Published in final edited form as:

Stat Med. 2018 July 10; 37(15): 2321–2337. doi:10.1002/sim.7672.

Secondary outcome analysis for data from an outcome-dependent sampling design

Yinghao Pan[1], **Jianwen Cai**[1], **Matthew P. Longnecker**[2], and **Haibo Zhou**[1]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[2]Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA

Abstract

Outcome-dependent sampling (ODS) scheme is a cost-effective way to conduct a study. For a study with continuous primary outcome, an ODS scheme can be implemented where the expensive exposure is only measured on a simple random sample and supplemental samples selected from 2 tails of the primary outcome variable. With the tremendous cost invested in collecting the primary exposure information, investigators often would like to use the available data to study the relationship between a secondary outcome and the obtained exposure variable. This is referred as secondary analysis. Secondary analysis in ODS designs can be tricky, as the ODS sample is not a random sample from the general population. In this article, we use the inverse probability weighted and augmented inverse probability weighted estimating equations to analyze the secondary outcome for data obtained from the ODS design. We do not make any parametric assumptions on the primary and secondary outcome and only specify the form of the regression mean models, thus allow an arbitrary error distribution. Our approach is robust to second- and higher-order moment misspecification. It also leads to more precise estimates of the parameters by effectively using all the available participants. Through simulation studies, we show that the proposed estimator is consistent and asymptotically normal. Data from the Collaborative Perinatal Project are analyzed to illustrate our method.

Keywords

biased sampling; estimating equation; missing data; secondary analysis; semiparametric estimation; validation sample

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

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

CONFLICT OF INTEREST

None declared.

ORCID

Yinghao Pan http://orcid.org/0000-0002-4022-1815

1 | INTRODUCTION

In many epidemiology studies, the primary outcome variable is easy to obtain, while some exposure variables are expensive or difficult to measure. This motivates statisticians to develop outcome-dependent sampling (ODS) designs, in which the selection probability depends on the primary outcome variable. The main idea of such ODS designs is to concentrate resources on those participants that are more informative in explaining the outcome/exposure relationship. The case-control design has been widely used for studies with a binary primary outcome.[1] Prentice[2] proposed a case-cohort study design for failure time regression analysis. Zhou et al[3] considers an ODS design for data with a continuous primary outcome: In their design, in addition to a simple random sample (SRS) from the full cohort, 2 supplemental SRSs are drawn from 2 tails of the outcome distribution. The initial SRS from the entire cohort provides information about the overall population, and supplemental samples allow investigators to oversample those participants that are more informative about the exposure-response relationship. One example of such ODS design is the Collaborative Perinatal Project (CPP).[3,4] The main purpose of CPP is to study the relationship between in utero exposure to polychlorinated biphenyls (PCBs) and multiple neurological outcomes, including children's IQ performance. As PCB level is expensive to ascertain, an ODS scheme is adopted: An SRS is taken, and 2 supplemental samples are chosen from 2 tails of the IQ distribution. Related works on ODS to evaluate the association between expensive exposure and the primary outcome variable include Zhou et al,[3,10,11] Weaver and Zhou,[5] Wang and Zhou,[6,8] Song et al,[7] and Qin and Zhou.[9]

In any real studies, it is typical that there are more than 1 endpoint of interest. As such, investigators would like to reuse the ODS data to study the association between a secondary outcome and the obtained exposure variable. For example, in the CPP data, investigators are also interested in examining the relationship between PCB level and children's birth weight. Many prior studies have tried to assess the association between these 2 measures, and yet so far have failed to reach a consistent conclusion.[12–18] With CPP data collected in the first place using an ODS design to evaluate children's IQ and PCB level, we are interested in adding some evidence to this research problem by developing a valid and precise method for secondary analysis under ODS designs.

In this paper, we develop a method for conducting secondary analysis under continuous outcome ODS design described by Zhou et al.[3] As the data obtained from ODS design is not a random sample of the overall population, performing secondary analysis is not straightforward. Ignoring the biased sampling nature of the data could yield an invalid estimate of the true parameters in the general population. The analysis restricted to the participants in the SRS portion is clearly inefficient as it underuses the available data. A significant amount of work was done on secondary analysis in case-control data. This includes the likelihood-based methods,[19–21] inverse probability weighting (IPW),[22,23] and estimating equation.[24,25] However, to the best of our knowledge, there has been no research conducted on the secondary regression analysis in the continuous outcome ODS design framework.

We propose estimating equation approaches to analyze a secondary outcome for data obtained from an ODS design with a continuous primary outcome. The advantage of our approach is that no additional model assumptions are specified. The augmented estimating equation utilizes the available information in the full cohort, and hence increases estimation precision. In addition, our method is computationally stable and fast. The organization of the paper is as follows. In Section 2, we present some notations, data structure and our model under ODS designs. In Section 3, we propose two estimating equations, IPW estimating equation and augmented IPW estimating equation. We give the corresponding asymptotic properties in Section 4. In Section 5, we present the simulation results that compare our proposed estimator to other competing estimators. In Section 6, we apply our methods to CPP data to study the relationship between children's birth weight and maternal PCB level. We conclude this paper by a brief discussion in Section 7.

2 | DATA STRUCTURE AND MODEL

To fix notation, let Y_1 be the primary continuous outcome variable that the ODS sampling scheme is based on. Let X be the expensive exposure, which is only observed for some participants, and Z be the vector of other covariates that are easy to obtain. Furthermore, let Y_2 denote a continuous secondary response. Our interest lies in inference of the secondary response Y_2 with respect to X adjusting for other covariates Z for data obtained from continuous outcome ODS design.

We partition the domain of Y_1 into a union of 3 mutually exclusive intervals: $A_1 \cup A_2 \cup A_3 = (-\infty, a] \cup (a, b] \cup (b, +\infty)$. We assume that the underlying data $\{(Y_1, Y_2, X, Z), i = 1, \ldots, N\}$ are independent and identically distributed random vectors, with N be the size of the full cohort. The ODS design proposed by Zhou et al[3] can be regarded as a 2-phase design: In the first phase, information on primary outcome, secondary outcome, and inexpensive covariates are observed for each member of the full cohort. That is, we observe $\{(Y_{1i}, Y_{2i}, Z_i), i = 1, \ldots, N\}$. In the second phase, the expensive exposure X is measured on an SRS of size n_0 from the full cohort and 2 supplemental SRSs drawn from 2 tails of the distribution of Y_1, ie, supplemental sample of size n_1 from $\{Y_1 \in A_1\}$ and supplemental sample of size n_3 from $\{Y_1 \in A_3\}$. Let V_0, V_1, V_3 be the index set of SRS, supplemental sample taken from $\{Y_1 \leq a\}$, and supplemental sample taken from $\{Y_1 > b\}$, respectively. That is to say, we observe $\{X_i, i \in V_0 \cup V_1 \cup V_3\}$ in the second phase. Here, the sample sizes n_0, n_1, and n_3 are fixed by design. Note that we use fixed-size sampling (sampling without replacement) for both the initial SRS and the supplemental samples. When stratum sizes (ie, number of participants in A_1 and A_3) are very large, it is equivalent to independent Bernoulli sampling, as the stratum specific sampling probabilities are effectively fixed.

Let $V = V_0 \cup V_1 \cup V_3$, and let n_V be the size of V. Then, $n_V = n_0 + n_1 + n_3$. Using terminology from measurement error literature, these n_V observations are called validation sample. In addition, we let $n_{\bar{V}} = N - n_V$. We refer to the $n_{\bar{V}}$ observations as the nonvalidation sample because expensive exposure X is not measured for these individuals. Let $\bar{V}$ represent the index set of the nonvalidation sample, and r_i be the indicator variable of observing X for participant i, then $V = \{i : r = 1\}$ and $\bar{V} = \{i : r_i = 0\}$.

The data structure for the ODS design can be summarized as the following:

$$\begin{array}{lll} \text{First phase :} & & \{Y_{1i}, Y_{2i}, Z_i\}, \quad i = 1, \cdots, N; \\ \text{Second phase :} & \{\text{SRS}\} & \{X_i\}, \quad i \in V_0; \\ & \{\text{supplemental sample 1}\} & \{X_i | Y_{1i} \in A_1\}, \ i \in V_1; \\ & \{\text{supplemental sample 2}\} & \{X_i | Y_{1i} \in A_3\}, \ i \in V_3 . \end{array} \quad (1)$$

Let $\mu_{1i} = \mathrm{E}(Y_{1i} | X_i, Z_i)$ and $\mu_{2i} = \mathrm{E}(Y_{2i} | X_i, Z_i)$ denote the conditional expectation of Y_{1i} and Y_{2i} given the covariates, respectively. In most problems, we are interested in estimating the regression coefficients (β, γ) from the following 2 models:

$$\mu_i = \begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix} = \begin{pmatrix} \mathrm{E}(Y_{1i} | X_i, Z_i) \\ \mathrm{E}(Y_{2i} | X_i, Z_i) \end{pmatrix} = \begin{pmatrix} g_1^{-1}(\beta_0 + \beta_1 X_i + \beta_2 Z_i) \\ g_2^{-1}(\gamma_0 + \gamma_1 X_i + \gamma_2 Z_i) \end{pmatrix}, \quad (2)$$

where $g_1(\cdot)$ and $g_2(\cdot)$ are specified link functions, such as $g(x) = x$ for linear regression. Without loss of generality, we use the identity link $g_1(x) = g_2(x) = x$ to illustrate our ideas throughout the paper. It is also worth mentioning that no distributional assumptions are made about y_{1i} and y_{2i}. Since analysis on secondary outcome is our primary goal, we focus on developing an inference procedure for $(\gamma_0, \gamma_1, \gamma_2)$.

