IMS Collections
From Probability to Statistics and Back: High-Dimensional Models and Processes
Vol. 9 (2013) 65–77

DOI: 10.1214/12-IMSCOLL906

Semiparametric models and two-phase samples: Applications to Cox regression

Norman E. Breslow[*,‡] **and Thomas Lumley**[†,§]

University of Washington, Seattle and University of Auckland

Abstract: A standard estimation method when fitting parametric models to data from two-phase stratified samples is inverse probability weighting of the estimating equations. In previous work we applied this approach to likelihood equations for both Euclidean and non-Euclidean parameters in semiparametric models. We proved weak convergence of the inverse probability weighted empirical process and derived an asymptotic expansion for the estimator of the Euclidean parameter. We also showed how adjustment of the sampling weights by their calibration to known totals of auxiliary variables, or their estimation using these same variables, could markedly improve efficiency.

Here we consider joint estimation of Euclidean and non-Euclidean parameters. Our asymptotic expansion for the non-Euclidean parameter is apparently new even in the special case of simple random sampling. The results are applied to estimation of survival probabilities for individual subjects using the regression coefficients (log hazard ratios) and baseline cumulative hazard function of the Cox proportional hazards model. Expressions derived for the variances of regression coefficients and cumulative hazards estimated after calibration of the weights aid construction of the auxiliary variables used for adjustment. We demonstrate empirically the improvement offered by calibration or estimation of the weights via simulation of two-phase stratified samples using publicly available data from the National Wilms Tumor Study and data analysis with the R survey package.

1. Introduction

Two-phase stratified sampling designs are useful for selecting informative subjects for ascertainment of expensive covariate information. They are particularly valuable for clinical medicine and epidemiology when a large cohort, the Phase I sample, is followed forward in time for the occurrence of a disease event and substantial information is already available for cohort members. Judicious selection of the Phase II sample, combined with efficient methods of analysis, can substantially lower costs associated with precise estimation of covariate effects. Paradigms include stratified versions of the case-control [7] and case-cohort [4, 21] designs.

A standard method of analysis of data from two-phase stratified samples is inverse probability (of sampling) weighting (IPW) of the estimating equations. In

[‡]Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195-7232, USA, e-mail: norm@uw.edu, url: http://faculty.washington.edu/norm

[§]Department of Statistics, Science Centre Building 303, 38 Princes Street, Auckland 1010, NZ, e-mail: t.lumley@auckland.ac.nz

[*]Supported in part by USPHS Grant 2 R01 CA054498

[†]Supported in part by the Marsden Fund

AMS 2000 subject classifications: Primary 60F05, 60F17; secondary 60J65, 60J70

Keywords and phrases: Asymptotic distributions, asymptotic efficiency, calibration, empirical processes, survival analysis, stratified sampling, two-phase

previous work [10, 11] using this approach with likelihood equations for semiparametric models [25, §12.25], we derived the asymptotic distribution of estimators of the Euclidean parameters by proving weak convergence of the IPW empirical process. Subsequently [8] we derived asymptotic expansions for the estimators when weights were adjusted by their calibration [14] to Phase I totals of auxiliary variables known for all cohort members. We adopted a "plug in" method of constructing near optimal calibration variables and evaluated its performance by repeated drawing of stratified Phase II samples from a cohort of National Wilms Tumor Study (NWTS) patients [9]. Similar results were obtained when the weights were adjusted via estimation, *i.e.*, by fitted values from a logistic model regressing the Phase II sampling indicators on the calibration variables [22].

After reviewing this earlier work, we consider joint estimation of Euclidean and non-Euclidean parameters with data from two-phase samples. An asymptotic expansion for the non-Euclidean parameter estimator based on calibrated weights is derived and used to motivate a further suggestion for calibration variables. Simulations based on NWTS data are extended to investigate the gains from use of calibrated or estimated weights to estimate the survival probabilities. The mathematical exposition is informal, with minimal attention paid to assumptions needed for a rigorous development. Lumley's R survey package [17] was used for all calculations.

2. Background and notation

2.1. The model

Following [25] consider a model $P_{\theta,\eta}(X)$ for a random variable X where $\theta \in \Theta \subset R^p$ and $\eta \in H \subset \mathcal{B}$, with $\mathcal{B}$ typically a normed space of functions or measures. $P_0 = P_{\theta_0,\eta_0}$ is the distribution from which X is actually sampled. Expectations are denoted $P_0 f = \int f(x)\, dP_0(x)$ for real or vector-valued functions f. Let $\dot{\ell}_{\theta,\eta}$ denote the p-dimensional likelihood score for θ and $B_{\theta,\eta}$ the score operator [3] that maps directions $h \in \mathcal{H}$ from which one dimensional sub-models η_t approach η into the corresponding likelihood scores. Let $(X_1, \ldots, X_N)$ denote a simple random sample from P_0 and denote by $\mathbb{P}_N$ the corresponding empirical measure: $\mathbb{P}_N f = (1/N)\sum_{i=1}^N f(X_i)$. We assume sufficient regularity to guarantee $\sqrt{N}$ consistency and asymptotic Gaussianity for maximum likelihood (ML) estimators $(\tilde{\theta}_N, \tilde{\eta}_N)$ [10, 11].

X is not completely observed for the simple random sample, however. We denote by $\tilde{X} = \tilde{X}(X)$ the portion that is observed and by U a vector of auxiliary variables. $W = (X, U)$ denotes data *potentially available* for all N subjects and $V = (\tilde{X}, U) \in \mathcal{V}$ data *actually observed* for all of them.

2.2. The sampling design

Consider two-phase sampling in which the Phase I sample (main cohort) is drawn by simple random sampling from an infinite *super-population* specified by P_0, which we *redefine* to be the distribution of W. Σ_N denotes the sigma field generated by $(W_1, \ldots, W_N)$, also known as the *complete data*. Our goal is to use $(V_1, \ldots, V_N)$ together with the additional data collected at Phase II to estimate (θ, η), coming as close as possible to the ML estimates $(\tilde{\theta}_N, \tilde{\eta}_N)$ that would have been obtained had complete data been observed.

