doi: 10.1111/j.1467-9469.2006.00523.x

Weighted Likelihood for Semiparametric Models and Two-phase Stratified Samples, with Application to Cox Regression

NORMAN E. BRESLOW

Department of Biostatistics, University of Washington

JON A. WELLNER

Departments of Statistics and Biostatistics, University of Washington

ABSTRACT. We consider semiparametric models for which solution of Horvitz–Thompson or inverse probability weighted (IPW) likelihood equations with two-phase stratified samples leads to $\sqrt{N}$ consistent and asymptotically Gaussian estimators of both Euclidean and non-parametric parameters. For Bernoulli (independent and identically distributed) sampling, standard theory shows that the Euclidean parameter estimator is asymptotically linear in the IPW influence function. By proving weak convergence of the IPW empirical process, and borrowing results on weighted bootstrap empirical processes, we derive a parallel asymptotic expansion for finite population stratified sampling. Several of our key results have been derived already for Cox regression with stratified case–cohort and more general survey designs. This paper is intended to help interpret this previous work and to pave the way towards a general Horvitz–Thompson approach to semiparametric inference with data from complex probability samples.

Key words: case–cohort, estimated weights, failure time, inverse probability weights, missing data

1. Introduction

Two-phase stratified sampling, also known as double sampling, was introduced by Neyman (1938) to estimate the population mean of a target variable that is costly or difficult to measure. At phase 1 a relatively large random sample is drawn and measurements are made on an auxiliary variable that is correlated with the target variable but easier to measure. At phase 2 measurements on the target variable are made for a subsample drawn randomly, without replacement, from within strata defined by the auxiliary variable. Neyman showed that the optimal, design unbiased linear estimator of the population mean is the Horvitz & Thompson (1952) estimator that weights each observation by the inverse of the probability of its inclusion in the phase 2 sample.

Two-phase stratified sampling designs can dramatically reduce the costs of regression modelling when the strata depend on (correlates of) both outcome and explanatory variables. A common method of estimation is 'weighted exogenous sampling maximum likelihood', here simply weighted likelihood or WL, in which one maximizes the inverse probability weighted (IPW) sum of log-likelihood contributions from the phase 2 observations (Manski & Lerman, 1977; Kalbfleisch & Lawless, 1988). Equivalently, one may solve an IPW version of the score equations (Skinner *et al.*, 1989, section 3.4). Although easy to implement, WL estimators are sometimes seriously inefficient (Robins *et al.*, 1994). Survey statisticians may still advocate their use, however, because even when the model is wrong they consistently estimate the finite population parameters that would be obtained by fitting the model to complete phase

1 data (Xie & Manski, 1989; Binder, 1992). Fully efficient estimators are available for logistic and other parametric regression models in situations where the phase 1 data consist only of stratum frequencies; see, e.g. Breslow *et al.*, 2003 and the references cited therein.

The asymptotic properties of WL estimators of Euclidean parameters in parametric models follow readily from standard results for M- and Z-estimators (van der Vaart, 1998, chapter 5). WL may also be used for estimation of both Euclidean and infinite dimensional parameters in semiparametric models, for which the paradigm is Cox (1972) proportional hazards regression. Lin (2000) developed asymptotic results for both regression coefficients and baseline cumulative hazard when fitting the Cox model to survey data including those obtained using two-phase sampling. Borgan *et al.* (2000) obtained the same results for the regression parameters when fitting the Cox model to data from exposure stratified case–cohort studies, in which all subjects who have a failure event (the cases) are sampled at phase 2. One purpose of this paper is to develop a modern theory of WL estimation in semiparametric models that encompasses these previous results, helps to interpret them and paves the way towards further applications. We also explore the relationship between results based on finite population stratified sampling at phase 2 and those based on independent and identically distributed (i.i.d.) variable probability sampling with sampling weights estimated using information from phase 1.

2. Notation, assumptions and problem statement

Suppose $P_{\theta,\eta}$ denotes a probability distribution in a semiparametric model for a random variable $X \in \mathcal{X}$, where $\theta \in \Theta \subset \mathbb{R}^p$ is the Euclidean parameter and η, taking values in a subset H of some Banach space $\mathcal{B}$, is the non-parametric one. Let $P_0 = P_{\theta_0,\eta_0}$ denote the distribution from which X is actually sampled. Following closely section 25.12 of van der Vaart (1998), suppose maximum likelihood (ML) estimators $(\hat{\theta}, \hat{\eta})$ are obtained by solving the system

$$\begin{aligned} &\Psi_{N1}(\theta,\eta) = \mathbb{P}_N \dot{\ell}_{\theta,\eta} = 0 \\ &\Psi_{N2}(\theta,\eta)h = \mathbb{P}_N B_{\theta,\eta} h - P_{\theta,\eta} B_{\theta,\eta} h = 0 \quad \forall\, h \in \mathcal{H}. \end{aligned} \tag{1}$$

Here $\dot{\ell}_{\theta,\eta}$ is the p-dimensional likelihood score for θ, $B_{\theta,\eta}$ is the score operator (Begun *et al.*, 1983) working on an infinite dimensional class $\mathcal{H}$ of directions h from which paths of one-dimensional submodels for η may approach η_0, and $\mathbb{P}_N$ is the empirical measure based on the i.i.d. sequence $X_1, \ldots, X_N$. Set $\dot{\ell}_0 = \dot{\ell}_{\theta_0,\eta_0}$ and $B_0 = B_{\theta_0,\eta_0}$. Often $\mathcal{H}$ is selected to be the unit ball in $\mathcal{B}$.

Suppose the following assumptions, which slightly strengthen the hypotheses of van der Vaart (1998, theorem 25.90), are satisfied so that $\sqrt{N}(\hat{\theta} - \theta_0, \hat{\eta} - \eta_0)$ is asymptotically Gaussian:

A1 for (θ,η) in a δ-neighbourhood of (θ_0,η_0) the functions $\dot{\ell}_{\theta,\eta}$ and $\{B_{\theta,\eta}h, h \in \mathcal{H}\}$ are contained in a P_0-Donsker class $\mathcal{F}$;

A2 $P_0\|\dot{\ell}_{\theta,\eta} - \dot{\ell}_0\|^2$ and $\sup_{h\in\mathcal{H}} P_0|B_{\theta,\eta}h - B_0 h|^2$ converge to 0 as $(\theta,\eta) \to (\theta_0,\eta_0)$;

A3 the map $\Psi = (\Psi_1, \Psi_2) : \Theta \times H \mapsto \mathbb{R}^p \times \ell^\infty(\mathcal{H})$ with components

$$\begin{aligned} &\Psi_1(\theta,\eta) = P_0 \dot{\ell}_{\theta,\eta} \\ &\Psi_2(\theta,\eta)h = P_0 B_{\theta,\eta} h - P_{\theta,\eta} B_{\theta,\eta} h, \quad h \in \mathcal{H}, \end{aligned} \tag{2}$$

which is the expectation of the random map $\Psi_N = (\Psi_{N1}, \Psi_{N2})$ in (1), has a Fréchet derivative $\dot{\Psi}_0$ at (θ_0,η_0) that is continuously invertible on its range.