3 | ESTIMATING EQUATION APPROACH

3.1 | Inverse probability weighted estimating equation

Let $\xi = (\beta, \gamma)$. Since we do not make any parametric assumptions about Y_1 and Y_2, no likelihood-based approaches are available. Let $e_i = (e_{1i}, e_{2i})' = (Y_{1i} - \mu_{1i}, Y_{2i} - \beta_2)'$. Following the ideas from Horvitz and Thompson, Liang and Zeger, and Zhao et al,[26–28] we first propose an IPW estimating equation that uses the validation sample only:

$$S_1(\xi, Q, \pi) = \sum_{i \in V} s_{1i}(\xi, Q, \pi) = \sum_{i \in V} \frac{1}{\pi_i} D_i^T Q^{-1} e_i = \sum_{i=1}^{N} \frac{r_i}{\pi_i} D_i^T Q^{-1} e_i = 0, \quad (3)$$

where Q is the covariance matrix of (Y_1, Y_2), ie, $Q = \mathrm{Cov}(Y_1, Y_2)$, π_i is the probability of being selected into the validation sample for each participant i, and

$$D_i = \frac{\partial \mu_i}{\partial (\beta, \gamma)^T} = \begin{pmatrix} 1 & X_i & Z_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & X_i & Z_i \end{pmatrix}.$$

The selection probability π_i is a function of the observed outcome value Y_{1i}. Let $V_{0,k}$ be the index set of the observations in the SRS that belongs to the kth stratum A_k. That is, $V_0 = V_{0,1} \cup V_{0,2} \cup V_{0,3}$. Then, π_i can be expressed as follows:

$$\pi_i = Pr\left(r_i = 1 \middle| Y_{1i}\right) = \begin{cases} Pr\left(i \text{ in } V_{0,1} \text{ or } V_1\right) & \text{if } Y_{1i} \le a, \\ Pr\left(i \text{ in } V_{0,2}\right) & \text{if } a < Y_{1i} \le b, \\ Pr\left(i \text{ in } V_{0,3} \text{ or } V_3\right) & \text{if } Y_{1i} > b. \end{cases}$$

For complex sampling designs, such as the ODS design described in this paper, it is difficult to express π_i in explicit forms. We cannot directly solve Equation 3 as the covariance matrix Q, and the selection probability π is unknown. Hence, the general idea is to plug in consistent estimators of Q and π into $S_1(\xi, Q, \pi)$ to get $\hat{S}_1(\xi) = S_1\left(\xi, \hat{Q}, \hat{\pi}\right)$ and then solve the equation $\hat{S}_1(\xi) = 0$.

Because Y_1 and Y_2 are observed for each member of the full cohort, a consistent estimator of the covariance matrix Q is the sample covariance derived from the full cohort. That is,

$$\hat{Q} = \begin{pmatrix} \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)^2 & \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)\left(Y_{2i} - \bar{Y}_2\right) \\ \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)\left(Y_{2i} - \bar{Y}_2\right) & \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{2i} - \bar{Y}_2\right) \end{pmatrix},$$

where $\bar{Y}_1$ and $\bar{Y}_2$ are sample means for Y_1 and Y_2, respectively.

Let N_k, $k = 1, 2, 3$ be the number of observations in the full cohort that belong to the kth stratum A_k. Similarly, let $n_{0,k}$, $k = 1, 2, 3$ be the number of observations in the SRS that belong to stratum A_k. That is, $N = N_1+N_2+N_3$, $n_0 = n_{0,1}+n_{0,2}+n_{0,3}$. Then for each participant, the observed probability of being sampled within its respective strata A_k can be written as

$$\hat{\pi}_i = \begin{cases} \left(n_{0,1} + n_1\right)/N_1 & \text{if } Y_{1i} \le a, \\ n_{0,2}/N_2 & \text{if } a < Y_{1i} \le b, \\ \left(n_{0,3} + n_3\right)/N_3 & \text{if } Y_{1i} > b. \end{cases}$$

It is straightforward to show that $\hat{\pi}_i$ is a consistent estimator for π_i. Hence, our first proposed estimator $\hat{\xi}_{IPW}$ satisfies the following estimating Equation 4 and can be obtained using Newton-Raphson algorithm.

$$\hat{S}_1(\xi) = \sum_{i \in V} \hat{s}_{1i}(\xi) = \sum_{i \in V} \frac{1}{\hat{\pi}_i} D_i^T \hat{Q}^{-1} e_i = \sum_{i=1}^{N} \frac{r_i}{\hat{\pi}_i} D_i^T \hat{Q}^{-1} e_i = 0. \quad (4)$$

3.2 | Augmented inverse probability weighted estimating equation

The weighted estimating equation 4 described above may not be precise as it uses only the information contained in the validation sample where the expensive exposure X is observed.

Following the ideas from Robins et al,[29] an augmented estimating equation can be used. Let u_i be any kernel function, and let $h(y_{1i}, y_{2i}, z_i)$ be any function, then

$$\sum_{i=1}^{N}\left[\frac{r_i}{\pi_i}u_i+\left(1-\frac{r_i}{\pi_i}\right)h\left(y_{1i},y_{2i},z_i\right)\right]=0$$

is an augmented estimating equation. Any choice of $h(\cdot)$ would lead to a consistent estimate of the parameters. This comes from the fact that

$$E\left[\left(1-\frac{r_i}{\pi_i}\right)h\left(y_{1i},y_{2i},z_i\right)\right]=E_{y_{1i},y_{2i},z_i}\left[h\left(y_{1i},y_{2i},z_i\right)E_{r_i|y_{1i},y_{2i},z_i}\left(1-\frac{r_i}{\pi_i}\right)\right]=0.$$

However, an optimal choice of $h(\cdot)$ would improve the estimation precision. From Robins et al,[29] it is shown that the optimal $h(\cdot)$ should be the conditional expectation of the kernel function given the observed data, ie, $h(y_{1i}, y_{2i}, z_i) = \mathrm{E}(u_i|y_{1i}, y_{2i}, z_i)$.

Following this line of reasoning, we propose the following augmented IPW (AIPW) estimating equation based on estimating equation 3:

$$S_2(\xi,Q,\pi)=\sum_{i=1}^{N}s_{2i}(\xi,Q,\pi)=\sum_{i=1}^{N}\left\{\frac{r_i}{\pi_i}D_i^TQ^{-1}e_i+\left(1-\frac{r_i}{\pi_i}\right)E_{X_i|Y_{1i},Y_{2i},Z_i}\left[D_i^TQ^{-1}e_i\right]\right\}=0. \quad (5)$$

Notice that the augmented estimating equation 5 incorporates all the information available in the full cohort, including those observations in the nonvalidation sample. To use estimating equation 5, one needs to assume a form of the conditional moments, ie, $\mathrm{E}(X_i|\ Y_{1i},\ Y_{2i},\ Z_i)$ and $\mathrm{E}\left(X_i^2\middle|Y_{1i},Y_{2i},Z_i\right)$. As the expensive exposure variable is often on a continuous scale, it is reasonable to assume that

$$\mathrm{E}\left(X_i\middle|Y_{1i},Y_{2i},Z_i\right)=\phi_0+\phi_1Y_{1i}+\phi_2Y_{2i}+\phi_3Z_i$$

and

$$\mathrm{Var}\left(X_i\middle|Y_{1i},Y_{2i},Z_i\right)=\sigma^2.$$

Then, the second-order moment can be expressed through the conditional mean and conditional variance as $\mathrm{E}\left(X_i^2\middle|Y_{1i},Y_{2i},Z_i\right)=\mathrm{Var}\left(X_i\middle|Y_{1i},Y_{2i},Z_i\right)+\left[\mathrm{E}\left(X_i\middle|Y_{1i},Y_{2i},Z_i\right)\right]^2$.

Let $\boldsymbol{\phi}=(\phi_0, \phi_1, \phi_2, \phi_3)$. We notice that $S_2(\xi, Q, \pi, \boldsymbol{\phi}, \sigma^2)$ has several components, in which ξ is the parameter of interest and $(Q, \pi, \boldsymbol{\phi}, \sigma^2)$ are nuisance parameters. The following summarizes the steps on how to conduct the analysis:

1. Fit a linear regression to obtain parameter estimates for $(\boldsymbol{\phi}, \sigma^2)$ based on the SRS portion of the data.
2. As outlined in Section 3.1, obtain the consistent estimator $\widehat{Q}, \hat{\pi}$ for Q and $\boldsymbol{\pi}$.
3. Plug $\left(\widehat{Q}, \hat{\pi}, \hat{\phi}, \hat{\sigma}^2\right)$ into S_2 to obtain $\hat{S}_2(\xi) = S_2\left(\xi, \widehat{Q}, \hat{\pi}, \hat{\phi}, \hat{\sigma}^2\right)$. Our second proposed estimator $\hat{\xi}_{AIPW}$ is the solution to the following augmented IPW estimating equation:

$$\hat{S}_2(\xi) = \sum_{i=1}^{N} \hat{s}_{2i}(\xi) = \sum_{i=1}^{N} \left\{ \frac{r_i}{\hat{\pi}_i} D_i^T \widehat{Q}^{-1} e_i + \left(1 - \frac{r_i}{\hat{\pi}_i}\right) \hat{E}_{X_i \mid Y_{1i}, Y_{2i}, Z_i} \left[D_i^T \widehat{Q}^{-1} e_i \right] \right\} = 0. \quad (6)$$

Note that how to estimate $(Q, \boldsymbol{\pi}, \boldsymbol{\phi}, \sigma^2)$ does not influence the asymptotic distribution of $\hat{\xi}_{AIPW}$ as long as the nuisance parameter estimates are root-N consistent. In Appendix A, we use a lemma from Yuan and Jennrich[30] to show why this is the case.

