Published in final edited form as:
Ann Stat. 2013 February 1; 41(1): 269–295. doi:10.1214/12-AOS1073.

WEIGHTED LIKELIHOOD ESTIMATION UNDER TWO-PHASE SAMPLING

Takumi Saegusa[1] and **Jon A. Wellner**[2]
Department of Biostatistics, University of Washington, Seattle, Washington 98195-7232, USA, tsaegusa@uw.edu

Department of Statistics, University of Washington, Seattle, Washington 98195-4322, USA, jaw@stat.washington.edu

Abstract

We develop asymptotic theory for weighted likelihood estimators (WLE) under two-phase stratified sampling without replacement. We also consider several variants of WLEs involving estimated weights and calibration. A set of empirical process tools are developed including a Glivenko–Cantelli theorem, a theorem for rates of convergence of *M*-estimators, and a Donsker theorem for the inverse probability weighted empirical processes under two-phase sampling and sampling without replacement at the second phase. Using these general results, we derive asymptotic distributions of the WLE of a finite-dimensional parameter in a general semiparametric model where an estimator of a nuisance parameter is estimable either at regular or nonregular rates. We illustrate these results and methods in the Cox model with right censoring and interval censoring. We compare the methods via their asymptotic variances under both sampling without replacement and the more usual (and easier to analyze) assumption of Bernoulli sampling at the second phase.

Key words and phrases

Calibration; estimated weights; weighted likelihood; semiparametric model; regular; nonregular

1. Introduction

Two-phase sampling is a sampling technique that aims at cost reduction and improved efficiency of estimation. At phase I, a large sample is drawn from a population, and information on variables that are easier to measure is collected. These phase I variables may be important variables such as exposure in a regression model, or simply may be auxiliary variables that are correlated with unavailable variables at phase I. The sample space is then stratified based on these phase I variables. At phase II, a subsample is drawn without replacement from each stratum to obtain phase II variables that are costly or difficult to measure. Strata formation seeks either to oversample subjects with important phase I variables, or to effectively sample subjects with targeted phase II variables, or both. This way, two-phase sampling achieves effective access to important variables with less cost.

[1] Supported by NIH/NIAID R01 AI089341.

[2] Supported in part by NSF Grant DMS-11-04832, NI-AID Grant 2R01 AI291968-04 and the Alexander von Humboldt Foundation.

Supplementary material for "Weighted likelihood estimation under two-phase sampling" (DOI: 10.1214/12-AOS1073SUPP;.pdf). Due to space constraints, the proofs and technical details have been given in the supplementary document [25]. References here beginning with "A." refer to [25].

While two-phase sampling was originally introduced in survey sampling by Neyman [20] for estimation of the "finite population mean" of some variable, it has become increasingly important in a variety of areas of statistics, biostatistics and epidemiology, especially since [22, 33] and [27].

The setting treated here is as follows:

- We begin with a semiparametric model $\mathscr{P}$ for a vector of variables X with values in $\mathscr{X}$ [The prime examples which we treat in detail in Section 4 are the Cox proportional hazards regression model with (a) right censoring, and (b) interval censoring.]
- Let $W = (X, U) \in \mathscr{X} \times \mathscr{U} \equiv \mathscr{W}$ where U is a vector of "auxiliary variables," not involved in the model $\mathscr{P}$ Suppose that $W \sim \tilde{P}_0$ and $X \sim P_0$. Now suppose that $V \equiv (\tilde{X}, U) \in \mathscr{V}$ where $\tilde{X} \equiv \tilde{X}(X)$ is a coarsening of X.
- At phase I we observe $V_1, \ldots, V_N$ i.i.d. as V, and then use the phase I data to form strata, that is, disjoint subsets $\mathscr{V}_1, \ldots, \mathscr{V}_J$ of $\mathscr{V}$ with $\sum_{j=1}^{J} \mathscr{V}_j = \mathscr{V}$. We let $N_j = \#\{i \leq N : V_i \in \mathscr{V}_j\}$.
- Next, a phase II sample is drawn by sampling without replacement $n_j \leq N_j$ items from stratum j. For the items selected we observe X_i. Thus for the selection indicators ξ_i we have $\tilde{P}_0 (\xi_i = 1|V_i) = (n_j/N_j)1_{\mathscr{V}_j} (V_i) \equiv \pi_0 (V_i)$.
- Finally weighted likelihood (or inverse probability weighted) estimation methods based on all the observed data are used to estimate the parameters of the model $\mathscr{P}$ and to make further inferences about the model.

It is now well known that the classical Horvitz–Thompson estimators [9] use only the phase II data and are inefficient, sometimes quite severely so; see, for example, [2, 3, 14, 23] and [34]. Improvements in efficiency of estimation can be achieved by "adjusting" the weights by use of the phase I data (even though the sampling probabilities are known). Two basic methods of adjustment are:

1. Estimated weights, a method originating in the missing data literature [23], and with significant further developments since in connection with many models in which the missing-ness mechanism is not known, in contrast to our current setting in which the missing-ness is by design.
2. Calibration, a method originating in the sample survey literature [8]; see also [13, 14].

One of our goals here is to study existing methods for adjustment of the weights of weighted likelihood methods and to introduce several new methods: modified calibration as suggested by Chan [6] and centered calibration as proposed here in Section 2.

A second goal is to give a systematic treatment of estimators based on sampling without replacement at phase II in the setting of general semiparametric models and to make comparisons with the behavior of estimators based on Bernoulli (or independent) sampling at phase II, thus continuing and strengthening the comparisons made in [4, 5], and [2, 3] for a particular sub-class of semiparametric models and adjustments via estimated weights and ordinary calibration. Many studies of the theoretical properties of procedures based on two-phase design data have been made for the case of Bernoulli sampling; see, for example, [11] and the review of case-cohort sampling given there. On the other hand, while statistical practice continues to involve phase II data sampled without replacement, most available theory in this case (other than [4, 5]) has developed on a model-by-model basis. As has become clear from [4, 5], sampling without replacement results in smaller asymptotic variances, and hence inference based on asymptotic variances derived from Bernoulli sampling will often be conservative. Our treatment here provides theory and tools for dealing directly with the sampling without replacement design. We do this by providing the relevant theory both for semiparametric models in which the infinite-dimensional nuisance parameters can be estimated at a regular rate ($\sqrt{n}$) with complete data, and semiparametric models in which the infinite-dimensional nuisance parameters can only be estimated at slower (nonregular) rates.

The main contributions of our paper are three-fold: First, we establish two Z-theorems giving weak sufficient conditions for asymptotic distributions of the WLEs in general semiparametric models. The first theorem covers the case where the nuisance parameter is estimable at a regular rate; this yields rigorous justification of [2, 3] under weaker conditions. The second theorem covers the case of general semiparametric models with nonregular rates for estimators of the nuisance parameters. The conditions of our theorems, formulated in terms of complete data, are almost identical to those for the MLE with complete data. This formulation allows us to use tools from empirical process theory together with the new tools developed here in a straightforward way. Second, we propose centered calibration, a new calibration method. This new calibration method is the only one guaranteed to yield improved efficiency over the plain WLE under both Bernoulli sampling and sampling without replacement, while other methods are warranted only for Bernoulli sampling. Third, we establish general results for the inverse probability weighted (IPW) empirical process. Some results such as a Glivenko–Cantelli theorem (Theorem 5.1) and a Donsker theorem (Theorem 5.3) are of interest in their own right. These results accounting for dependence due to the sampling design are useful in verifying the conditions of Z-theorems in applications. For instance, Theorems 5.1 and 5.2 easily establish consistency and rates of convergence under our “without replacement” sampling scheme. We illustrate application of the general results with examples in Section 4.

The rest of the paper is organized as follows. In Section 2, we introduce our estimation procedures in the context of a general semiparametric model. The WLE and methods involving adjusted weights are discussed. Two Z-theorems are presented in Section 3; these yield asymptotic distributions of the WLEs of finite-dimensional parameters of the model. All estimators are compared under Bernoulli sampling and sampling without replacement with different methods of adjusting weights. In Section 4 we apply our Z-theorems to the Cox model with both right censoring and interval censoring. Section 5 consists of general results for IPW empirical processes. Several open problems are briefly discussed in Section 6. All proofs, except those in Section 4 and auxiliary results, are collected in [25].

2. Sampling, models and estimators

We use the basic notation introduced in the previous section. After stratified sampling, X is fully observed for n_j subjects in the jth stratum at phase II. The observed data is $(V, X\xi, \xi)$ where ξ is the indicator of being sampled at phase II. We use a doubly subscripted notation: for example, $V_{j,i}$ denotes $\underset{\sim}{V}$ for the ith subject in stratum j. We denote the stratum probability for the jth stratum by $\nu_j \equiv P_0(\underset{\sim}{V} \in \mathscr{V}_j)$, and the conditional expectation given membership in the jth stratum by $P_{0|j}\,(\cdot) \equiv P_0(\cdot|V \in \mathscr{V}_j)$.

The sampling probability is $P(\xi = 1|V_i) = \pi_0\,(V_i) = n_j/N_j$ for $V_i \in \mathscr{V}_j$. These sampling probabilities are assumed to be strictly positive; that is, there is a constant $\sigma > 0$ such that $0 < \sigma \le \pi_0(\upsilon) \le 1$ for $\upsilon \in \mathscr{V}$ We assume that $n_j/N_j \to p_j > 0$ for $j = 1, \ldots, J$ as $N \to \infty$. Although dependence is induced among the observations $(V_i, \xi_i\, X_i, \xi_i)$ by the sampling indicators, the vector of sampling indicators ($\xi_{j1}, \ldots, \xi_{jN_j}$) within strata, are exchangeable for each $j = 1, \ldots, J$, and the J random vectors ($\xi_{j1}, \ldots, \xi_{jN_j}$) are independent.

The empirical measure is one of the most useful tools in empirical process theory. Because the X_i's are observed only for a sub-sample at phase II, we define, instead, the IPW empirical measure $\mathbb{P}_N^\pi$ by

$$\mathbb{P}_N^\pi = \frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_0(V_i)}\delta_{X_i} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j}\frac{\xi_{j,i}}{n_j/N_j}\delta_{X_{j,i}},$$

where δ_{X_i} denotes a Dirac measure placing unit mass on X_i. The identity in the last display is justified by the arguments in Appendix A of [4]. We also define the IPW empirical process by $\mathbb{G}_N^\pi = \sqrt{N}(\mathbb{P}_N^\pi - P_0)$ and the phase II empirical process for the jth stratum by $\mathbb{G}_{j,N_j}^\xi \equiv \sqrt{N_j}(\mathbb{P}_{j,N_j}^\xi - (n_j/N_j)\mathbb{P}_{j,N_j})$, $j = 1, \ldots, J$, $\mathbb{P}_{j,N_j}^\xi \equiv N_j^{-1}\sum_{i=1}^{N_j}\xi_{j,i}\delta_{X_{j,i}}$ is the phase II empirical measure for the jth stratum, and $\mathbb{P}_{j,N_j} \equiv N_j^{-1}\sum_{i=1}^{N_j}\delta_{X_{j,i}}$ is the empirical measure for all the data in the jth stratum; note that the latter empirical measure is not observed. Then, following [4], we decompose $\mathbb{G}_N^\pi$ as follows:

$$\mathbb{G}_N^\pi = \mathbb{G}_N + \sum_{j=1}^{J}\sqrt{\frac{N_j}{N}}\left(\frac{N_j}{n_j}\right)\mathbb{G}_{j,N_j}^\xi, \quad (2.1)$$

where $\mathbb{P}_N = N^{-1}\sum_{j=1}^{J} N_j\mathbb{P}_{j,N_j}$ and $\mathbb{G}_N = \sqrt{N}(\mathbb{P}_N - P_0)$. Notice that $\mathbb{G}_{j,N_j}^\xi$ correspond to "exchangeably weighted bootstrap" versions of the stratum-wise complete data empirical processes $\mathbb{G}_{j,N_j} \equiv \sqrt{N_j}(\mathbb{P}_{j,N_j} - P_{0|j})$. This observation allows application of the "exchangeably weighted bootstrap" theory of [21] and [32], Section 3.6.

2.1. Improving efficiency by adjusting weights

Efficiency of estimators based on IPW empirical processes can be improved by adjusting weights, either by estimated weights [23] or by calibration [8] via use of the phase I data; see also [14]. Besides these, we discuss two variants of calibration, modified calibration [6], and our proposed new method, centered calibration.