A4 $(\hat{\theta}, \hat{\eta})$ is consistent for (θ_0,η_0) and satisfies $\Psi_N(\hat{\theta}, \hat{\eta}) = 0$.

Assumption A3 is typically established by showing that the information operator $B_0^* B_0$ is continuously invertible and thus that η is estimable at a $\sqrt{N}$ rate. This is the most restrictive assumption, but one that leads quickly to our main result.

With two-phase sampling, however, X is not observed for all N subjects. At phase 1 we observe only a coarsening $\tilde{X} = \tilde{X}(X)$ of X plus auxiliary variables $U \in \mathcal{U}$ that serve to determine the sampling strata. X is fully observed for subjects sampled at phase 2. Let $W = (X, U) \in \mathcal{W} = \mathcal{X} \times \mathcal{U}$ denote the variables potentially available for everyone, but in fact fully observed only for those in the phase 2 sample, and $V = (\tilde{X}, U) \in \mathcal{V} = \tilde{\mathcal{X}} \times \mathcal{U}$ denote the variables actually observed for everyone. We write $\tilde{P}_0$ for the distribution of $W = (X, U)$ and denote by $\Sigma_N = \sigma[W_1, \ldots, W_N]$ the sigma field of information, also referred to as the complete data, potentially available for N subjects. A sequence of binary indicators $(\xi_1, \ldots, \xi_N)$ shows which subjects are selected ($\xi_i = 1$) at phase 2 for observation of X_i. We consider two probability models for the indicators ξ_i. In the first, known as Bernoulli or Manski & Lerman (1997) sampling, each phase 1 subject is examined in succession for the value of V_i and the indicator ξ_i is independently generated with $\Pr(\xi_i = 1 \mid W_i) = \Pr(\xi_i = 1 \mid V_i) = \pi_0(V_i)$, where π_0 is a known sampling function. This preserves the i.i.d. structure for the observations $(\xi_i, V_i, \xi_i X_i)$. Note the crucial missing at random (MAR) assumption: π_0 depends only on what is observed at phase 1. We write Q_0 for the distribution of (W_i, ξ_i). If $\mathcal{V}$ is partitioned into J strata $\mathcal{V}_1 \cup \ldots \cup \mathcal{V}_J$, stratified Bernoulli sampling corresponds to the special case where $\pi_0(v) = p_j$ for $v \in \mathcal{V}_j$. We assume that all J strata are sampled with positive probability or more generally that

$$0 < \sigma \leq \pi_0(v) \leq 1 \quad \text{for} \quad v \in \mathcal{V}. \tag{3}$$

Our derivation of the asymptotic properties of the WL estimator for Bernoulli sampling applies to any known sampling function π_0 that satisfies (3), not just to stratified Bernoulli sampling. Even though π_0 is known, however, it is advisable to estimate it using a correct parametric model so as to increase efficiency (Pierce, 1982; Robins *et al.*, 1994). WL estimating equations with known sampling weights only involve data for subjects sampled at phase 2; estimation of the weights allows incorporation of phase 1 data available for all subjects. Provided the weights are estimated efficiently, this increases the efficiency of the WL estimators of the Euclidean parameters of interest (Henmi & Eguchi, 2004). When fitting logistic regression models by WL to phase 2 case–control samples, for example, use of the empirical (estimated) weights leads to fully efficient estimates of odds ratio parameters whereas use of the *a priori* sampling weights may be seriously inefficient (Scott & Wild, 1986). We consider parametric estimation of π_0 in section 6.

The second stratified sampling model corresponds to Neyman's original design and is usually closer to actual practice. Here, we observe the entire phase 1 sample at once and record the stratum frequencies

$$N_j = \sum_{i=1}^{N} \mathbf{1}_{\mathcal{V}_j}(V_i) \quad \text{for } j = 1, \ldots, J.$$

At phase 2 samples of size $n_j \leq N_j$ are drawn at random, without replacement, from each of the J finite phase 1 strata. Using now a doubly subscripted notation where $\xi_{j,i}$ denotes the indicator variable for ith subject in stratum j, the essential features of this design are that, conditionally on Σ_N: (i) for $j = 1, \ldots, J$ the random variables $(\xi_{j1}, \ldots, \xi_{jN_j})$ are exchangeable with $\Pr(\xi_{j,i} = 1 \mid \Sigma_N) = n_j / N_j$; and (ii) the J random vectors $(\xi_{j1}, \ldots, \xi_{jN_j})$ are independent. Our problem is to estimate (θ, η) using the incomplete observations V_i on everyone and the complete observations X_i on subjects sampled at phase 2.

3. Weighted likelihood estimator

WL estimates are obtained by solving Horvitz–Thompson (IPW) versions of the likelihood equations. Define the IPW *empirical measure* 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} denotes Dirac measure placing unit mass on X_i and

$$\pi_i = \begin{cases} \pi_0(V_i) & \text{for Bernoulli sampling} \\ \dfrac{n_j}{N_j} \text{ if } V_i \in \mathcal{V}_j & \text{for finite population stratified sampling.} \end{cases} \tag{5}$$

Then, instead of (1) we solve

$$\begin{aligned} &\Psi_{N1}^{\pi}(\theta,\eta) = \mathbb{P}_N^{\pi}\dot{\ell}_{\theta,\eta} = 0 \\ &\Psi_{N2}^{\pi}(\theta,\eta)h = \mathbb{P}_N^{\pi}B_{\theta,\eta}h - P_{\theta,\eta}B_{\theta,\eta}h = 0 \quad \text{for all } h \in \mathcal{H}. \end{aligned} \tag{6}$$

In view of the MAR assumption, for any integrable function $f : \mathcal{X} \mapsto \mathbb{R}$ and under either Bernoulli or finite population stratified sampling,

$$E\frac{\xi_i}{\pi_i}f(X_i) = E\left[E\left(\frac{\xi_i}{\pi_i}\middle|\Sigma_N\right)f(X_i)\right] = Ef(X_i), \quad i = 1,\ldots,N,$$

so that $E\mathbb{P}_N^{\pi}f = E\mathbb{P}_N f = P_0 f$. Consequently, the random map $\Psi_N^{\pi} = (\Psi_{N1}^{\pi}, \Psi_{N2}^{\pi})$ defined by (6) has the same expectation as the random map Ψ_N in (1), namely $\Psi = (\Psi_1, \Psi_2)$ as in (2). The implication is that the assumptions A1–A4 made to guarantee the asymptotic normality of the ML estimator based on complete phase 1 data are also the assumptions needed to guarantee the asymptotic normality of the WL estimator based on two-phase data.

