Published in final edited form as:
Stat Biosci. 2009 May 1; 1(1): 32. doi:10.1007/s12561-009-9001-6.

Improved Horvitz-Thompson Estimation of Model Parameters from Two-phase Stratified Samples: Applications in Epidemiology

Norman E. Breslow,
Department of Biostatistics, University of Washington, Seattle, WA, USA, Tel.: +1-206-543-2035, Fax: +1-206-616-2724

Thomas Lumley,
Department of Biostatistics, University of Washington, Seattle, WA, USA

Christie M Ballantyne,
Department of Medicine, Baylor College of Medicine, Houston, TX, USA

Lloyd E. Chambless, and
Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA

Michal Kulich
Department of Probability and Mathematical Statistics, Charles University, Prague, CZ

Norman E. Breslow: norm@u.washington.edu; Thomas Lumley: tlumley@u.washington.edu; Christie M Ballantyne: cmb@bcm.edu; Lloyd E. Chambless: Lloyd.Chambless@mail.cscc.unc.edu; Michal Kulich: kulich@karlin.m3.cuni.cz

Abstract

The case-cohort study involves two-phase sampling: simple random sampling from an infinite super-population at phase one and stratified random sampling from a finite cohort at phase two. Standard analyses of case-cohort data involve solution of inverse probability weighted (IPW) estimating equations, with weights determined by the known phase two sampling fractions. The variance of parameter estimates in (semi)parametric models, including the Cox model, is the sum of two terms: (i) the model based variance of the usual estimates that would be calculated if full data were available for the entire cohort; and (ii) the design based variance from IPW estimation of the unknown cohort total of the efficient influence function (IF) contributions. This second variance component may be reduced by adjusting the sampling weights, either by calibration to known cohort totals of auxiliary variables correlated with the IF contributions or by their estimation using these same auxiliary variables. Both adjustment methods are implemented in the R `survey` package. We derive the limit laws of coefficients estimated using adjusted weights. The asymptotic results suggest practical methods for construction of auxiliary variables that are evaluated by simulation of case-cohort samples from the National Wilms Tumor Study and by log-linear modeling of case-cohort data from the Atherosclerosis Risk in Communities Study. Although not semiparametric efficient, estimators based on adjusted weights may come close to achieving full efficiency within the class of augmented IPW estimators.

Keywords

Calibration; Case-cohort; Estimation; Log-linear model; Semiparametric

1 Introduction

Two phase stratified sampling designs were proposed by Neyman (1938) for estimation of the finite population mean of a target variable that was difficult to measure. The average amount of money spent on food by each family residing in a given district was mentioned as a possible

target of inference. At the first phase, a large sample is drawn from the population. Information on an auxiliary variable, easier to measure but correlated with the target variable, is collected and used to stratify the sample. At phase two, random sub-samples are drawn without replacement from each stratum for measurement of the target variable. The technique is widely used in survey sampling to reduce costs.

Two phase designs have also been proposed for use in epidemiology. They are particularly valuable when a large cohort (the phase one sample) is under surveillance for a disease event of interest and sampling from the cohort is required to obtain information on additional covariates. White (1982) proposed stratifying the second phase of sampling on both disease status and exposure when more information was needed on confounding factors. She argued that, when both disease and exposure were rare, this was more efficient than the standard case-control design that stratified on disease status alone. Borgan et al. (2000) considered exposure stratified versions of the case-cohort study (Prentice, 1986).

In these epidemiologic applications, the target of inference is not a finite population mean but rather parameters in a probability model – an infinite "superpopulation" from which the cohort is regarded as constituting a phase one random sample. For the stratified case-control study the parameters of interest are odds ratios, exponentiated coefficients in a logistic regression model. For the stratified case-cohort study they are hazard ratios, exponentiated coefficients in the Cox (1972) proportional hazards model, and often also the baseline hazard function.

Current standard practice for estimation of regression coefficients in the Cox model is solution of a Horvitz and Thompson (1952) inverse probability weighted (IPW) version of the Cox (1975) partial likelihood equations (Barlow, 1994; Barlow et al., 1999; Borgan et al., 2000). Survey statisticians advocate this approach on grounds that when the model is misspecified, as is generally the case, it consistently estimates the parameters that would be estimated by fitting the "wrong" model to the cohort were complete data available fot it (Binder, 1992). The Horvitz-Thompson approach is known to be inefficient, however, sometimes seriously so (Robins et al., 1994). One reason is that it often ignores much of the information available for the cohort. The Atherosclerosis Risk in Communities (ARIC) investigators, for example, conduct numerous stratified case-control and case-cohort studies nested in their cohort of 15,972 subjects sampled from four U.S. communities (The ARIC Investigators 1989). Information on standard risk factors for cardiovascular disease is available for nearly the entire cohort from interviews, bioassays and imaging studies conducted at baseline. Additional information on candidate genes or biomarkers is collected for a phase two cohort random sample (CRS), also known as a sub-cohort, that is stratified on demographic factors and sometimes also on carotid wall thickness. Even when enriched by disease cases that occur outside the CRS, the phase two sample typically contains no more than 10–15% of the cohort. ARIC analyses have ignored information on adjustment factors available for the great majority of potential controls.

Recently improved communication between biostatisticians and survey methodologists has led to a better understanding of the two phase sampling designs used in epidemiology and to improved methods of analysis. Statistical efficiency may be enhanced by adjustment of the standard Horvitz-Thompson sampling weights, either by calibrating the weights to cohort totals of auxiliary variables (Deville and Särndal, 1992) or by using these variables to estimate the weights (Robins et al., 1994). We describe some theory and methods for calibration and estimation of weights when fitting semiparametric models that apply *a fortiori* to the fitting of parametric models. The resulting improvements in precision are illustrated by fitting a log-linear model to data from an ARIC study of coronary heart disease (Ballantyne et al., 2004). A companion paper for epidemiologists (Breslow et al., 2009) investigates the corresponding gains in efficiency of hazard ratios estimated from the ARIC data. By simulation of case-cohort

samples from the National Wilms Tumor Study (NWTS) (D'Angio et al., 1989; Green et al., 1998), we also illustrate the problems that may occur when one attempts to use too many auxiliary variables for calibration or estimation of the weights.

A proposal is made for construction of auxiliary variables with the goal of achieving approximate optimality within the class of augmented IPW estimators (Wang and Chen, 2001). We do not consider semiparametric efficient methods of estimation that make greater use of the assumptions of the fitted model, usually at the price of introducing bias if the model does not hold. Such methods are reasonably well developed for logistic regression analysis of stratified case-control data when the covariates available for the entire cohort are all discrete (Scott and Wild, 1997; Breslow and Holubkov, 1997). Semiparametric efficient methods for estimation of hazard ratios from stratified case-cohort samples, which pose greater computational problems, are currently under development; see, for example, Nan (2004); Scheike and Martinussen (2004); Zeng and Lin (2007).