Let $Z_i \equiv g(V_i)$ be the auxiliary variables for the ith subject for a known transformation g. For estimated weights with binary regression, Z_i contains the membership indicators for the strata $I_{\mathscr{V}_j}(V_i)$, $j = 1, \ldots, J$. Observations with $\pi_0(V) = 1$ are dropped from binary regression, and the original weight 1 is used. For notational simplicity, we write Z_i for either method, and assume that sampling probabilities are strictly less than 1 for all strata.

2.1.1. Estimated weights

2.1.1. Estimated weights—The method of estimated weights adjusts weights through binary regression on the phase I variables. The sampling probability for the ith subject is modeled by $p_\alpha(\xi_i|Z_i) = G_e(Z_i^T\alpha)^{\xi_i}(1 - G_e(Z_i^T\alpha))^{1-\xi_i} \equiv \pi_\alpha(V_i)^{\xi_i}\{1 - \pi_\alpha(V_i)\}^{1-\xi_i}$, where $\alpha \in \mathscr{A}_e \subset \mathbb{R}^{J+k}$ is a regression parameter and $G_e : \mathbb{R} \mapsto [0, 1]$ is a known function. If $G_e(x) = e^x/(1+e^x)$, for instance, then the adjustment simply involves logistic regression. Let $\hat{\alpha}_N$ be the estimator of α that maximizes the pseudo- (or composite) likelihood

$$\prod_{i=1}^{N} p_\alpha(\xi_i|Z_i)=\prod_{i=1}^{N} G_e(Z_i^T\alpha)^{\xi_i}\left(1-G_e(Z_i^T\alpha)\right)^{1-\xi_i}. \qquad (2.2)$$

We define the IPW empirical measure with estimated weights by

$$\mathbb{P}_N^{\pi,e}=\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_{\widehat{\alpha}_N}(V_i)}\delta_{X_i}=\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_0(V_i)}\frac{\pi_0(V_i)}{G_e(Z_i^T\widehat{\alpha}_N)}\delta_{X_i},$$

and the IPW empirical process with estimated weights by $\mathbb{G}_N^{\pi,e}=\sqrt{N}(\mathbb{P}_N^{\pi,e}-P_0)$.

2.1.2. Calibration—Calibration adjusts weights so that the inverse probability weighted average from the phase II sample is equated to the phase I average, whereby the phase I information is taken into account for estimation. Specifically, we find an estimator $\widehat{\alpha}_N$ that is the solution for $\alpha \in \mathscr{A}_c \subset \mathbb{R}^k$ of the following calibration equation:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i G_c(V_i;\alpha)}{\pi_0(V_i)}Z_i=\frac{1}{N}\sum_{i=1}^{N}Z_i, \qquad (2.3)$$

where $G_c(V;\alpha) \equiv G(g(V)^T\alpha) = G(Z^T\alpha)$ for known G with $G(0) = 1$ and $\dot{G}(0) > 0$. We call $\pi_\alpha(V) \equiv \pi_0(V)/G_c(V;\alpha)$ the calibrated sampling probability. We define the *calibrated IPW empirical measure* by

$$\mathbb{P}_N^{\pi,c}=\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_{\widehat{\alpha}_N}(V_i)}\delta_{X_i}=\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_0(V_i)}G(Z_i^T\widehat{\alpha}_N)\delta_{X_i}$$

and the *calibrated IPW empirical process* by $\mathbb{G}_N^{\pi,c}=\sqrt{N}(\mathbb{P}_N^{\pi,c}-P_0)$.

Examples for G in the definition of G_c are listed in [8] (F in their notation). For $G(x)=1+x, \mathbb{P}_N^{\pi,c}X$ is a well-known regression estimator of the mean of X. Since we assume boundedness of G later, we may want to consider truncated versions of these examples instead. Note that choice of G in (variants of) calibration does not affect asymptotic results on WLEs.

As noted in [13], there are several different approaches to calibration. Here, and in introducing variants of calibration below, we adopt the view that calibration proceeds by making the smallest possible change in weights in order to match an estimated phase II average with the corresponding phase I average. Another approach proceeds via regression modeling of the variable X of interest and the auxiliary variables V, leading to a robustness discussion on effects of the validity of the model on estimation for X. We prefer the former view because we do not assume a model for X and V throughout this paper. In fact, our results are independent of such a modeling assumption.

2.1.3. Modified calibration—Modifying the function G_c in calibration so that individuals with higher sampling probabilities $\pi(V_i)$ receive less weight was proposed by [6] in a missing response problem where observations are i.i.d. (see, e.g., [28] for recent developments in this area and [14] for their connections with calibration methods). An interpretation of this method within the framework of [8] is discussed in [26]. In modified calibration, we find the estimator $\hat{\alpha}_N$ that is the solution for $\alpha \in \mathscr{A}_{\mathrm{mc}} \subset \mathbb{R}^k$ of the following calibration equation:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i G_{\mathrm{mc}}(V_i;\alpha)}{\pi_0(V_i)}Z_i=\frac{1}{N}\sum_{i=1}^{N}Z_i, \quad (2.4)$$

where $G_{\mathrm{mc}}(V;\alpha) \equiv G((\pi_0(V)^{-1} - 1)Z^T\alpha)$ for known G with $G(0) = 1$ and $\dot{G}(0) > 0$. We call $\pi_\alpha(V) \equiv \pi_0(V)/G_{\mathrm{mc}}(V;\alpha)$ the *calibrated sampling probability with modified calibration*. We define the *IPW empirical measure with modified calibration* by

$$\mathbb{P}_N^{\pi,\mathrm{mc}}=\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_{\hat{\alpha}_N}(V_i)}\delta_{X_i}=\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_0(V_i)}G(\frac{1-\pi_0(V_i)}{\pi_0(V_i)}Z_i^T\hat{\alpha}_N)\delta_{X_i}$$

and the corresponding IPW empirical process by $\mathbb{G}_N^{\pi,\mathrm{mc}}=\sqrt{N}(\mathbb{P}_N^{\pi,\mathrm{mc}}-P_0)$.

2.1.4. Centered calibration—We propose a new method, centered calibration, that calibrates on centered auxiliary variables with modified calibration. This method improves the plain WLE under our sampling scheme, while retaining the good properties of modified calibration. See Section 3.4 for a discussion of its advantage and connections to other methods.

In centered calibration, we find the estimator $\hat{\alpha}_N$ that is the solution for $\alpha \in \mathscr{A}_{\mathrm{cc}} \subset \mathbb{R}^k$ of the following calibration equation:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i G_{\mathrm{cc}}(V_i;\alpha)}{\pi_0(V_i)}(Z_i-\overline{Z}_N)=0, \quad (2.5)$$

where $G_{\mathrm{cc}}(V;\alpha) \equiv G((\pi_0(V)^{-1} - 1)(Z - \overline{Z_N})^T\alpha)$ for known G with $G(0) = 1$ and $\dot{G}(0) > 0$ and $\overline{Z}_N=N^{-1}\sum_{i=1}^{N}Z_i$. We call $\pi_\alpha(V) \equiv \pi_0(V)/G_{\mathrm{cc}}(V;\alpha)$ the *calibrated sampling probability with centered calibration*. We define the *IPW empirical measure with centered calibration* by

$$\mathbb{P}_N^{\pi,\mathrm{cc}}=\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_{\hat{\alpha}_N}(V_i)}\delta_{X_i}=\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_0(V_i)}G_{\mathrm{cc}}(V_i;\hat{\alpha}_N)\delta_{X_i}$$

and the corresponding IPW empirical process by $\mathbb{G}_N^{\pi,\mathrm{cc}}=\sqrt{N}(\mathbb{P}_N^{\pi,\mathrm{cc}}-P_0)$.

2.2. Estimators for a semiparametric model $\mathscr{P}$

We study the asymptotic distribution of the weighted likelihood estimator of a finite-dimensional parameter θ in a general semiparametric model $\mathscr{P}=\{P_{\theta,\eta}:\theta \in \Theta, \eta \in H\}$ where $\Theta \subset \mathbb{R}^p$ and the nuisance parameter space H is a subset of some Banach space $\mathscr{B}$. Let $P_0 = P_{\theta_0,\eta_0}$ denote the true distribution.

The MLE for complete data is often obtained as a solution to the infinite-dimensional likelihood equations. In such models, the WLE under two-phase sampling is obtained by solving the corresponding infinite-dimensional inverse probability weighted likelihood equations. Specifically, the WLE $(\hat{\theta}_N, \hat{\eta}_N)$ is a solution to the following weighted likelihood equations:

$$\Psi^{\pi}_{N,1}(\theta,\eta)=\mathbb{P}^{\pi}_{N}\dot{\ell}_{\theta,\eta}=o_{P*}(N^{-1/2}), \| \Psi^{\pi}_{N,2}(\theta,\eta)h \|_{\mathscr{H}}=\| \mathbb{P}^{\pi}_{N}(B_{\theta,\eta}h-P_{\theta,\eta}B_{\theta,\eta}h) \|_{\mathscr{H}}=o_{P*}(N^{-1/2}), \quad (2.6)$$

where $\dot{\ell}_{\theta,\eta}\in\mathscr{L}^{0}_{2}(P_{\theta,\eta})^{p}$ is the score function for θ, and the score operator $B_{\theta,\eta}:\mathscr{H}\mapsto\mathscr{L}^{0}_{2}(P_{\theta,\eta})$ is the bounded linear operator mapping a direction h in some Hilbert space $\mathscr{H}$ of one-dimensional submodels for η along which $\eta \to \eta_0$. The WLE with estimated weights $(\hat{\theta}_{N,e}, \hat{\eta}_{N,e})$, the calibrated WLE $(\hat{\theta}_{N,c}, \hat{\eta}_{N,c})$, the WLE with modified calibration $(\hat{\theta}_{N,\mathrm{mc}}, \hat{\eta}_{N,\mathrm{mc}})$ and the WLE with centered calibration $(\hat{\theta}_{N,\mathrm{cc}}, \hat{\eta}_{N,\mathrm{cc}})$ are obtained by replacing $\mathbb{P}^{\pi}_{N}$ by $\mathbb{P}^{\pi,\#}_{N}$ with $\# \in \{e, c, \mathrm{mc}, \mathrm{cc}\}$ in (2.6), respectively. Let $\dot{\ell}_0 = \dot{\ell}_{\theta_0,\eta_0}$, and $B_0 = B_{\theta_0,\eta_0}$.

3. Asymptotics for the WLE in general semiparametric models

We consider two cases: in the first case the nuisance parameter η is estimable at a regular (i.e., $\sqrt{n}$) rate, and for ease of exposition, η is assumed to be a measure. In the second case η is only estimable at a nonregular (slower than $\sqrt{n}$) rate. Our theorem (Theorem 3.2) concerning the second case nearly covers the former case, but requires slightly more smoothness and a separate proof of the rate of convergence for an estimator of η. On the other hand, our theorem (Theorem 3.1) concerning the former case includes a proof of the (regular) ($\sqrt{n}$) rate of convergence, and hence is of interest by itself.

3.1. Conditions for adjusting weights

We assume the following conditions for estimators of α for adjusted weights. Throughout this paper, we may assume both Conditions 3.1 and 3.2 at the same time, but it should be understood that the former condition is used exclusively for the estimators regarding estimated weights and the latter condition is imposed only for estimators regarding (variants of) calibration. Also, it should be understood that Conditions 3.2(a)(i) and (d)(i), Conditions 3.2(a)(ii) and (d)(ii) and Conditions 3.2(a)(iii) and (d)(iii) are assumed for estimators defined via calibration, modified calibration and centered calibration, respectively.

Condition 3.1 (Estimated weights). (a) The estimator $\hat{\alpha}_N$ is a maximizer of the pseudo-likelihood (2.2).