At Phase II a subsample is drawn with sampling probabilities dependent on V and the remainder of X is observed for subjects in the subsample. We consider stratified random samples where $\mathcal{V}$ is partitioned into J strata, $\mathcal{V} = \mathcal{V}_1 \bigcup \cdots \bigcup \mathcal{V}_J$. Let R_i denote a random binary indicator of whether or not the i^{th} main cohort subject is sampled at Phase II and define the corresponding sampling probability by $\pi_i = \Pr(R_i = 1|V_i)$. There are two possibilities. With *Bernoulli* sampling [15, 18] the N variables V_i are inspected sequentially and the R_i are generated independently with $\pi_i = p_j$ for $V_i \in \mathcal{V}_j$. The $p_j > 0$ are known sampling probabilities. This setup preserves the i.i.d. structure of the observed data $(V_i, R_i, R_i X_i)$ and simplifies the theoretical development. The *finite population stratified sampling* (FPSS) design, which is closer to actual practice, records the stratum frequencies $N_j = \sum_{i=1}^{N} \mathbf{1}_{\mathcal{V}_j}(V_i)$ at Phase I and selects into the Phase II sample $n_j \leq N_j$ of the observations in the j^{th} stratum by random sampling *without replacement*. With FPSS the sampling indicators R_i are no longer mututally independent. However, those corresponding to different strata are independent and within each stratum the indicators are exchangeable. The Phase II sample size is $n = \sum_{j=1}^{J} n_j = \sum_{i=1}^{N} R_i$. The sampling fractions are assumed to converge: $n_j/N_j \to p_j$ as $N \uparrow \infty$.

2.3. IPW empirical measure and estimating equations

Define the discrete measure $\mathbb{P}_N^\pi$ by putting masses $1/(N\pi_i)$ on each of the n selected $(R_i = 1)$ observations and 0 mass on the remaining $(N - n)$. $\mathbb{P}_N^\pi$ is analogous to the bootstrap in that it involves sampling from $\mathbb{P}_N$. The IPW estimating equations are

$$\mathbb{P}_N^\pi \dot{\ell}_{\theta,\eta} = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i}{\pi_i} \dot{\ell}_{\theta,\eta}(X_i) = 0, \tag{2.1}$$

$$\mathbb{P}_N^\pi B_{\theta,\eta} h = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i}{\pi_i} B_{\theta,\eta} h(X_i) = 0 \qquad \forall h \in \mathcal{H}, \tag{2.2}$$

whose solution we denote by $(\widehat{\theta}_N, \widehat{\eta}_N)$. Were complete data available for all main cohort subjects, the maximum likelihood estimates $(\tilde{\theta}_N, \tilde{\eta}_N)$ would be obtained by solving the same equations with $\mathbb{P}_N$ replacing $\mathbb{P}_N^\pi$. The only Phase I information used in this process are the stratum frequencies $(N_1, \ldots, N_J)$ that determine the sampling weights π_i in (2.1) and (2.2). In Section 3.3 we show how to utilize more of this information by adjusting the weights.

3. IPW estimation of Euclidean parameters

3.1. Weak convergence of the IPW empirical process

Asymptotic properties of the *IPW empirical process* $\mathbb{G}_N^\pi = \sqrt{N}(\mathbb{P}_N^\pi - P_0)$ under Bernoulli sampling follow from van der Vaart's infinite dimensional Z-estimation theorem [25, Thm 19.26]. Using results of Præstgaard and Wellner [20] on weak convergence of the exchangeably weighted bootstrap, Breslow and Wellner [10] showed for the FPSS design that, with $\mathbb{G}_N = \sqrt{N}(\mathbb{P}_N - P_0)$ the usual empirical process,

$$\mathbb{G}_N^\pi = \mathbb{G}_N + \sqrt{N}\big(\mathbb{P}_N^\pi - \mathbb{P}_N\big) \rightsquigarrow \mathbb{G} + \sum_{j=1}^{J} \sqrt{\nu_j} \sqrt{\frac{1 - p_j}{p_j}} \mathbb{G}_j, \tag{3.1}$$

where $\nu_j = P_0(V \in \mathcal{V}_j)$ is the "size" of the j^{th} stratum, $\mathbb{G}$ is the P_0-Brownian bridge and, with $P_{0|j}$ denoting P_0 restricted to stratum j[1], $\mathbb{G}_j$ is the $P_{0|j}$-Brownian bridge. The limiting Gaussian processes $(\mathbb{G}, \mathbb{G}_1, \ldots, \mathbb{G}_J)$, which are mutually independent, are each indexed by a P_0-Donsker class of functions $\mathcal{F}$ and $\rightsquigarrow$ denotes weak convergence in $\ell^\infty(\mathcal{F})$.

3.2. Asymptotic distribution of $\widehat{\theta}_N$

Under stated regularity conditions A1-A4, we previously [10, §5] derived the expansion

$$\sqrt{N}\dot{\Psi}_0 \begin{pmatrix} \widehat{\theta}_N - \theta_0 \\ \widehat{\eta}_N - \eta_0 \end{pmatrix} = -\mathbb{G}_N^\pi \begin{pmatrix} \dot{\ell}_0 \\ B_0 h \end{pmatrix} + o_p(1), \tag{3.2}$$

where $\dot{\Psi}_0$ denotes the Fréchet derivative at (θ_0, η_0) of the map $\Psi = (\Psi_1, \Psi_2) : \Theta \times H \mapsto R^p \times \ell^\infty(\mathcal{H})$ with components $\Psi_1(\theta, \eta) = P_0 \dot{\ell}_{\theta,\eta}$ and $\Psi_2(\theta, \eta) = P_0 B_{\theta,\eta} h - P_{\theta,\eta} B_{\theta,\eta}$, $h \in \mathcal{H}$. We further assumed (A5) that η could be considered a measure and that $\dot{\Psi}_0$ admitted a partition as in [25, Eq. 25.91]. With $I_0 = P_0(\dot{\ell}_0 \dot{\ell}_0^{\mathrm{T}})$ the ordinary information for θ, B_0^* the adjoint of $B_0 = B_{\theta_0,\eta_0}$ and $(B_0^* B_0)$ the *information operator* [3], this allowed us to write (3.2) as