2 Properties of Horvitz-Thompson Estimators

This section reviews results of Breslow and Wellner (2007). Suppose N subjects are sampled at random from an infinite population and indexed by $i = 1, \ldots, N$. These subjects constitute the (main) cohort following the first phase of sampling. Denote by $V \in \mathcal{V}$ a vector of random variables that is observed for all cohort members. Suppose $\mathcal{V}$ is partitioned $\mathcal{V} = V_1 \cup \cdots \cup \mathcal{V}_J$ and the cohort is divided correspondingly into J strata, with the i^{th} subject in stratum j if $V_i \in \mathcal{V}_j$. Let N_j denote the number of subjects in the j^{th} stratum, $j = 1, \ldots, J$, so $N = N_1 + \ldots + N_J$. For the epidemiologic designs, one stratum typically consists of the disease cases while the remaining strata contain the controls. At the second phase of sampling $n_j \le N_j$ subjects are sampled at random without replacement from the j^{th} stratum, with sampling for different strata conducted independently. Additional variables are observed for the $n = n_1 + \ldots + n_J$ subjects sampled at phase two. Let $W \in \mathcal{W}$ denote the vector of random variables that are *potentially* available for the cohort, namely, V plus the additional variables known only for the n subjects sampled at phase two. We denote by $\Sigma_N = \sigma[W_1, \ldots, W_N]$ the sigma field of information potentially available for everyone. An important aspect of this formulation for the case-cohort study is that it accomodates random sampling of both cases and controls. Often cases (and controls) are absent from the phase two sample because some data are missing by chance rather than by design, for example, because of uninformative biological samples (Mark and Katki, 2006). The methods described herein still apply provided that such absence is random within defined strata.

Let ξ_i be a binary indicator of whether ($\xi_i = 1$) or not ($\xi_i = 0$) the i^{th} subject is sampled at phase two and let $\pi_i = \Pr(\xi_i = 1)$ be the probability of such sampling. Thus, if $j(i)$ denotes the stratum of the i^{th} subject, $\pi_i = n_{j(i)}/N_{j(i)}$. Furthermore, for any pair $i \ne i'$,

$$\pi_{i,i'} \overset{\text{def}}{=} \Pr(\xi_i = \xi_{i'} = 1) = \begin{cases} \pi_i \pi_{i'} = \frac{n_{j(i)}}{N_{j(i)}} \cdot \frac{n_{j(i')}}{N_{j(i')}} & \text{if } j(i) \ne j(i') \\ \frac{n_{j(i)}}{N_{j(i)}} \cdot \frac{n_{j(i)}-1}{N_{j(i)}-1} & \text{if } j(i) = j(i') \end{cases} . \tag{1}$$

The goal of the investigation is to make inferences about parameters in a (superpopulation) model for a random variable $X = X(W) \in \mathcal{X}$ which is completely observed for subjects in the CRS but in general only partially observed for those in the cohort. The model may be parametric or semiparametric. In the latter case it is specified by probability distributions $P_{\theta,\eta}$ for X that are indexed by a Euclidean parameter $\theta \in \Theta \subset \mathbb{R}^p$ and an infinite dimensional parameter $\eta \in \Xi$. The paradigm is the Cox model, where θ denotes the vector of regression coefficients and

η the baseline cumulative hazard function. We denote by $P_0 = P_{\theta_0,\eta_0}$ the "true" model and use operator notation $Pf = \mathrm{E}_P f(X)$ for expectations.

Let $\dot{\ell}_{\theta,\eta}$ denote the usual parametric likelihood score for θ and $B_{\theta,\eta}$ denote the *score operator* (Begun et al., 1983) that maps one dimensional submodels for η, indexed by directions $h \in \mathcal{H}$ from which η's in the submodel approach η_0, into the corresponding score functions. If X was known for all N cohort subjects, parameters could be estimated by solving likelihood equations (van der Vaart, 1998, Sect. 25.12)

$$\mathbb{P}_N \dot{\ell}_{\theta,\eta} = 0 \tag{2}$$

$$\mathbb{P}_N B_{\theta,\eta} h = 0 \forall h \in \mathcal{H} \tag{3}$$

for estimators $(\tilde{\theta}_N, \tilde{\eta}_N)$, where $\mathbb{P}_N$ denotes empirical measure based on $X_1, \ldots, X_N$. Assumptions are made at least sufficient to guarantee that $\sqrt{N}(\tilde{\theta}_N - \theta_0, \tilde{\eta}_N - \eta_0)$ is asymptotically Gaussian. For two phase data, $\mathbb{P}_N$ in (2) and (3) is replaced by IPW empirical measure defined by

$$\mathbb{P}_N^{\pi} = \frac{1}{N} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} \delta_{X_i}, \tag{4}$$

where δ_{X_i} is Dirac measure that puts unit mass on X_i, and the solution is denoted by $(\hat{\theta}_N, \hat{\eta}_N)$. When applied with the Cox model this leads to an IPW version of the so called Breslow (1974) estimator of the cumulative hazard function and to the same IPW version of the partial-likelihood equations that has been proposed by Binder (1992) and Lin (2000), among others.

Denote the *efficient information* by $\tilde{I}_0 = P_0 \left[\left(I - B_0 (B_0^* B_0)^{-1} B_0^* \right) \dot{\ell}_0 \dot{\ell}_0^T \right]$ and the *efficient influence function* by $\tilde{\ell}_0 = \tilde{I}_0^{-1} \left(I - B_0 (B_0^* B_0)^{-1} B_0^* \right) \dot{\ell}_0$, where B_0^* is the adjunct of B_0. Invertibility of the *information operator* $B_0^* B_0$ is implicit in the assumption that η is estimable at a $\sqrt{N}$ rate. The principal result of Breslow and Wellner (2007, p. 94) is

$$\begin{aligned} \sqrt{N}(\hat{\theta}_N - \theta_0) &= \sqrt{N}(\tilde{\theta}_N - \theta_0) + \sqrt{N}(\hat{\theta}_N - \tilde{\theta}_N) \\ &= \sqrt{N} \mathbb{P}_N \tilde{\ell}_0 + \sqrt{N}(\mathbb{P}_N^{\pi} - \mathbb{P}_N) \tilde{\ell}_0 + o_p(1). \end{aligned} \tag{5}$$

The first term on the right hand side of (5) represents the usual asymptotic expansion for the unobservable estimator $\tilde{\theta}_N$. It converges in distribution to $\mathbb{G}\tilde{\ell}_0$, where $\mathbb{G}$ is the P_0-Brownian bridge (van der Vaart and Wellner, 1996, Sect. 2.1). Conditionally on Σ_N, and hence considered as a random function only of the sampling indicators ξ_i, the second term converges similarly by virtue of the weak convergence

$$\sqrt{N}(\mathbb{P}_N^{\pi} - \mathbb{P}_N) \rightsquigarrow \sum_{j=1}^{J} \sqrt{\nu_j} \sqrt{\frac{1-p_j}{p_j}} \mathbb{G}_j \quad \text{in} \quad \ell^{\infty}(\mathcal{F}). \tag{6}$$

Here $\mathcal{F}$ is a (Donsker) class of functions that contains the likelihood scores for (θ, η) in a neighborhood of (θ_0, η_0); $\nu_j = \lim N_j/N$ is the fraction of the population in stratum j; $p_j = \lim n_j/N_j$ is the limiting samping fraction; and $\mathbb{G}_j$ denotes the Brownian bridge process restricted to stratum j, *i.e.*, based on the distribution $P_{0|j}(\cdot) = \mathrm{E}(\cdot|V \in \mathcal{V}_j)$. Furthermore, the processes $\{\mathbb{G}, \mathbb{G}_1, \ldots, \mathbb{G}_J\}$ are mutually independent.