- **(b)** $Z \in \mathbb{R}^{J+k}$ is not concentrated on a $(J + k)$-dimensional affine space of $\mathbb{R}^{J+k}$ and has bounded support.
- **(c)** $G_e : \mathbb{R} \mapsto [0, 1]$ is a twice continuously differentiable, monotone function.
- **(d)** $S_0 \equiv P_0[\{\dot{G}_e(Z^T \alpha_0)\}^2\{\pi_0(V)(1-\pi_0(V))\}^{-1} Z^{\otimes 2}]$ is finite and nonsingular, where $\dot{G}_e$ is a derivative of G_e.
- **(e)** The “true” parameter $\alpha_0 = (\alpha_{0,1}, \ldots, \alpha_{0,J+k})$ is given by $\alpha_{0,j}=G_e^{-1}(p_j)$ for $j = 1, \ldots, J$ and $\alpha_{0,j} = 0$, for $j = J + 1, \ldots, J + k$. The parameter α is identifiable, that is, $p_\alpha = p_{\alpha_0}$ almost surely implies $\alpha = \alpha_0$.
- **(f)** For a fixed $p_j \in (0, 1)$, n_j satisfies $n_j = [N_j p_j]$ for $j = 1, \ldots, J$.

Condition 3.2 (Calibrations). (a) (i) The estimator $\hat{\alpha}_N=\hat{\alpha}^{c}_{N}$ is a solution of calibration equation (2.3). (ii) The estimator $\hat{\alpha}_N=\hat{\alpha}^{\mathrm{mc}}_{N}$ is a solution of calibration equation (2.4). (iii) The estimator $\hat{\alpha}_N=\hat{\alpha}^{\mathrm{cc}}_{N}$ is a solution of calibration equation (2.5).

- **(b)** $Z \in \mathbb{R}^k$ is not concentrated at 0 and has bounded support.

(c) G is a strictly increasing continuously differentiable function on $\mathbb{R}$ such that $G(0) = 1$ and for all x, $-\infty < m_1 \leq G(x) \leq M_1 < \infty$ and $0 < \dot{G}(x) \leq M_2 < \infty$, where $\dot{G}$ is the derivative of G.

(d) (i) $P_0 Z^{\otimes 2}$ is finite and positive definite. (ii) $P_0[\pi_0(V)^{-1}(1 - \pi_0(V))Z^{\otimes 2}]$ is finite and positive definite. (iii) $P_0[\pi_0(V)^{-1}(1 - \pi_0(V))(Z - \mu_Z)^{\otimes 2}]$ is finite and positive definite where $\mu_Z = PZ$.

(e) (e) The "true" parameter $\alpha_0 = 0$.

Condition 3.1(f) may seem unnatural at first, but in practice the phase II sample size n_j can be chosen by the investigator so that the sampling probability p_j can be understood to be automatically chosen to satisfy $n_j = [N_j p_j]$. The other parts of Condition 3.1 are standard in binary regression, and Condition 3.2 is similar to Condition 3.1.

Asymptotic properties of $\hat{\alpha}_N$ for all methods are proved in [25].

3.2. Regular rate for a nuisance parameter

We assume the following conditions.

Condition 3.3 (Consistency). The estimator $(\hat{\theta}_N, \hat{\eta}_N)$ is consistent for (θ_0, η_0) and solves the weighted likelihood equations (2.6), where $\mathbb{P}_N^{\pi}$ may be replaced by $\mathbb{P}_N^{\pi,\#}$ with $\# \in \{e, c, \text{mc}, \text{cc}\}$ for the estimators with adjusted weights.

Condition 3.4 (Asymptotic equicontinuity). Let $\mathscr{F}_1(\delta) = \{\dot{\ell}_{\theta,\eta} : |\theta - \theta_0| + \|\eta - \eta_0\| < \delta\}$ and $\mathscr{F}_2(\delta) = \{B_{\theta,\eta}h - P_{\theta,\eta}B_{\theta,\eta}h : h \in \mathscr{H}, |\theta - \theta_0| + \|\eta - \eta_0\| < \delta\}$. There exists a $\delta_0 > 0$ such that (1) $\mathscr{F}_k(\delta_0)$, $k = 1, 2$, are P_0-Donsker and $\sup_{h\in\mathscr{H}} P_0|f_j - f_{0,j}|^2 \to 0$, as $|\theta - \theta_0| + \|\eta - \eta_0\| \to 0$, for every $f_j \in \mathscr{F}_j(\delta_0)$, $j = 1, 2$, where $f_{0,1} = \ell_{\theta_0,\eta_0}$ and $f_{0,2} = B_0 h - P_0 B_0 h$, (2) $\mathscr{F}_k(\delta_0)$, $k = 1, 2$, have integrable envelopes.

Condition 3.5. The map $\Psi = (\Psi_1, \Psi_2) : \Theta \times H \mapsto \mathbb{R}^p \times \ell^\infty(\mathscr{H})$ with components

$$\Psi_1(\theta,\eta) \equiv P_0\Psi_{N,1}(\theta,\eta) = P_0\dot{\ell}_{\theta,\eta}, \;\; \Psi_2(\theta,\eta)h \equiv P_0\Psi_{N,2}(\theta,\eta) = P_0B_{\theta,\eta}h - P_{\theta,\eta}B_{\theta,\eta}h, h \in \mathscr{H},$$

has a continuously invertible Fréchet derivative map $\dot{\Psi}_0 = (\dot{\Psi}_{11}, \dot{\Psi}_{12}, \dot{\Psi}_{21}, \dot{\Psi}_{22})$ at (θ_0, η_0) given by $\dot{\Psi}_{ij}(\theta_0, \eta_0)h = P_0(\dot{\psi}_{i,j,\theta_0,\eta_0,h})$, $i, j \in \{1, 2\}$ in terms of $L_2(P_0)$ derivatives of $\psi_{1,\theta,\eta,h} = \ell_{\theta,\eta}$ and $\psi_{2,\theta,\eta,h} = \mathrm{B}_{\theta,\eta}h - P_{\theta,\eta}B_{\theta,\eta}h$; that is,

$$\sup_{h\in\mathscr{H}}\left[P_0\{\psi_{i,\theta,\eta_0,h} - \psi_{i,\theta_0,\eta_0,h} - \dot{\psi}_{i1,\theta_0,\eta_0,h}(\theta - \theta_0)\}^2\right]^{1/2} = o(\| \theta - \theta_0 \|),$$

$$\sup_{h\in\mathscr{H}}\left[P_0\{\psi_{i,\theta_0,\eta,h} - \psi_{i,\theta_0,\eta_0,h} - \dot{\psi}_{i2,\theta_0,\eta_0,h}(\eta - \eta_0)\}^2\right]^{1/2} = o(\| \eta - \eta_0 \|).$$

Furthermore, $\dot{\Psi}_0$ admits a partition

$$(\theta - \theta_0, \eta - \eta) \mapsto \begin{pmatrix} \dot{\Psi}_{11} & \dot{\Psi}_{12} \\ \dot{\Psi}_{21} & \dot{\Psi}_{22} \end{pmatrix} \begin{pmatrix} \theta - \theta_0 \\ \eta - \eta_0 \end{pmatrix},$$

where

$$\begin{aligned}\dot{\Psi}_{11}(\theta-\theta_0)&=-P_{\theta_0,\eta_0}\dot{\ell}_{\theta_0,\eta_0}\dot{\ell}^T_{\theta_0,\eta_0}(\theta-\theta_0),\\ \dot{\Psi}_{12}(\eta-\eta_0)&=-\int B^*_{\theta_0,\eta_0}\dot{\ell}_{\theta_0,\eta_0}d(\eta-\eta_0),\\ \dot{\Psi}_{21}(\theta-\theta_0)h&=-P_{\theta_0,\eta_0}B_{\theta_0,\eta_0}h\dot{\ell}^T_{\theta_0,\eta_0}(\theta-\theta_0),\\ \dot{\Psi}_{22}(\eta-\eta_0)h&=-\int B^*_{\theta_0,\eta_0}B_{\theta_0,\eta_0}hd(\eta-\eta_0)\end{aligned}$$

and $B^*_{\theta_0,\eta_0}B_{\theta_0,\eta_0}$ is continuously invertible.

Let $\tilde{I}_0=P_0[(I-B_0(B_0^*B_0)^{-1}B_0^*)\dot{\ell}_0\dot{\ell}_0^T]$ be the efficient information for θ and $\tilde{\ell}_0=\tilde{I}_0^{-1}(I-B_0(B_0^*B_0)^{-1}B_0^*)\dot{\ell}_0$ be the efficient influence function for θ for the semiparametric model with complete data.

Theorem 3.1. *Under Conditions* 3.1–3.5,

$$\sqrt{N}(\widehat{\theta}_N-\theta_0)=\sqrt{N}\mathbb{P}_N^{\pi}\tilde{\ell}_0+o_{P*}(1)\rightsquigarrow Z\sim N_p(0,\Sigma),\quad \sqrt{N}(\widehat{\theta}_{N,\#}-\theta_0)=\sqrt{N}\mathbb{P}_N^{\pi,\#}\tilde{\ell}_0+o_{P*}(1)\rightsquigarrow Z_\#\sim N_p(0,\Sigma_\#),$$

where # ∈ {*e*, *c*, mc, cc},

$$\Sigma\equiv I_0^{-1}\sum_{j=1}^J\nu_j\frac{1-p_j}{p_j}\mathrm{Var}_{0|j}(\tilde{\ell}_0),\quad (3.1)$$

$$\Sigma_\#\equiv I_0^{-1}+\sum_{j=1}^J\nu_j\frac{1-p_j}{p_j}\mathrm{Var}_{0|j}((I-Q_\#)\tilde{\ell}_0)\quad (3.2)$$

and (recall Conditions 3.1 *and* 3.2)

$$\begin{aligned}Q_ef&\equiv P_0[\pi_0^{-1}(V)f\dot{G}_e(Z^T\alpha_0)Z^T]S_0^{-1}(1-\pi_0(V))^{-1}\dot{G}_e(Z^T\alpha_0)Z,\\ Q_cf&\equiv P_0[fZ^T]\{P_0Z^{\otimes 2}\}^{-1}Z,\\ Q_{\mathrm{mc}}f&\equiv P_0[(\pi_0^{-1}(V)-1)fZ^T]\{P_0[(\pi_0^{-1}(V)-1)Z^{\otimes 2}]\}^{-1}Z,\\ Q_{\mathrm{cc}}f&\equiv P_0[(\pi_0^{-1}(V)-1)f(Z-\mu_Z)^T]\{P_0[(\pi_0^{-1}(V)-1)(Z-\mu_Z)^{\otimes 2}]\}^{-1}\times(Z-\mu_Z).\end{aligned}$$

Remark 3.1. Our conditions in Theorem 3.1 are the same as those in [5] except the integrability condition. Our Condition 3.4(2) requires existence of integrable envelopes for class of scores while the condition (A1*) in [5] requires square integrable envelopes. Note that this integrability condition is required only for the WLE with adjusted weights, as in [4].

Remark 3.2. As can be seen from the definition of $Q_\#$, the choice of G in calibration does not affect the asymptotic variances while G_e in the method of estimated weights does affect the asymptotic variance.

3.3. Nonregular rate for a nuisance parameter

For $\underline{h}=(h_1,\dots,h_p)^T$ with $h_k\in H$, $k=1,\dots,p$, let $B_{\theta,\eta}[\underline{h}]=(B_{\theta,\eta}h_1,\dots,B_{\theta,\eta}h_p)^T$. We assume the following conditions.

Condition 3.6 (Consistency and rate of convergence). An estimator $(\hat{\theta}_N,\hat{\eta}_N)$ of (θ_0,η_0) satisfies $|\hat{\theta}_N - \theta_0| = o_P(1)$, and $\|\hat{\eta}_N - \eta_0\| = O_P(N^{-\beta})$ for some $\beta > 0$.

Condition 3.7 (Positive information). There is an $\underline{h}^*=(h_1^*,\ldots,h_p^*)$, where $h_k^* \in H$ for $k = 1, \ldots, p$, such that

$$P_0\{(\dot{\ell}_0 - B_0[\underline{h}^*])B_0h\}=0\text{for }\ \text{all}h \in H.$$

The efficient information $I_0 \equiv P_0(\dot{\ell}_0 - B_0[\underline{h}^*])^{\otimes 2}$ for θ for the semiparametric model with complete data is finite and nonsingular. Denote the efficient influence function for the semiparametric model with complete data by $\tilde{\ell}_0 \equiv I_0^{-1}(\dot{\ell}_0 - B_0[\underline{h}^*])$.