$$-I_0 \sqrt{N}(\widehat{\theta}_N - \theta_0) - \sqrt{N}(\widehat{\eta}_N - \eta_0) B_0^* \dot{\ell}_0 = -\mathbb{G}_N^\pi \dot{\ell}_0 + o_p(1), \tag{3.3}$$

$$-P_0\big[(B_0 h)\dot{\ell}_0^{\mathrm{T}}\big]\sqrt{N}(\widehat{\theta}_N - \theta_0) - \sqrt{N}(\widehat{\eta}_N - \eta_0) B_0^* B_0 h = -\mathbb{G}_N^\pi B_0 h + o_p(1). \tag{3.4}$$

Choosing $h = (B_0^* B_0)^{-1} B_0^* \dot{\ell}_0$ and subtracting (3.3) from (3.4) as in [25, p. 424] led to the key result

$$\sqrt{N}(\widehat{\theta}_N - \theta_0) = \mathbb{G}_N^\pi \tilde{\ell}_0 + o_p(1) \tag{3.5}$$

which may also be written in the form

$$\begin{aligned} \sqrt{N}(\widehat{\theta}_N - \theta_0) &= \sqrt{N}(\tilde{\theta}_N - \theta_0) + \sqrt{N}(\widehat{\theta}_N - \tilde{\theta}_N) \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^N \tilde{\ell}_0(X_i) + \frac{1}{\sqrt{N}} \sum_{i=1}^N \left(\frac{R_i}{\pi_i} - 1 \right) \tilde{\ell}_0(X_i) + o_p(1), \end{aligned} \tag{3.6}$$

valid for both Bernoulli and FPSS designs. Here $\tilde{\ell}_0 = \tilde{\mathcal{I}}_0^{-1} \ell_0^*$ denotes the semiparametric *efficient influence function* whose components

$$\ell_0^* = \big(I - B_0 \big(B_0^* B_0\big)^{-1} B_0^*\big)\dot{\ell}_0 \quad \text{and} \tag{3.7}$$

$$\tilde{\mathcal{I}}_0 = P_0\big[\big(I - B_0 \big(B_0^* B_0\big)^{-1} B_0^*\big)\dot{\ell}_0 \dot{\ell}_0^{\mathrm{T}}\big] \tag{3.8}$$

are the *efficient score* and *efficient information*, respectively.

The expansion (3.5) together with (3.1) shows that $\sqrt{N}(\widehat{\theta}_N - \theta_0)$ has under FPSS an asymptotically Gaussian distribution with mean zero and variance

$$\mathrm{Var}_{\mathrm{A}} \sqrt{N}(\widehat{\theta}_N - \theta_0) = \tilde{\mathcal{I}}_0^{-1} + \sum_{j=1}^J \nu_j \frac{1 - p_j}{p_j} \mathrm{Var}_j(\tilde{\ell}_0). \tag{3.9}$$

[1] $P_{0|j}(A) = P_0(A\mathcal{V}j)/P_0(\mathcal{V}_j)$

This equals the sum of the *Phase I variance* $\tilde{\mathcal{I}}_0^{-1}$, the variance of the unobserved ML estimator $\tilde{\theta}_N$ based on complete data, and the *Phase II variance*, which captures the error in the normalized difference between $\widehat{\theta}_N$ and $\tilde{\theta}_N$. Here Var_j denotes the within stratum variance based on $P_{0|j}$. The analogous expression for Bernoulli sampling replaces $\mathrm{Var}_j(\tilde{\ell}_0)$ with $\mathrm{E}_j(\tilde{\ell}_0^{\otimes 2})$, the within stratum second moment. FPSS has a clear advantage if the sampling strata are correlated with $\tilde{\ell}_0(X)$.

3.3. Adjustment of the weights

The asymptotic distribution of the second term on the RHS of equation (3.6) is almost surely the same whether it is considered unconditionally or conditionally given Σ_N, the sigma-field for all the study data both observed and unobserved [10]. Conditional inference, which considers randomness only in the sampling indicators R_i, is called *design based* inference by survey samplers [17]. From this perspective, the asymptotic Phase II variance is the almost sure limit of the design based variance of the IPW estimator of the unknown *finite population total* $\tilde{\ell}_{\mathrm{Tot}} = \sum_{i=1}^N \tilde{\ell}_0(X_i)$. This insight provides the key to using sample survey methods to improve estimation efficiency through adjustment of the weights so as to bring in more of the Phase I information.

One approach is calibration [14] of the *design weights* $d_i = 1/\pi_i$ using a q-vector $C = C(V)$ of calibration variables that are correlated with $\tilde{\ell}_0$. New weights $w_i = g_i d_i$ are selected to be as close as possible to the d_i in terms of a distance measure $G(w, d)$ and yet to exactly estimate the Phase I totals of the C_i. Mathematically the problem is to minimize, as a function of the design weight multipliers g_i, the sum $\sum_{i=1}^N R_i G(w_i, d_i)$ subject to constraints

$$\sum_{i=1}^N R_i w_i C_i = \sum_{i=1}^N C_i \tag{3.10}$$

known as the *calibration equations*. The optimization problem involves a q-vector $\lambda = \hat{\lambda}_N$ of Lagrange multipliers for the constraints (3.10). Choosing $G(w, d) = (w - d)^2/2d$, one finds $g_i = 1 - \hat{\lambda}_N^{\mathrm{T}} C_i$ in a procedure known as *least squares* calibration. Choosing the Poisson deviance $G(w, d) = w \log(w/d) - w + d$ for the distance measure, in a procedure known as *raking*, yields $g_i = \exp(-\hat{\lambda}_N^{\mathrm{T}} C_i)$ so the weights $w_i = g_i d_i$ are always positive. Under standard regularity conditions for design based inference, and mild conditions on the distance measure G, Deville and Särndal [14] showed that solutions to the optimization problem satisfied

$$\widehat{\lambda}_N = \widehat{D}_N^{-1}\big(\mathbb{P}_N^{\pi} - \mathbb{P}_N\big)C + O_p\big(n^{-1}\big), \qquad \text{where} \tag{3.11}$$

$$\frac{1}{N}\widehat{D}_N = \mathbb{P}_N^{\pi} C C^{\mathrm{T}} = P_0(CC^{\mathrm{T}}) + o_p(1), \tag{3.12}$$

whatever G was chosen.