The asymptotic variance of $\hat{\theta}_N$ is thus the sum of two components, one corresponding to each phase of sampling. The first component is the usual variability of an estimator based on random sampling from an infinite population (model) assuming no missing data. It is not amenable to modification using methods described here. The second component, which is amenable to improvement, represents the additional variability stemming from the fact that some components of X are observable only at phase two. It is the normalized *design based* variance of the standard Horvitz-Thompson estimator of an unknown finite population total, namely, the total of the efficient influence function (IF) contributions for all N phase one subjects. Similar results have appeared in the sample survey literature for estimation of Euclidean parameters by solving estimating equations that contain no nuisance functions (Rao et al., 2002; Rubin-Bleuer and Kratina, 2005). Equations (5) and (6) provide the extension to semiparametric models under two-phase stratified sampling, with the semiparametric efficient influence function $\tilde{\ell}_0$ playing the role of the ordinary parametric IF. For phase two subjects the IF contributions $\tilde{\ell}_0(x_i)$ may be approximated with negligible error from the observed x_i by *dfbeta*'s (Therneau and Grambsch, 2000, p. 155) defined by

$$dfbeta_i \overset{\text{def}}{=} \tilde{\ell}_{(\hat{\theta}_N, \hat{\eta}_N)}(x_i). \tag{7}$$

Explicit formulae are available for the Cox model (Cain and Lange, 1984; van der Vaart, 1998, Sect. 25.12.1). Values may be obtained from standard statistical packages as a type of residual following a model fit. In R, for example, they are obtained with the command `db<-resid(model.fit, type='dfbeta')`.

3 Calibration and Estimation of the Weights

Survey statisticians are adept at improving estimates of finite population (here, cohort or phase one) totals when *auxiliary variables*, closely correlated with the target variable, are available for the entire population. We consider $\tilde{\ell}_0(x_i)$ to be the vector of target variables, denote by $z_i = z(v_i)$ a q-vector of auxiliary variables and set $z_{\text{tot}} = \sum_{i=1}^{N} z_i$. Some suggestions for choice of the z_i are given later. The idea behind calibration is to modify the design weights $d_i = \pi_i^{-1}$ to new weights $w_i = g_i d_i$ such that the w_i and d_i are as close as possible yet the phase one totals z_{tot} of the auxiliary variables are exactly estimated. If $G(w, d)$ denotes a distance measure, the problem is thus to minimize $\sum_{i=1}^{N} \xi_i G(w_i, d_i)$ subject to the constraints

$$\widehat{z_{\text{tot}}} \overset{\text{def}}{=} \sum_{i=1}^{N} \xi_i w_i z_i = z_{\text{tot}} \tag{8}$$

known as the *calibration equations*. Here we consider $G(w, d) = (w-d)^2/2d$ and $G(w, d) = w \log(w/d) - w + d$, the Poisson deviance. See Deville and Särndal (1992) for other possibilities.

Let λ be a q-vector of Lagrange multipliers corresponding to the constraints (8). For $G(w, d) = (w-d)^2/2d$, solution of the constrained minimization problem by standard calculus yields $g_i = 1 - \lambda^{\mathrm{T}} z_i$ which, when substituted into (8), leads to an explicit solution for λ:

$$\widehat{\lambda}_N = \left(\sum_{i=1}^{N} \xi_i d_i z_i z_i^{\mathrm{T}} \right)^{-1} \left(\sum_{i=1}^{N} \xi_i d_i z_i - z_{\text{tot}} \right). \tag{9}$$

The estimator obtained with the resulting weights is known as the generalized regression or GREG estimator (Särndal et al., 1989). When used to estimate a finite population total $Y = \sum_{i=1}^{N} y_i$ via

$$\widehat{Y}_{\text{GREG}} = \sum_{i=1}^{N} \xi_i \frac{g_i}{\pi_i} y_i = \sum_{i=1}^{N} z_i^{\mathrm{T}} \widehat{\beta} + \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} \left(y_i - z_i^{\mathrm{T}} \widehat{\beta} \right), \quad \text{where}$$
$$\widehat{\beta} = \left(\sum_{i=1}^{N} \frac{\xi_i}{\pi_i} z_i z_i^{\mathrm{T}} \right)^{-1} \left(\sum \frac{\xi_i}{\pi_i} z_i y_i \right),$$

the GREG estimator is the finite population sum of the fitted values plus the Horvitz-Thompson estimator of the sum of the residuals.

When $G(w, d)$ is the Poisson deviance, solution of the minimization problem gives $g_i = \exp(-\lambda^{\mathrm{T}} z_i)$ and the calibration equations, now solved iteratively for λ, become

$$\sum_{i=1}^{N} \xi_i \exp(-\lambda^{\mathrm{T}} z_i) d_i z_i = z_{\text{tot}}.$$

Under standard regularity assumptions (Isaki and Fuller, 1982) applicable to design based inference, Deville and Särndal (1992, p. 379) show that the solution generally satisfies

$$\widehat{\lambda}_N = \widehat{B}_N^{-1} \left(\sum_{i=1}^{N} \xi_i d_i z_i - z_{\text{tot}} \right) + O_p(n^{-1}), \quad \text{where} \tag{10}$$

$$\frac{1}{N}\widehat{B}_N=\frac{1}{N}\left(\sum_{i=1}^{N}\xi_i d_i z_i z_i^{\mathrm{T}}\right)\xrightarrow{\mathrm{p}} P_0(ZZ^{\mathrm{T}}). \tag{11}$$

When $g_i=\exp(-\widehat{\lambda}_n^{\mathrm{T}} z_i)$ the calibrated weights are always positive (with GREG they may not be) and the resulting estimator is known as the (generalized) raking estimator. Since under mild regularity conditions (see Sect. 7, Appendix) $\widehat{\lambda}_N=O_p\left(\frac{1}{\sqrt{N}}\right)$, the g_i converge to 1 and hence the estimated w_i converge to the design weights d_i.

The asymptotic properties of the estimator $\hat{\theta}_N(\hat{\lambda}_N)$ based on calibrated weights may be derived from results in Breslow and Wellner (2008) by letting the Lagrange multipliers λ play the role of "nuisance" parameters α. Regardless of which distance function is used for calibration (Deville and Särndal, 1992), one finds

$$\begin{aligned}\sqrt{N}\left(\widehat{\theta}_N(\widehat{\lambda}_N)-\theta_0\right)\rightsquigarrow\ &\mathbb{G}\tilde{\ell}_o\\ &+\sum_{j=1}^{J}\sqrt{\nu_j}\sqrt{\frac{1-p_j}{p_j}}\mathbb{G}_j\left(\tilde{\ell}_0-QZ\right).\end{aligned} \tag{12}$$

Here $QZ=P_0\left(\tilde{\ell}_0 Z^{\mathrm{T}}\right)P_0^{-1}(ZZ^{\mathrm{T}})Z$ is the least squares projection of each component of $\tilde{\ell}_0$ onto the linear subspace of $L_2(P_0)$ spanned by components of Z. Hence the effect of calibration on the (asymptotic) phase two component of variance is to replace $\tilde{\ell}_0$ by the residual after its (population) least squares regression on Z. The Appendix contains a brief derivation of (12).