Indeed, for Bernoulli sampling, van der Vaart's (1998) theorem 25.90, or more precisely his theorem 19.26 of which it is a restatement, applies virtually without change. The Donsker class $\mathcal{F}$ in A1 is modified to $\tilde{\mathcal{F}} = \{[\xi/\pi_0(V)]f(X), f \in \mathcal{F}\}$. As under the hypothesis (3) it is the product of a fixed bounded function with the Donsker class $\mathcal{F}$, the fact that $\tilde{\mathcal{F}}$ is Donsker for the joint distribution Q_0 of (W, ξ) follows from van der Vaart & Wellner (1996, example 2.10.10). The random map Ψ_N corresponding to the estimating functions (6) is the ordinary empirical measure $\mathbb{Q}_N$ for $\{(W_i, \xi_i), i = 1,\ldots,N\}$ applied to the unbiased estimating functions $(\xi/\pi_0)\dot{\ell}_{\theta,\eta}$ and $(\xi/\pi_0)B_{\theta,\eta}h$. A4 will generally follow from (3) and, together with (6), the arguments used to establish consistency for the complete data ML estimator. A2 and A3 are unchanged.

For finite population stratified sampling, however, the more general theorem 3.3.1 of van der Vaart & Wellner (1996) is needed to deal with the non-i.i.d. data. To verify its hypotheses, we must first establish weak convergence of the empirical process based on $\mathbb{P}_N^{\pi}$ for stratified sampling.

4. Weak convergence of the IPW empirical process under finite population stratified sampling

Two-phase stratified sampling resembles the bootstrap in that it involves random sampling from the finite, albeit incompletely observed, population $\{X_1,\ldots,X_N\}$. Here, we use results on weighted bootstrap empirical processes from Præstgaard & Wellner (1993, theorem 2.2), as incorporated in van der Vaart & Wellner (1996, theorem 3.6.13), to demonstrate weak convergence of the IPW empirical process $\mathbb{G}_N^{\pi} = \sqrt{N}(\mathbb{P}_N^{\pi} - P_0)$ for finite population stratified sampling. First note that, with the subscript j, i denoting the ith of N_j observations in stratum j and with π_i defined by the second expression in (5),

$$\mathbb{P}_N^{\pi} = \frac{1}{N}\sum_{j=1}^{J}\frac{N_j}{n_j}\sum_{i=1}^{N_j}\xi_{j,i}\delta_{X_{j,i}} = \frac{1}{N}\sum_{j=1}^{J}\frac{N_j^2}{n_j}\mathbb{P}_{j,N_j}^{\xi}, \tag{7}$$

where

$$\mathbb{P}_{j,N_j}^{\xi} = \frac{1}{N_j}\sum_{i=1}^{N_j}\xi_{j,i}\delta_{X_{j,i}}$$

is a 'finite sampling empirical measure' for the jth stratum. Similarly one can express the ordinary empirical measure as

$$\mathbb{P}_N = \frac{1}{N}\sum_{j=1}^{J}N_j\mathbb{P}_{j,N_j}, \tag{8}$$

where

$$\mathbb{P}_{j,N_j} = \frac{1}{N_j}\sum_{i=1}^{N}\delta_{X_i}\mathbf{1}_{\mathcal{V}_j}(V_i) = \frac{1}{N_j}\sum_{i=1}^{N_j}\delta_{X_{j,i}} \tag{9}$$

denotes the empirical measure for the jth stratum. Justification of the second (doubly indexed) form is given in appendix A.

Combining (7) and (8), and letting $\mathbb{G}_N = \sqrt{N}(\mathbb{P}_N - P_0)$ denote the standard empirical process, we have

$$\begin{aligned}
\mathbb{G}_N^{\pi} &= \sqrt{N}(\mathbb{P}_N^{\pi} - P_0),\\
&= \sqrt{N}(\mathbb{P}_N - P_0) + \sqrt{N}(\mathbb{P}_N^{\pi} - \mathbb{P}_N),\\
&= \mathbb{G}_N + \frac{1}{\sqrt{N}}\sum_{j=1}^{J}\left(\frac{N_j^2}{n_j}\right)\left(\mathbb{P}_{j,N_j}^{\xi} - \frac{n_j}{N_j}\mathbb{P}_{j,N_j}\right),\\
&= \mathbb{G}_N + \sum_{j=1}^{J}\sqrt{\frac{N_j}{N}}\left(\frac{N_j}{n_j}\right)\mathbb{G}_{j,N_J}^{\xi},
\end{aligned} \tag{10}$$

where

$$\mathbb{G}_{j,N_j}^{\xi} = \sqrt{N_j}\left(\mathbb{P}_{j,N_j}^{\xi} - \frac{n_j}{N_j}\mathbb{P}_{j,N_j}\right) \tag{11}$$

is the 'finite sampling empirical process' for stratum j.

The first term in (10) converges to the P_0-Brownian bridge process $\mathbb{G}$ indexed by the Donsker class $\mathcal{F}$ mentioned in A1. Let $P_{0|j}(\cdot) = E(\cdot|V \in \mathcal{V}_j)$ denote $\tilde{P}_0$ conditional on membership in stratum j, i.e. for measurable $A \subset \mathcal{X}$, $P_{0|j}(A) = \tilde{P}_0[A\mathbf{1}_{\mathcal{V}_j}(V)]/v_j$ with $v_j = \tilde{P}_0\mathbf{1}_{\mathcal{V}_j}(V)$, and let $\mathbb{G}_j$ denote the $P_{0|j}$-Brownian bridge, also indexed by $\mathcal{F}$. Our aim was to establish the weak convergence of the remaining terms on the right-hand side of (10). If as $N \to \infty$ the sampling fractions converge with $n_j/N_j \to p_j$, the assumption on the exchangeable 'weights' $(\xi_{j,1}, \ldots, \xi_{j,N_j})$ in equation (3.6.8) of van der Vaart & Wellner (1996) holds trivially with

$$\frac{1}{N_j}\sum_{i=1}^{N_j}(\xi_{j,i} - \bar{\xi}_{j\cdot})^2 \xrightarrow{\mathrm{p}} p_j(1-p_j).$$

Furthermore, with $\rightsquigarrow$ denoting weak convergence in $\ell^{\infty}(\mathcal{F})$, $\sqrt{N_j}(\mathbb{P}_{j,N_j} - P_{0|j}) \rightsquigarrow \mathbb{G}_j$; see appendix B for the proof. Thus their theorems 3.6.13 and 1.12.4 imply that, for almost every sequence of complete data, $\mathbb{G}_{j,N_j}^{\xi} \rightsquigarrow \sqrt{p_j(1-p_j)}\mathbb{G}_j$. Conditionally on Σ_N, the processes $\mathbb{G}_{j,N_j}^{\xi}$ are mutually independent because of the independence of the $\{\xi_{j,i}\}$ in different strata. Furthermore, by virtue of the fact that they also are (unconditionally) uncorrelated with

$\mathbb{G}_N = \sqrt{N}(\mathbb{P}_N - P_0)$, which follows along the lines of van der Vaart & Wellner (1996, corollary 2.9.3) or that (conditionally) they have the same limiting distributions for almost all sequences of data, the vector of processes $(\mathbb{G}_N, \mathbb{G}^{\xi}_{1,N_1}, \ldots, \mathbb{G}^{\xi}_{J,N_J})$ converges weakly to the vector of independent Brownian bridge processes $(\mathbb{G}, \mathbb{G}_1, \ldots, \mathbb{G}_J)$. The continuous mapping theorem yields

$$\mathbb{G}^{\pi}_N \rightsquigarrow \mathbb{G} + \sum_{j=1}^{J} \sqrt{v_j}\sqrt{\frac{1-p_j}{p_j}}\,\mathbb{G}_j. \tag{12}$$

This result formalizes and extends proposition 1 of Self & Prentice (1988) and the arguments in section 4 of Borgan *et al.* (2000).