In [8] we used these results, together with (3.2) and [11, Thm. 1], to derive the asymptotic distribution for $\widehat{\theta}_N(\widehat{\lambda}_N)$, the estimator obtained using calibrated weights in place of design weights. The first step was to write the conclusion of [11, Thm 1] in the form

$$\begin{aligned} &-I_0\sqrt{N}\big(\widehat{\theta}_N(\widehat{\lambda}_N) - \theta_0\big) - \sqrt{N}\big(\widehat{\eta}_N(\widehat{\lambda}_N) - \eta_0\big)B_0^*\dot{\ell}_0 \\ &\quad = -I_0\sqrt{N}\big(\widehat{\theta}_N(0) - \theta_0\big) - \sqrt{N}\big(\widehat{\eta}_N(0) - \eta_0\big)B_0^*\dot{\ell}_0 \\ &\qquad + P_0(\dot{\ell}_0 C^{\mathrm{T}})\sqrt{N}\widehat{\lambda}_N + o_p(1) \end{aligned} \tag{3.13}$$

and

$$
\begin{aligned}
&-P_0\big[(B_0h)\dot{\ell}_0^T\big]\sqrt{N}\big(\widehat{\theta}_N(\widehat{\lambda}_N)-\theta_0\big)-\sqrt{N}\big(\widehat{\eta}_N(\widehat{\lambda}_N)-\eta_0\big)B_0^*B_0h\\
&\quad=-P_0\big[(B_0h)\dot{\ell}_0^T\big]\sqrt{N}\big(\widehat{\theta}_N(0)-\theta_0\big)-\sqrt{N}\big(\widehat{\eta}_N(0)-\eta_0\big)B_0^*B_0h\\
&\qquad+P_0\big[(B_0h)C^{\mathrm{T}}\big]\sqrt{N}\widehat{\lambda}_N+o_p(1).
\end{aligned}
\tag{3.14}
$$

From (3.11), (3.12) and the fact that $n\uparrow\infty$ faster than $\sqrt{N}$ we concluded

$$
\sqrt{N}\widehat{\lambda}_N=\big(\mathbb{G}_N^{\pi}-\mathbb{G}_N\big)\big[P_0(CC^{\mathrm{T}})\big]^{-1}C+o_p(1).
\tag{3.15}
$$

Choosing $h=(B_0^*B_0)^{-1}B_0^*\dot{\ell}_0$ in (3.14) and subtracting this equation from (3.13) yielded

$$
\sqrt{N}\big(\widehat{\theta}_N(\widehat{\lambda}_N)-\theta_0\big)=\sqrt{N}\big(\widehat{\theta}_N(0)-\theta_0\big)-P_0(\tilde{\ell}_0C^{\mathrm{T}})\sqrt{N}\widehat{\lambda}_N+o_p(1).
\tag{3.16}
$$

Combining (3.6), (4.7) and (3.16) led to the conclusion that $\sqrt{N}(\widehat{\theta}_N(\widehat{\lambda}_N)-\theta_0)$ had a limiting mean zero Gaussian distribution with variance

$$
\mathrm{Var}_{\mathrm{A}}\sqrt{N}\big(\widehat{\theta}_N(\widehat{\lambda}_N)-\theta_0\big)=\tilde{\mathcal{I}}_0^{-1}+\sum_{j=1}^{J}\nu_j\frac{1-p_j}{p_j}\mathrm{Var}_j(\tilde{\ell}_0-QC),
\tag{3.17}
$$

where $QC=P_0(\tilde{\ell}_0C^{\mathrm{T}})(P_0CC^{\mathrm{T}})^{-1}C$ is the projection in $L_2(P_0)$ of each component of $\tilde{\ell}_0$ onto the linear subspace spanned by components of C. Under Bernoulli sampling the optimal choice for C is $C^{\mathrm{opt}}=\mathrm{E}(\tilde{\ell}_0|V)$. Estimators $\widehat{\theta}_N(\widehat{\lambda}_N)$ with weights calibrated to C^{opt} have asymptotic variance equal to that of the optimal member of the class of augmented IPW estimators considered by Robins *et al.* [22] and others [8].

3.4. Applications to Cox regression

For Cox's [12] model $X=(T,\Delta,Z)$, where T is the observed failure time, Δ a censoring indicator and Z a vector of covariates. The Euclidean parameter θ is the vector of regression coefficients. The non-Euclidean parameter η consists of the baseline hazard function Λ, the conditional (given Z) distribution of censoring times and the marginal distribution of Z [19, 24]. Efficient estimation of θ with incomplete data is seriously complicated by the presence of the three infinite dimensional parameters. With complete data, however, the likelihood factors so that ML estimation of (θ,Λ) need not consider the censoring and covariate distributions [25, §25.12.1]. Since our interest is in IPW versions of the standard ML equations, we follow this latter approach.