Condition 3.8 (Asymptotic equicontinuity). (1) For any $\delta_N \downarrow 0$ and $C > 0$,

$$\sup_{|\theta-\theta_0|\le\delta_N,\|\eta-\eta_0\|\le CN^{-\beta}}|\mathbb{G}_N(\dot{\ell}_{\theta,\eta}-\dot{\ell}_0)|=o_P(1), \quad \sup_{|\theta-\theta_0|\le\delta_N,\|\eta-\eta_0\|\le CN^{-\beta}}|\mathbb{G}_N(B_{\theta,\eta}-B_0)[\underline{h}^*]|=o_P(1).$$

(2) There exists a $\delta > 0$ such that the classes $\{\dot{\ell}_{\theta,\eta} : |\theta - \theta_0| + \|\eta - \eta_0\| \le \delta\}$ and $\{B_{\theta,\eta}[\underline{h}^*] : |\theta - \theta_0| + \|\eta - \eta_0\| \le \delta\}$ are P_0-Glivenko–Cantelli and have integrable envelopes. Moreover, $\dot{\ell}_{\theta,\eta}$ and $B_{\theta,\eta}[\underline{h}^*]$ are continuous with respect to (θ, η) either pointwise or in $L_1(P_0)$.

Condition 3.9 (Smoothness of the model). For some $\alpha > 1$ satisfying $\alpha\beta > 1/2$ and for (θ, η) in the neighborhood $\{(\theta, \eta) : |\theta - \theta_0| \le \delta_N, \|\eta - \eta_0\| \le CN^{-\beta}\}$,

$$|P_0\{\dot{\ell}_{\theta,\eta} - \dot{\ell}_0+\dot{\ell}_0(\dot{\ell}_0^T(\theta-\theta_0)+B_0[\eta-\eta_0])\}|=o(|\theta-\theta_0|)+O(\|\ \eta-\eta_0\ \|^{\alpha}),$$
$$|P_0\{(B_{\theta,\eta}-B_0)[\underline{h}^*]+B_0[\underline{h}^*](\dot{\ell}_0^T(\theta-\theta_0)+B_0[\eta-\eta_0])\}|=o(|\theta-\theta_0|)+O(\|\ \eta-\eta_0\ \|^{\alpha}).$$

In the previous section, we required that the WLE solves the weighted likelihood equations (2.6) for all $h \in \mathscr{H}$. Here, we only assume that the WLE $(\hat{\theta}_N, \hat{\eta}_N)$ satisfies the weighted likelihood equations

$$\Psi_{N,1}^{\pi}(\theta,\eta,\alpha)=\mathbb{P}_N^{\pi}\dot{\ell}_{\theta,\eta}=o_{P*}(N^{-1/2}),\ \ \Psi_{N,2}^{\pi}(\theta,\eta,\alpha)[\underline{h}^*]=\mathbb{P}_N^{\pi}B_{\theta,\eta}[\underline{h}^*]=o_{P*}(N^{-1/2}). \quad (3.3)$$

The corresponding WLEs with adjusted weights, $(\hat{\theta}_{N,\#}, \hat{\eta}_{N,\#})$ with $\# \in \{e, c, \text{mc}, \text{cc}\}$ satisfy (3.3) with $\mathbb{P}_N^{\pi}$ replaced by $\mathbb{P}_N^{\pi,\#}$.

Theorem 3.2. *Suppose that the WLE is a solution of* (3.3) *where* $\mathbb{P}_N^{\pi}$ *may be replaced by* $\mathbb{P}_N^{\pi,\#}$ with $\# \in \{e, c, \text{mc}, \text{cc}\}$ *for the estimators with adjusted weights. Under Conditions* 3.1, 3.2 *and* 3.6–3.9,

$$\sqrt{N}(\hat{\theta}_N-\theta_0)=\sqrt{N}\mathbb{P}_N^{\pi}\tilde{\ell}_0+o_{P*}(1)\rightsquigarrow Z\sim N_p(0,\Sigma),\ \ \sqrt{N}(\hat{\theta}_{N,\#}-\theta_0)=\sqrt{N}\mathbb{P}_N^{\pi,\#}\tilde{\ell}_0+o_{P*}(1)\rightsquigarrow Z_{\#}\sim N_p(0,\Sigma_{\#}),$$

where Σ *and* $\Sigma_{\#}$ *are as defined in* (3.1) *and* (3.2) *of Theorem* 3.1, *but now* I_0 *and* $\tilde{\ell}_0$ *are defined in Condition* 3.7, *and* $Q_{\#}$ *are defined in Theorem* 3.1.

Remark 3.3. Our conditions are identical to those of the Z-theorem of [10] except Condition 3.8(2). This additional condition is not stringent for the following reasons. First, the Glivenko–Cantelli condition is usually assumed to prove consistency of estimators before deriving asymptotic distributions. Second, a stronger $L_2(P_0)$-continuity condition is standard as is seen in Condition 3.4 (see also Section 25.8 of [31]). Note that the $L_1(P_0)$-continuity condition is only required for the WLEs with adjusted weights.

3.4. Comparisons of methods

We compare asymptotic variances of five WLEs in view of improvement by adjusting weights and change of designs. We also include in comparison special cases of adjusting weights involving stratum-wise adjustment.

3.4.1. Stratified Bernoulli sampling—We first give a statement of the result corresponding to Theorem 3.1 for stratified Bernoulli sampling where all subjects are independent with the sampling probability p_j if $V \in \mathscr{V}_j$ and $\hat{\theta}_N^{\text{Bern}}and\hat{\theta}_{N,\#}^{\text{Bern}}$ with # ∈ {*e*, *c*, mc, cc} are the corresponding WLE and WLEs with adjusted weights.

Theorem 3.3. *Suppose Conditions* 3.1 [*except Condition* 3.1(f)] *and* 3.2 *hold. Let* $\xi_i \in \{0, 1\}$ *and* ξ *be i.i.d. with* $E[\xi|V]=\pi_0(V)=\sum_{j=1}^{J} p_j I(V \in \mathscr{V}_j)$.

1. *Suppose that the WLE is a solution of* (3.3) where $\mathbb{P}_N^{\pi}$ *may be replaced by* $\mathbb{P}_N^{\pi,\#}$ *with* # ∈ {*e*, *c*, mc, cc} *for the estimators with adjusted weights. Under the same conditions as in Theorem* 3.1,

$$\sqrt{N}(\hat{\theta}_N^{\text{Bern}} - \theta_0)=\sqrt{N}\mathbb{P}_N^{\pi}\tilde{\ell}_0+o_{P*}(1) \rightsquigarrow Z^{\text{Bern}}\sim N_p(0,\Sigma^{\text{Bern}}),$$
$$\sqrt{N}(\hat{\theta}_{N,\#}^{\text{Bern}} - \theta_0)=\sqrt{N}\mathbb{P}_N^{\pi,\#}\tilde{\ell}_0+o_{P*}(1) \rightsquigarrow Z_{\#}^{\text{Bern}}\sim N_p(0,\Sigma_{\#}^{\text{Bern}}),$$

where

$$\Sigma^{\text{Bern}} \equiv I_0^{-1}+\sum_{j=1}^{J}\nu_j\frac{1-p_j}{p_j}P_{0|j}(\tilde{\ell}_0)^{\otimes 2}, \quad (3.4)$$

$$\Sigma_{\#}^{\text{Bern}} \equiv I_0^{-1}+\sum_{j=1}^{J}\nu_j\frac{1-p_j}{p_j}P_{0|j}((I-Q_{\#})\tilde{\ell}_0)^{\otimes 2}, \quad (3.5)$$

where $Q_{\#}$ *with* # ∈ {*e*, *c*, mc, cc} *are defined in Theorem* 3.1.

2. *Under the same conditions as in Theorem* 3.2, *the same conclusions in* (1) *hold with* I_0 *and* $\tilde{\ell}_0$ *replaced by those defined in Condition* 3.7.

Comparing the variance–covariance matrices in Theorem 3.3 to those in Theorems 3.1 and 3.2, we obtain the following corollary comparing designs. All estimators have smaller variances under sampling without replacement.

Corollary 3.1. *Under the same conditions as in Theorem* 3.3,

$$\Sigma=\Sigma^{\text{Bern}}-\sum_{j=1}^{J}\nu_j\frac{1-p_j}{p_j}\{P_{0|j}\tilde{\ell}_0\}^{\otimes 2}, \ \Sigma_{\#}=\Sigma_{\#}^{\text{Bern}}-\sum_{j=1}^{J}\nu_j\frac{1-p_j}{p_j}\{P_{0|j}(I-Q_{\#})\tilde{\ell}_0\}^{\otimes 2}, \# \in \{e,c,\text{mc},\text{cc}\}.$$

Variance formulas (3.5) with # ∈ {*e*, mc, cc} except for the ordinary calibration have the following alternative representations which show the efficiency gains over the plain WLE under Bernoulli sampling.

Corollary 3.2. *Under the same conditions as in Theorem* 3.3,

$$\Sigma_{\#}^{\mathrm{Bern}}=\Sigma^{\mathrm{Bern}}-\mathrm{Var}\left(\frac{\xi-\pi_0(V)}{\pi_0(V)}Q_{\#}\tilde{\ell}_0\right),\#\in\{e,\mathrm{mc},\mathrm{cc}\}.$$

3.4.2. Within-stratum adjustment of weights—Adjusting weights can be carried out in every stratum. This is proposed by Breslow et al. [2, 3] for ordinary calibration. Consider calibration on $\tilde{Z}$ where $\tilde{Z} \equiv (Z^{(1)}, \ldots, Z^{(J)})^T$ with $Z^{(j)} \equiv I(V \in \mathscr{V}_j)Z^T$. The calibration equation (2.3) becomes

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i G_c(\tilde{Z}_i;\alpha)}{\pi_0(V_i)}Z_i I(V_i\in\mathscr{V}_j)=\frac{1}{N}\sum_{i=1}^{N}Z_i I(V_i\in\mathscr{V}_j), j=1,\ldots,J,$$

where $\alpha \in \mathbb{R}^{Jk}$. We call this special case *within-stratum calibration*. We define *within-stratum modified and centered calibration* analogously.

We also call estimated weights carried out within stratum *within-stratum estimated weights*. Recall that *Z* in estimated weights contains the membership indicators for the strata and the rest are other auxiliary variables, say $Z^{[2]}$. Within-stratum estimated weights uses $\tilde{Z} \equiv (Z^{(1)}, \ldots, Z^{(J)})^T$ where $Z^{(j)} \equiv I(V \in \mathscr{V}_j)(Z^{[2]})^T$ with 1 included in $Z^{[2]}$. The "true" parameter $\tilde{\alpha}_0$ has zero for all elements except having $G_e^{-1}(p_j)$ for the element corresponding to $I(V \in \mathscr{V}_j)$, $j = 1, \ldots, J$.

The following corollary summarizes within-stratum adjustment of weights under stratified Bernoulli sampling and sampling without replacement. All methods achieve improved efficiency over the plain WLE under Bernoulli sampling while centered calibration is the only method to yield a guaranteed improvement under sampling without replacement. This is because centering yields the $L_2^0(P_{0|j})$-projection suitable for the conditional variances in (3.2) while noncentering results in the $L_2(P_{0|j})$-projection for the conditional expectations in (3.5).

Corollary 3.3. (1) (*Bernoulli*) *Under the same conditions as in Theorem* 3.3 *with Z replaced by* $\tilde{Z}$ *and* α_0 *replaced by* $\tilde{\alpha}_0$ *for within-stratum estimated weights,*

$$\Sigma_{\#}^{\mathrm{Bern}}=\Sigma^{\mathrm{Bern}}-\sum_{j=1}^{J}\nu_j\frac{1-p_j}{p_j}P_{0|j}(Q_{\#}^{(j)}\tilde{\ell}_0)^{\otimes 2}, \quad (3.6)$$

where # ∈ {*e*, *c*, mc, cc} *and*

$$\begin{aligned}
Q_e^{(j)}f &\equiv P_{0|j}[f\dot{G}_e(\tilde{Z}^T\tilde{\alpha}_0)(Z^{[2]})^T]\{P_{0|j}\dot{G}_e^2(\tilde{Z}^T\tilde{\alpha}_0)(Z^{[2]})^{\otimes 2}\}^{-1}\times\dot{G}_e(\tilde{Z}^T\tilde{\alpha}_0)I(V\in\mathscr{V}_j)Z^{[2]},\\
Q_c^{(j)}f &\equiv P_{0|j}[fZ^T]\{P_{0|j}[Z^{\otimes 2}]\}^{-1}I(V\in\mathscr{V}_j)Z,\\
Q_{\mathrm{mc}}^{(j)}f &\equiv Q_c^{(j)}f,\\
Q_{\mathrm{cc}}^{(j)}f &\equiv P_{0|j}[f(Z-\mu_{Z,j})^T]\{P_{0|j}[(Z-\mu_{Z,j})^{\otimes 2}]\}^{-1}I(V\in\mathscr{V}_j)(Z-\mu_{Z,j})
\end{aligned}$$

with $\mu_{Z,j} \equiv E[I\,(V \in \mathscr{V}_j)Z]$ *for* $j = 1, \ldots, J$.