5. Asymptotic distributions of the WL estimator

We apply theorem 19.26 of van der Vaart (1998) to conclude that, under Bernoulli sampling,

$$\sqrt{N}\dot{\Psi}_0 \begin{pmatrix} \hat{\theta}-\theta_0 \\ \hat{\eta}-\eta_0 \end{pmatrix} h = -\mathbb{G}_N \frac{\xi}{\pi_0} \begin{pmatrix} \dot{\ell}_0 \\ B_0 h \end{pmatrix} + o_p(1) \text{ uniformly for } h \in \mathcal{H}. \tag{13}$$

Similarly, using theorem 3.3.1 of van der Vaart & Wellner (1996) together with the development of the previous section, we conclude that for finite population stratified sampling

$$\sqrt{N}\dot{\Psi}_0 \begin{pmatrix} \hat{\theta}-\theta_0 \\ \hat{\eta}-\eta_0 \end{pmatrix} h = -\mathbb{G}^{\pi}_N \begin{pmatrix} \dot{\ell}_0 \\ B_0 h \end{pmatrix} + o_p(1) \text{ uniformly for } h \in \mathcal{H}. \tag{14}$$

We have already argued that the hypotheses of the first theorem follow from appropriately modified versions of A1–A4. Together with the weak convergence of $\mathbb{G}^{\pi}_N$ just established, they also suffice for the second theorem. In particular, the stochastic condition (3.3.2) of van der Vaart & Wellner (1996) follows from A1 and A2 together with the proof of their lemma 3.3.5 applied to each of $\mathbb{G}_N, \mathbb{G}^{\xi}_{1,N_1}, \ldots, \mathbb{G}^{\xi}_{1,N_1}$.

In practice, attention is usually focused on inferences for the Euclidean parameter θ. To derive a general expression for the asymptotic variance of $\hat{\theta}$ we further assume

A5 $\dot{\Psi}_0$ admits a partition as in equation (25.91) of van der Vaart (1998) where the information operator $B_0^* B_0$ is continuously invertible.

Following closely the arguments in section 25.12 of van der Vaart, we calculate from (13) that under Bernoulli sampling

$$\sqrt{N}(\hat{\theta}-\theta_0) = \mathbb{G}_N \frac{\xi}{\pi_0}\tilde{\ell}_0 + o_p(1) \tag{15}$$

whereas from (14) under finite population stratified sampling

$$\sqrt{N}(\hat{\theta}-\theta_0) = \mathbb{G}^{\pi}_N \tilde{\ell}_0 + o_p(1), \tag{16}$$

where in both cases $\tilde{\ell}_0$ denotes the efficient influence function

$$\tilde{\ell}_0 = \tilde{I}_0^{-1}(I - B_0\left(B_0^* B_0\right)^{-1} B_0^*)\dot{\ell}_0 \tag{17}$$

and

$$\tilde{I}_0 = P_0[(I - B_0(B_0^* B_0)^{-1} B_0^*)\dot{\ell}_0 \dot{\ell}_0^T] \tag{18}$$

is the efficient information. As $P_0\tilde{\ell}_0 = 0$, moreover, both (15) and (16) may be expressed as

$$\sqrt{N}(\hat{\theta}-\theta_0) = \sqrt{N}\mathbb{P}^{\pi}_N \tilde{\ell}_0 + o_p(1) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \frac{\xi_i}{\pi_i}\tilde{\ell}_0(X_i) + o_p(1), \tag{19}$$

which expansion constitutes the principal result of this paper.

Under Bernoulli sampling with known π_0 the asymptotic variance is therefore

$$\begin{aligned}\mathrm{var}_A\sqrt{N}(\hat\theta-\theta_0) &= \mathrm{var}\left(\frac{\xi}{\pi_0}\tilde\ell_0\right)\\ &= \mathrm{var}\,E\left(\frac{\xi}{\pi_0}\tilde\ell_0\,\middle|\,X\right)+E\,\mathrm{var}\left(\frac{\xi}{\pi_0}\tilde\ell_0\,\middle|\,X\right)\\ &= \mathrm{var}(\tilde\ell_0)+\mathrm{E}\left[\frac{\tilde\ell_0^{\otimes 2}}{\pi_0^2}\mathrm{var}(\xi|X)\right]\\ &= \tilde I_0^{-1}+\tilde P_0\left(\frac{1-\pi_0}{\pi_0}\tilde\ell_0^{\otimes 2}\right).\end{aligned} \tag{20}$$

In the special case of stratified Bernoulli sampling, with $\pi_i=\pi_0(V_i)=p_j$ for $V_i\in\mathcal{V}_j$, (20) becomes

$$\tilde I_0^{-1}+\sum_{j=1}^{J} v_j\frac{1-p_j}{p_j}P_{0|j}\left(\tilde\ell_0^{\otimes 2}\right) \tag{21}$$

by averaging over the stratum-specific conditonal expectations. On the other hand, from (12) and (16) directly, the asymptotic variance under finite population stratified sampling is

$$\tilde I_0^{-1}+\sum_{j=1}^{J} v_j\frac{1-p_j}{p_j}\mathrm{var}_j(\tilde\ell_0), \tag{22}$$

where $\mathrm{var}_j(f)=P_{0|j}(f^{\otimes 2})-P_{0|j}^{\otimes 2}(f)$. Comparing the expressions in (21) and (22) shows the substantial potential gain from keeping track of the stratum frequencies for the phase 1 data.