4 | ASYMPTOTIC RESULTS

In this section, we will present theorems regarding the consistency and asymptotic normality for our proposed estimators $\hat{\xi}_{IPW}$ and $\hat{\xi}_{AIPW}$. Let ξ_* be the true values of the parameters of interest, and let $(Q_*, \boldsymbol{\pi}_*, \boldsymbol{\phi}_*, \sigma_*^2)$ denote the true values of the nuisance parameters $(Q, \boldsymbol{\pi}, \boldsymbol{\phi}, \sigma^2)$. In addition, we let E_k denote the conditional expectation given $Y_1 \in A_k$. That is, for any function $f(\cdot)$, $E_k[f(Y_1, Y_2, X, Z)] = E[f(Y_1, Y_2, X, Z) \mid Y_1 \in A_k]$. Under regularity conditions outlined in Appendix A, assuming that $n_0/n_V \to \rho_0 > 0$ and $n_k/n_V \to \rho_k \geq 0$ for $k = 1, 3$, the following theorems hold for $\hat{\xi}_{IPW}$ and $\hat{\xi}_{AIPW}$:

Theorem 1

$\hat{\xi}_{IPW}$ and $\hat{\xi}_{AIPW}$ converge in probability to ξ_*.

Theorem 2

Let θ denote the nuisance parameters (Q, π), then $\hat{\xi}_{IPW}$ has the following asymptotic distributional properties:

$$\sqrt{n_V}\left(\hat{\xi}_{IPW} - \xi_*\right) \xrightarrow{D} N\left(0, I_1^{-1}(\xi_*, \theta_*) \sum\nolimits_1 (\xi_*, \theta_*) I_1^{-1}(\xi_*, \theta_*)\right), \quad (7)$$

where

$$I_1(\xi,\theta) = -\rho_0 E\left[\frac{\partial s_1(\xi,\theta)}{\partial \xi^T}\right] - \rho_1 E_1\left[\frac{\partial s_1(\xi,\theta)}{\partial \xi^T}\right] - \rho_3 E_3\left[\frac{\partial s_1(\xi,\theta)}{\partial \xi^T}\right]$$

and

$$\sum_1(\xi,\theta) = \rho_0 E\left[s_1(\xi,\theta)s_1(\xi,\theta)^T\right] + \rho_1 E_1\left[s_1(\xi,\theta)s_1(\xi,\theta)^T\right] + \rho_3 E_3\left[s_1(\xi,\theta)s_1(\xi,\theta)^T\right].$$

Replacing the population quantities with the sample quantities, a consistent estimator for the asymptotic variance-covariance matrix can be obtained as

$$\hat{I}_1^{-1}\left(\hat{\xi}_{IPW},\hat{\theta}\right)\widehat{\sum}_1\left(\hat{\xi}_{IPW},\hat{\theta}\right)\hat{I}_1^{-1}\left(\hat{\xi}_{IPW},\hat{\theta}\right),$$

where $\hat{I}_1(\xi,\theta) = -\frac{1}{n_V}\sum_{i\in V}\frac{\partial s_{1i}(\xi,\theta)}{\partial \xi^T}$ and $\hat{\sum}_1(\xi,\theta) = \frac{1}{n_V}\sum_{i\in V} s_{1i}(\xi,\theta)s_{1i}(\xi,\theta)^T$.

Theorem 3

Let $n = (Q, \boldsymbol{\pi}, \boldsymbol{\phi}, \sigma^2)$ denote all the nuisance parameters. $\hat{\xi}_{AIPW}$ has the following asymptotic distributional properties:

$$\sqrt{N}\left(\hat{\xi}_{AIPW} - \xi_*\right) \xrightarrow{D} N\left(0, I_2^{-1}(\xi_*,\eta_*)\sum_2(\xi_*,\eta_*)I_2^{-1}(\xi_*,\eta_*)\right), \quad (8)$$

where $I_2(\xi,\eta) = -E\left[\frac{\partial}{\partial \xi^T}s_2(\xi,\eta)\right]$, and $\sum_2(\xi,\eta) = E\left[s_2(\xi,\eta)s_2(\xi,\eta)^T\right]$.

Replacing the population quantities with the sample quantities, a consistent estimator for the asymptotic variance-covariance matrix can be obtained as:

$$\hat{I}_2^{-1}\left(\hat{\xi}_{AIPW},\hat{\eta}\right)\widehat{\sum}_2\left(\hat{\xi}_{AIPW},\hat{\eta}\right)\hat{I}_2^{-1}\left(\hat{\xi}_{AIPW},\hat{\eta}\right),$$

where $\hat{I}_2(\xi,\eta) = -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial \xi^T}s_{2i}(\xi,\eta)$ and $\hat{\sum}_2(\xi,\eta) = \frac{1}{N}\sum_{i=1}^{N} s_{2i}(\xi,\eta)s_{2i}(\xi,\eta)^T$.

Outline of the proofs are in Appendix A. We will apply a result of Foutz[31] to prove the consistency of our estimator and use Taylor expansion together with Slutsky theorem to prove asymptotic normality.

5 | SIMULATION STUDIES

In this section, we conduct extensive simulation studies to evaluate the finite sample performance of our proposed estimators. There are 5 competing estimators: (1) $\hat{\xi}_{SRS}$ denotes

the regression estimator based on SRS portion of validation sample. (2) $\hat{\xi}_R$ denotes the regression estimator from an SRS of the same size as the validation sample. Notice that this estimator is not available in practice, because when the existing data are obtained from an ODS design, it is impossible to obtain a SRS of the same size of the ODS sample. We include $\hat{\xi}_R$ in the table for comparison purpose only. (3) $\hat{\xi}_{IPW}$ denotes the estimate from our inverse probability weighted estimating equation proposed in Section 3.1. (4) $\hat{\xi}_{AIPW}$ denotes the estimate from our proposed augmented IPW estimating equation in Section 3.2. (5) $\hat{\xi}_{SPML}$ denotes a semiparametric maximum likelihood estimator similar to Jiang et al.[20] In deriving $\hat{\xi}_{SPML}$, we assume that (Y_1, Y_2) is bivariate normal and use estimated likelihood technique to deal with the nuisance functions. The details are shown in Appendix B.

The data are generated from the following models:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X + \beta_2 Z + e, \\ Y_2 &= \gamma_0 + \gamma_1 X + \gamma_2 Z + \varepsilon, \end{aligned} \quad (9)$$

where $X \sim N(0, 1)$, $Z \sim Bernoulli(0.45)$ and (e, ε) follows a bivariate normal distribution with $var(e) = \sigma_1^2$, $var(\varepsilon) = \sigma_2^2$, $cov(e,\varepsilon) = \rho\sigma_1\sigma_2$. The true parameter values are $\beta_0 = 1$, $\beta_2 = -0.5$, $\gamma_0 = 1$, $\gamma_2 = -0.5$, $\sigma_1 = \sigma_2 = 1$, and $\rho = 0.8$. We allow β_1 and γ_1 to take value 0 or 0.5.

Let μ_{Y_1} and σ_{Y_1} be the sample mean and standard deviation for the primary outcome Y_1 observed in the full cohort. In continuous outcome ODS design, we first select an SRS of size n_0, then a supplemental sample of size n_1 is chosen from $\left\{Y_1 \le \mu_{Y_1} - a\sigma_{Y_1}\right\}$ and a supplemental sample of size n_3 chosen from $\left\{Y_1 \ge \mu_{Y_1} + a\sigma_{Y_1}\right\}$. The validation sample has size $n_V = n_0 + n_1 + n_3$. We consider the following 2 settings: (1) $n_0 = 200$, $n_1 = n_3 = 100$. (2) $n_0 = 300$, $n_1 = n_3 = 50$. In addition, we also vary the cutoff points of the strata by considering different values of a, ie, $a = 1$ or 1.5.

Tables 1 and 2 show the simulation results based on 1000 independent replications. The full cohort size is $N = 3000$. In Table 1, (n_0, n_1, n_3) = (200, 100, 100). In Table 2, (n_0, n_1, n_3) = (300, 50, 50). The average of parameter estimates (Mean), empirical variance of parameter estimates across all simulations (VAR), average of the variance estimator $\left(\widehat{\text{VAR}}\right)$, and 95% CI coverage are reported. In addition, we show the sample relative efficiency of all estimators relative to $\hat{\xi}_{IPW}$ in terms of estimating γ_1. The sample relative efficiency is defined as the ratio of empirical variance, ie, $SRE_{AIPE:IPW} = var\left(\hat{\xi}_{IPW}\right)/var\left(\hat{\xi}_{AIPW}\right)$.

