0.005	0.046	0.045	0.940	1.000	0.026	0.027	0.953
		$\beta_{P:\text{ model (2)}}$	0.057	0.081	0.081	0.883	0.997	0.028	0.028	0.952
1	0.5	β_W	0.501	0.102	0.102	0.935	1.004	0.066	0.064	0.933
		β_Y	0.498	0.081	0.080	0.942	1.001	0.049	0.049	0.947
		β_R	0.509	0.081	0.079	0.953	1.001	0.056	0.056	0.948
		β_P	0.505	0.064	0.064	0.947	1.002	0.028	0.029	0.940
		$\beta_{P:\text{ no }W}$	0.502	0.080	0.085	0.961	1.007	0.059	0.059	0.958
		$\beta_{P:\text{ model (1)}}$	0.519	0.040	0.051	0.951	1.001	0.020	0.024	0.960
		$\beta_{P:\text{ model (2)}}$	0.465	0.070	0.066	0.913	1.122	0.030	0.029	0.011
	0	β_W	–0.001	0.106	0.100	0.929	1.004	0.062	0.063	0.937
		β_Y	–0.002	0.087	0.086	0.945	1.002	0.044	0.045	0.949

			$\hat{\beta}_1$				$\hat{\beta}_2$			
a	β_1	**Methods**	**Mean**	**SE**	$\widehat{\mathbf{SE}}$	**C.I.**	**Mean**	**SE**	$\widehat{\mathbf{SE}}$	**C.I.**
		β_R	–0.003	0.077	0.077	0.950	1.002	0.056	0.056	0.950
		β_P	0.005	0.061	0.063	0.947	1.002	0.029	0.029	0.940
		$\beta_{P:\text{ no } W}$	0.001	0.094	0.091	0.939	0.997	0.063	0.065	0.960
		$\beta_{P:\text{ model (1)}}$	0.000	0.044	0.043	0.958	1.002	0.026	0.025	0.942
		$\beta_{P:\text{ model (2)}}$	0.091	0.078	0.074	0.754	0.995	0.027	0.026	0.946

NOTE: Results are based on the model $Y = \beta_0 + X\beta_1 + W^2\beta_2 + \varepsilon$; the true parameter values are $\beta_0 = 0$ and 0.5 respectively, and $\beta_2 = 1$. The outpoints for the ODS design were $\mu_Y \pm a\sigma_Y$. The second-stage proportion for all strata are all 0.2. Explanations of estimators are given in Table 1.

Table 3

Empirical relative efficiencies of the simulation studies.

Cut points	Second-stage proportion	Methods	Model			
			$\beta_1 = 0$		$\beta_1 = 0.5$	
			$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
0.6	0.5	β_V	0.573	0.097	0.566	0.104
		β_W	0.758	0.643	0.758	0.659
		β_Y	0.810	0.587	0.855	0.604
		β_R	0.904	0.692	0.887	0.763
	0.2	β_V	0.607	0.080	0.653	0.087
		β_W	0.646	0.415	0.647	0.448
		β_Y	0.703	0.614	0.825	0.612
		β_R	0.780	0.435	0.786	0.484
1	0.5	β_V	0.531	0.092	0.543	0.094
		β_W	0.698	0.650	0.714	0.634
		β_Y	0.786	0.590	0.849	0.542
		β_R	0.917	0.722	0.882	0.722
	0.2	β_V	0.428	0.057	0.482	0.063
		β_W	0.575	0.468	0.627	0.424
		β_Y	0.701	0.659	0.790	0.571
		β_R	0.792	0.518	0.790	0.500
1.2	0.5	β_V	0.516	0.087	0.536	0.093
		β_W	0.667	0.600	0.700	0.619
		β_Y	0.778	0.571	0.830	0.578
		β_R	0.875	0.706	0.880	0.743
	0.2	β_V	0.414	0.070	0.617	0.077
		β_W	0.514	0.375	0.612	0.394
		β_Y	0.655	0.533	0.778	0.522
		β_R	0.733	0.453	0.778	0.473

NOTE: The relative efficiency is defined as the ratio of the empirical standard error of $\hat{\beta}_P$ over the corresponding estimator. For other settings see footnote of Table 2.

Table 4

Sample size needed for testing $H_0 : \beta_1 = 0$ for a given power for models in the simulation studies.

Second-stage proportion	Power	True β_1	Sample size					
			β_Y			β_P		
			$a = 0.6$	$a = 1$	$a = 1.2$	$a = 0.6$	$a = 1$	$a = 1.2$
0.5	0.8	0.05	5351	4978	4573	3468	3039	2769
		0.15	573	527	507	385	338	308
	0.85	0.05	5968	5378	4973	3967	3476	3168
		0.15	681	619	585	441	386	352
0.2	0.8	0.05	12507	11163	10356	6430	5841	4749
		0.15	1324	1257	1094	714	649	528
	0.85	0.05	13485	13047	11732	7355	6682	5432
		0.15	1576	1406	1337	817	742	604

Table 5

The analysis of CPP data.

Covariate	β_V	β_R	β_P
Intercept	67.948/5.593*	69.991/4.325*	70.302/2.732*
PCB	0.134/0.472	0.104/0.348	0.191/0.327
SSES	1.078/0.364*	1.349/0.268*	1.550/0.169*
EDU	3.606/1.342*	3.520/0.997*	2.728/0.604*
AGE	0.038/0.120	0.040/0.089	0.019/0.055
WHITE	16.957/3.867*	11.669/2.909*	10.010/1.770*
BLACK	7.207/3.806	2.029/2.869	0.130/1.749
SEX	−0.409/1.454	−0.738/1.094	−0.390/0.674

NOTE: The estimates and the standard error of the CPP data are recorded in the form “estimate/standard error”. We mark “*” to mean that the corresponding parameter estimate is significant at 5% level.