Let $N(t)=\Delta\cdot\mathbf{1}[T\le t]$ and $Y(t)=\mathbf{1}[T\ge t]$ denote counting and "at risk" processes for $t\in[0,\tau]$ and let

$$
M(t)=N(t)-\int_0^t e^{Z^{\mathrm{T}}\theta_0}Y(s)\,d\Lambda_0(s)
\tag{3.18}
$$

denote the usual martingale process [1, §2]. Define $S_0^{(0)}(t)=P_0(e^{Z^{\mathrm{T}}\theta_0}Y(t))$, $S_0^{(1)}(t)=P_0(Ze^{Z^{\mathrm{T}}\theta_0}Y(t))$ and $m(t)=S_0^{(1)}/S_0^{(0)}(t)=P_0(Z|T=t,\Delta=1)$. Cox

regression admits simple, explicit expressions for the scores:

$$\dot{\ell}_0(X) = \Delta Z - Z e^{Z^{\mathrm{T}}\theta_0}\Lambda_0(T) = \int_0^\tau Z\, dM, \tag{3.19}$$

$$B_0 h(X) = \Delta h(T) - e^{Z^{\mathrm{T}}\theta_0}\int_0^T h\, d\Lambda_0 = \int_0^\tau h\, dM \qquad \forall h \in \mathcal{H}, \tag{3.20}$$

where $\mathcal{H} = \mathrm{BV}[0,\tau]$ is the set of bounded functions of bounded variation on $[0,\tau]$, corresponding to one-dimensional submodels of the form $d\Lambda_t = (1+ht)\,d\Lambda$. Solution of the IPW likelihood equations (2.1, 2.2) leads to IPW versions of the Cox [13] "partial likelihood" equations for θ and the "Breslow" [5] estimator of the baseline hazard [10].

van der Vaart [25, §25.12] derived the adjoint operator evaluated at the θ score, the information operator for continuous Λ_0 and its inverse as

$$B_0^*\dot{\ell}_0 = S_0^{(1)},\ B_0^* B_0 h = h S_0^{(0)} \quad \text{and} \quad \bigl(B_0^* B_0\bigr)^{-1} h = h/S_0^{(0)} \tag{3.21}$$

and thus obtained the efficient score and information

$$\ell_0^* = \bigl[I - B_0\bigl(B_0^* B_0\bigr)^{-1} B_0^*\bigr]\dot{\ell}_0 = \int_0^\tau \bigl[Z - m(t)\bigr]\, dM(t), \tag{3.22}$$

$$\tilde{\mathcal{I}}_0 = P_0\bigl(\ell_0^* \ell_0^{*\mathrm{T}}\bigr) = P_0 e^{Z^{\mathrm{T}}\theta_0}\int_0^\tau \bigl[Z - m(t)\bigr]^{\otimes 2} Y(t)\, d\Lambda_0(t) \tag{3.23}$$

which implied a limiting distribution for $\tilde{\theta}_N$ in agreement with Cox [3]. The asymptotic variances for $\widehat{\theta}_N$ and $\widehat{\theta}_N(\widehat{\lambda}_N)$ under FPSS are obtained by using these expressions to replace $\tilde{\mathcal{I}}_0$ and $\tilde{\ell}_0 = \tilde{\mathcal{I}}_0^{-1}\ell_0^*$ in formulas (3.9) and (3.17), respectively.

4. IPW estimation of the non-Euclidean parameter

4.1. Asymptotic distribution of $\widehat{\eta}_N$ and $\widehat{\eta}_N(\widehat{\lambda}_N)$

Define the operator $A : \mathcal{H} \mapsto L_2(P_0)$ by

$$Ah = B_0\bigl(B_0^* B_0\bigr)^{-1} h - P_0\bigl[B_0\bigl(B_0^* B_0\bigr)^{-1} h\dot{\ell}_0^{\mathrm{T}}\bigr]\tilde{\ell}_0. \tag{4.1}$$

Substituting $(B_0^* B_0)^{-1}h$ for h in (3.4), using (3.5) and rearranging we have

$$\sqrt{N}(\widehat{\eta}_N - \eta_0)h = \mathbb{G}_N^\pi Ah + o_p(1). \tag{4.2}$$

This explicit expansion for $\hat{\eta}_N$ is apparently a new result, even for the case of simple random sampling where $\mathbb{G}_N$ is substituted for $\mathbb{G}_N^\pi$.

Again substituting $(B_0^* B_0)^{-1}h$ for h in (3.14), using (3.16) and rearranging yields another expansion for the estimator obtained using calibrated weights:

$$\begin{aligned}&\sqrt{N}\bigl(\widehat{\eta}_N(\widehat{\lambda}_N) - \eta_0\bigr)h \\ &\quad = \sqrt{N}\bigl[\widehat{\eta}_N(0) - \eta_0\bigr]h + P_0\bigl[B_0\bigl(B_0^* B_0\bigr)^{-1} h\dot{\ell}_0^{\mathrm{T}}\bigr]\sqrt{N}\bigl(\widehat{\theta}_N(0) - \widehat{\theta}_N(\widehat{\lambda}_N)\bigr) \\ &\qquad - P_0\bigl[B_0\bigl(B_0^* B_0\bigr)^{-1} h C^{\mathrm{T}}\bigr]\sqrt{N}\widehat{\lambda}_N + o_p(1).\end{aligned} \tag{4.3}$$

Writing (3.16) in the form

$$\sqrt{N}\bigl(\widehat{\theta}_N(0) - \widehat{\theta}_N(\widehat{\lambda}_N)\bigr) = P_0(\tilde{\ell}_0 C^{\mathrm{T}})\sqrt{N}\widehat{\lambda}_N + o_p(1),$$

and using (3.15) and (4.2) we find

$$\begin{aligned}(4.4)\qquad &\sqrt{N}\big(\widehat{\eta}_N(\widehat{\lambda}_N)-\eta_0\big)h\\ &\quad=\mathbb{G}_N Ah+\big(\mathbb{G}_N^{\pi}-\mathbb{G}_N\big)\big\{Ah-P_0(AhC^{\mathrm{T}})\big[P_0(CC^{\mathrm{T}})\big]^{-1}C\big\}+o_p(1).\end{aligned}$$

It follows that $\sqrt{N}(\widehat{\eta}_N(\widehat{\lambda}_N)-\eta_0)$ has a limiting mean zero Gaussian distribution indexed by $h\in\mathcal{H}$ such that

$$\begin{aligned}(4.5)\qquad &\mathrm{Var}_{\mathrm{A}}\sqrt{N}\big(\widehat{\eta}_N(\widehat{\lambda}_N)-\eta_0\big)h\\ &\quad=\mathrm{Var}_0(Ah)+\sum_{j=1}^{J}\nu_j\frac{1-p_j}{p_j}\mathrm{Var}_j\big[Ah-\Pi(Ah|C)\big]\end{aligned}$$

where Var_0 is the variance under P_0 and $\Pi(Ah|C)$ denotes the least squares projection of Ah onto the linear subspace spanned by the calibration variables C.