From Tables 1 and 2, we have the following observations: (1) All 5 estimators yield virtually unbiased estimates. (2) ξ_{SPML} has the most precise estimate. However, as later shown in the simulation studies, the validity of $\hat{\xi}_{SPML}$ depends heavily on the correctness of bivariate

normal assumption and is hence not robust. (3) Among $\hat{\xi}_{SRS}$, $\hat{\xi}_{IPW}$, and $\hat{\xi}_{AIPW}$, the augmented estimating equation estimator $\hat{\xi}_{AIPW}$ is the most precise in all settings except for when $\beta_1 = \gamma_1 = 0$, where the performance of $\hat{\xi}_{IPW}$ and $\hat{\xi}_{AIPW}$ are similar. For example, when $(n_0, n_1, n_3) = (200, 100, 100)$, $a = 1$, $\beta_1 = 0$, $\gamma_1 = 0.5$, the empirical variance estimating γ_1 is 0.0016 for $\hat{\xi}_{AIPW}$, which is smaller than 0.0026 for $\hat{\xi}_{IPW}$ and 0.0049 for $\hat{\xi}_{SRS}$. The precision gain comes from the fact that $\hat{\xi}_{IPW}$ and $\hat{\xi}_{AIPW}$ use more participants than $\hat{\xi}_{SRS}$. (4) For all estimators, averages of the variance estimator is very close to the empirical variance (ie, $\widehat{\mathrm{VAR}}$ is close to VAR). (5) The 95% CI coverage is close to 0.95, which implies that the asymptotic normal approximation works well in these finite sample size settings. (6) When the cutoff points are further out (ie, $a = 1$ versus $a = 1.5$), the precision gains of $\hat{\xi}_{IPW}$ and $\hat{\xi}_{AIPW}$ over $\hat{\xi}_{SRS}$ are slightly lower (Table 1). The ODS sample is more enriched with a=1.5. However, this more enriched study design is offset by the highly variable IPW weighting distribution that results in precision loss. (7) Comparing the results across Tables 1 and 2 in terms of estimating γ_1, we find that, for a given validation sample size ($n_V = 400$), when SRS sample size is larger (ie, $n_0 = 200$ versus $n_0 = 300$), the variance of $\hat{\xi}_{IPW}$ and $\hat{\xi}_{AIPW}$ decreases. However, the variance of $\hat{\xi}_{SRS}$ has a faster decreasing rate. That is, the precision gains of $\hat{\xi}_{IPW}$ and $\hat{\xi}_{AIPW}$ over $\hat{\xi}_{SRS}$ is smaller when the SRS sample size is larger.

We investigate the scenario where the error term ε is not normally distributed. The simulation set up is the same as Table 1 except for the error term. We assume that $e \sim N(0, 1)$, ε is a gamma distribution with shape parameter 2, rate parameter 1, then normalized to have mean 0 and variance 1. This error term is right skewed. From Table 3, we see that our proposed estimators are more robust to model misspecification than $\hat{\xi}_{SPML}$. When true $\gamma_1 = 0.5$, the 95% CI coverage rate is poor for $\hat{\xi}_{SPML}$. $\hat{\xi}_{SPML}$ also has larger empirical variance than $\hat{\xi}_{IPW}$ and $\hat{\xi}_{AIPW}$ as it misspecified the distribution of Y_2. Another limitation of $\hat{\xi}_{SPML}$ is that it is subject to “curse of dimensionality” as nonparametric method is used to estimate the nuisance function. Therefore, the method cannot be directly applied when the dimension of Z is relatively high, ie, > 3. In addition, the method does not have a natural extension when Z has both discrete and continuous components.

We further investigate the performance of our proposed estimators in estimating γ_1 under different combinations of the SRS sample and supplemental samples. The simulation set up is as follows: $a = 1.0, \beta_0 = 1, \beta_1 = 0.5, \beta_2 = -0.5, \gamma_0 = 1, \gamma_1 = 0.5, \gamma_2 = -0.5, \sigma_1 = \sigma_2 = 1, \rho = 0.8$. The full cohort has size N=3000. We fix the validation sample to have size n_V=400 and vary (n_0, n_1, n_3). Figure 1 shows the sample relative efficiency of $\hat{\xi}_{AIPW}$ and $\hat{\xi}_{SRS}$ relative to $\hat{\xi}_{IPW}$ over a wide range of proportion of the SRS sample in the validation sample (n_0/n_V). We confirm that when SRS size is larger, there is larger precision gain of $\hat{\xi}_{AIPW}$ over $\hat{\xi}_{IPW}$, while the precision gain of $\hat{\xi}_{IPW}$ and $\hat{\xi}_{AIPW}$ over $\hat{\xi}_{SRS}$ is smaller.

We also evaluate the precision gain of our proposed estimators for different values of ρ, where ρ is the correlation coefficient between 2 error terms e and e in (9). Figure 2 shows the corresponding sample relative efficiency. When correlation changes from 0 to 1, the precision gain of $\hat{\xi}_{IPW}$ over $\hat{\xi}_{SRS}$ is relatively stable for different ρ values. On the other hand, the precision gain of $\hat{\xi}_{AIPW}$ over $\hat{\xi}_{IPW}$ is decreasing when $\gamma_1 > 0$, and increasing when $\gamma_1 < 0$.

6 | COLLABORATIVE PERINATAL PROJECT DATA

In this section, we applied our method to analyze the CPP data set. The CPP is originally conducted as a cohort study to evaluate the risk factors for birth defects and other neurological disorders of childhood.[32] The study involved 12 hospitals/universities located across the United States. In all, 55 908 pregnancies were registered, representing the experience about 44 000 women. The children born during the study were followed up until 8 years old. One hypothesis is that maternal PCB levels are related to children's IQ performance at 7 years of age.[4] Because the exposure variable PCB levels are very expensive to measure, ODS design is conducted on the basis of CPP data. An SRS of 849 individuals is selected, and then 2 supplemental samples are selected based on the children's IQ score. One supplemental sample of size 81 is selected from the lower tail of the IQ distribution, defined by 1 standard deviation below the mean IQ score in the CPP population. The other supplemental sample of size 108 is chosen from 1 standard deviation above the mean IQ score.

Many researchers have made efforts to assess whether there is association between PCB level and child's birth weight, but the findings from these studies are not consistent. Some indicate an inverse association,[12,16,18] while other indicates a positive association[13] or no association at all.[14,15,17] We use our available CPP data to perform the secondary outcome analysis. In our analysis, we use the Weschler Intelligence Scale for children at 7 years old (IQ) as the primary outcome, and child's birth weight (in grams) to be the secondary outcome. Other confounding variables include parent's education level (EDU), social economic status of the child's family (SES), race ethnicity of the child (RACE), and gender of the child (GENDER).

Table 4 shows the parameter estimates, standard error, and 95% CIs for the secondary outcome model, which regresses birth weight over PCB level adjusting for other covariates. Simple random sample denotes the regression analysis based on the SRS portion of the data. Inverse probability weighted denotes the inverse probability weighted estimating equation we proposed in Section 3.1, which uses the validation sample only. Augmented IPW denotes the augmented inverse probability weighted estimating equation proposed in Section 3.2. The semiparametric maximum likelihood estimator is excluded from the analysis, as it does not have a natural extension when the covariates contain both discrete and continuous components.

All 3 analyses confirm that maternal PCB concentration is not significantly associated with child's birth weight. However, the proposed IPW and AIPW estimator provide more precise estimates of the effects, evidenced by the reduced standard error and narrower CI. For example, the standard error for the PCB effect is 8.41 for IPW and 8.46 for AIPW, which is

smaller than 9.20 for the standard regression analysis based on the SRS portion of the data. All 3 analyses confirm that being White has a positive impact on child's birth weight, while AIPW analysis shows that social economic status has a positive impact on birth weight. In addition, girls have lower birth weight compared with boys as we expected. Note that there seems to be some difference between AIPW and IPW estimator for the EDU effect. We conducted additional simulation studies to show that such difference is merely due to sampling variability, and the observed difference is not that large if we take into account of the magnitude of the standard error. The details are shown in the Supporting Information.

7 | DISCUSSION

Investigators would all like to use the availability of expensive exposure that is already measured in a previous study. Most studies have multiple endpoints beside the primary outcome. This means that we often need to reuse the already collected data to analyze a secondary outcome in relation to the expensive exposure. However, when the original data are collected via an outcome-dependent fashion, secondary outcome analysis can be challenging as the ODS sample is no longer an SRS of the general population. As more studies are conducted using ODS designs, there is ever increasing needs for performing secondary analysis correctly and precisely for data from these studies. Our research is intended to fill these gaps and is the first attempt to develop precise inference procedure for a secondary outcome under the continuous outcome ODS design. We proposed IPW and AIPW estimating equations, in which only the form of the regressions are specified. Our proposed approach has the advantage of making no parametric distribution assumptions on (Y_1, Y_2) and thus is robust to model misspecification. Yet our proposed estimators are able to improve estimation precision relative to the naive analysis using SRS sample only.

Recall that r is the indicator variable of being selected into the validation sample. The expensive exposure X and r are conditionally independent given Y_1. This means that in the Step 1 of AIPW analysis, we could actually use the whole validation sample (n_V participants) to estimate the nuisance parameters (ϕ, σ^2). As we mentioned before, how to estimate the nuisance parameter (ϕ, σ^2) does not influence the asymptotic distribution of $\hat{\xi}_{AIPW}$ as long as the nuisance parameter estimates are consistent. However, in small sample scenarios, there could be some difference. For instance, when full cohort has size 3000, (n_0, n_1, n_3) = (50, 175, 175), the proposed estimator $\hat{\xi}_{AIPW}$ could have some bias (mean of parameter estimates is 0.515 while the true value for γ_1 is 0.5), if we only used SRS (50 participants) to estimate the nuisance parameters. On the other hand, the bias is reduced to 0.07, if we use the validation sample (400 participants) to estimate the nuisance parameters and then solve the AIPW estimating equation. When SRS sample size is larger, such as (n_0, n_1, n_3) = (200, 100, 100), the results between these 2 approaches are almost identical.

We found that our proposed estimators have the same asymptotic variance regardless of whether we are using fixed-size sampling (sampling without replacement) or independent Bernoulli sampling in choosing the initial SRS and the supplemental samples. This is because even though (r_1, ⋯, r_N) are correlated under fixed-size sampling, it can be proved

that $cov\left(\frac{r_i}{\pi_i}D_i^T Q^{-1} e_i, \frac{r_j}{\pi_j}D_j^T Q^{-1} e_j\right) = 0$ for different participants i and j. For case-cohort design, the difference between 2 subcohort sampling methods have been studied.[33,34] They found that the variance under fixed-size SRS is always smaller than or equal to the variance under Bernoulli sampling. There is equivalence when the corresponding covariance parts equal to 0 (see the comparison between $\sum_H^*$ and Σ_H in Kulich and Lin[33]).