(2) (*without replacement*) *Under the same conditions as in Theorems* 3.1 *or* 3.2 *with* Z *is replaced by* $\tilde{Z}$

$$\Sigma_{\mathrm{cc}}=\Sigma-\sum_{j=1}^{J}\nu_j\frac{1-p_j}{p_j}\mathrm{Var}_{0|j}(Q_{\mathrm{cc}}^{(j)}\tilde{\ell}_0). \quad (3.7)$$

.

3.4.3. Comparisons—We summarize Corollaries 3.1–3.3. Every method of adjusting weights improves efficiency over the plain WLE in a certain design and with a certain range of adjustment of weights (within-stratum or “across-strata” adjustment). However, particularly notable among all methods is centered calibration. While other methods gain efficiency only under Bernoulli sampling, centered calibration improves efficiency over the plain WLE under both sampling schemes. There is no known method of “across-strata” adjustment that is guaranteed to gain efficiency over the plain WLE under stratified sampling without replacement.

There are close connections among all methods. When the auxiliary variables have mean zero, centered and modified calibrations are essentially the same. The ordinary and modified calibrations give the same asymptotic variance when carried out stratum-wise. For Z and α_0 defined for estimated weights, estimated weights and modified calibration based on $(1 - \pi_0(V))^{-1}\,\dot{G}_e(Z^T\alpha_0)Z$ performs the same way. Similarly within-stratum estimated weights with $\tilde{Z}$ and $\tilde{\alpha}_0$ is as good as within-stratum calibration based on $\dot{G}_e(\tilde{Z}^T\tilde{\alpha}_0)\tilde{Z}$.

As seen in the relationship among methods, there is no single method superior to others in each situation. In fact, performance depends on choice and transformation of auxiliary variables, the true distribution P_0 and the design. For our “without replacement” sampling scheme, within-stratum centered calibration is the only method guaranteed to gain efficiency while other methods may perform even worse than the plain WLE.

4. Examples

For asymptotic normality of WLEs, consistency and rate of convergence need to be established first to apply our Z-theorems in Section 3. To this end, general results on IPW empirical processes discussed in the next section will be useful. We illustrate this in the Cox models with right censoring and interval censoring under two-phase sampling.

Let $T \sim F$ be a failure time, and X be a vector of covariates with bounded supports in the regression model. The Cox proportional hazards model [7] specifies the relationship

$$\Lambda(t|x)=\exp(\theta^T x)\Lambda(t),$$

where $\theta \in \Theta \subset \mathbb{R}^p$ is the regression parameter, $\Lambda \in H$ is the (baseline) cumulative hazard function. Here the space H for the nuisance parameter Λ is the set of nonnegative, nondecreasing cadlag functions defined on the positive line. The true parameters are θ_0 and Λ_0.

In addition to X, let U be a vector of auxiliary variables collected at phase I which are correlated with the covariate X. For simplicity of notation, we assume that the covariates X are only observed for the subject sampled at phase II. Thus, if some of the coordinates of X

are available at phase I, then we include an identical copy of those coordinates of X in the vector of U.

4.1. Cox model with right censored data

Under right censoring, we only observe the minimum of the failure time T and the censoring time $C \sim G$. Define the observed time $Y = T \wedge C$ and the censoring indicator $\Delta = I(T \leq C)$. The phase I data is $V = (Y, \Delta, U)$, and the observed data is $(Y, \Delta, \xi X, U, \xi)$ where ξ is the sampling indicator. With right censored data and complete data, the theory for maximum likelihood estimators in the Cox model has received several treatments; the one we follow most closely here is that of [31]. For the Cox model with case-cohort data, see [27] and for treatments with even more general designs [1] and [12]. Here, for both sampling without replacement and Bernoulli sampling, we continue the developments of [4, 5]. We assume the following conditions:

Condition 4.1. The finite-dimensional parameter space Θ is compact and contains the true parameter θ_0 as an interior point.

Condition 4.2. The failure time T and the censoring time C are conditionally independent given X, and that there is $\tau > 0$ such that $P(T > \tau) > 0$ and $P(C \geq \tau) = P(C = \tau) > 0$. Both T and C have continuous conditional densities given the covariates $X = x$.

Condition 4.3. The covariate X has bounded support. For any measurable function h, $P(X \neq h(Y)) > 0$.

Let $\lambda(t) = (d/dt)\Lambda(t)$ be the baseline hazard function. With complete data, the density of (Y, Δ, X) is

$$p_{\theta,\Lambda}(y,\delta,x)=\left\{\lambda(y)e^{\theta^T x-\Lambda(y)e^{\theta^T x}}(1-G)(y|x)\right\}^{\delta}\left\{e^{-\Lambda(y)e^{\theta^T x}}g(y|x)\right\}^{1-\delta}p_X(x),$$

where p_X is the marginal density of X and $g(\cdot|x)$ is the conditional density of C given $X = x$. The score for θ is given by $\dot{\ell}_{\theta,\Lambda}(y, \delta, x) = x\{\delta - e^{\theta^T x}\Lambda(y)\}$, and the score operator $B_{\theta,\Lambda} : \mathscr{H} \mapsto L_2(P_{\theta,\Lambda})$ is defined on the unit ball $\mathscr{H}$ in the space $BV[0, \tau]$ such that $B_{\theta,\Lambda}h(y, \delta, x) = \delta h(y) - e^{\theta^T x}\int_{[0,y]} h\, d\Lambda$. Because the likelihood based on the density above does not yield the MLE for complete data, we define the log likelihood for one observation for complete data by $\ell_{\theta,\Lambda}(y, \delta, x) = \log\{(e^{\theta^T x}\Lambda\{y\})^{\delta} e^{-\Lambda(y)e^{\theta^T x}}\}$ where $\Lambda\{t\}$ is the (point) mass of Λ at t. Then maximizing the weighted log likelihood $\mathbb{P}_N^{\pi}\ell_{\theta,\Lambda}$ reduces to solving the system of equations $\mathbb{P}_N^{\pi}\dot{\ell}_{\theta,\Lambda}=0\mathrm{and}\mathbb{P}_N^{\pi}B_{\theta,\Lambda}h=0$ for every $h \in \mathscr{H}$. The efficient score for θ for complete data is given by

$$\ell^*_{\theta_0,\Lambda_0}(y,\delta,x)=\delta(x-(M_1/M_0)(y))-e^{\theta_0^T x}\int_{[0,y]}\delta(x-(M_1/M_0)(t))d\Lambda_0(t),$$

, and the efficient information for θ for complete data is

$$\tilde{I}_{\theta_0,\Lambda_0}=E[(\ell^*_{\theta_0,\Lambda_0})^{\otimes 2}]=Ee^{\theta_0^T X}\int_0^{\tau}\left(X-\frac{M_1}{M_0}(y)\right)^{\otimes 2}(1-G)(y|X)d\Lambda_0(y),$$

where $M_k(s)=P_{\theta_0,\Lambda_0}[X^k e^{\theta_0^T X}I(Y \geq s)], k=0, 1.$

Theorem 4.1 (Consistency). *Under Conditions* 3.1, 3.2, 4.1–4.3, *the WLEs are consistent for* (θ_0, Λ_0).

Proof. This proof follows along the lines of the proof given by [29], but with the usual empirical measure replaced by the IPW empirical measure (with adjusted weights), and by use of Theorem 5.1. For details see [25].

Our *Z*-theorem (Theorem 3.1) yields asymptotic normality of the WLEs.

Theorem 4.2 (Asymptotic normality). *Under Conditions* 3.1, 3.2, 4.1–4.3,

$$\sqrt{N}(\widehat{\theta}_N - \theta_0) = \sqrt{N}\mathbb{P}_N^{\pi}\tilde{\ell}_{\theta_0,\Lambda_0} + o_{P*}(1) \to_d N(0, \Sigma), \quad \sqrt{N}(\widehat{\theta}_{N,\#} - \theta_0) = \sqrt{N}\mathbb{P}_N^{\pi,\#}\tilde{\ell}_{\theta_0,\Lambda_0} + o_{P*}(1) \to_d N(0, \Sigma_{\#}),$$

where # ∈ {*e*, *c*, mc, cc}, $\tilde{\ell}_{\theta_0,\Lambda_0} = I^{-1}_{\theta_0,\Lambda_0}\ell^*_{\theta_0,\Lambda_0}$ *is the efficient influence function for* θ *for complete data, and* Σ *and* $\Sigma_{\#}$ *are given in Theorem* 3.1.

Proof. We verify the conditions of Theorem 3.1. Condition 3.3 holds by Theorem 4.1. Conditions 3.4 and 3.5 hold under the present hypotheses as was shown in [31], Section 25.12.

For variance estimation regarding $\widehat{\theta}_N, \widehat{I}_N \equiv \mathbb{P}_N^{\pi}(\ell^*_{\widehat{\theta}_N,\widehat{\Lambda}_N})^{\otimes 2}$ can be used to estimate I_0. Letting $\widehat{\tilde{\ell}}_0 \equiv \widehat{I}_N^{-1}\ell^*_{\widehat{\theta}_N,\widehat{\Lambda}_N}$, we can estimate $\mathrm{Var}_{0|j}\tilde{\ell}_0$ by $\widehat{P}_j\tilde{\ell}_0^{\otimes 2} - \{\widehat{P}_j\tilde{\ell}_0\}^{\otimes 2}$ where $\widehat{P}_j\tilde{\ell}_0 \equiv \mathbb{P}_N^{\pi}\widehat{\tilde{\ell}}_0 I(V \in \mathscr{V}_j)$ and $\widehat{P}_j\tilde{\ell}_0^{\otimes 2} \equiv \mathbb{P}_N^{\pi}\widehat{\tilde{\ell}}_0^{\otimes 2} I(V \in \mathscr{V}_j)$. The other four cases are similar.

4.2. Cox model with interval censored data

Let *Y* be a censoring time that is assumed to be conditionally independent of a failure time *T* given a covariate vector *X*. Under the case 1 interval censoring, we do not observe *T* but (*Y*, Δ) where Δ ≡ *I* (*T* ≤ *Y*). The phase I data is *V* = (*Y*, Δ, *U*) and the observed data is (*Y*, Δ, ξ *X*, *U*, ξ) where ξ is the sampling indicator. In the case of complete data, maximum likelihood estimates for this model were studied by Huang [10]. For a generalized version of this model and two-phase data with Bernoulli sampling, weighted likelihood estimates with and without estimated weights have recently been studied by Li and Nan [11]. Here we treat two-phase data under sampling without replacement at phase II and with both estimated weights and calibration.

With complete data, the log likelihood for one observation is given by

$$\ell(\theta, F) \equiv \delta\log\{1 - \overline{F}(y)^{\exp(\theta^T x)}\} + (1-\delta)\log\overline{F}(y)^{\exp(\theta^T x)} \equiv \delta\log\{1 - e^{-\Lambda(y)\exp(\theta^T x)}\} - (1-\delta)e^{\theta^T x}\Lambda(y) \equiv \ell(\theta, \Lambda),$$

where $\overline{F} \equiv 1 - F = e^{-\Lambda}$. The score for θ and the score operator $B_{\theta,\Lambda}$ for Λ for complete data are $\dot{\ell}_{\theta,\Lambda} = x\exp(\theta^T x)\Lambda(y)(\delta r(y, x; \theta, \Lambda) - (1-\delta))$ and $B_{\theta,\Lambda}[h] = \exp(\theta^T x)h(y)\{\delta r(y, x; \theta, \Lambda) - (1-\delta)\}$ where $r(y, x; \theta, \Lambda) = \exp(-e^{\theta^T x}\Lambda(y))/\{1 - \exp(-e^{\theta^T x}\Lambda(y))\}$. The efficient score for θ for complete data is given by

$$\ell^*_{\theta_0,\Lambda_0} = e^{\theta_0^T x}Q(y, \delta, x; \theta_0, \Lambda_0)\Lambda_0(y)\left\{x - \frac{E[Xe^{2\theta_0^T X}O(Y|X)|Y=y]}{E[e^{2\theta_0^T X}O(Y|X)|Y=y]}\right\},$$

where $Q(y, \delta, x; \theta, \Lambda) = \delta r(y, x; \theta, \Lambda) - (1 - \delta)$ and $O(y|x) = r(y, x; \theta_0, \Lambda_0)$. The efficient information for θ for complete data $\tilde{I}_{\theta_0,\Lambda_0}=E[(\ell^*_{\theta_0,\Lambda_0})^{\otimes 2}]$ is given by $\tilde{I}_{\theta_0, \Lambda_0} = E[R(Y, X)\{X - E[X R(Y, X)|Y]/E[R(Y, X)|Y]\}]$ where $R(Y, X)=\Lambda_0^2(Y|X)O(Y|X)$. See [10] for further details.