4.2. Application to Cox regression

From (3.21) we find for $h\in\mathcal{H}=\mathrm{BV}[0,\tau]$

$$(4.6)\qquad Ah=\int_0^{\tau}\frac{h}{S_0^{(0)}}\,dM-P_0\bigg(\int\frac{h}{S_0^{(0)}}\,dM\dot{\ell}_0^{\mathrm{T}}\bigg)\tilde{\ell}_0,$$

where the two terms on the RHS are uncorrelated by construction: ℓ_0^* is the least squares projection of $\dot{\ell}_0$ on the orthogonal complement of the *nuisance tangent space*, which is readily seen from (3.20) to equal $\overline{\{\int_0^{\tau}h\,dM:h\in\mathrm{BV}[0,\tau]\}}$.

4.2.1. Results for complete data

Substituting $\mathbb{G}_N$ for $\mathbb{G}_N^{\pi}$ in (4.2), we find the asymptotic distribution for the ML estimator from

$$(4.7)\qquad \sqrt{N}(\tilde{\Lambda}_N-\Lambda_0)h=\mathbb{G}_N\bigg[\int_0^{\tau}\frac{h}{S_0^{(0)}}\,dM-P_0\bigg(\int_0^{\tau}\frac{h}{S_0^{(0)}}\,dM\dot{\ell}_0^{\mathrm{T}}\bigg)\tilde{\ell}_0\bigg]+o_p(1).$$

Set $h_t=\mathbf{1}[0,t]$ and let $\langle M\rangle(t)=e^{Z^{\mathrm{T}}\theta_0}\int_0^t Y\,d\Lambda_0$ denote the predictable variation process for the counting process martingale M [1, Eq. 2.43]. Using standard results for the predictable covariation processes of martingale integrals [1, Eq. 2.31], we conclude that the limiting process $\mathbb{Z}_*=\mathbb{Z}_*(t)$ of $\sqrt{N}(\tilde{\Lambda}_N-\Lambda_0)(t)$ is the sum of a mean zero Gaussian process $\mathbb{Z}=\mathbb{Z}(t)$, with

$$\mathrm{Cov}\big(\mathbb{Z}(t),\mathbb{Z}(s)\big)=P_0\int_0^{\tau}\frac{h_th_s}{[S_0^{(0)}]^2}\,d\langle M\rangle=\int_0^{t\wedge s}\frac{1}{S_0^{(0)}}\,d\Lambda_0,$$

and an independent mean zero p-dimensional Gaussian variable Z_*, with covariance equal to $\tilde{I}_0^{-1}$, premultiplied by the function

$$P_0\bigg(\int_0^t\frac{1}{S_0^{(0)}}\,dM\int_0^{\tau}Z\,dM\bigg)=\int_0^t\frac{S_0^{(1)}}{S_0^{(0)}}\,d\Lambda_0=\int_0^t m\,d\Lambda_0,$$

in agreement with Begun *et al.* [3, p. 450]. In the special case $\theta=0$, so that the survival times are sampled from a homogeneous population, $S_0^{(0)}=P_0Y$ and

the baseline cumulative hazard estimator reduces to the Nelson-Aalen estimator [1, p. 72]. The second term in square brackets in (4.7) disappears and $\sqrt{N}(\tilde{\Lambda}_N - \Lambda_0)(t)$ converges to a mean zero Gaussian process with independent increments and variance function

$$\operatorname{Var}_{\mathrm{A}} \sqrt{N}(\tilde{\Lambda}_N - \Lambda_0)(t) = \operatorname{Var}\left(\int_0^t \frac{1}{P_0 Y}\, dM \right) = \int_0^t \frac{1}{P_0 Y}\, d\Lambda_0$$

as established by Breslow and Crowley [6, Theorem 4].

4.2.2. Results for finite population stratified sampling

Asymptotic properties of $\widehat{\Lambda}_N$ under FPSS follow by substituting (4.6) for Ah in (4.2) and using the basic weak convergence result (3.1). Deriving explicit expressions for the variances of the limiting Gaussian process when $h = h_t = \mathbf{1}[0,t]$ is more complicated than for complete data since the random variables $\tilde{\ell}_0$ and $\int_0^t (h/S_0^{(0)})\, dM$ need not be uncorrelated under $P_{0|j}$ as they are for P_0.