In our paper, when we implemented our AIPW estimator, we specified the form of the conditional moments, $E(X \mid Y_1, Y_2, Z)$ and $E(X^2 \mid Y_1, Y_2, Z)$ using a linear regression model. In practice, the expensive exposure X might be discrete, then a generalized linear model could be adopted. We could also use some nonparametric techniques to estimate these conditional moments. Interest rises to see if there exists any difference between parametric and non-parametric methods. In addition, we used the observed selection probability $\hat{\pi}_i$ in both IPW and AIPW estimators. In some ODS designs, the true selection probability is known based on the design structure and can be calculated. It would be interesting to see whether using the true probability or observed probability makes any difference in statistical efficiency. Furthermore, we are looking for a flexible parametric family to jointly model the primary and secondary outcome. Copula seems to be a natural way to achieve the goal. This is another possible area of future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research is part of Yinghao Pan's PhD dissertation. It is partially supported by grants R01 ES021900 and P30 ES010126 from the National Institute of Environmental Health Sciences, P01 CA142538 from the National Cancer Institute, and the Intramural Research Program of the NIEHS/NIH. The authors thank editors and 2 referees for their suggestions to improve the paper.

Funding information

National Cancer Institute, Grant/Award Number: P01 CA142538; National Institute of Environmental Health Sciences, Grant/Award Numbers: R01 ES021900 and P30 ES010126

APPENDIX A

PROOFS

We provide the outline of the proofs for the augmented IPW estimator $\hat{\xi}_{AIPW}$. The proof for $\hat{\xi}_{IPW}$ follows the similar arguments and thus omitted here. Let $\eta = (Q, \pi, \phi, \sigma^2)$ denote all the nuisance parameters. We make the following regularity conditions:

1. The parameter space of ξ, Ξ, is a compact subspace of $\mathbf{R}^p$, and that the true underlying value ξ_* lies in the interior of the parameter space; the covariate space, $\mathcal{X}$ is a compact subset of $\mathbf{R}$; and the covariate space, $\mathcal{Z}$, is a compact subset of $\mathbf{R}^q$ for some $q \geq 1$.

2. For all (Y_1, Y_2, X, Z), $s_2(\xi, \eta)$ is continuous for all $\xi \in \Xi$; the partial derivatives $\frac{\partial s_2(\xi,\eta)}{\partial \xi_i}$, for $i = 1, \ldots, p$, exist and are continuous for all $\xi \in \Xi$.

3. Interchanges of differentiation and integration are valid for $s_2(\xi, \eta)$ and its first-order partial derivatives with respect to ξ.

4. The following expected value matrix is finite and negative definite at (ξ_*, η_*):

$$E\left[\frac{\partial}{\partial \xi^T} s_2(\xi,\eta)\right].$$

5. The supremum of $\frac{\partial s_2(\xi,\eta)}{\partial \xi^T}$ in the neighborhood of ξ_* is bounded by a function g that has finite expectation. (Outline of the proof for consistency):

By law of larger numbers, it is straightforward to show that

$$\frac{1}{N}\sum_{i=1}^{N} s_{2i}(\xi_*, \eta_*) \xrightarrow{p} E\left[s_2(\xi_*, \eta_*)\right] = 0 \text{ as } N \to \infty. \quad \text{(A1)}$$

Using the fact that $\hat{\eta} = \left(\hat{Q}, \hat{\pi}, \hat{\phi}, \hat{\sigma}^2\right)$ is a consistent estimator of η_*, and $\hat{s}_{2i}(\xi) = s_{2i}(\xi, \hat{\eta})$, it can be shown that

$$\frac{1}{N}\sum_{i=1}^{N} \hat{s}_{2i}(\xi_*) - \frac{1}{N}\sum_{i=1}^{N} s_{2i}(\xi_*, \eta_*) \xrightarrow{p} 0 \text{ as } N \to \infty. \quad \text{(A2)}$$

Combining (A1) and (A2), we know that

$$\frac{1}{N}\hat{S}_2(\xi_*) = \frac{1}{N}\sum_{i=1}^{N} \hat{s}_{2i}(\xi_*) \xrightarrow{p} 0 \text{ as } N \to \infty.$$

Furthermore, using Assumptions 1-3 and consistency of $\hat{\eta}$, we can show that

$$\frac{1}{N}\left[\frac{\partial \hat{S}_2(\xi)}{\partial \xi^T} - \frac{\partial S_2(\xi, \eta_*)}{\partial \xi^T}\right] \xrightarrow{p} 0 \text{ as } N \to \infty. \quad \text{(A3)}$$

holds uniformly for ξ in parameter space. Also, by law of large numbers,

$$\frac{1}{N}\frac{\partial S_2(\xi, \eta_*)}{\partial \xi^T} = \frac{1}{N}\sum_{i=1}^{N} s_{2i}(\xi, \eta_*) \xrightarrow{p} E\left[\frac{\partial}{\partial \xi^T} s_2(\xi, \eta_*)\right] \text{ as } N \to \infty. \quad \text{(A4)}$$

Combining (A3) and (A4), we have

$$\frac{1}{N}\frac{\partial \hat{S}_2(\xi)}{\partial \xi^T} \xrightarrow{p} E\left[\frac{\partial}{\partial \xi^T} s_2(\xi, \eta_*)\right] \text{ as } N \to \infty .$$

uniformly for $\xi \in \Xi$. From Assumption 4, we know that $E\left[\frac{\partial}{\partial \xi^T} s_2(\xi, \eta)\right]$ is negative definite and hence invertible at (ξ_*, η_*). Hence, we can apply a result of Foutz[31] that uses the inverse function theorem to prove that our proposed estimator is a consistent and unique solution to the estimating equations.

(outline of the proof for asymptotic normality):

Let us denote

$$A_N = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \xi^T} s_{2i}(\xi_*, \eta_*), \quad B_N = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \eta^T} s_{2i}(\xi_*, \eta_*).$$

Using lemma 2 from Yuan and Jennrich,[30] we have

$$\sqrt{N}\left[\frac{1}{N} S_2(\hat{\xi}, \hat{\eta}) - \frac{1}{N} S_2(\xi_*, \eta_*)\right] = A_N \sqrt{N}(\hat{\xi} - \xi_*) + B_N \sqrt{N}(\hat{\eta} - \eta_*). \quad \text{(A5)}$$

Let A and B be the limit of A_N and B_N, respectively. By law of large numbers, we know that

$$A_N = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \xi^T} s_{2i}(\xi_*, \eta_*) \xrightarrow{p} E\left[\frac{\partial s_2(\xi_*, \eta_*)}{\partial \xi^T}\right] = -I_2(\xi_*, \eta_*)$$
$$B_N = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \eta^T} s_{2i}(\xi_* \eta_*) \xrightarrow{p} E\left[\frac{\partial s_2(\xi_* \eta_*)}{\partial \eta^T}\right] = 0.$$

That is, $A = -I_2(\xi_*, \eta_*)$ and $B = 0$. Rearrange (A5) and using the fact that $S_2(\hat{\xi}, \hat{\eta}) = 0$, we have

$$-A_N \sqrt{N}(\hat{\xi} - \xi_*) = \frac{1}{\sqrt{N}} S_2(\xi_*, \eta_*) + B\sqrt{N}(\hat{\eta} - \eta_*) + (B_N - B)\sqrt{N}(\hat{\eta} - \eta_*).$$

$B = 0$, $\sqrt{N}(\hat{\eta} - \eta_*)$ is bounded in probability. Hence, we know that

$$-A_N \sqrt{N}(\hat{\xi} - \xi_*) = \frac{1}{\sqrt{N}} S_2(\xi_*, \eta_*) + o_p(1).$$

By the proposition in appendix 1 of Kulich and Lin,[33]

$$-A_N\sqrt{N}\left(\hat{\xi}-\xi_*\right)\xrightarrow{D}N\left(0,\sum_2\left(\xi_*,\eta_*\right)\right),$$

where $\Sigma_2(\xi, \eta) = E[s_2(\xi, \eta)s_2(\xi, \eta)^T$. Then, using Slutsky theorem, we know that

$$\sqrt{N}\left(\hat{\xi}-\xi_*\right)\xrightarrow{D}N\left(0,I_2^{-1}\left(\xi_*,\eta_*\right)\sum_2\left(\xi_*,\eta_*\right)I_2^{-1}\left(\xi_*,\eta_*\right)\right).$$

Notice that B=0 implies that the asymptotic distribution of $\hat{\eta}$ does not influence the asymptotic distribution of $\hat{\xi}$ as long as $\hat{\eta}$ is root-N consistent.

APPENDIX B: BRIEF DESCRIPTION OF THE ESTIMATED LIKELIHOOD APPROACH

To develop the semiparametric maximum likelihood estimator $\hat{\xi}_{SPML}$, we need to assume that (Y_1, Y_2) is bivariate normal. That is,

$$\begin{aligned}Y_{1i}&=\beta_0+\beta_1X_i+\beta_2Z_i+e_i,\\ Y_{2i}&=\gamma_0+\gamma_1X_i+\gamma_2Z_i+\varepsilon_i,\end{aligned}$$

where (e_i, ε_i) follows a bivariate normal distribution with $var(e_i) = \sigma_1^2$, $var(\varepsilon_i) = \sigma_2^2$, $cov(e_i, \varepsilon_i) = \rho\sigma_1\sigma_2$.