6. Bernoulli sampling with estimated weights

Let $\mathcal{V}_0$ denote an additional stratum, possibly null, such that $\xi_i=1$ for $V_i\in\mathcal{V}_0$. Introduction of this special stratum with $p_0=1$ does not affect the previous development; in particular, equations (19)–(22) continue to hold. For $V_i\notin\mathcal{V}_0$ suppose

$$\Pr(\xi_i=1|X_i,V_i;\alpha)=\Pr(\xi_i=1|V_i;\alpha)=\pi_\alpha(V_i)<1, \tag{23}$$

where $\alpha\in\Xi\subset\mathbb{R}^q$ is a parameter to be estimated by ML from the phase 1 observations $\{V_i, i=1,\ldots,N\}$ not in $\mathcal{V}_0$. We assume sufficient regularity in the model for α, e.g. to satisfy the hypotheses of theorem 5.21 of van der Vaart (1998), so that the ML estimator $\hat\alpha$ is consistent and asymptotically normal with influence function

$$\tilde\ell_0^\alpha=\mathbf{1}_{\mathcal{V}_0^c}\left(\tilde P_0\mathbf{1}_{\mathcal{V}_0^c}\frac{\dot\pi_0^{\otimes 2}}{\pi_0(1-\pi_0)}\right)^{-1}\dot\pi_0\frac{\xi-\pi_0}{\pi_0(1-\pi_0)}. \tag{24}$$

Here for $V\in\mathcal{V}_0^c$, the complement of $\mathcal{V}_0$, $\pi_0(V)=\pi_{\alpha_0}(V)$ is the true sampling function while $\dot\pi_0(V)$ denotes the q-vector of partial derivatives of $\pi_\alpha(V)$ with respect to α evaluated at $\alpha=\alpha_0$. If $\hat\theta(\alpha)$ denotes the WL estimator under two-phase Bernoulli sampling with 'known' sampling function $\pi_\alpha(V)$, then from (24) and (19) we have

$$\sqrt{N}\begin{pmatrix}\hat\theta(\alpha_0)-\theta_0\\ \hat\alpha-\alpha_0\end{pmatrix}=\sqrt{N}\begin{pmatrix}\mathbb{P}_N^\pi\tilde\ell_0\\ \mathbb{Q}_N\tilde\ell_0^\alpha\end{pmatrix}+o_p(1). \tag{25}$$

Furthermore, with $\hat\pi_i=\pi_{\hat\alpha}(V_i)$ for $V_i\in\mathcal{V}_0^c$ otherwise $\hat\pi_i=1$, we show in appendix C that under some further mild assumptions regarding $\pi_\alpha(V)$

$$\sqrt{N}(\mathbb{P}_N^{\hat\pi}-\mathbb{P}_N^{\pi_0})\tilde\ell_0=-\tilde P_0\left(\mathbf{1}_{\mathcal{V}_0^c}\frac{\tilde\ell_0\dot\pi_0^T}{\pi_0}\right)\sqrt{N}(\hat\alpha-\alpha_0)+o_p(1). \tag{26}$$

The joint asymptotic normality of $(\hat{\theta}(\alpha_0), \hat{\alpha})$ that follows from (25), together with the Taylor expansion (26), are precisely the hypotheses used by Pierce (1982) to deduce the asymptotic properties of $\hat{\theta}(\hat{\alpha})$. His results lead to the conclusion that $\sqrt{N}[\hat{\theta}(\hat{\alpha}) - \theta_0] \rightsquigarrow Z$, where $Z \in \mathbb{R}^p$ is mean zero Gaussian with covariance matrix

$$\mathrm{var}_A \sqrt{N}\left(\hat{\theta}(\hat{\alpha}) - \theta_0\right) = \mathrm{var}\left(\frac{\xi}{\pi_0}\tilde{\ell}_0\right) - \tilde{P}_0 \mathbf{1}_{\mathcal{V}_0^c} \frac{\tilde{\ell}_0 \dot{\pi}_0^T}{\pi_0} \left(\tilde{P}_0 \mathbf{1}_{\mathcal{V}_0^c} \frac{\dot{\pi}_0^{\otimes 2}}{\pi_0(1-\pi_0)}\right)^{-1} \tilde{P}_0 \mathbf{1}_{\mathcal{V}_0^c} \frac{\dot{\pi}_0 \tilde{\ell}_0^T}{\pi_0}. \quad (27)$$

A matrix calculation shows that, when (27) is evaluated for stratified Bernoulli sampling

$$\pi_\alpha = \pi_\alpha(V) = \begin{cases} 1, & V \in \mathcal{V}_0 \\ \alpha_j, & V \in \mathcal{V}_j, j = 1, \ldots, J, \end{cases}$$

the asymptotic variance for the WL estimator $\hat{\theta}$ with *estimated* sampling probabilities $\hat{\alpha}_j = n_j/N_j$ is identical to the finite population sampling variance (22) with $p_j = \alpha_{j,0} = \lim n_j/N_j$.

Two possibilities present themselves for estimation of the terms in (27). Using (20), we could estimate the first term by

$$\widehat{\mathrm{var}\left(\frac{\xi}{\pi_0}\tilde{\ell}_0\right)} = \tilde{I}^{-1}_{\hat{\theta},\hat{\eta}} + \frac{1}{N}\sum_{i=1}^N \frac{\xi_i(1-\hat{\pi}_i)}{\hat{\pi}_i^2} \tilde{\ell}^{\otimes 2}_{\hat{\theta},\hat{\eta}}(X_i),$$

the expression in the middle of the second term by

$$\widehat{\tilde{P}_0 \mathbf{1}_{\mathcal{V}_0^c} \frac{\dot{\pi}_0^{\otimes 2}}{\pi_0(1-\pi_0)}} = \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\mathcal{V}_0^c}(V_i) \frac{\dot{\pi}^{\otimes 2}_{\hat{\alpha}}(V_i)}{\hat{\pi}_i(1-\hat{\pi}_i)}$$

and similarly for $\tilde{P}_0(\tilde{\ell}_0 \dot{\pi}_0^T/\pi_0)$. A more empirical approach, however, would be to use the θ and α influence function contributions themselves to estimate these terms as in

$$\widehat{\mathrm{var}\left(\frac{\xi}{\pi_0}\tilde{\ell}_0\right)} = \frac{1}{N}\sum_{i=1}^N \left(\frac{\xi_i}{\hat{\pi}_i}\tilde{\ell}_{\hat{\theta},\hat{\eta}}(X_i)\right)^{\otimes 2},$$

$$\begin{aligned} \widehat{\tilde{P}_0 \mathbf{1}_{\mathcal{V}_0^c} \frac{\tilde{\ell}_0 \dot{\pi}_0^T}{\pi_0}} &= \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\mathcal{V}_0^c}(V_i) \frac{\xi_i}{\hat{\pi}_i} \frac{\tilde{\ell}_{\hat{\theta},\hat{\eta}}(X_i)}{\hat{\pi}_i} \dot{\pi}_{\hat{\alpha}}(V_i)^T \\ &= \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\mathcal{V}_0^c}(V_i) \left(\frac{\xi_i \tilde{\ell}_{\hat{\theta},\hat{\eta}}(X_i)}{\hat{\pi}_i}\right) \left(\frac{\dot{\pi}_{\hat{\alpha}}(V_i)^T(\xi_i - \hat{\pi}_i)}{\hat{\pi}_i(1-\hat{\pi}_i)}\right) \end{aligned}$$

and

$$\widehat{\tilde{P}_0 \mathbf{1}_{\mathcal{V}_0^c} \frac{\dot{\pi}_0^{\otimes 2}}{\pi_0(1-\pi_0)}} = \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\mathcal{V}_0^c}(V_i) \left(\frac{\dot{\pi}_{\hat{\alpha}}(V_i)(\xi_i - \hat{\pi}_i)}{\hat{\pi}_i(1-\hat{\pi}_i)}\right)^{\otimes 2}.$$