4.3. Limit law for estimator of individual hazards

The limiting distribution for the estimated cumulative hazard function $e^{z_0 \widehat{\theta}_N} \widehat{\Lambda}_N(t)$ for a subject with covariates z_0 may be derived from equations (3.5) and (4.2) by inserting the expressions shown earlier (3.22), (3.23), (4.6) for the efficient influence function $\tilde{\ell}_0 = \tilde{\mathcal{I}}^{-1} \ell_0^*$ and the operator A under Cox regression, and using the delta method. Suppressing the subscripts N, we have

$$\begin{aligned}
&\sqrt{N}\big[e^{z_0\widehat{\theta}}\widehat{\Lambda}(t) - e^{z_0\theta_0}\Lambda_0(t)\big] \\
&\quad = \sqrt{N}\big[e^{z_0\theta_0}(\widehat{\Lambda} - \Lambda_0)(t) + \big(e^{z_0\widehat{\theta}} - e^{z_0\theta_0}\big)\Lambda_0(t)\big] \\
&\qquad + \sqrt{N}\big[\big(e^{z_0\widehat{\theta}} - e^{z_0\theta_0}\big)(\widehat{\Lambda} - \Lambda_0)(t)\big] \\
&\quad = \sqrt{N}\big[e^{z_0\theta_0}(\widehat{\Lambda} - \Lambda_0)(t) + e^{z_0\theta^*} z_0(\widehat{\theta} - \theta_0)\Lambda_0(t)\big] + o_p(1) \\
&\quad = e^{z_0\theta_0}\sqrt{N}\big[(\widehat{\Lambda} - \Lambda_0)(t) + z_0(\widehat{\theta} - \theta_0)\Lambda_0(t)\big] + o_p(1) \\
&\quad = e^{z_0\theta_0}\mathbb{G}_N^{\pi}\left[\int_0^t \frac{dM}{S_0^{(0)}} + \left(z_0\Lambda_0(t) - \int_0^t m\, d\Lambda_0\right)\tilde{\ell}_0\right] + o_p(1),
\end{aligned} \tag{4.8}$$

where in the third line $|\theta^* - \theta_0| \le |\widehat{\theta} - \theta_0|$. For complete data, substituting $\mathbb{G}_N$ for $\mathbb{G}_N^{\pi}$ and again exploiting the orthogonality of the two terms within square brackets, this converges to the Gaussian process

$$e^{z_0\theta_0}\left[\mathbb{Z}(t) + \int_0^t (z_0 - m)\, d\Lambda_0 \cdot Z_*\right].$$

Compare with Begun *et al.* [3, p. 451], who worked with the survival function instead of the cumulative hazard and thus had an additional term $\exp(-e^{z_0^{\mathrm{T}}\theta_0}\Lambda_0(t))$ multiplying this expression.

The expansion (4.8) suggests that, for estimation of the cumulative hazard at time t, we take as an additional calibration variable $\mathrm{E}(\int_0^t dM/S_0^{(0)}|V)$.

5. Simulations

In view of (3.17) and the ensuing discussion, a good choice for the calibration variables C for θ estimation would be an approximation to $C^{\rm opt} = {\rm E}(\tilde{\ell}_0|V)$. In [8], following Kulich and Lin [16], we proposed an approximation that involved five steps: (i) fitting a parametric model using IPW to the Phase II data to predict each of the partially missing components of X from V; (ii) imputing values for the partially missing components of X for all Phase I subjects using the prediction model; (iii) fitting the main model $P_{\theta,\eta}(X)$ to the Phase I subjects using the imputed data; (iv) taking for the C_i the estimated influence function contributions that are routinely supplied by standard programs; and (v) using the C_i to calibrate the weights for IPW fitting of the model $P_{\theta,\eta}$ using (2.1) and (2.2). An example of a *model assisted* survey sampling technique [23], the prediction model in (i) need not be correct for the procedure to yield valid inferences.

Simulations to assess the improvement in θ estimation with calibration or estimation of the weights were reported in [8]. These results are extended here to estimation of Λ and of survival probabilities. Briefly, a Cox model for time to relapse (or death without relapse) as a function of unfavorable (UH) vs. favorable (FH) histologic type, age at diagnosis (yr.), stage (III/IV vs. I/II) and tumor diameter (cm.) was fitted to data for 3,915 NWTS patients.[2] Histologic type as evaluated by central pathology was treated as the partially missing variable, with histologic type as evaluated by the patient's institution, and thus known for all cohort members, used as a surrogate. Sixteen strata were defined based on outcome (relapsed cases vs. controls), institutional histology (UH vs. FH), stage and age ($<$ vs. ≥ 1). All cases and all those with UH were sampled at 100%. Among FH controls the Phase II sample contained 120/452 of those with stage I/II disease < 1 year, 160/1,620 of those with stage I/II disease ≥ 1 year, all 40 of those with stage III/IV disease $<$ 1 year and 208/914 of those with stage III/IV disease ≥ 1 year, for a total Phase II sample size of 1,329 patients.

Ten thousand Phase II samples were drawn and each used first to fit a logistic regression model by IPW to estimate UH (central pathology) with UH (institutional histology) and other Phase I variables as predictors. This model was used to impute UH (central pathology) for all Phase I subjects and the Cox model was first fitted using the imputed data. Calibration variables included the estimated influence function contributions ("delta-betas") described above to approximate $C^{\rm opt} = {\rm E}(\tilde{\ell}_0|V)$. A further calibration variable

$$\int_0^t \frac{d\widehat{M}_i}{\widehat{S}_0^{(0)}} = \frac{\Delta_i \mathbf{1}[T_i \leq t]}{\widehat{S}_0^{(0)}(T_i)} - e^{Z_i^{\rm T}\widehat{\theta}} \int_0^{t\wedge T_i} \frac{d\widehat{\Lambda}}{\widehat{S}_0^{(0)}},$$

where $\widehat{\theta}$, $\widehat{\Lambda}$ and $\widehat{S}_0^{(0)}$ denote quantities estimated from the imputed data fit, was employed at each time $t = 1, 2, 5, 10$ to improve the efficiency of estimation of $\Lambda(t)$ as suggested following (4.8). Finally, the Cox model was fitted to the two-phase data using IPW score equations with either standard (design) or adjusted (calibrated and estimated) weights. We ignored the additional variability stemming from the use of Phase II data to construct the calibration variables. Further work is needed to determine if this actually increases the asymptotic variance.