For a participant in V(validation sample), the contribution to the likelihood is (Y_1, Y_2, X, Z). For a participant in $\bar{V}$ (nonvalidation sample), the contribution is (Y_1, Y_2, Z). Hence, the likelihood corresponding to (1) is proportional to

$$L(\xi)=\prod_{i\in V}f_\xi\left(Y_{1i},Y_{2i}\middle|X_i,Z_i\right)\times\prod_{j\in\bar{V}}\int_x f_\xi\left(Y_{1j},Y_{2j}\middle|x,Z_j\right)dG_{X|Z}\left(x\middle|Z_j\right),$$

where $f_\xi(Y_1, Y_2 | X, Z)$ is the density function of a bivariate normal distribution and $G_{X|Z}(\cdot|\cdot)$ represents the conditional distribution function of X given Z. The log-likelihood is

$$l(\xi)=\sum_{i\in V}\log f_\xi\left(Y_{1i},Y_{2i}\middle|X_i,Z_i\right)+\sum_{j\in\bar{V}}\log\left\{\int_x f_\xi\left(Y_{1j},Y_{2j}\middle|x,Z_j\right)dG_{X|Z}\left(x\middle|Z_j\right)\right\}.$$

Notice that $G_{X|Z}$ is a nuisance function. Using ideas from Weaver and Zhou,[5] we propose to work with the following estimated log-likelihood function:

$$\hat{l}(\xi)=\sum_{i\in V}\log f_\xi\left(Y_{1i},Y_{2i}\middle|X_i,Z_i\right)+\sum_{j\in\bar{V}}\log\left\{\int_x f_\xi\left(Y_{1j},Y_{2j}\middle|x,Z_j\right)d\hat{G}_{X|Z}\left(x\middle|Z_j\right)\right\},$$

where we nonparametrically estimate $G_{X|Z}$ using the SRS sample. For discrete Z, let

$$\hat{G}_{X|Y}(x|z) = \sum_{i \in V_0} I(X_i \le x, Z_i = z) / \sum_{i \in V_0} I(Z_i = z).$$

For continuous Z, we use the kernel method, ie,

$$\hat{G}_{X|Z}(x|z) = \sum_{i \in V_0} I(X_i \le x) K_H(Z_i - z) / \sum_{i \in V_0} K_H(Z_i - z),$$

where $K_H(\cdot) = |H|^{-1/2} K(H^{-1/2}\cdot)$ is a kernel with a bandwidth matrix H. Then, the proposed semiparametric maximum likelihood estimator $\hat{\xi}_{SPML}$ is the solution to the following estimating equation:

$$\frac{1}{N} \frac{\partial \hat{l}(\xi)}{\partial \xi} = 0.$$

References

1. Cornfield J. Method of estimating comparative rates from clinical data. Application to cancer of the lung, breast, and cervix. J Natl Cancer Inst. 1951; 11(6):1269–1275. [PubMed: 14861651]
2. Prentice R. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986; 73(1):1–11.
3. Zhou H, Weaver M, Qin J, Longnecker M, Wang M. A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. Biometrics. 2002; 58(2):413–421. [PubMed: 12071415]
4. 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 Male Birth Defects and Neurodevelopmental Outcomes in the Collaborative Perinatal Project. Washington, DC: Study Proposal, National Institute of Environmental Health Sciences; 1997.
5. Weaver M, Zhou H. An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. J Am Stat Assoc. 2005; 100(470):459–469.
6. Wang X, Zhou H. A semiparametric empirical likelihood method for biased sampling schemes with auxiliary covariates. Biometrics. 2006; 62(4):1149–1160. [PubMed: 17156290]
7. 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(1):221–228. [PubMed: 20107493]
8. Wang X, Zhou H. Design and inference for cancer biomarker study with an outcome and auxiliary-dependent subsampling. Biometrics. 2010; 66(2):502–511. [PubMed: 19508239]
9. Qin G, Zhou H. Partial linear inference for a 2-stage outcome-dependent sampling design with a continuous outcome. Biostatistics. 2011; 12(3):506–520. [PubMed: 21156990]
10. Zhou H, Qin G, Longnecker MP. A partial linear model in the outcome-dependent sampling setting to evaluate the effect of prenatal PCB exposure on cognitive function in children. Biometrics. 2011; 67(3):876–885. [PubMed: 21039397]
11. Zhou H, Wu Y, Liu Y, Cai J. Semiparametric inference for a 2-stage outcome-auxiliary-dependent sampling design with continuous outcome. Biostatistics. 2011; 12(3):521–534. [PubMed: 21252082]
12. Fein G, Jacobson J, Jacobson S, Schwartz P, Dowler J. Prenatal exposure to polychlorinated biphenyls: effects on birth size and gestational age. J Pediatr. 1984; 105(2):315–320. [PubMed: 6431068]

13. Dar E, Kanarek M, Anderson H, Sonzogni W. Fish consumption and reproductive outcomes in Green Bay, Wisconsin. Environ Res. 1992; 59(1):189–201. [PubMed: 1425509]
14. Vartiainen T, Jaakkola JJ, Saarikoski S, Tuomisto J. Birth weight and sex of children and the correlation to the body burden of PCDDs/PCDFs and PCBs of the mother. Environ Health Perspect. 1998; 106(2):61–66. [PubMed: 9432971]
15. Grandjean P, Bjerve K, Weihe P, Steuerwald U. Birthweight in a fishing community: significance of essential fatty acids and marine food contaminants. Int J Epidemiol. 2001; 30(6):1272–1278. [PubMed: 11821327]
16. Karmaus W, Zhu X. Maternal concentration of polychlorinated biphenyls and dichlorodiphenyl dichlorethylene and birth weight in Michigan fish eaters: a cohort study. Environ Health. 2004; 3(1):1. [PubMed: 14748928]
17. Longnecker M, Klebanoff M, Brock J, Guo X. Maternal levels of polychlorinated biphenyls in relation to preterm and small-for-gestational-age birth. Epidemiology. 2005; 16(5):641–647. [PubMed: 16135940]
18. Murphy L, Gollenberg A, Louis G, Kostyniak P, Sundaram R. Maternal serum preconception polychlorinated biphenyl concentrations and infant birth weight. Environ Health Perspect. 2010; 118(2):297–302. [PubMed: 20123616]
19. Lee A, Mcmurchy L, Scott AJ. Re-using data from case-control studies. Stat Med. 1997; 16(12): 1377–1389. [PubMed: 9232759]
20. Jiang Y, Scott A, Wild CJ. Secondary analysis of case-control data. Stat Med. 2006; 25(8):1323–1339. [PubMed: 16220494]
21. Lin D, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. Genet Epidemiol. 2009; 33(3):256–265. [PubMed: 19051285]
22. Richardson D, Rzehak P, Klenk J, Weiland S. Analyses of case-control data for additional outcomes. Epidemiology. 2007; 18(4):441–445. [PubMed: 17473707]
23. Monsees G, Tamimi R, Kraft P. Genome-wide association scans for secondary traits using case-control samples. Genet Epidemiol. 2009; 33(8):717–728. [PubMed: 19365863]
24. Wei J, Carroll RJ, Muller UU, Keilegom IV, Chatterjee N. Robust estimation for homoscedastic regression in the secondary analysis of case-control data. J R Stat Soc Series B Stat Methodol. 2013; 75(1):185–206. [PubMed: 23637568]
25. Ma Y, Carroll RJ. Semiparametric estimation in the secondary analysis of case-control studies. J R Stat Soc Series B Stat Methodol. 2016; 78(1):127–151. [PubMed: 26834506]
26. Horvitz D, Thompson D. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc. 1952; 47(260):663–685.
27. Liang K, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73(1):13–22.
28. Zhao L, Lipsitz S, Lew D. Regression analysis with missing covariate data using estimating equations. Biometrics. 1996; 52:1165–1182. [PubMed: 8962448]
29. Robins J, Rotnitzky A, Zhao L. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc. 1994; 89(427):846–866.
30. Yuan K, Jennrich R. Estimating equations with nuisance parameters: theory and applications. Ann Inst Stat Math. 2000; 52(2):343–350.
31. Foutz R. On the unique consistent solution to the likelihood equations. J Am Stat Assoc. 1977; 72(357):147–148.
32. Niswander K, Gordon M. The women and their pregnancies. Washington, D. C.: U.S. Government Printing Office; 1972. U.S. Department of Health, Education, and Welfare Publication (NIH) 73-379
33. Kulich M, Lin DY. Additive hazards regression for case-cohort studies. Biometrika. 2000; 87(1): 73–87.
34. Nan B, Yu M, Kalbfleisch J. Censored linear regression for case-cohort studies. Biometrika. 2006; 93(4):747–762.

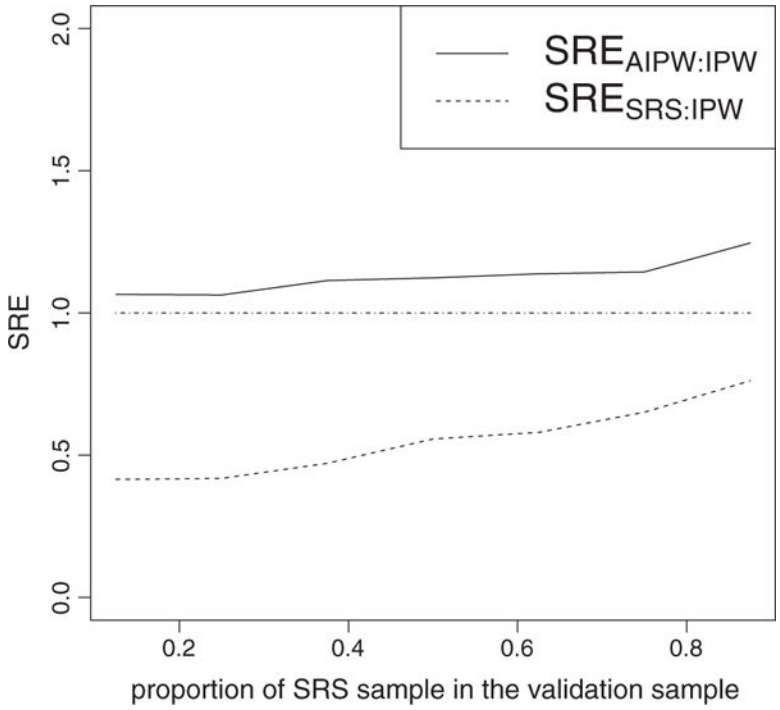

FIGURE 1.