The asymptotic variance of the calibration estimator could be further reduced if we replaced Q in the j^{th} summand of (12) by $Q_j=P_{0|j}(\tilde{\ell}_0 Z^{\mathrm{T}})P_{0|j}^{-1}(ZZ^{\mathrm{T}})$, the projection with respect to the *conditional* distribution of the data in stratum. This is because, with $\mathrm{Var}_j f=P_{0|j}f^{\otimes 2}-P_{0|j}^{\otimes 2}f$, Q_jZ minimizes $\mathrm{Var}_j(\tilde{\ell}_0-AZ)$ among all linear functions of Z. Consequently, instead of calibrating to the overall total, one might consider calibrating to the subtotals of z within each stratum. This is accomplished by defining a new $q\times J$ vector of calibration variables by $\tilde{Z}^{\mathrm{T}}=[Z_1^{\mathrm{T}}Z_2^{\mathrm{T}}\cdots Z_J^{\mathrm{T}}]$ where $Z_j=\mathbf{1}(V\in\mathcal{V}_j)Z$ and setting $\tilde{Q}=P_0\left(\tilde{\ell}_0\tilde{Z}^{\mathrm{T}}\right)P_0^{-1}(\tilde{Z}\tilde{Z}^{\mathrm{T}})$. Then $\tilde{Q}\tilde{Z}=\sum_{j=1}^{J}\mathbf{1}(V\in\mathcal{V}_j)Q_jZ$ and $\mathbb{G}_j(\tilde{\ell}-\tilde{Q}\tilde{Z})=\mathbb{G}_j(\tilde{\ell}-Q_jZ)$ as desired. However, since this may greatly increase the number of calibration variables, the increased variability in the weights may result in an estimator with increased rather than reduced finite sample variance in all but the largest samples. See Sect. 5 below.

Biostatisticians (Robins et al., 1994) have also suggested adjustment of the sampling weights to improve efficiency, namely, by estimating the weights using a correct parametric model $\pi(z_i;\alpha)=\Pr(\xi_i=1|Z_i=z_i)$. For two-phase stratified sampling the model is rendered correct by including the J binary stratum indicators among the adjustment variables z. Logistic regression is typically used for $\pi(z_i;\alpha)$, in which case the estimating equations for α become

$$\sum_{i=1}^{N}\xi_i z_i=\sum_{i=1}^{N} z_i/w_i \tag{13}$$

where $w_i = p(z_i; \hat{\alpha}_N)^{-1}$ denotes the weight estimated for subject i. Note the parallel with the calibration equations (8): the weights appear on opposite sides. Results of Breslow and Wellner (2008) may again be used to derive the asymptotic distribution of $\hat{\theta}_N$ ($\hat{\alpha}_N$) under two-phase stratified sampling. See the Appendix for a brief derivation of the weak convergence

$$\sqrt{N}\left(\widehat{\theta}_N(\widehat{\alpha}_N)-\theta_0\right) \rightsquigarrow \mathbb{G}\tilde{\ell}_0+\sum_{j=1}\sqrt{\nu_j}\sqrt{\tfrac{1-p_j}{p_j}}\mathbb{G}_j\left(\tilde{\ell}_0-p_jRZ\right) \quad \text{where}$$
$$R=P_0(1-\pi_0)\tilde{\ell}_0Z^{\mathrm{T}}(P_0\pi_0(1-\pi_0)ZZ^{\mathrm{T}})^{-1}. \tag{14}$$

Note that R is also a projection matrix, now of components of $\tilde{\ell}_0$ onto the linear span of $\pi_0 Z$ with respect to the distribution P_π defined by $P_\pi f=P_0\left(\frac{1-\pi_0}{\pi_0}f\right)/P_0\left(\frac{1-\pi_0}{\pi_0}\right)$.

Previous discussions of Horvitz-Thompson methods with estimated weights have derived their properties for *Bernoulli* sampling, whereby the N phase one subjects are selected independently for membership in the phase two sample with known sampling probabilities $\pi_0(v)$ (Robins et al., 1994; Mark and Katki, 2006). This facilitates comparison of asymptotic properties in terms of influence functions. Using the asymptotic expansions in the Appendix, which hold for both sampling schemes, the influence function under Bernoulli sampling for the estimator based on calibrated weights is readily shown to equal

$$\frac{\xi}{\pi_0(v)}\tilde{\ell}_0(x)-\frac{\xi-\pi_0(v)}{\pi_0(v)}Qz(v) \tag{15}$$

where $\pi_0(v) = p_j$ for $v \in \mathcal{V}_j$. The asymptotic phase two variance under Bernoulli sampling is therefore $\sum_{j=1}^{J}\sqrt{\nu_j}\sqrt{(1-p_j)/p_j}P_{0|j}\left(\tilde{\ell}_0-QZ\right)^{\otimes 2}$. The limit laws under the two sampling schemes, however, are subtly different. The asymptotic phase two variance implied by (12) for two phase stratified sampling, namely, $\sum_{j=1}^{J}\sqrt{\nu_j}\sqrt{(1-p_j)/p_j}\mathrm{Var}_j\left(\tilde{\ell}_0-QZ\right)$, is less than that for Bernoulli sampling.

Expression (15) shows immediately how to choose z. Selecting $z^{\mathrm{opt}} = \mathrm{E}(\tilde{\ell}_0|V = v)$, we find $Qz^{\mathrm{opt}} = z^{\mathrm{opt}}$ and conclude that the class of calibrated IPW estimators contains the optimal member of the class of augmented IPW estimators considered by Robins et al. (1994),Wang and Chen (2001) and Mark and Katki (2006). Influence functions for this class take the form (15) with $Qz(v)$ replaced by $\varphi(v)$, an arbitrary function of v, and the choice $\varphi = z^{\mathrm{opt}}$ minimizes the asymptotic variance. Similarly, the influence function when weights are estimated using auxiliary variables z has the form (15) where Qz is replaced by $R\pi_0 z$. To obtain the optimal influence function, therefore, one would include the variables $z^{\mathrm{opt}}/\pi_0(v)$ together with the stratum indicators as predictors of the weights.

Calculation of $z^{\mathrm{opt}} = \mathrm{E}(\tilde{\ell}_0|V = v)$ requires knowledge of the conditional distribution $[X|V]$, which is generally unavailable. One approach to approximating z^{opt} is to specify parametric regression models for the components of ℓ_0 on V (Robins et al., 1994, Sect. 2.7) and to estimate the

parameters by IPW regression of the *dfbeta* on V using the second phase data. Kulich and Lin (2004) proposed an alternative "plug in" method for approximating the conditional expectation that we adopt here. It is likely to be most useful when there are only one or two partially missing variables. The steps are as follows:

1. Using (linear or logistic) regression models fitted by IPW to the phase two data, develop and fit parametric models to predict each partially missing variable (known only at phase two) given variables v known for all.
2. Use the models in step 1 to impute values $\hat{x}_i$ for everyone in the phase one sample; variables already known for everyone are used in their original form.
3. Fit the interest model $P_{\theta,\eta}(X)$ to the phase one sample using the imputed values $\hat{x}_i$.
4. Construct auxiliary variables z as imputed *dfbetas* (7) from the model fitted in step 3; these are estimates of z^{opt}.
5. Estimate θ using weights adjusted to the z_i.

Examples of this approach are given in the companion paper (Breslow et al., 2009) and in Sect. 5 below.

An interesting special case arises when V is discrete with $K > J$ levels, say $V \in \{1, \ldots, K\}$. Define $Z^{\mathrm{T}} = (\mathbf{1}(V = 1), \ldots, \mathbf{1}(V = K))$. Then both calibration and estimation lead to the same adjusted weights, namely, the inverse sampling fractions within each of the K strata defined by levels of V. This is a finer stratification than that actually used for sampling. Since

$$QZ = \sum_{k=1}^{K} P_{0|k}(\tilde{\ell}_0)\mathbf{1}(V=k) = \mathrm{E}(\tilde{\ell}_0|V),$$

where $P_{0|k}$ is defined analogously to $P_{0|j}$, it follows that Horwitz-Thompson estimation based on the finer stratification yields the most efficient estimate within the augmented IPW class. Survey statisticians refer to this method as *post-stratification*.