The resulting asymptotic variance for $\hat{\theta}$ may be recognized as comprising the residual sums of squares and of cross-products from the least squares regressions of each of the p components of the $\hat{\theta}$ influence function contributions $\xi_i \tilde{\ell}_{\hat{\theta},\hat{\eta}}(X_i)/\hat{\pi}_i$, to which subjects not in the phase 2 sample contribute 0, on the q components of the estimated $\hat{\alpha}$ influence function contributions (24), to which subjects having $V_i \in \mathcal{V}_0$ contribute 0. See Henmi & Eguchi (2004) for a recent discussion and interpretation. This suggests the following estimation procedure:

1 Estimate α from the phase 1 data and compute the estimated sampling fractions $\hat{\pi}_i$.
2 Estimate θ and η from the phase 2 data by WL, using the inverse $\hat{\pi}_i$ as 'known' weights.
3 Regress each component of the influence function contributions for $\hat{\theta}$ on those for $\hat{\alpha}$.
4 Estimate $\mathrm{var}_A(\hat{\theta})$ as the matrix comprising the residual sums of squares and of cross-products from these regressions.

Therneau & Grambsch (2000, p. 166), who cited earlier work by Pugh *et al.* (1994), suggested this procedure for the special case of Cox regression, to which we now direct our attention.

7. Application to the Cox proportional hazards model

Our development of the Cox model follows closely that of van der Vaart (1998, section 25.12) where $X=(\Delta, T, Z)$ with $T=\min(\tilde{T}, C)$ a censored failure time, $\Delta=\mathbf{1}_{[\tilde{T}\le C]}$ the failure indicator and $Z\in\mathbb{R}^p$ a vector of covariates. The Euclidean parameter is the p-vector of regression coefficients θ in the linear predictor $z\theta$. The non-parametric parameter $\eta=(\Lambda, G, G_Z)$ has three infinite dimensional components: $\Lambda(\cdot)=\int_0^{\cdot}\lambda(s)\,\mathrm{d}s$ the baseline cumulative hazard function, assumed differentiable; $G(t\,|\,z)=\Pr(C\le t\,|\,Z=z)$ the conditional distribution of the censoring time; and G_Z, the marginal distribution of the covariates. We introduce the usual notation for the 'at-risk' process $Y(t)=\mathbf{1}_{[T\ge t]}$ and the event counting process $N(t)=\Delta\mathbf{1}_{[T\le t]}$ and we make the standard assumptions: (i) the true failure time $\tilde{T}$ and the censoring time C are independent given Z; and (ii) there is a finite maximum censoring time τ such that $\Pr[Y(\tau)=1]>0$. van der Vaart (1998) makes some further 'partly unnecessary' assumptions to simplify his development, namely that the covariates Z are bounded, that G and G_Z have densities as indicated below and especially that $\Pr(C\ge\tau)=\Pr(C=\tau)>0$ (see discussion in section 8). Writing the density for $x=(\delta, t, z)$, with z a row vector, as

$$\exp(-e^{z\theta}\Lambda(t))[\mathrm{e}^{z\theta}\lambda(t)(1-G(t-\,|\,z))]^{\delta}[g(t\,|z)]^{1-\delta}g_Z(z), \tag{28}$$

and noting that G and G_Z factor out of the complete data likelihood, van der Vaart (1998) considers ML estimation for (θ, Λ) only. With $\mathcal{H}$ denoting the unit ball in the space $\mathcal{B}=\mathrm{BV}[0,\tau]$ of functions of bounded variation on $[0,\tau]$, he develops the following explicit expressions for the θ score vector, the Λ score operator that maps functions $h\in\mathcal{H}$ to functions of the data, its adjoint (but only evaluated for the θ scores) and the information operator that maps $\mathcal{H}$ onto itself:

$$\begin{aligned}\dot{\ell}_{\theta,\Lambda}(x)&=\delta z - z\,\mathrm{e}^{z\theta}\Lambda(t)\\ B_{\theta,\Lambda}h(x)&=\delta h(t)-\mathrm{e}^{z\theta}\int_0^t h\,\mathrm{d}\Lambda\end{aligned} \tag{29}$$

$$\begin{aligned}B^*_{\theta,\Lambda}\dot{\ell}_{\theta,\Lambda}(t)&=P_{\theta,\Lambda}Y(t)Z\,\mathrm{e}^{Z\theta}\\ B^*_{\theta,\Lambda}B_{\theta,\Lambda}h(t)&=h(t)P_{\theta,\Lambda}Y(t)\,\mathrm{e}^{Z\theta}.\end{aligned} \tag{30}$$

These are used to calculate the efficient scores

$$\begin{aligned}\ell^*_{\theta,\Lambda}(x)&=\dot{\ell}_{\theta,\Lambda}-B_{\theta,\Lambda}(B^*_{\theta,\Lambda}B_{\theta,\Lambda})^{-1}B^*_{\theta,\Lambda}\dot{\ell}_{\theta,\Lambda}\\ &=\delta[z-m(t;\theta)]-\mathrm{e}^{z\theta}\int_0^t[z-m(s;\theta)]\,\mathrm{d}\Lambda(s)\end{aligned} \tag{31}$$

and efficient information

$$\begin{aligned}\tilde{I}_0&=I_0-P_0B_0\left(B_0^*B_0\right)^{-1}B_0^*\dot{\ell}_0\\ &=P_0\left(\mathrm{e}^{Z\theta_0}\int_0^{\tau}[Z-m(t;\theta_0)]^{\otimes 2}\Pr(T\ge t\,|\,Z)\,\mathrm{d}\Lambda_0(t)\right),\end{aligned} \tag{32}$$

respectively, where $I_0=P_0\dot{\ell}_0\dot{\ell}_0^T$ and $m(t;\theta)=S^{(1)}(t;\theta)/S^{(0)}(t;\theta)$ with

$$\begin{aligned}S^{(0)}(t;\theta)&=P_0\,\mathrm{e}^{Z\theta}Y(t)\\ S^{(1)}(t;\theta)&=P_0Z\,\mathrm{e}^{Z\theta}Y(t).\end{aligned}$$