Sample relative efficiencies (SREs) comparing $\hat{\xi}_{AIPW}$ and $\hat{\xi}_{SRS}$ to $\hat{\xi}_{IPW}$ in terms of estimating γ_1, under various combinations of simple random sample (SRS) and supplemental samples. The SRE is defined as $SRE_{AIPW:IPW} = var(\hat{\xi}_{IPW})/var(\hat{\xi}_{AIPW})$. The X-axis is the fraction of SRS in the validation sample: n_0/n_V

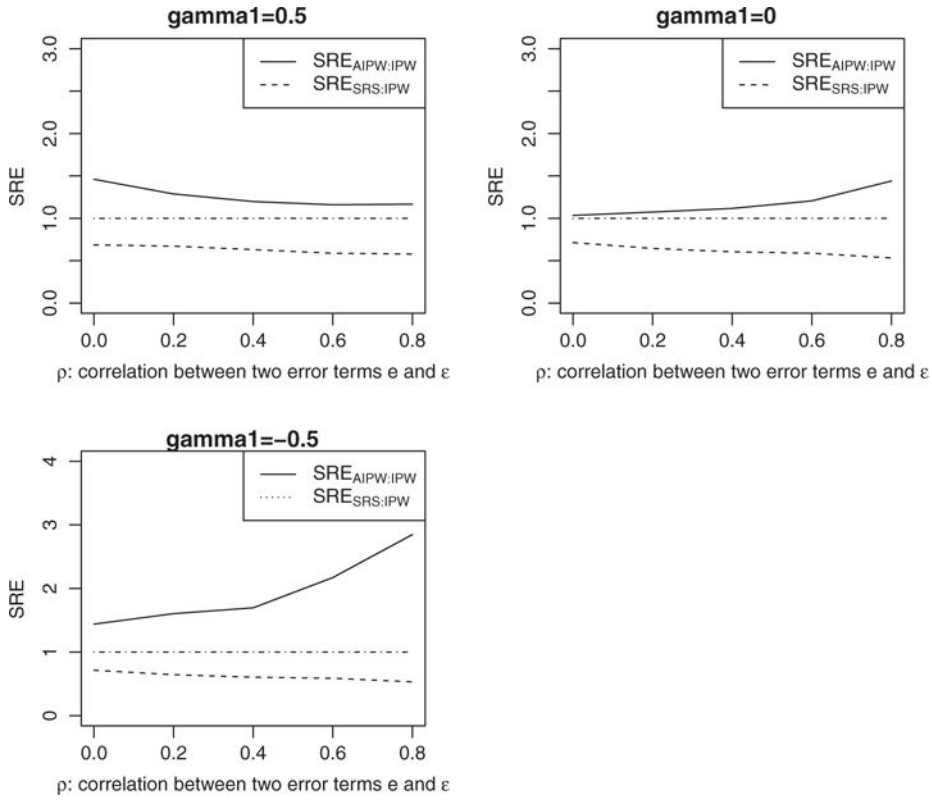

FIGURE 2.

Sample relative efficiencies (SREs) comparing $\hat{\xi}_{AIPW}$ and $\hat{\xi}_{SRS}$ to $\hat{\xi}_{IPW}$ in terms of estimating γ_1, under different values of ρ. The sample relative efficiency is defined as $SRE_{AIPW:IPW} = var(\hat{\xi}_{IPW})/var(\hat{\xi}_{AIPW})$. AIPW, augmented IPW; IPW, inverse probability weighted

TABLE 1

Simulation results based on 1000 simulations with $n_0 = 200$, $n_1 = n_3 = 100$, the validation sample size is $n_V = 400$, the full cohort size is $N = 3000$

				$\hat{\beta}_1$				$\hat{\gamma}_1$				
α	β_1	γ_1	Methods	Mean	VAR	$\widehat{\text{VAR}}$	CI	Mean	VAR	$\widehat{\text{VAR}}$	CI	SRE
1.0	0	0	ξ_{SRS}	−0.002	0.0049	0.0050	0.948	−0.001	0.0046	0.0050	0.966	0.58
			ξ_R	0.000	0.0024	0.0025	0.958	0.001	0.0026	0.0025	0.950	1.04
			ξ_{IPW}	−0.001	0.0020	0.0021	0.947	0.002	0.0027	0.0027	0.947	1.00
			ξ_{AIPW}	0.000	0.0021	0.0021	0.946	0.002	0.0027	0.0027	0.945	1.00
			ξ_{SPML}	0.001	0.0015	0.0014	0.954	0.001	0.0019	0.0018	0.940	1.42
		0.5	ξ_{SRS}	0.002	0.0048	0.0050	0.956	0.502	0.0049	0.0050	0.955	0.53
			ξ_R	0.001	0.0026	0.0025	0.949	0.501	0.0026	0.0025	0.950	1.00
			ξ_{IPW}	0.003	0.0020	0.0020	0.952	0.502	0.0026	0.0026	0.947	1.00
			ξ_{AIPW}	0.003	0.0014	0.0014	0.956	0.504	0.0016	0.0016	0.938	1.63
			ξ_{SPML}	0.002	0.0010	0.0010	0.950	0.500	0.0013	0.0013	0.955	2.00
	0.5	0	ξ_{SRS}	0.501	0.0053	0.0050	0.944	0.001	0.0049	0.0050	0.952	0.55
			ξ_R	0.501	0.0026	0.0025	0.943	0.001	0.0026	0.0025	0.939	1.04
			ξ_{IPW}	0.503	0.0023	0.0025	0.958	0.003	0.0027	0.0028	0.947	1.00
			ξ_{AIPW}	0.503	0.0018	0.0018	0.936	0.000	0.0018	0.0018	0.942	1.50
			ξ_{SPML}	0.499	0.0017	0.0017	0.942	0.002	0.0014	0.0014	0.944	1.93
		0.5	ξ_{SRS}	0.499	0.0052	0.0050	0.949	0.499	0.0050	0.0050	0.948	0.56
			ξ_R	0.500	0.0025	0.0025	0.947	0.498	0.0026	0.0025	0.950	1.08
			ξ_{IPW}	0.501	0.0022	0.0025	0.960	0.500	0.0028	0.0028	0.952	1.00
			ξ_{AIPW}	0.502	0.0020	0.0020	0.951	0.502	0.0025	0.0023	0.943	1.12
			ξ_{SPML}	0.500	0.0016	0.0017	0.958	0.500	0.0017	0.0018	0.955	1.65
1.5	0.5	0	ξ_{SRS}	0.500	0.0048	0.0052	0.968	0.002	0.0046	0.0052	0.960	0.67
			ξ_R	0.498	0.0024	0.0025	0.956	−0.001	0.0024	0.0025	0.940	1.29

				$\hat{\beta}_1$				$\hat{\gamma}_1$				
a	β_1	γ_1	**Methods**	**Mean**	**VAR**	$\widehat{\text{VAR}}$	**CI**	**Mean**	**VAR**	$\widehat{\text{VAR}}$	**CI**	**SRE**
			ξ_{IPW}	0.504	0.0027	0.0029	0.955	0.004	0.0031	0.0033	0.953	1.00
			ξ_{AIPW}	0.505	0.0022	0.0022	0.947	0.001	0.0022	0.0022	0.949	1.41
			ξ_{SPML}	0.504	0.0017	0.0017	0.948	0.006	0.0014	0.0014	0.947	2.21
		0.5	ξ_{SRS}	0.496	0.0050	0.0050	0.947	0.500	0.0049	0.0050	0.956	0.61
			ξ_R	0.502	0.0025	0.0025	0.953	0.502	0.0025	0.0025	0.950	1.20
			ξ_{IPW}	0.501	0.0028	0.0029	0.954	0.504	0.0030	0.0032	0.955	1.00
			ξ_{AIPW}	0.503	0.0025	0.0025	0.944	0.505	0.0027	0.0027	0.947	1.11
			ξ_{SPML}	0.501	0.0017	0.0018	0.950	0.501	0.0018	0.0018	0.953	1.67

Abbreviation: SRE, sample relative efficiency. Results are based on the model $Y_1=\beta_0+\beta_1 X+\beta_2 Z+e$, $Y_2=Y_0+Y_1 X+Y_2 Z+\varepsilon$, where $X\sim N(0, 1)$, $Z\sim Bernoulli(0.45)$ and (e, ε) follow a bivariate normal distribution with $var(e)=\sigma_1^2$, $var(\varepsilon)=\sigma_2^2$, $cov(e, \varepsilon)=\rho\sigma_1\sigma_2$; the true parameter values are $\beta_0=1$, $\beta_2=-0.5$, $\gamma_0=1$, $\gamma_2=-0.5$, $\sigma_1=\sigma_2=1$, $\rho=0.8$. The cutoff points for the outcome-dependent sampling design are $\mu_{Y_1}-a\sigma_{Y_1}$ and $\mu_{Y_1}+a\sigma_{Y_1}$. ξ_{SRS} denotes the regression estimator based on simple random sample portion of the validation sample. ξ_R denotes the regression estimator from a simple random sample of the same size as the validation sample. ξ_{IPW} denotes the estimate from our inverse probability weighted (IPW) estimating equation. ξ_{aipw} denotes the estimate from augmented IPW (AIPW) estimating equation. ξ_{SPWL} is a semiparametric maximum likelihood (SPML) estimator similar to Jiang et al,[20] which models (Y_1, Y_2) parametrically using a bivariate normal distribution.