4 Log-linear Modeling

We first consider log-linear modeling of multinomial outcomes δ corresponding to occupancy of one of L cells ($\ell = 1, \ldots, L$) in a multidimensional table, with calibration to marginal totals rather than to IF contributions in a regression model. The outcome for subject i is $\delta_i^{\mathrm{T}} = (\delta_{i,1}, \ldots, \delta_{i,L})$ where $\delta_{i,\ell} = 1$ if the subject occupies cell ℓ, otherwise $\delta_{i,\ell} = 0$. Redefine X to denote a $L \times p$ design matrix ($p < L$), without the intercept to maintain identifiability, and let $\theta \in \Theta \subset R^p$ be a parameter vector such that, with x_ℓ denoting the ℓ^{th} row of X,

$$p_\ell(\theta) = \Pr(\text{occupy cell } \ell|\theta) = \frac{\exp(x_\ell\theta)}{\sum_{m=1}^{L}\exp(x_m\theta)}. \tag{16}$$

The fully parametric model $P_\theta(\delta)$ thus has likelihood

$$\mathrm{lik}(\theta) = \prod_{i=1}^{N}\prod_{\ell=1}^{L}[\,p_{i,\ell}(\theta)]^{\delta_{i,\ell}}.$$

Our results for semiparametric models and two-phase stratified samples apply to this situation with $\tilde{\ell}_\theta$ denoting the ordinary (parametric) influence function rather than the semiparametric efficient one. Details have been worked out by Kovacevic and Rai (2002) for joint model-design based inference for parameters in log-linear models under more general sampling designs. With $p(\theta)$ defined by $p^{\mathrm{T}}(\theta) = (p_1(\theta), \ldots, p_L(\theta))$ denoting the vector of cell occupancy probabilites under the model, the likelihood score is $\ell_\theta = X^{\mathrm{T}}[\delta - p(\theta)]$. With $O = \sum_{i=1}^{N} \delta_i$ denoting the (unobserved) vector of table frequencies for the phase one sample, the MLE $\hat{\theta}_N$ would be obtained from (2) as the solution to $\mathbb{P}_N\, \ell_\theta(\delta_i) = X^{\mathrm{T}}[O - p(\theta)N] = 0$. The equation to be solved for $\hat{\theta}_N$ is instead $\mathbb{P}_N^{\pi} \tilde{\ell}_\theta = X^{\mathrm{T}} \left[O^{\pi} - p(\theta)\widehat{N} \right]$ where $O^{\pi} = \sum_{i=1}^{N} \xi_i \pi_i^{-1} \delta_i$ denotes the estimated vector of frequencies and $\widehat{N} = \sum_{i=1}^{N} \xi_i \pi_i^{-1}$ (= N for two-phase stratified sampling) is the estimated phase one total. Thus one simply fits the log-linear model to the estimated table of frequencies.

$$I_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^{\mathrm{T}}$$

With $= X^{\mathrm{T}} \mathrm{diag}\{p_\ell(\theta)[1 - p_\ell(\theta)]\} X$ denoting the parametric information and $\tilde{\ell}_\theta = I_\theta^{-1} X^{\mathrm{T}}[\delta - p(\theta)]$ the parametric influence function, one can use equations (5) and (6) to determine the asymptotic variance of $\hat{\theta}_N$. The phase one component of variance is estimated by $(N I_{\hat{\theta}_N})^{-1}$ or more robustly

by

$$\widehat{\mathrm{Var}}_{\mathrm{PHS-I}} = \frac{1}{N(N-1)} I_{\hat{\theta}_N}^{-1} X^{\mathrm{T}} \left[\sum_{\ell=1}^{L} O_\ell^{\pi} (e_\ell - \hat{p})(e_\ell - \hat{p})^{\mathrm{T}} \right] X I_{\hat{\theta}_N}^{-1}$$

where $\hat{p} = p(\hat{\theta}_N)$ and e_ℓ denotes the L-vector with one in the ℓ^{th} position and zeroes elsewhere. Similarly, if $O_{j,\ell}$ denotes the number of phase two subjects in stratum j observed to occupy cell ℓ of the table, then the phase two component of variance is estimated by

$$\widehat{\mathrm{Var}}_{\mathrm{PHS-II}} = \frac{1}{N^2} I_{\hat{\theta}_N}^{-1} X^{\mathrm{T}} \left[\sum_{j=1}^{J} N_j \frac{1 - n_j/N_j}{n_j/N_j} S_j^2 \right] X I_{\hat{\theta}_N}^{-1} \quad \text{where}$$

$$S_j^2 = \frac{1}{n_j - 1} \sum_{\ell=1}^{L} O_{\ell,j} \left(e_\ell - \hat{p}^{(j)} \right) \left(e_\ell - \hat{p}^{(j)} \right)^{\mathrm{T}}$$

and $\hat{p}^{(j)}$ is the observed vector of cell occupancy fractions in stratum j.

As an illustration, we studied the association between high density lipoprotein-cholesterol (HDL-C) and lipoprotein-associated phospholipase A_2 (Lp-PLA_2) using data from the ARIC case-cohort study of Ballantyne et al. (2004). Measurements on HDL-C were available for the entire cohort of 12,345 participants who had plasma collected at their second follow-up visit and were free of coronary heart disease (CHD) at that time. Levels of Lp-PLA_2 were determined by assay of the stored plasma samples for the 604 who developed CHD during the 6–8 years of additional follow-up and for 732 controls who had been sampled for the cohort random sample after stratification on gender, age and ethnicity (two levels each). The ARIC investigators were interested in whether their data would support the finding by Persson et al. (2007) of a negative association between Lp-PLA_2, a new biomarker of inflammation, and the

well known risk factor HDL-C. Sampling weights for the nine sampling strata ranged from 1.2 for the stratum consisting of the CHD cases to 32 for white female controls under age 55. (The CHD cases were effectively sampled at less than 100% since some plasma samples were uninformative.)

All statistical analyses were carried out using the R `survey` package (Lumley, 2004) available from http://cran.r-project.org/. Table 1A shows the frequencies for the cross-classification of the two risk factors, estimated together with their standard errors using standard sampling weights. They were obtained using IPW estimates of population means that are implemented in the `svytotal` function. Note that the marginal HDL-C totals of the frequencies estimated using standard weights do *not* agree with the actual (integral) HDL-C totals for the cohort.

To formally assess the statistical significance of the clearly negative association, we fit a log-linear model with a term for a linear×linear interaction. The design matrix was specified as

$$X^{\mathrm{T}}=\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}.$$

The first four columns of X (rows of X^{T}) describe the standard independence model, a log-linear model with main effects for rows (HDL-C) and columns (Lp-PLA_2). The interaction term in the last column corresponds to the single degree of freedom linear by linear association test for dependence in a two-way table. Fitting the model using `svyloglin`, the interaction coefficient was −0.5323 with standard error 0.0734 yielding a test statistic $Z^2 = 52.6$, $p = 4 \times 10^{-13}$. Expanding the model to saturation and testing for dependence using a Wald statistic yielded $\chi^2 = 60.1$, DF = 4, $p = 3 \times 10^{-12}$. Most of the association was explained by the linear interaction.