To fit the Cox model by WL to two-phase stratified samples, first define IPW estimators of the two quantities just considered by $\hat{S}^{(0)}(t;\theta)=\mathbb{P}_N^{\pi}\,\mathrm{e}^{Z\theta}Y(t)$ and $S^{(1)}(t;\theta)=\mathbb{P}_N^{\pi}Z\,\mathrm{e}^{Z\theta}Y(t)$. By definition the WL estimators solve

$$\Psi_{N1}^{\pi}(\theta, \Lambda) = \mathbb{P}_N^{\pi} \dot{\ell}_{\theta, \Lambda} = 0 \tag{33}$$

$$\Psi_{N2}^{\pi}(\theta, \Lambda)h = \mathbb{P}_N^{\pi} B_{\theta, \Lambda} h = 0 \quad \text{for all} \quad h \in \mathcal{H}, \tag{34}$$

where we have used the fact that $P_{\theta, \Lambda} B_{\theta, \Lambda} h = 0$. Substituting h_t defined by

$$h_t(s) = \frac{\mathbf{1}_{[s \le t]}}{\hat{S}^{(0)}(s, \theta)},$$

for h in (34) and solving using (30) gives

$$\hat{\Lambda}_\theta(t) = \mathbb{P}_N^{\pi} \frac{\Delta \mathbf{1}[T \le t]}{\hat{S}^{(0)}(T; \theta)} = \frac{1}{N} \sum_{i=1}^{N} \int_0^t \frac{\xi_i}{\pi_i} \frac{\mathrm{d}N_i(s)}{\hat{S}^{(0)}(s; \theta)}, \tag{35}$$

which may be recognized as an IPW version of the so-called Breslow (1974) estimator. For fixed θ this satisfies $\mathbb{P}_N^{\pi} B_{\theta, \hat{\Lambda}_\theta} h = 0$ for all $h \in \mathcal{H}$, as is easily checked, and maximizes the WL, as in van der Vaart (1998, example 25.69). Substituting $\hat{\Lambda}_\theta$ for Λ in (33) and evaluating using (29) yields the IPW Cox 'partial score' equation

$$\Psi_{N1}^{\pi}(\theta, \hat{\Lambda}_\theta) = \mathbb{P}_N^{\pi} \Delta \left[Z - \hat{m}(T; \theta) \right] = \frac{1}{N} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} \Delta_i \left[Z_i - \frac{\hat{S}^{(1)}(T_i; \theta)}{\hat{S}^{(0)}(T_i; \theta)} \right] = 0,$$

whose joint solution with (35) is the WL estimator $(\hat{\theta}, \hat{\Lambda}_{\hat{\theta}})$. According to the results of this paper, its large sample properties follow from those already developed for the ML estimator with complete data, which solves the same equations with $\xi_i = \pi_i = 1$, $i = 1, \ldots, N$. The joint asymptotic distribution of $(\hat{\theta}, \hat{\Lambda}_{\hat{\theta}})$ under Bernoulli sampling is obtained by applying the inverse map $\dot{\Psi}_0^{-1}$ defined in assumption A2 to both sides of equation (13), where $\dot{\ell}_0$ and B_0 are given by (29) and (30). The joint asymptotic distribution under finite population stratified sampling is obtained in the same manner from (14). In either case $\sqrt{N}(\hat{\theta} - \theta_0)$ has the asymptotic linear expansion (19), where $\tilde{\ell}_0 = \tilde{I}^{-1} \ell_0^*$ is the well-known influence function for the Cox model, here given explicitly by equations (31) and (32).

These are the estimators proposed for Cox regression by Binder (1992), Pugh *et al.* (1994), Borgan *et al.* (2000, estimator II), Lin (2000) and others for a variety of complex sampling and missing data problems.

8. Discussion

The key step in deriving asymptotic properties of a regular, asymptotically linear estimator of a Euclidean parameter θ is determination of its influence function. The principal result of this paper states that the estimator that solves IPW versions of estimating equations has an asymptotic expansion involving the IPW empirical measure of the usual influence function for simple random sampling. For fully parametric models, where both θ and η are of finite dimension, this result is a straightforward consequence of the theory of unbiased estimating equations. We provide the extension to semiparametric models for the special case where the estimating equations are derived from likelihood scores for the Euclidean parameter and from the score operator for the general parameter, and where the latter is estimable at the standard $\sqrt{N}$ rate. Our principal result, however, is likely to hold much more generally.

A first generalization would involve replacing the likelihood equation (1) with unbiased estimating equations of the form $\Psi_{N1}(\theta, \eta) = \mathbb{P}_N \psi_{1;\theta, \eta}(X) = 0$ and $\Psi_{N2}(\theta, \eta) = \mathbb{P}_N \psi_{2;\theta, \eta}(X)h = 0 \, \forall \, h \in \mathcal{H}$, where $\psi_{1;\theta, \eta} \in \mathbb{R}^p$ and $\psi_{2;\theta, \eta} \in \ell^\infty(\mathcal{H})$ both have expectation 0 under $P_{\theta, \eta}$. A1–A4 would be modified to satisfy the conditions, appropriately stated for the partitioned parameter, of van der Vaart (1998, theorem 19.26) for Bernoulli sampling or of van der Vaart & Wellner (1996, theorem 3.3.1) for finite population stratified sampling. This approach would

require continuous invertibility of $\dot{\Psi}_{22}$, the Fréchet derivative of $\Psi_2 = P_0\psi_{2;\theta,\eta}$ with respect to η. Conclusions would be given by equation (13) or (14), replacing $\dot{\ell}_0$ with $\psi_{1;\theta_0,\eta_0}$ and $B_0 h$ with $\psi_{2;\theta_0,\eta_0}h$. As a simple application, using the same notation as in section 7, consider the additive hazards regression model (Lin & Ying, 1994) where $\lambda(t\,|\,z) = \lambda(t) + z\theta$ and where θ and $\eta(t) = PZY(t)/PY(t)$ are jointly estimated from simple random samples by solving

$$\Psi_{N1}(\theta,\eta) = \mathbb{P}_N\left\{\Delta[Z-\eta(T)] - \int_0^\tau [Z-\eta(t)]Z\theta Y(t)\,\mathrm{d}t\right\} = 0$$

$$\Psi_{N2}(\theta,\eta)(t) = \mathbb{P}_N[Z-\eta(t)]Y(t) = 0 \quad \forall\, t\in[0,\tau].$$

Solution of the IPW version of these equations yields the estimator for θ suggested by Kulich & Lin (2000) for case–cohort sampling, namely

$$\hat{\theta} = \left\{\mathbb{P}_N^\pi \int_0^\tau [Z-\hat{\eta}(t)]^{\otimes 2} Y(t)\,\mathrm{d}t\right\}^{-1} \mathbb{P}_N^\pi \Delta[Z-\hat{\eta}(T)],$$

where $\hat{\eta}(t) = \mathbb{P}_N^\pi ZY(t)/\mathbb{P}_N^\pi Y(t)$. Calculation of the inverse of the operator $\dot{\Psi} = (\dot{\Psi}_1, \dot{\Psi}_2)$ in partitioned form is facilitated by the fact that

$$\Psi_1(\theta,\eta) = P_0\psi_{1;\theta,\eta} = P_0\int_0^\tau [Z-\eta(t)][\lambda_0(t) + Z(\theta_0-\theta)]Y(t)\,\mathrm{d}t$$

and $\Psi_2(\theta,\eta)(t) = P_0[Z-\eta(t)]Y(t)$ are linear in θ and η, that $\dot{\Psi}_{22}^{-1}(\eta-\eta_0)(t) = -(\eta-\eta_0)(t)/P_0\,Y(t)$ and that $\dot{\Psi}_{21} = 0$. We leave details to the interested reader.