TABLE 2

Simulation results based on 1000 simulations with $n_0 = 300$, $n_1 = n_3 = 50$, the validation sample size is $n_V = 400$, the full cohort size is N=3000

				$\hat{\beta}_1$			$\hat{\gamma}_1$			
α	β_1	γ_1	Methods	Mean	VAR	$\widehat{\text{VAR}}$	Mean	VAR	$\widehat{\text{VAR}}$	SRE
1.0	0	0	ξ_{SRS}	0.000	0.0034	0.0034	−0.001	0.0036	0.0034	0.69
			ξ_R	0.000	0.0024	0.0025	0.000	0.0025	0.0025	1.00
			ξ_{IPW}	0.001	0.0021	0.0020	0.000	0.0025	0.0023	1.00
			ξ_{AIPW}	0.001	0.0021	0.0020	0.000	0.0025	0.0023	1.00
			ξ_{SPML}	0.001	0.0018	0.0018	0.000	0.0021	0.0021	1.19
		0.5	ξ_{SRS}	−0.004	0.0034	0.0034	0.496	0.0034	0.0034	0.65
			ξ_R	−0.001	0.0023	0.0025	0.500	0.0024	0.0025	0.92
			ξ_{IPW}	−0.001	0.0020	0.0020	0.498	0.0022	0.0023	1.00
			ξ_{AIPW}	−0.001	0.0013	0.0014	0.500	0.0014	0.0014	1.57
			ξ_{SPML}	0.001	0.0012	0.0012	0.498	0.0013	0.0014	1.69
	0.5	0	ξ_{SRS}	0.501	0.0032	0.0034	0.001	0.0034	0.0034	0.62
			ξ_R	0.499	0.0024	0.0025	0.000	0.0025	0.0025	0.84
			ξ_{IPW}	0.502	0.0018	0.0022	0.001	0.0021	0.0024	1.00
			ξ_{AIPW}	0.502	0.0014	0.0015	0.000	0.0016	0.0015	1.31
			ξ_{SPML}	0.500	0.0014	0.0015	0.001	0.0014	0.0014	1.50
		0.5	ξ_{SRS}	0.500	0.0032	0.0034	0.498	0.0032	0.0033	0.66
			ξ_R	0.499	0.0025	0.0025	0.499	0.0025	0.0025	0.84
			ξ_{IPW}	0.502	0.0018	0.0022	0.501	0.0021	0.0024	1.00
			ξ_{AIPW}	0.503	0.0017	0.0017	0.502	0.0019	0.0018	1.11
			ξ_{SPML}	0.500	0.0016	0.0016	0.499	0.0016	0.0018	1.31
1.5	0.5	0	ξ_{SRS}	0.500	0.0036	0.0034	0.001	0.0035	0.0034	0.74
			ξ_R	0.498	0.0025	0.0025	−0.001	0.0026	0.0025	1.00

				$\hat{\beta}_1$			$\hat{\gamma}_1$			
a	**β_1**	**γ_1**	**Methods**	**Mean**	**VAR**	$\widehat{\text{VAR}}$	**Mean**	**VAR**	$\widehat{\text{VAR}}$	**SRE**
			ξ_{IPW}	0.501	0.0023	0.0022	0.000	0.0026	0.0024	1.00
			ξ_{AIPW}	0.502	0.0018	0.0016	0.000	0.0018	0.0017	1.44
			ξ_{SPML}	0.500	0.0014	0.0014	0.002	0.0014	0.0014	1.86
		0.5	ξ_{SRS}	0.499	0.0035	0.0034	0.499	0.0034	0.0034	0.68
			ξ_R	0.502	0.0026	0.0025	0.501	0.0027	0.0025	0.85
			ξ_{IPW}	0.500	0.0021	0.0022	0.500	0.0023	0.0024	1.00
			ξ_{AIPW}	0.501	0.0018	0.0018	0.501	0.0019	0.0019	1.21
			ξ_{SPML}	0.499	0.0014	0.0015	0.499	0.0017	0.0017	1.35

Abbreviation: SRE, sample relative efficiency. Results are based on the model $Y_1=\beta_0+\beta_1 X+\beta_2 Z+e$, $Y_2=\gamma_0+\gamma_1 X+\gamma_2 Z+\varepsilon$, where $X\sim N(0, 1)$, $Z\sim Bernoulli(0.45)$ and (e, ε) follow a bivariate normal distribution with $var(e)=\sigma_1^2$, $var(\varepsilon)=\sigma_2^2$, $cov(e,\varepsilon)=\rho\sigma_1\sigma_2$; the true parameter values are $\beta_0=1$, $\beta_2=-0.5$, $\gamma_0=1$, $\gamma_2=-0.5$, $\sigma_1=\sigma_2=1$, $\rho=0.8$. The cutoff points for the outcome-dependent sampling design are $\mu_{Y_1}-a\sigma_{Y_1}$ and $\mu_{Y_1}+a\sigma_{Y_1}$ ξ_{SRS} denotes the regression estimator based on simple random sample (SRS) portion of the validation sample. ξ_R denotes the regression estimator from an SRS of the same size as the validation sample. ξ_{IPW} denotes the estimate from our inverse probability weighted (IPW) estimating equation. ξ_{AIPW} denotes the estimate from augmented IPW (AIPW) estimating equation. ξ_{SPMI} is a semiparametric maximum likelihood (SPML) estimator similar to Jiang et al,[20] which models (Y_1, Y_2) parametrically using a bivariate normal distribution.

TABLE 3

Simulation results when ε is not normal. The full cohort size is N=3000, (n_0, n_1, n_3)=(200, 100, 100)

				$\hat{\gamma}_1$			
α	β_1	γ_1	Methods	Mean	VAR	$\widehat{\text{VAR}}$	CI
1.0	0	0	ξ_{IPW}	0.003	0.0038	0.0036	0.941
			ξ_{AIPW}	0.003	0.0039	0.0036	0.926
			ξ_{SPML}	0.001	0.0027	0.0026	0.960
		0.5	ξ_{IPW}	0.499	0.0038	0.0037	0.946
			ξ_{AIPW}	0.502	0.0037	0.0035	0.932
			ξ_{SPML}	0.493	0.0055	0.0041	0.871
	0.5	0	ξ_{IPW}	0.002	0.0036	0.0034	0.933
			ξ_{AIPW}	0.001	0.0032	0.0030	0.930
			ξ_{SPML}	−0.001	0.0021	0.0021	0.953
		0.5	ξ_{IPW}	0.501	0.0031	0.0032	0.954
			ξ_{AIPW}	0.505	0.0026	0.0027	0.940
			ξ_{SPML}	0.494	0.0038	0.0029	0.899

Results are based on the model $Y_1=\beta_0+\beta_1 X+\beta_2 Z+e$, $Y_2=\gamma_0+\gamma_1 X+\gamma_1 Z+\varepsilon$, where $X\sim N(0, 1)$, $Z\sim Bernoulli(0.45)$, and $e\sim N(0, 1)$, ε is a gamma distribution with shape parameter 2, rate parameter 1, normalized to have mean 0 and variance 1. The true parameter values are β_0=1, β_2=−0.5, γ_0=1, γ_2=−0.5. The cutoff points for the outcome-dependent sampling design are $\mu_{Y_1} - a\sigma_{Y_1}$ and $\mu_{Y_1} + a\sigma_{Y_1}$. ξ_{IPW} denotes the estimate from our inverse probability weighted (IPW) estimating equation; ξ_{aipw} denotes the estimate from augmented IPW (AIPW) estimating equation; ξ_{SPML} is a semiparametric maximum likelihood (SPML) estimator similar to Jiang et al,[20] which models (Y_1, Y_2) parametrically using a bivariate normal distribution.

TABLE 4

Analysis for a secondary outcome: child's birth weight in CPP study

Variables		SRS	IPW	AIPW
Int	Estimate	3208.14	3209.44	3089.24
	SE	86.54	79.49	21.97
	95% CI	(3038.52, 3377.76)	(3053.64, 3365.24)	(3046.18, 3132.30)
PCB	Estimate	−0.67	−5.44	−5.49
	SE	9.20	8.41	8.46
	95% CI	(−18.70, 17.36)	(−21.92, 11.04)	(−22.07, 11.09)
EDU	Estimate	−8.97	−8.81	1.70
	SE	9.10	8.39	1.36
	95% CI	(−26.81, 8.87)	(−25.25, 7.63)	(−0.97, 4.37)
SES	Estimate	13.94	16.06	13.58
	SE	10.77	10.33	2.16
	95% CI	(−7.17, 35.05)	(−4.19, 36.31)	(9.35, 17.81)
RACE	Estimate	204.10	206.16	189.84
(WHITE = 1)	SE	38.13	37.93	6.29
	95% CI	(129.37, 278.83)	(131.82, 280.50)	(177.51, 202.17)
GENDER	Estimate	−147.69	−145.65	−119.25
(FEMALE = 1)	SE	34.72	31.96	5.33
	95% CI	(−215.74, −79.64)	(−208.29, −83.01)	(−129.70, −108.80)

The response variable is child's birth weight in grams. The expensive exposure is mother's polychlorinated biphenyl (PCB) level. Other confounding variables include: parent's education level (EDU), social economic status of the child's family (SES), race ethnicity of the child (RACE), and gender of the child (GENDER). The results under simple random sample (SRS) are the regression analysis using the SRS portion of the outcome-dependent sampling sample. Inverse probability weighted (IPW) is the inverse probability weighted estimating equation, and augmented IPW (AIPW) is the augmented inverse probability weighted estimating equation we proposed. SE, standard error.