Table 1B shows the tabular frequencies estimated after calibration of the weights to the marginal totals of HDL-C using the generalized raking procedure described in Sect. 3. These are the classical raking estimates that are proportionally adjusted to the known marginal totals (Deming and Stephan, 1940) so that the estimated and actual marginal (HDL-C) totals do now agree. The standard errors of the estimated marginal totals reflect only the binomial sampling variability at phase one; the phase two component is zero. With one minor exception where the estimated frequency also increased, the standard errors of frequencies in the cross-classification decreased markedly. There was no meaningful change in the precision of estimation of the marginal totals of Lp-PLA_2, reflecting the fact that no phase one information about this margin was utilized. Fitting the interaction model with calibrated weights led to a slightly more significant linear by linear association test, $Z^2 = 54.0$, $p = 2 \times 10^{-13}$, but there was virtually no change in the 4 DF test ($\chi^2 = 60.2$).

Results of estimating the table frequencies using estimated weights (not shown) were similar to those obtainned with calibrated weights in terms of precision. The estimated marginal HDL-C totals were 3412.9, 6014.9 and 2915.2, however, and hence differed from the actual totals shown in the margin of Table 2B. Nonetheless, the phase two contributions to the SEs of the marginal totals were still neglible if not absolutely zero. For the three levels of HDL-C, the SEs reported by `svytotal` were 49.693, 55.534 and 47.190, whereas the SEs based only on phase one variability were 49.687, 55.527 and 47.184.

See the companion paper for Cox regression analyses of the ARIC data.

5 Simulation Study of Cox Regression Modeling

To study the effect of different methods of adjustment of sampling weights on the performance of Cox regression coefficients estimated with the `svycoxph` function, we used data on $N = 3,915$ NWTS patients that were also considered in the the companion paper. The data and sample R code are available from the website http://faculty.washington.edu/norm/software.html. These data were also used by Kulich and Lin (2004), whom we follow in choice of models, to illustrate their "combined, doubly weighted" estimator. The goal was to estimate hazards associated with prognostic variables in a failure time analysis, with event-free survival (time from diagnosis to disease progression or death) as the endpoint. The most important prognostic factor was histologic subtype, classified as favorable (FH) or unfavorable histology (UH). Information on histology was available both from the central pathology reference laboratory and from the institution where the patient was treated. This made possible the simulation of exposure stratified case-cohort studies (Borgan et al., 2000), treating the central pathology measurement as known only for the phase two sample and the institutional pathology measurement as one of the variables used for stratification of the sample and adjustment of the weights. The Cox model was first fit to the entire cohort, yielding the normally unobservable $\tilde{\theta}_N$ for comparison with the two-phase estimates $\hat{\theta}_N$. Mean squared differences $(\tilde{\theta}_N - \hat{\theta}_N)^2$ averaged over 10,000 phase two samples provided an empirical estimate of the phase two variance component.

Phase two sampling was stratified on the basis of institutional histology, stage of disease (low=I, II *vs* high=III, IV), age at diagnosis (babies *vs* 1+ year olds) and whether (cases) or not (controls) the patient had relapsed or died before the end of follow-up. All cases and all but the three largest control strata were included in their entirety in the phase two sample: 120 of 452 FH control babies with low stage disease; 160 of 1,620 1+ year old controls with low stage disease; and 120 of 914 1+ year old controls with high stage disease were sampled at random for phase two. This provided 660 phase two controls for comparison with the 669 cases and a phase two sample size of $n = 1,329$. The Cox model contained terms for histology (UH), age as a linear spline with separate slopes for babies (Age_0) and older children (Age_1), a binary indicator of high *vs* low stage, tumor diameter (cm) and interactions between stage and diameter and between histology and the two age terms.

Auxiliary variables for calibration and estimation of the weights were constructed using the procedure described at the end of Sect. 3. For each of the 10,000 simulated phase two samples we first used the `svyglm` function with standard sampling weights to predict histology (central pathology) based on a logistic regression model containing terms for institutional (local) histology (the main predictor), stage, age greater or less than 10 years, study (NWTS-3 *vs* NWTS-4) and an interaction between local UH and stage. The logistic regression coefficients were then used in conjunction with phase one data to yield a predicted probability of having UH for each of the 3,915 phase one subjects. The Cox model was fit using the ordinary `coxph` procedure with this predicted probability in place of the binary UH covariate. Estimated IF contributions (*dfbetas*) were extracted, augmented by 1 to ensure that they were positive, and used as the auxiliary variables z for calibration. For estimation the *dfbetas* were divided by the known sampling probabilities before adding 1, and combined with the stratum indicators in the prediction equation.

Here is a brief summary of results of the initial simulation study as reported in the companion paper (Breslow et al., 2009). Calibration again used the raking procedure. Mean values of the regression coefficients were nearly identical for the three case-cohort estimation methods (standard, calibrated and estimated weights) and close to those estimated for the entire cohort. Both calibration and estimation reduced the phase two variance components, sometimes to negligible levels. For example, the phase two standard error of the Age_0 coefficient was reduced

from 0.162 with standard weights to 0.037 with calibrated and to 0.061 with estimated weights. The phase two SE for the UH main effect was reduced by 29% with calibration and to 28% with estimation. There were lesser reductions for the interaction with Age_0.

Additional simulations based on only 1,000 replications were carried out to determine whether the modification of the auxiliary variables suggested in the paragraph following equation (12) in Sect. 3, intended to reduce the asymptotic phase two variance further under two phase stratified sampling, would actually improve estimates in finite samples. This increased the number of adjustment variables from 8, the number of terms in the Cox model, to 32 in view of the four sampling strata. Substantial numerical difficulties were experienced with the larger number of adjustment variables, particularly for calibration. The `survey` option to force convergence of the iterative solution of the calibration equations, by returning the weights after 200 iterations, was required to prevent otherwise frequent failure of the algorithm but led to outliers in the estimated regression coefficients and the failure of the `svycoxph` procedure in 2 samples. This increased the phase two variability in comparison with calibration based on a smaller number of variables.

Results are shown in Table 2. The column labeled Cox SE shows the robust (Lin and Wei, 1989) standard errors for the Cox model fit to the entire cohort. These were noticeably larger than the model based standard errors, reflecting a lack of proportionality for several covariates. The phase one variance components estimated by `svycoxph` also are the robust versions, emphasizing the sample survey view that the goal is to approximate the results of fitting a possibly misspecified model to the entire cohort. The columns labeled SE for the various two-phase case-cohort methods are averages of the total estimated standard errors, incorporating both phase one and phase two variability. The columns labeled RMSE contain the empirical phase two standard errors. For most coefficients, in stark contrast to results obtained with just 8 adjustment variables, the RMSE using calibrated weights were larger even than with standard sampling weights. The `svycoxph` procedure returned regression coefficients for all 1,000 replicates when weights were determined by estimation, but they too had slightly larger RMSE than those obtained with fewer adjustment variables. On the other hand, they were more accurate than estimates obtained using standard weights. Qualitatively similar results were obtained when the adjustment variables were limited to the 24 corresponding to the three strata sampled at less than 100%.

6 Discussion and Conclusions

This paper has reviewed statistical properties of Horvitz-Thompson estimators of Euclidean parameters in semiparametric models with two-phase stratified samples, derived the corresponding properties for modified estimators where the sampling weights are adjusted by calibration or estimation, and illustrated the methods by log-linear and Cox regression modeling of stratified case-cohort data. A limitation of the results of van der Vaart (1998) for the Cox model, which were used by Breslow and Wellner (2007) and apply therefore to those reported here, is that time-dependent covariates were excluded from consideration. This is currently also a limitation of the `svycoxph` function in the `survey` package, at least so far as correct estimation of the variance in concerned. Work is in progress to modify `svycoxph` so that it will calculate the variances correctly when multiple time slice records are included in the survival analysis for each phase two subject. This will provide the extension needed to deal with time-dependent covariates. We think it likely that the asymptotic results hold for this situation and in fact much more generally. We are hopeful that a general theory for estimation of parameters in semiparametric models can be developed for sample survey designs along the lines of what Lin (2000) has provided for the Cox model.