We impose the same assumptions made for complete data in [10].

Condition 4.4. The finite-dimensional parameter space Θ is compact and contains the true parameter θ_0 as its interior point.

Condition 4.5. (a) The covariate X has bounded support; that is, there exists x_0 such that $|X| \le x_0$ with probability 1. (b) For any $\theta \ne \theta_0$, the probability $P(\theta^T X \neq \theta_0^T X)>0$.

Condition 4.6. $F_0(0) = 0$. Let $\tau_{F_0} = \inf\{t : F_0(t) = 1\}$. The support of Y is an interval $S[Y] = [l_Y, u_Y]$ and $0 < l_Y \le u_Y < \tau_{F_0}$.

Condition 4.7. The cumulative hazard function Λ_0 has strictly positive derivative on $S[Y]$, and the joint function $G(y, x)$ of (Y, X) has bounded second order (partial) derivative with respect to y.

4.2.1. Consistency—The characterization of WLEs $(\hat{\theta}_N, \hat{\Lambda}_N)$ and $(\hat{\theta}_{N,\#}, \hat{\Lambda}_{N,\#})$ with $\# \in \{e, c, \text{mc}, \text{cc}\}$ maximizing $\mathbb{P}_N^\pi \ell(\theta,\Lambda) \text{or} \mathbb{P}_N^{\pi,\#}\ell(\theta,\Lambda)$ is given in [25], Lemma A.5. We prove consistency of the WLEs in the metric given by $d((\theta_1, \Lambda_1), (\theta_2, \Lambda_2)) \equiv \|\theta_1 - \theta_2\| + \|\Lambda_1 - \Lambda_2\|_{P_Y}$, where $\|\cdot\|$ is the Euclidean metric and $\| \Lambda_1 - \Lambda_2\|_{P_Y}^2 = \int (\Lambda_1(y) - \Lambda_2(y))^2 dP_Y$, and P_Y is the marginal probability measure of the censoring variable Y.

Theorem 4.3 (Consistency). *Under Conditions* 3.1, 3.2, 4.4–4.7, *the WLEs are consistent in the metric d*.

Proof. We only prove consistency for the WLE. Proofs for the other four estimators are similar.

Let $\tilde{H}$ be the set of all subdistribution functions defined on $[0, \infty]$. We denote the WLE of F as $\hat{F}_N = 1 - e^{-\hat{\Lambda}_N}$. Define the set $\mathscr{F}$ of functions by

$$\mathscr{F} \equiv \{f(\theta,F)=\delta(1-\overline{F}(y)^{\exp(\theta^T x)})+(1-\delta)\overline{F}(y)^{\exp(\theta^T x)}:\theta \in \Theta, F \in \tilde{H}\}.$$

Boundedness of X and compactness of $\Theta \subset \mathbb{R}^p$ imply that the set $\{e^{\theta^T x} : \theta \in \Theta\}$ is Glivenko–Cantelli. The set $\tilde{H}$ is also Glivenko–Cantelli since it is a subset of the set of bounded monotone functions. Thus, it follows from boundedness of functions in $\mathscr{F}$ and the Glivenko–Cantelli preservation theorem [30] that $\mathscr{F}$ is Glivenko–Cantelli.

Let $0 < \alpha < 1$ be a fixed constant. It follows by concavity of the function $u \mapsto \log u$ and Jensen's inequality that

$$P_0[\log\{1+\alpha(f(\theta,F)/f(\theta_0,F_0)-1)\}] \le \log(P_0[1+\alpha(f(\theta,F)/f(\theta_0,F_0)-1)])=\log(1-\alpha+\alpha P_0[f(\theta,F)/f(\theta_0,F_0)]) \le 0,$$

where the first equality holds if and only if $1 + \alpha(f(\theta, F)/f(\theta_0, F_0) - 1)$ is constant on $S[Y]$, in other words, $(\theta, F) = (\theta_0, F_0)$ on $S[Y]$ by the identifiability Condition 4.5. Note also that by monotonicity of the logarithm

$$P_0[\log\{1+\alpha(f(\theta,F)/f(\theta_0,F_0)-1)\}] \geq P_0[\log\{1+\alpha(0-1)\}]=\log(1-\alpha).$$

Thus, the set $\mathscr{G}$= {log{1 + α(f (θ, F)/f (θ_0, F_0) − 1)} : f (θ, F) ∈ $\mathscr{F}$} has an integrable envelope. To see this, form a sequence (θ_n, F_n) such that

$$g_n \equiv \log\{1+\alpha(f(\theta_n,F_n)/f(\theta_0,F_0)-1)\} \nearrow \sup_{\theta\in\Theta,F\in\tilde{H}} \log\{1+\alpha(f(\theta,F)/f(\theta_0,F_0)-1)\} \equiv G.$$

Then $\{g_n - \log(1-\alpha)\}_{n\in\mathbb{N}}$ is a monotone increasing sequence of nonnegative functions. By the monotone convergence theorem, $P_0 g_n - \log(1-\alpha) \to P_0 G - \log(1-\alpha) \leq -\log(1-\alpha)$. Thus we choose $G \vee -\log(1-\alpha)$ as an integrable envelope. Also, the set $\mathscr{G}$ is Glivenko–Cantelli by a Glivenko–Cantelli preservation theorem [30].

Now, by the concavity of the map $u \mapsto \log u$, and the definition of the WLE, we have

$$\begin{aligned}
&\mathbb{P}_N^\pi \log\{1+\alpha(f(\widehat{\theta}_N,\widehat{F}_N)/f(\theta_0,F_0)\\
&\quad -1)\} \geq \mathbb{P}_N^\pi\{(1\\
&\quad\ -\alpha)\log(1)\\
&\quad\ +\alpha\log\{f(\widehat{\theta}_N,\widehat{F}_N)/f(\theta_0,F_0)\}\}\\
&\quad =\alpha\{\mathbb{P}_N^\pi \log f(\widehat{\theta}_N,\widehat{F}_N)\\
&\quad\ -\mathbb{P}_N^\pi \log f(\theta_0,F_0)\} \geq 0.
\end{aligned}$$

Since Θ and $\tilde{H}$ are compact, there is a subsequence of $(\hat{\theta}_N, \hat{F}_N)$ converging to $(\theta_\infty, F_\infty) \in \Theta \times \tilde{H}$. Along this subsequence, Theorem 5.1 implies that

$$0 \leq \mathbb{P}_N^\pi \log\{1+\alpha(f(\widehat{\theta}_N,\widehat{F}_N)/f(\theta_0,F_0)-1)\} \to_{P*} P_{\theta_0,F_0}[\log\{1+\alpha(f(\theta_\infty,F_\infty)/f(\theta_0,F_0)-1)\}] \leq 0,$$

so that P_{θ_0,F_0} log{1 + α(f (θ_∞, F_∞)/f (θ_0, F_0) − 1)} = 0. This is possible when (θ_∞, F_∞) = (θ_0, F_0) because (θ, F) ↦ P[log{1 + α(f (θ, F)/f (θ_0, F_0) − 1)}] attains its maximum only at (θ_0, F_0). Hence we conclude that $(\hat{\theta}_N, \hat{F}_N)$ converges to (θ_0, F_0) in the sense of Kullback–Leibler divergence. Since the Kullback–Leibler divergence bounds the Hellinger distance, it follows by Lemma A5 of [17] that $d((\hat{\theta}_N, \hat{\Lambda}_N), (\theta_0, \Lambda_0)) = o_{P*}(1)$.

4.2.2. Rate of convergence—We prove the rate of convergence of the WLE is $N^{1/3}$ by applying the rate theorem (Theorem 5.2) in Section 5. Since we proved the consistency of $(\hat{\theta}_N, \hat{\Lambda}_N)$ to (θ_0, Λ_0) on $S[Y]$, under Condition 4.6 we can restrict a parameter space of Λ to $H_M \equiv \{\Lambda \in H : M^{-1} \leq \Lambda \leq M$, on $S[Y]\}$, where M is a positive constant such that $M^{-1} \leq \Lambda_0 \leq M$ on $S[Y]$. Define $\mathscr{M} \equiv \{\ell(\theta, \Lambda): \theta \in \Theta, \Lambda \in H_M\}$.

Theorem 4.4 (Rate of convergence). *Under Conditions* 4.4–4.7,

$$d((\widehat{\theta}_N,\widehat{\Lambda}_N),(\theta_0,\Lambda_0))=O_{P*}(N^{-1/3}).$$

This holds if we replace the WLE by the WLEs with adjusted weights assuming Conditions 3.1 *and* 3.2.

Proof. Since the rate of convergence for the WLE is easier to verify than the other four estimators, we only prove the theorem for the WLE with modified calibration. The cases for the WLEs with adjusted weights.

We proceed by verifying the conditions in Theorem 5.2. Bound (5.4) follows by Lemma 5.2 in Section 5 and Lemma A5 of [17]. For bound (5.5), we follow the proof of (5.3) in [10]. Since $\hat{\alpha}_N$ is consistent, we can specify the small neighborhood $\mathscr{A}_{\mathrm{mc},\,0}$ of a zero vector such that $G_{\mathrm{mc}}(z;\alpha)$ is contained in a small interval that contains 1 and consists of strictly positive numbers. Thus, multiplying the log likelihood by a uniformly bounded quantity $G_{\mathrm{mc}}(z;\alpha)$ only requires a slight modification of Huang's proof of his Lemma 3.1 to obtain $\sup_Q \log N_{[\cdot]}(\varepsilon, G\mathscr{M}, L_2(Q)) \lesssim \varepsilon^{-1}$ for ε small enough where the supremum is taken over the all discrete probability measures and $G\mathscr{M} = \{G_{\mathrm{mc}}(\cdot;\alpha)\ell(\theta,\Lambda) : \alpha \in \mathscr{A}_{\mathrm{mc},\,0}, \ell(\theta,\Lambda) \in \mathscr{M}\}$. Let $G\mathscr{M}_\delta = \{m(\theta,\Lambda,\alpha) - m(\theta_0,\Lambda_0,\alpha) : m(\theta,\Lambda,\alpha) \in G\mathscr{M}, d((\theta,\Lambda),(\theta_0,\Lambda_0)) \le \delta\}$. It follows by Lemma 3.2.2 of [32] that $E^*\| \mathbb{G}_N \|_{G\mathscr{M}_\delta} \lesssim \delta^{1/2}\{1+(\delta^{1/2}/\delta^2\sqrt{N})M\} \equiv \phi_N(\delta)$. Apply Theorem 5.2 to conclude $r_N = N^{1/3}$.

4.2.3. Asymptotic normality of the estimators

4.2.3. Asymptotic normality of the estimators—We apply Theorem 3.2 to derive the asymptotic distributions of the WLEs.

Theorem 4.5 (Asymptotic normality). *Under Conditions* 3.1, 3.2, 4.4–4.7,

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = \sqrt{N}\mathbb{P}_N^{\pi}\tilde{\ell}_{\theta_0,\Lambda_0} + o_{P*}(1) \rightsquigarrow N(0,\Sigma), \quad \sqrt{N}(\hat{\theta}_{N,\#} - \theta_0) = \sqrt{N}\mathbb{P}_N^{\pi,\#}\tilde{\ell}_{\theta_0,\Lambda_0} + o_{P*}(1) \rightsquigarrow N(0,\Sigma_\#),$$

where # ∈ {*e*, *c*, mc, cc}, $\tilde{\ell}_{\theta_0,\Lambda_0} = I^{-1}_{\theta_0,\Lambda_0}\ell^*_{\theta_0,\Lambda_0}$ *is the efficient influence function for complete data and* Σ *and* $\Sigma_\#$ *are given in Theorem* 3.2.