Figure 1 shows survival curves estimated from the Cox model fit to the entire cohort for four covariate configurations: A) FH, age = 1, stage = I/II, diameter =

[2]Data available at http://faculty.washington.edu/norm/software.html

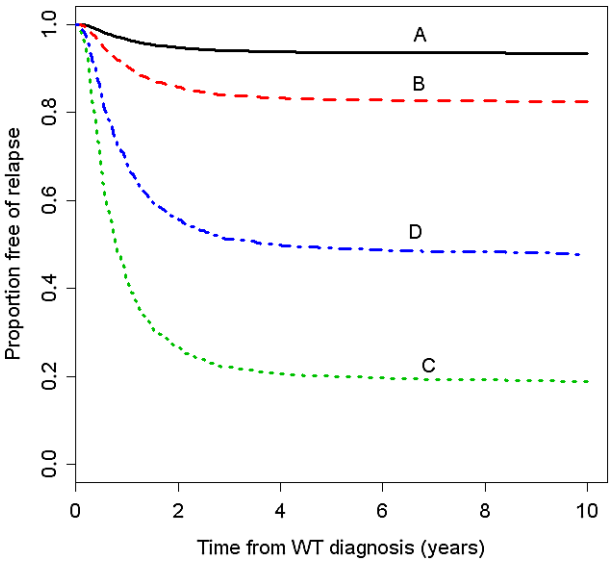

FIG 1. *Predicted relapse-free survival curves (complete data).*

8; B) FH, age = 4, stage = III/IV, diameter = 10; C) UH, age = 0.5, stage = I/III, diameter = 10; and D) UH, age = 7, stage = III/IV, diameter = 16. Averages of the curves estimated by IPW from the 10,000 replications of the two-phase sampling design were virtually identical to those shown except for configuration D. Here the estimated percentages of relapse-free survival at 1, 2, 5 and 10 years past diagnosis (Dx) were 68.2, 55.7, 49.2 and 47.8 for the fit to the entire cohort, but only 67.1, 54.4, 47.8 and 46.4 for the averages based on standard weights. Averages for adjusted weights were within 0.2 percentage points of those for standard weights.

Table 1 shows the root mean squared errors (RMSE) of estimation using the two-phase design of the survival probabilities shown in Figure 1. These are empirical Phase II standard deviations. Adjustment of the weights reduced the Phase II error, with the relative gains substantial for Configurations A and B.

6. Discussion

This paper has used tools of semiparametric inference and empirical processes, as developed over the years by Jon Wellner, his students and his colleagues, to at-

TABLE 1
Phase II RMSE of estimation of relapse-free survival probabilities

	Configuration A				Configuration B			
Years from Dx:	1	2	5	10	1	2	5	10
Standard	0.24	0.36	0.43	0.44	0.39	0.56	0.65	0.67
Calibrated	0.10	0.14	0.17	0.18	0.26	0.34	0.40	0.41
Estimated	0.13	0.20	0.23	0.24	0.24	0.33	0.38	0.39
	Configuration C				Configuration D			
Years from Dx:	1	2	5	10	1	2	5	10
Standard	2.76	2.71	2.51	2.45	4.71	5.85	6.24	6.31
Calibrated	2.50	2.44	2.26	2.21	4.49	5.55	5.93	5.99
Estimated	2.47	2.44	2.25	2.20	4.58	5.70	6.09	6.15

tempt to solve a problem of substantial practical interest and importance: how to improve estimation of "survival" probabilities with data from two-phase stratified sampling designs that are increasingly used in epidemiology and clinical medicine. Little attention was paid to assumptions needed to justify the formal calculations and substantial work will be needed to clearly delineate the boundaries of application. We relied heavily on van der Vaart's [25, §25.12.1] treatment of the Cox model, which involved several "partly unnecessary" assumptions including fixed, bounded covariates and restrictions on censoring. Assumptions needed for the Z-estimation theorem with estimated nuisance parameters [11] likewise include bounded calibration variables and covariates. We anticipate that the results can be shown to be valid under conditions such as those imposed by Andersen and Gill [2] in their classic treatment of the Cox model with time-dependent covariates. We are hopeful that the basic weak convergence result (3.1) can be extended from FPSS to other complex sampling designs. In addition to the theoretical developments, extensive simulation studies, ideally based on actual data such as those conducted in §5, will be needed to engender confidence in the proposed methods.

References

[1] Aalen, O. O., Borgan, O. and Gjessing, H. K. (2008). *Survival and Event History Analysis.* Springer, New York.

[2] Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics* **10** 1100–1120.

[3] Begun, J. M., Hall, W. J., Huang, W.-M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Annals of Statistics* **11** 452-452.

[4] Borgan, O., Langholz, B., Samuelsen, S. O., Goldstein, L. and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Analysis* **6** 39–58.

[5] Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics* **30** 89–99.

[6] Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics* **2** 437–453.

[7] Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association* **91** 14–28.

[8] Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E. and Kulich, M. (2009). Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistics in Biosciences* **1** 32–49.

[9] Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E. and Kulich, M. (2009). Using the whole cohort in the analysis of case-cohort Data. *American Journal of Epidemiology* **169** 1398–1405.

[10] Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics* **34** 86–102.

[11] Breslow, N. E. and Wellner, J. A. (2008). 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". *Scandinavian Journal of Statistics* **35** 186–192.

[12] Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society (Series B)* **34** 187–220.
[13] Cox, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276.
[14] Deville, J. C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87** 376–382.
[15] Kalbfleisch, J. D. and Lawless, J. F. (1988). Likelihood analysis of multistate models for disease incidence and mortality. *Statistics in Medicine* **7** 149–160.
[16] Kulich, M. and Lin, D. Y. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association* **99** 832–844.
[17] Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. John Wiley & Sons, Hoboken, New Jersey.
[18] Manski, C. F. and Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica* **45** 1977–1988.
[19] Nan, B., Emond, M. and Wellner, J. A. (2004). Information bounds for Cox regression models with missing data. *Annals of Statistics* **32** 723–753.
[20] Præstgaard, J. and Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Annals of Probability* **21** 2053–2086.
[21] Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73** 1–11.
[22] Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89** 846–866.
[23] Särndal, C. E., Swensson, B. and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer, New York.
[24] Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.
[25] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.