The asymptotic properties of estimators with weights adjusted using calibration and estimation are very similar. Finite sample performance of the two methods of adjustment was also very similar in the initial simulation results for the NWTS cohort as reported in the companion paper. An advantage of calibration for the survey statistician is that information on the auxiliary variables is not needed for everyone in the cohort (phase one sample). Only the totals over the cohort are required. The more detailed information is required for estimation of the sampling probabilities.

Our asymptotic results suggested the possibility of improvement by using stratum specific versions of the imputed IF contributions for calibration of the weights. Robins et al. (1994) likewise noted that, so far as asymptotic variances are concerned, increasing the number of auxiliary variables used to estimate the weights can only reduce (or at least never increase) the asymptotic variance of the estimators. As the simulations reported here illustrate, however, actual performance in finite samples may be quite different. Adjustment based on calibration was particularly susceptible to numerical problems caused by a large number of auxiliary variables. In practical work, standard or estimated sampling weights should be used when the calibration algorithm fails to converge. Options for calibration that bound the discrepancy between design and calibrated weights have been implemented in the R `survey` package and further work exploring the finite sample properties of estimators using such weights would be desirable. It would also be of interest to compare our "plug in" method of approximating the optimal auxiliary variables z^{opt} to other methods of estimating the conditional mean.

Methods for improving the analysis of data from two phase stratified samples have now been implemented in the flexible `survey` package of the freely available R statistical system. There is no longer any excuse for epidemiologists and statisticians to waste valuable information by inefficient analysis of data from stratified case-control and case-cohort studies.

Acknowledgments

The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by the National Heart, Lung and Blood Institute contracts N01-HC-55-15, N01-NC-55016, N01-HC-55019, N01-HC-55020, N01-HC-55021 and N01-HC-55022. The National Wilms Tumor Study and Norman Breslow received financial support from grants R01-CA-054498, R01-CA-40644 and earlier grants from the National Cancer institute. Michal Kulich received financial support from Research Project MSM 0021620839 from Ministry of Education, Czech Republic. The authors thank the staff and participants of the ARIC and NWTS studies for their important contributions.

References

Ballantyne CM, Hoogeveen RC, Bang H, et al. Lipoprotein-associated phospholipase A(2), high-sensitivity C-reactive protein, and risk for incident coronary heart disease in middle-aged men and women in the Atherosclerosis Risk in Communities (ARIC) study. Circulation 2004;109:837–842. [PubMed: 14757686]

Barlow WE. Robust variance estimation for the case-cohort design. Biometrics 1994;50:1064–1072. [PubMed: 7786988]

Barlow WE, Ichikawa L, Rosner D, Izumi S. Analysis of case-cohort designs. J Clin Epidemiol 1999;52:1165–1172. [PubMed: 10580779]

Begun JM, Hall WJ, Huang W-M, Wellner JA. Information and asymptotic efficiency in parametric-nonparametric models. Ann Stat 1983;11:432–452.

Binder DA. Fitting Cox's proportional hazards model from survey data. Biometrika 1992;79:139–147.

Borgan O, Langholz B, Samuelsen SO, et al. Exposure stratified case-cohort designs. Lifetime Data Anal 2000;6:39–58. [PubMed: 10763560]

Breslow N. Covariance analysis of censored survival data. Biometrics 1974;30:89–99. [PubMed: 4813387]

Breslow NE, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. J Roy Stat Soc B 1997;59:447–461.

Breslow NE, Lumley T, Ballantyne CM, et al. Using the whole cohort in the analysis of case-cohort data. Am J Epidemiol 2009;192 (in press).

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.

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'. Scand J Stat 2008;35:186–192.

Cain KC, Lange NT. Approximate case influence for the proportional hazards regression model with censored data. Biometrics 1984;40:493–499. [PubMed: 6386066]

Cox DR. Regression models and life-tables (with discussion). J Roy Stat Soc B 1972;34:187–220.

Cox DR. Partial likelihood. Biometrika 1975;62:269–276.

D'Angio GJ, Breslow N, Beckwith JB, et al. Treatment of Wilms' tumor: Results of the third National Wilms' Tumor Study. Cancer 1989;64:349–360. [PubMed: 2544249]

Deming WE, Stephan FF. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann Math Stat 1940;11:427–444.

Deville JC, Särndal C-E. Calibration estimators in survey sampling. J Am Stat Assoc 1992;87:376–382.

Green DM, Breslow NE, Beckwith JB, et al. Comparison between single-dose and divided-dose administration of dactinomycin and doxorubicin for patients with Wilms' tumor: A report from the National Wilms' Tumor Study Group. J Clin Oncol 1998;16:237–245. [PubMed: 9440748]

Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 1952;47:663–685.

The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) study: design and objectives. Am J Epidemiol 1989;129:687–702. [PubMed: 2646917]

Isaki CT, Fuller WA. Survey design under the regression superpopulation model. J Am Stat Assoc 1982;77:89–96.

Kovacevic MS, Rai SN. Log-linear modelling of change using longitudinal survey data. Commun Stat Theory Methods 2002;31:1815–1835.

Kulich M, Lin DY. Improving the efficiency of relative-risk estimation in case-cohort studies. J Am Stat Assoc 2004;99:832–844.

Lin DY. On fitting Cox's proportional hazards models to survey data. Biometrika 2000;87:37–47.

Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. J Am Stat Assoc 1989;84:1074–1078.

Lumley T. Analysis of complex survey samples. J Stat Software 2004;9:1–19.

Mark SD, Katki HA. Specifying and implementing nonparametric and semiparametric survival estimators in two-stage (nested) cohort studies with missing case data. J Am Stat Assoc 2006;101:460–471.

Nan B. E3cient estimation for case-cohort studies. Can J Stat 2004;32:403–419.

Neyman J. Contribution to the theory of sampling human populations. J Am Stat Assoc 1938;33:101–116.

Persson M, Nilsson JA, Nelson JJ, et al. The epidemiology of Lp-PLA(2): Distribution and correlation with cardiovascular risk factors in a population-based cohort. Atherosclerosis 2007;190:388–396. [PubMed: 16530769]

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

Rao JNK, Yung W, Hidiroglou M. Estimating equations for the analysis of survey data using poststratification information. Sankhya 2002;64:364–378.

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

Rubin-Bleuer S, Kratina IS. On the two-phase framework for joint model and design-based inference. Ann Stat 2005;33:2789–2810.

Särndal C-E, Swensson B, Wretman JH. The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. Biometrika 1989;76:527–537.

Scheike TH, Martinussen T. Maximum likelihood estimation for Cox's regression model under case-cohort sampling. Scand J Stat 2004;31:283–293.

Scott AJ, Wild CJ. Fitting regression models to case-control data by maximum likelihood. Biometrika 1997;84:57–71.

Therneau, TM.; Grambsch, PM. Modeling Survival Data: Extending the Cox Model. Springer-Verlag; New York: 2000.

van der Vaart, AW. Asymptotic Statistics. Cambridge University Press; Cambridge: 1998.

van der Vaart, AW.; Wellner, JA. Weak Convergence and Empirical Processes with Applications in Statistics. Springer; New York: 1996.