Proof. We proceed by verifying the conditions of Theorem 3.2 for the WLE with modified calibration. The other four cases are similar.

Condition 3.6 is satisfied with β = 1/3 by Theorems 4.3 and 4.4. Conditions 3.7–3.9 are verified by [10] with

$$\underline{h}^*(y) \equiv \Lambda_0(y)E[Xe^{2\theta_0^T X}O(Y|X)|Y{=}y]/E[e^{2\theta_0^T X}O(Y|X)|Y{=}y].$$

Since $\mathbb{P}_N^{\pi,\mathrm{mc}}\dot{\ell}_{\hat{\theta}_{N,\mathrm{mc}},\hat{\Lambda}_{N,\mathrm{mc}}} = 0$ by Lemma A.5, it remains to show that

$$\mathbb{P}_N^{\pi,\mathrm{mc}}B_{\hat{\theta}_{N,\mathrm{mc}},\hat{\Lambda}_{N,\mathrm{mc}}}[\underline{h}^*] = o_{P*}(N^{-1/2}).$$

Let $g_0 \equiv \underline{h}^* \circ \Lambda_0^{-1}$ be the composition of $\underline{h}^*$ and the inverse of Λ_0. Note that Λ_0 is a strictly increasing continuous function by our assumption. Since $g_0(\hat{\Lambda}_{N,\,\mathrm{mc}}(y))$ is a right continuous function and has exactly the same jump points as $\hat{\Lambda}_{N,\,\mathrm{mc}}(y)$, by Lemma A.5, $\mathbb{P}_N^{\pi,\mathrm{mc}} g_0(\hat{\Lambda}_{N,\mathrm{mc}}(Y))e^{\hat{\theta}^T_{N,\mathrm{mc}}X}Q(Y,\Delta,X;\hat{\theta}_{N,\mathrm{mc}},\hat{\Lambda}_{N,\mathrm{mc}}) = 0$. By Conditions 4.5–4.7, h^* has bounded derivative. This and the assumption that Λ_0 has strictly positive derivative by Condition 4.7 imply that g_0 has bounded derivative, too. So, noting that $\underline{h}^* = g_0 \circ \Lambda_0$, we have

$$
\begin{aligned}
&\mathbb{P}_N^{\pi,\mathrm{mc}} B_{\widehat{\theta}_{N,\mathrm{mc},}\widehat{\Lambda}_{N,\mathrm{mc}}}[\underline{h}^*] \\
&= \mathbb{P}_N^{\pi,\mathrm{mc}} \underline{h}^*(Y) e^{\widehat{\theta}_{N,\mathrm{mc}}^T X} Q(Y, \Delta, X; \widehat{\theta}_{N,\mathrm{mc}}, \widehat{\Lambda}_{N,\mathrm{mc}}) \\
&= \mathbb{P}_N^{\pi,\mathrm{mc}} \{ g_0 \circ \Lambda_0(Y) \\
&\quad - g_0(\widehat{\Lambda}_{N,\mathrm{mc}}(Y)) \} e^{\widehat{\theta}_{N,\mathrm{mc}}^T X} Q(Y, \Delta, X; \widehat{\theta}_{N,\mathrm{mc}}, \widehat{\Lambda}_{N,\mathrm{mc}}) \\
&= (\mathbb{P}_N^{\pi,\mathrm{mc}} - P_{\theta_0,\Lambda_0}) \{ g_0 \circ \Lambda_0(Y) - g_0(\widehat{\Lambda}_{N,\mathrm{mc}}(Y)) \} \\
&\ \times e^{\widehat{\theta}_{N,\mathrm{mc}}^T X} Q(Y, \Delta, X; \widehat{\theta}_{N,\mathrm{mc}}, \widehat{\Lambda}_{N,\mathrm{mc}}) \\
&+ P_{\theta_0,\Lambda_0} \{ g_0 \circ \Lambda_0(Y) \\
&\quad - g_0(\widehat{\Lambda}_{N,\mathrm{mc}}(Y)) \} \\
&\ \times e^{\widehat{\theta}_{N,\mathrm{mc}}^T X} Q(Y, \Delta, X; \widehat{\theta}_{N,\mathrm{mc}}, \widehat{\Lambda}_{N,\mathrm{mc}}).
\end{aligned}
$$

Huang [10] showed that the second term in the display is $o_{P*}(N^{-1/2})$. We show that the first term in the display is also $o_{P*}(N^{-1/2})$. Let $C > 0$ be an arbitrary constant. Define for a fixed constant $\eta > 0$, $\mathscr{D}\eta) \equiv \{\psi(y, x; \theta, \Lambda) : d((\theta, \Lambda), (\theta_0, \Lambda_0)) \leq \eta, \Lambda \in H_M\}$, where $\psi(y, \delta, x; \theta, \Lambda) \equiv \{g_0 \circ \Lambda_0(y) - g_0(\Lambda(y))\} \times e^{\theta^T x} Q(y, \delta, x; \theta, \Lambda)$. Because Huang [10] showed that $\mathscr{D}\eta)$ is Donsker for every $\eta > 0$ and that $\|\mathbb{G}_N\|_{\mathscr{D}(CN^{-1/3})} = o_{P*}(1)$, it follows by Lemma 5.4 with $\mathscr{F}_N$ replaced by $\mathscr{D}CN^{-1/3})$ that $\| \mathbb{G}_N^{\pi,\mathrm{mc}} \|_{\mathscr{D}(CN^{-1/3})} = o_{P*}(1)$. This completes the proof.

Unlike the previous example, $\ell_{\theta,\Lambda}^*$ depends on additional unknown functions, and the method of variance estimation used in the previous example does not apply to the present case. See the discussion in Section 6.

5. General results for IPW empirical processes

The IPW empirical measure and IPW empirical process inherit important properties from the empirical measure and empirical process, respectively. We emphasize the similarity between empirical processes and IPW empirical processes.

5.1. Glivenko–Cantelli theorem

The next theorem states that the Glivenko–Cantelli property for complete data is preserved under two-phase sampling.

Theorem 5.1. *Suppose that $\mathscr{F}$ is P_0-Glivenko–Cantelli. Then*

$$\| \mathbb{P}_N^{\pi} - P_0 \|_{\mathscr{F}} \rightarrow_{P*} 0, \quad (5.1)$$

where $\|\cdot\|_{\mathscr{F}}$ is the supremum norm. This also holds if we replace $\mathbb{P}_N^{\pi}$ by $\mathbb{P}_N^{\pi,\#}$ with $\# \in \{e, c,$ mc, cc$\}$ *assuming Conditions* 3.1 *and* 3.2.

5.2. Rate of convergence

The rate of convergence of an M-estimator for complete data is often established via maximal inequalities for the empirical processes. If we follow the same line of reasoning, it is natural to derive maximal inequalities for IPW empirical processes, though this may require some efforts. Fortunately, these maximal inequalities for empirical processes (or slight modifications of them) suffice to establish the same rate of convergence under two-phase sampling.

Theorem 5.2. *Let* $\mathscr{M} = \{m_\theta : \theta \in \Theta\}$ *be the set of criterion functions and define* $\mathscr{M}_\delta = \{m_\theta - m_{\theta_0} : d(\theta, \theta_0) < \delta\}$ *for some fixed* $\delta > 0$ *where d is a semimetric on the parameter space* Θ.

(1) *Suppose that for every* θ *in a neighborhood of* θ_0,

$$P_0(m_\theta - m_{\theta_0}) \lesssim -d^2(\theta,\theta_0); \quad (5.2)$$

here $a \lesssim b$ *means* $a \le K\, b$ *for some constant* $K \in (0, \infty)$. *Assume that there exists a function* φ_N *such that* $\delta \mapsto \varphi_N(\delta)/\delta^\alpha$ *is decreasing for some* $\alpha < 2$ *(not depending on N), and for every N,*

$$E^*\| \mathbb{G}_N \|_{\mathscr{M}_\delta} \lesssim \phi_N(\delta), \quad (5.3)$$

where $\mathbb{G}_N$ *is the empirical process. If an estimator* $\hat{\theta}_N$ *satisfying* $\mathbb{P}_N^\pi m_{\hat{\theta}_N} \ge \mathbb{P}_N^\pi m_{\theta_0} - O_{P*}(r_N^{-2})$ *converges in outer probability to* θ_0, *then* $r_N d(\hat{\theta}_N, \theta_0) = O_{P*}(1)$ *for every sequence* r_N *such that* $r_N^2 \phi_N(1/r_N) \le \sqrt{N}$ *for every N*.

(2) *Let* $\# \in \{e, c, \mathrm{mc}, \mathrm{cc}\}$ *be fixed. Suppose Condition* 3.2 *holds. Suppose also that for every* $\theta \in \Theta$ *in a neighborhood of* θ_0,

$$P_0\{\tilde{G}_\#(V;\alpha)(m_\theta - m_{\theta_0})\} \lesssim -d^2(\theta,\theta_0) + |\alpha - \alpha_0|^2, \quad (5.4)$$

where $\tilde{G_e} = \pi_0(V)/G_e$ *or* $\tilde{G_\#} = G_\#$ *with* $\# \in \{c, \mathrm{mc}, \mathrm{cc}\}$. *Assume that*

$$E^*\| \mathbb{G}_N \|_{\tilde{G}_\# \mathscr{M}_\delta} \lesssim \phi_N(\delta), \quad (5.5)$$

where $\tilde{G_\#}\mathscr{M}_\delta \equiv \{\tilde{G_\#}(\cdot;\alpha) f : |\alpha| \le \delta, \alpha \in \mathscr{A}_N, f \in \mathscr{M}_\delta\}$ *for some* $\mathscr{A}_N \subset \mathscr{A}_\#$. *Then an estimator* $\hat{\theta}_{N,\#}$ *satisfying* $\mathbb{P}_N^{\pi,\#} m_{\hat{\theta}_{N,\#}} \ge \mathbb{P}_N^{\pi,\#} m_{\theta_0} - O_{P*}(r_N^{-2})$ *has the same rate of convergence as* $\hat{\theta}_N$ *in part* (1) *if it is consistent*.

Remark 5.1. The key to establishing a general theorem for the rate of convergence is to make use of the boundedness of the weights in the IPW empirical process and also deal with the dependence of the weights. In treating independent bootstrap weights in the weighted bootstrap ([15], Lemmas 1–3), require the boundedness of bootstrap weights because the product of an unbounded weight and a bounded function is no longer bounded. Our theorem exploits the boundedness of sampling indicators in the IPW empirical processes by applying a multiplier inequality for the case of bounded weights (Lemma 5.1) to cover more general cases.

The following is a multiplier inequality for bounded exchangeable weights. Note that the sum of stochastic processes in the second term is divided by $n^{1/2}$ rather than $k^{1/2}$.

Lemma 5.1. *For i.i.d. stochastic processes* $Z_1, \ldots, Z_n$, *every bounded, exchangeable random vector* $(\xi_1, \ldots, \xi_n)$ *with each* $\xi_i \in [l, u]$ *that is independent of* $Z_1, \ldots, Z_n$, *and any* $1 \le n_0 \le n$,

$$E\left\| \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i Z_i \right\|_{\mathscr{F}}^* \le \frac{2(n_0-1)}{n}\sum_{i=1}^n E^*\| Z_i \|_{\mathscr{F}} E \max_{1\le i\le n}\frac{\xi_i}{\sqrt{n}} + 2(u-l)\max_{n_0\le k\le n} E\left\| \frac{1}{\sqrt{n}}\sum_{i=n_0}^k Z_i \right\|_{\mathscr{F}}^*.$$

Bound (5.5) is not difficult to verify in the presence of bound (5.3) since $G_\#(\cdot\,; \alpha)$ is a bounded monotone function indexed by a finite-dimensional parameter. Bound (5.4) may be verified through the lemma below for some applications including the Cox model with interval censoring.

Lemma 5.2. *Suppose Conditions* 3.1 *and* 3.2 *hold. Let* m_θ *be the log likelihood log* p_θ *where* p_θ *is the density with dominating measure* μ, *and d is the Hellinger distance. Then the bound* (5.4) *holds*.