Wang CY, Chen HY. Augmented inverse probability weighted estimator for Cox missing covariate regression. Biometrics 2001;57:414–419. [PubMed: 11414564]

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

Zeng D, Lin DY. Maximum likelihood estimation in semiparametric regression models with censored data. J Roy Stat Soc B 2007;69:507–536.

7 Appendix

Here we provide a brief outline of the derivation of (12) for the calibration estimator. For in a neighborhood of the limiting value $\lambda_0 = 0$ denote modified weights by $\pi_\lambda^{-1}(v)=\pi_0^{-1}(v)[1-\lambda^{\mathrm{T}}z]$, where $\pi_0^{-1}(v)$ are known sampling weights depending on v. Define classes of functions

$$\psi_{1;\theta,\eta,\lambda}(w,\xi)=\frac{\xi}{\pi_\lambda(v)}\dot{\ell}_{\theta,\eta}(x) \text{ and}$$
$$\psi_{2;\theta,\eta,\lambda,h}(w,\xi)=\frac{\xi}{\pi_\lambda(v)}B_{\theta,\eta}h(x)$$

for $\theta\in\Theta$, $\eta\in\Xi$, and $h\in\mathcal{H}$, and let $\Psi_{1;\theta,\eta,\lambda}=\mathrm{P}_0\,\psi_{1;\,\theta,\eta,\lambda}\,(W,\xi)$ and $\Psi_{2;\,\theta,\eta,\lambda,h}=\mathrm{P}_0\,\psi_{2;\,\theta,\eta,\lambda,h}(W,\xi)$ denote their expectation. Then

$$\dot{\Psi}_{1,\lambda}=\frac{\partial}{\partial\lambda^{\mathrm{T}}}P_0\psi_{1;\theta,\eta,\lambda,h}\Big|_{\lambda=0}=-P_0\left(\dot{\ell}_0Z^{\mathrm{T}}\right) \quad \text{and}$$
$$\dot{\Psi}_{2;\lambda}h=\frac{\partial}{\partial\lambda^{\mathrm{T}}}P_0\psi_{2;\theta,\eta,\lambda,h}\Big|_{\lambda=0}=-P_0(B_0hZ^{\mathrm{T}}),\quad h\in\mathcal{H}.$$

Substituting these expressions for the analogous quantities $\Psi_{1,\alpha}$ and $\Psi_{2;\alpha}h$ in Breslow and Wellner (2008, Sect. 3) and reworking the calculations there shows that

$$\begin{aligned}\sqrt{N}\left(\widehat{\theta}_N(\widehat{\lambda}_N)-\theta_0\right)&\\ =\sqrt{N}\left(\widehat{\theta}_N(0)-\theta_0\right)&\\ -P_0\left(\tilde{\ell}_0Z^{\mathrm{T}}\right)\sqrt{N}\widehat{\lambda}_N+o_p(1).&\end{aligned}$$

From (6), (9) and (11), assuming $n\uparrow\infty$ faster than $\sqrt{N}$ as $N\uparrow\infty$, we find

$$\sqrt{N}\widehat{\lambda}_N = B^{-1}\sqrt{N}(\mathbb{P}_N^{\pi} - \mathbb{P}_N)Z + o_p(1) \rightsquigarrow \sum_{j=1}^{J}\sqrt{v_j}\sqrt{\frac{1-p_j}{p_j}}\mathbb{G}_j B^{-1}Z$$

while from (5) and (6)

$$\sqrt{N}\left(\widehat{\theta}_N(0) - \theta_0\right) \rightsquigarrow \mathbb{G}\tilde{\ell}_0 + \sum_{j=1}^{J}\sqrt{v_j}\sqrt{\frac{1-p_j}{p_j}}\mathbb{G}_j\tilde{\ell}_0.$$

Combining the last three equations yields the desired result (12).

The limit law for $\hat{\theta}_N$ based on estimated weights is derived similarly. From Breslow and Wellner (2008, eq. 18) one concludes for either Bernoulli or two-phase stratified sampling

$$\sqrt{N}\left(\widehat{\theta}_N(\widehat{\alpha}_N) - \theta_0\right) = \sqrt{N}\left(\widehat{\theta}_N(\alpha_0) - \theta_0\right) - P_0\left(\tilde{\ell}_0(1-\pi_0)Z\right)\sqrt{N}(\widehat{\alpha}_N - \alpha_0) + o_p(1).$$

A Taylor expansion with remainder of $\hat{\alpha}_N$ about α_0 in (13) leads to

$$\sqrt{N}(\widehat{\alpha}_N - \alpha_0) = [P_0\pi_0(1-\pi_0)ZZ^{\mathrm{T}}]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}[\xi_i - \pi_0(v_i)]Z + o_p(1).$$

Combining these two equations with (5) we find

$$\sqrt{N}\left(\widehat{\theta}(\widehat{\alpha}_N) - \theta_0\right) = \sqrt{N}\mathbb{P}_N\tilde{\ell}_0 + \sqrt{N}(\mathbb{P}_N^{\pi} - \mathbb{P}_N)\left(\tilde{\ell}_0 - \pi_0 RZ\right) + o_p(1)$$

and the conclusion (14) follows from the arguments in Breslow and Wellner (2007, Sect. 4).

Table 1

Association between HDL-C and Lp-PLA$_2$: Standard *vs* Calibrated Weights

HDL-C (mg/L)	A. Standard weights Lp-PLA$_2$ (μG/L)				B. Calibrated weights Lp-PLA$_2$ (μG/L)			
	0–0.309	**0.31–0.42**	**0.421–1**	**Total**	**0–0.309**	**0.31–0.42**	**0.421–1**	**Total**
	Estimated frequencies				Estimated frequencies			
<40	701.4	938.6	1,561.3	3,201.3	739.0	988.9	1,645.0	3,373
40–59.0	1,764.4	2,310.2	2,175.3	6,249.9	1,665.0	2,180.1	2,052.9	5,898
≥60.0	1,569.6	909.4	414.8	2,893.8	1,667.4	966.0	440.6	3,074
Total	4,035.4	4,158.2	4,154.4	12,345	4,071.4	4,135.1	4,138.5	12,345
	Standard errors				Standard errors			
<40	105.7	111.3	138.9	185.4	99.7	105.0	117.7	49.5
40–59.0	166.9	187.5	170.0	234.1	144.4	155.9	146.7	55.5
≥60.0	164.2	124.1	81.9	197.7	128.0	117.0	83.1	48.0
Total	217.3	222.2	206.1		212.9	220.3	201.8	

Table 2

Average Total Standard Error (SE) and Empirical Phase Two Standard Error (RMSE). Adjustment to Within Stratum Totals of Estimated Influence Function Contributions

	Cox	Standard weights		Calibrated weights*		Estimated weights	
Model term	**SE**	**SE**	**RMSE**	**SE**	**RMSE**	**SE**	**RMSE**
UH	.503	.537	.188	.572	.157	.517	.146
Age_0 (yr)	.321	.360	.158	.429	.228	.323	.085
Age_1 (yr)	.015	.026	.021	.024	.030	.016	.016
Stage (III, IV *vs* I, II)	.259	.346	.231	.365	.265	.271	.116
Diameter (cm)	.015	.021	.015	.021	.018	.015	.008
Stage × Diameter	.020	.029	.020	.028	.020	.021	.010
UH×Age_0	.552	.612	.287	.668	.321	.587	.258
UH×Age_1	.033	.051	.048	.053	.060	.046	.050

*Based on 998 replications for which coefficients were obtained