5.3. Donsker theorem

The next theorem yields weak convergence of the IPW empirical processes under sampling without replacement.

Theorem 5.3. *Suppose that* $\mathscr{F}$ *with* $\|P_0\|_{\mathscr{F}} < \infty$ *is* P_0*-Donsker and Conditions* 3.1 *and* 3.2 *hold. Then*

$$\mathbb{G}_N^\pi \rightsquigarrow \mathbb{G}^\pi \equiv \mathbb{G}+\sum_{j=1}^J \sqrt{\nu_j}\sqrt{\frac{1-p_j}{p_j}}\mathbb{G}_j, \quad (5.6)$$

$$\mathbb{G}_N^{\pi,\#} \rightsquigarrow \mathbb{G}^{\pi,\#} \equiv \mathbb{G}+\sum_{j=1}^J \sqrt{\nu_j}\sqrt{\frac{1-p_j}{p_j}}\mathbb{G}_j(\cdot - Q_\#\cdot) \quad (5.7)$$

in $\ell^\infty(\mathscr{F})$ *where* $\# \in \{e, c, \text{mc}, \text{cc}\}$, *the* P_0*-Brownian bridge process*, $\mathbb{G}$ *indexed by* $\mathscr{F}$ *and the* $P_{0|j}$*-Brownian bridge processes*, $\mathbb{G}_j$, *indexed by* $\mathscr{F}$ *are all independent*.

Remark 5.2. The integrability hypothesis $\|P_0\|_{\mathscr{F}} < \infty$ is only required for the IPW empirical processes with adjusted weights.

For a Donsker set $\mathscr{F}$, it follows by Theorem 5.3 and Lemma 2.3.11 of [32] that asymptotic equicontinuity in probability and in mean follows for the metric that depends on the limit process. In applications, it is of interest to have these results for the original metric $\rho_{P_0}(f, g) = \sigma_{P_0}(f - g)$.

Theorem 5.4. *Let* $\mathscr{F}$ *be Donsker and define* $\mathscr{F}_\delta = \{f - g : f, g \in \mathscr{F}, \rho_{P_0}(f, g) < \delta\}$ *for some fixed* $\delta > 0$. *Then, for every sequence* $\delta_N \downarrow 0$,

$$E^*\| \mathbb{G}_N^\pi \|_{\mathscr{F}_{\delta_N}} \to 0$$

and consequently, $\| \mathbb{G}_N^\pi \|_{\mathscr{F}_{\delta_N}} = o_{P*}(1)$. *Moreover*, $\| \mathbb{G}_N^{\pi,\#} \|_{\mathscr{F}_{\delta_N}} = o_{P*}(1)$ *for* $\# \in \{e, c, \text{mc}, \text{cc}\}$ *assuming Conditions* 3.1 *and* 3.2.

We end this section with two important lemmas. The first lemma is an extension of Lemma 3.3.5 of [32] and will be used in our proof of Theorem 3.1 to verify asymptotic equicontinuity.

Lemma 5.3. *Suppose* $\mathscr{F} = \{\psi_{\theta,h} - \psi_{\theta_0,h} : \|\theta - \theta_0\| < \delta, h \in \mathscr{H}\}$ *is* P_0*-Donsker for some* $\delta > 0$ *and that* $\sup_{h\in\mathscr{H}} P_0(\psi_{\theta,h} - \psi_{\theta_0,h})^2 \to 0$, *as* $\theta \to \theta_0$. *If* $\hat{\theta}_N$ *converges in outer probability to* θ_0, *then*

$$\| \mathbb{G}_N^{\pi}(\psi_{\hat{\theta}_N,h} - \psi_{\theta_0,h}) \|_{\mathscr{H}} = o_{P*}(1).$$

This also holds if we replace $\mathbb{G}_N^{\pi}$ *by* $\mathbb{G}_N^{\pi,\#}$ *with* $\# \in \{e, c, \mathrm{mc}, \mathrm{cc}\}$ *assuming Conditions* 3.1 *and* 3.2 *hold and* $\|P_0\|_{\mathscr{F}} < \infty$.

The second lemma is used to verify asymptotic equicontinuity in the proof of Theorem 3.2, the first part for the IPW empirical process and the second part for the other four IPW empirical processes with adjusted weights.

Lemma 5.4. *Let* $\mathscr{F}_N$ *be a sequence of decreasing classes of functions such that* $\|\mathbb{G}_N\|_{\mathscr{F}_N} = o_{P*}(1)$. *Assume that there exists an integrable envelope for* $\mathscr{F}_{N_0}$ *for some* N_0. *Then* $E\|\mathbb{G}_N\|_{\mathscr{F}_N} \to 0$ *as* $N \to \infty$. *As a consequence,* $\| \mathbb{G}_N^{\pi} \|_{\mathscr{F}_N} = o_{P*}(1)$.

Suppose, moreover, that $\mathscr{F}_N$ *is* P_0*-Glivenko–Cantelli with* $\|P_0\|_{\mathscr{F}_{N_1}} < \infty$ *for some* N_1, *and that every* $f = f_N \in \mathscr{F}_N$ *converges to zero either point-wise or in* $L_1(P_0)$ *as* $N \to \infty$. *Then* $\| \mathbb{G}_N^{\pi,e} \|_{\mathscr{F}_N} = o_{P*}(1), \| \mathbb{G}_N^{\pi,\mathrm{c}} \|_{\mathscr{F}_N} = o_{P*}(1), \| \mathbb{G}_N^{\pi,\mathrm{mc}} \|_{\mathscr{F}_N} = o_{P*}(1) \, and \| \mathbb{G}_N^{\pi,\mathrm{cc}} \|_{\mathscr{F}_N} = o_{P*}(1)$, *assuming Conditions* 3.1 *and* 3.2.

6. Discussion

We developed asymptotic theory for weighted likelihood estimation under two-phase sampling, introduced and studied a new calibration method, centered calibration, and compared several WLE estimation methods involving adjusted weights. The methods of proof and general results for the IPW empirical process are applicable to other estimation procedures. For example, the weighted Kaplan–Meier estimator can be shown to be asymptotically Gaussian via our Donsker theorem (Theorem 5.3) together with the functional delta method. A particularly interesting application is to study asymptotic properties of estimators that are known to be efficient under Bernoulli sampling (e.g., estimator of [19]). Whether or not these estimators are “efficient” under our sampling scheme is an open problem; see [16] for a definition of efficiency with non-i.i.d. data.

There are several other open problems. Variance estimation under two-phase sampling has been restricted to the case where the asymptotic variance is a known function up to parameters as discussed in Section 4, while there are several methods available for complete data in a general case (e.g., [18]). In [24] the first author has proposed and studied nonparametric bootstrap variance estimation methods which remain valid even under model misspecification; these results will appear elsewhere. Another direction of research is to study (local and global) model misspecification under two-phase sampling where missingness is by design. An interesting open problem beyond our sampling scheme is to study other complex survey designs. Stratified sampling without replacement is sufficiently simple for the existing bootstrap empirical process theory to apply. Other complex designs may provide interesting theoretical challenges, perhaps in connection with extensions of bootstrap empirical process theory.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We owe thanks to Kwun Chuen Gary Chan for suggesting the modified calibration method introduced in Section 2.1.3. We also thank Norman Breslow for many helpful conversations concerning two-phase sampling, and two referees for their constructive comments and suggestions.

REFERENCES

1. Binder DA. Fitting Cox's proportional hazards models from survey data. Biometrika. 1992; 79:139–147. MR1158522.
2. Breslow NE, Lumley T, Ballantyne C, Chambless L, Kulich M. Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. Stat. Biosc. 2009; 1:32–49.
3. Breslow NE, Lumley T, Ballantyne C, Chambless L, Kulich M. Using the whole cohort in the analysis of case-cohort data. Am. J. Epidemiol. 2009; 169:1398–1405. [PubMed: 19357328]
4. Breslow NE, Wellner JA. .Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. Scand. J. Stat. 2007; 34:86–102. MR2325244.
5. Breslow NE, Wellner JA. A Z-theorem with estimated nuisance parameters and correction note for: "Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression" [ScandJStatist. 34 (2007), no. 1, 86-102; MR2325244]. Scand. J. Stat. 2008; 35:186–192. MR2391566.
6. Chan KCG. Uniform improvement of empirical likelihood for missing response problem. Electron. J. Stat. 2012; 6:289–302.
7. Cox DR. Regression models and life-tables (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 1972; 34:187–220. MR0341758.
8. Deville J-C, Särndal C-E. Calibration estimators in survey sampling. J. Amer. Statist. Assoc. 1992; 87:376–382. MR1173804.
9. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Assoc. 1952; 47:663–685. MR0053460.
10. Huang J. Efficient estimation for the proportional hazards model with interval censoring. Ann. Statist. 1996; 24:540–568. MR1394975.
11. Li Z, Nan B. Relative risk regression for current status data in case-cohort studies. Canad. J. Statist. 2011; 39:557–577. MR2860827.
12. Lin DY. On fitting Cox's proportional hazards models to survey data. Biometrika. 2000; 87:37–47. MR1766826.
13. Lumley, T. Complex Surveys:A Guide to Analysis Using R. Hoboken, NJ: Wiley; 2010.
14. Lumley T, Shaw PA, Dai JY. Connections between survey calibration estimators and semiparametric models for incomplete data. Int. Stat. Rev. 2011; 79:200–232. [PubMed: 23833390]
15. Ma S, Kosorok MR. Robust semiparametric M-estimation and the weighted bootstrap. J. Multivariate Anal. 2005; 96:190–217. MR2202406.
16. McNeney B, Wellner JA. Application of convolution theorems in semiparametric models with non-i.i.d. data. J. Statist. Plann. Inference. 2000; 91:441–480. Prague Workshop on Perspectives in Modern Statistical Inference: Parametrics, Semi-parametrics, Non-parametrics (1998) MR1814795.
17. Murphy SA, van der Vaart AW. Semiparametric likelihood ratio inference. Ann. Statist. 1997; 25:1471–1509. MR1463562.
18. Murphy SA, van der Vaart AW. Observed information in semi-parametric models. Bernoulli. 1999; 5:381–412. MR1693616.
19. Nan B. Efficient estimation for case-cohort studies. Canad. J. Statist. 2004; 32:403–419. MR2125853.
20. Neyman J. Contribution to the theory of sampling human populations. J. Amer. Statist. Assoc. 1938; 33:101–116.

21. Præstgaard J, Wellner JA. Exchangeably weighted bootstraps of the general empirical process. Ann. Probab. 1993; 21:2053–2086. MR1245301.

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

23. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J. Amer. Statist. Assoc. 1994; 89:846–866. MR1294730.

24. Saegusa, T. Ph.D. thesis. Seattle, WA: Univ. Washington; 2012. Weighted likelihood estimation under two-phase sampling.

25. Saegusa T, Wellner JA. Supplement to"Weighted likelihood estimation under two-phase sampling". 2012

26. Saegusa, T.; Wellner, JA. Technical Report 592. Seattle, WA: Dept. Statistics, Univ. Washington; 2012. Weighted likelihood estimation under two-phase sampling. Available at arXiv:1112.4951.

27. Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. Ann. Statist. 1988; 16:64–81. MR0924857.

28. Tan Z. Efficient restricted estimators for conditional mean models with missing data. Biometrika. 2011; 98:663–684. MR2836413.

29. van der Vaart, A. Lectures on Probability Theory and Statistics (Saint-Flour, 1999). Lecture Notes in Math. Vol. 1781. Berlin: Springer; 2002. Semiparametric statistics; p. 331-457.MR1915446

30. van der Vaart, A.; Wellner, JA. High Dimensional Probability. II (Seattle. WA, 1999). Progress in Probability. Vol. 47. Boston, MA: Birkhäuser; 2000. Preservation theorems for Glivenko– Cantelli and uniform Glivenko–Cantelli classes; p. 115-133.MR1857319

31. van der Vaart, AW. Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Vol. 3. Cambridge: Cambridge Univ. Press; 1998. MR1652247

32. van der Vaart, AW.; Wellner, JA. Weak Convergence and Empirical Processes :With Applications to Statistics. New York: Springer; 1996. . MR1385671

33. White JE. A two stage design for the study of the relationship between a rare exposure and and a rare disease. Am. J. Epidemiol. 1986; 115:119–128. [PubMed: 7055123]

34. Zheng H, Little RJA. Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. Survey Methodology. 2004; 30:209–218.