A second generalization would be to complex probability sampling designs for selecting phase 2 observations from the finite population obtained at phase 1. The Horvitz & Thompson (1952) theorem provides design-based variances of IPW sample means for a very general class of designs in terms of the first- and second-order inclusion probabilities: π_i the probability of including the ith of N phase 1 observations in the phase 2 sample, and $\pi_{ii'}$ the probability that the observations labelled i and i' are both included (Overton & Stehman, 1995). With finite population stratified sampling, for example, the first-order probabilities π_i are given by (5). If i and i' are in separate strata, $\pi_{ii'} = \pi_i\pi_{i'}$, whereas if both are in the jth stratum, $\pi_{ii'} = n_j(n_j-1)/[N_j(N_j-1)]$. Lin (2000), building on the work of Binder (1992), considered fitting the Cox model to survey data collected by means of probability sampling. Inserting the first- and second-order inclusion probabilities for finite population stratified sampling into the expression he derives for the asymptotic variance of the regression parameter θ leads to our equation (22), with the efficient information and influence function as defined in section 7. We envisage development of a theory for joint estimation of Euclidean and infinite dimensional parameters in semiparametric models fitted to two-phase data, where the second phase observations are selected using quite general complex probability sampling procedures.

Some investigators of two-phase designs for failure time data, e.g. Borgan *et al.* (2000) and Kulich & Lin (2000, 2004), have restricted attention to covariate stratified versions of the case–cohort design. This is a stratified sampling design whereby all subjects who fail are included at phase 2 ($\pi = 1$). Although this may well be an efficient design when the failure rate is low, the assumption that $\xi = 1$ whenever $\Delta = 1$ is often unnecessary and may sometimes be unduly restrictive. Not only does it limit application when the phase 1 population has large numbers of both failures and non-failures, but also does so when the sampling has been carried out for one failure type but it is of interest to evaluate another. When following patients enrolled in a clinical trial, for example, all deaths may be sampled as 'cases' but it may later be decided to analyse the data also in terms of 'event-free survival'. In other

contexts, biological samples may turn out to be non-informative so that data are still missing for substantial numbers of subjects, including failed cases, who are sampled at phase 2. Provided one is willing to make the standard MAR assumptions, the WL methods described herein may still be used by determining the stratum frequencies for subjects having complete data at phase 2 and using these to estimate the sampling weights. Thus the more general stratified sampling framework considered here provides a useful extension of previous results for covariate stratified case–cohort studies.

On the other hand, much of the previous work with Cox's model has used a counting process formulation that facilitates handling of time-dependent covariates, multiple failures per subject and staggered entry into the cohort. Assumptions A1–A4 of section 2 have been established for Cox's model only under the much more stringent and 'partly unnecessary' conditions imposed by van der Vaart (1998, section 25.12.1). The requirement that everyone still 'on-study' be censored at the common time τ would apply to situations in which t referred to calendar time, everyone was entered on study at $t=0$ and there was a common closing date at $t=\tau$. It would not apply, however, if subjects were entered on study at various calendar times but withdrawn on a common closing date, and t was taken to be 'time-on-study'. Nor would it apply if t was 'age' and subjects both entered and exited the study at various ages. We look forward to further work that eases these restrictions, in particular to a determination as to whether or not the general approach extends to Cox regression under standard assumptions (Andersen & Gill, 1982).

The major drawback of WL estimation is its lack of statistical efficiency. The efficiency loss has been determined for parametric regression models fitted to two-phase stratified data, where numerical evaluation of profile likelihoods and explicit calculation of information bounds are both feasible (Lawless *et al.*, 1999; Breslow *et al.*, 2003). It can be quite serious when fitting logistic models to stratified case–control data (Robins *et al.*, 1994; Scott & Wild, 1997; Breslow & Holubkov, 1997). Attempts to improve efficiency with Cox regression include use of time-dependent sampling weights (Borgan *et al.*, 2000; Kulich & Lin, 2004), and work is in progress to extend our theoretical approach to this situation. Other methods to improve or evaluate efficiency in particular contexts have been proposed by Chen (2001), Nan *et al.* (2004), Nan (2004), Scheike & Martinussen (2004) and many others. Most of these methods are relatively recent and involve sufficiently complex calculations, or sufficiently restrictive assumptions, that none have yet seen widespread use. These limitations are certain to decline with advances in computing hardware and software, making more efficient estimation methods more widely available. In the meantime, the WL estimation procedure outlined at the end of section 6 offers a relatively simple and robust alternative. It is likely to remain the method of choice for many survey statisticians for the reasons mentioned in the introduction, namely, their interest in finite population parameters defined as solutions to ML estimating equations. (See Scott & Wild, 2002, however, for a revised and contrary view.) As emphasized by Robins *et al.* (1994), in view of the interpretation of (27) as a residual sum of squares, enrichment of the model (23) for π can only enhance the efficiency of θ estimation. When the sampling probabilities vary, as in finite population stratified sampling, inclusion of the stratum factors in the model is essential to avoid bias. Finer stratification or the inclusion of auxiliary variables in the model for π serves the cause of efficiency. Equation (22) suggests that such additional variables would be most valuable if they could somehow be chosen to be highly correlated with the efficient scores. The doubly weighted estimator developed by Kulich & Lin (2004) for exposure stratified case–cohort studies is intriguing in that it uses a separate set of (time dependent) weights for each covariate. A preliminary analysis is conducted to estimate quantities that resemble within stratum conditional expectations of partial score contributions given the phase 1 data, and these are used to form the weights. An

extension of their approach to more general two-phase stratified or other complex sampling designs would be of considerable interest.

In his Appendix Lin (2000) remarks 'To our knowledge, there does not exist a general theory on the conditions required for the tightness and weak convergence of Horvitz–Thompson processes. However, the results of van der Vaart & Wellner (1996, sections 2.9, 3.6, 3.7) can be applied to possibly stratified simple random sampling and can potentially be extended to other survey designs.' One purpose of this paper has been to carry out in detail the program mentioned for stratified random sampling. We conjecture that weak convergence of the IPW empirical process and our fundamental equation (19) for Horvitz–Thompson estimators also hold with other complex sampling designs, and work is in progress to explore these extensions.

Acknowledgements

The second author owes thanks to Galen Shorack for a helpful discussion concerning the representation in appendix A. This work was supported in part by grants from the US National Institutes of Health and National Science Foundation.

References

Andersen, P. K. & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100–1120.

Begun, J. M., Hall, W. J., Huang, W.-M. & Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11**, 432–452.

Binder, D. A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* **79**, 139–147.

Borgan, Ø., Langholz, B., Samuelsen, S. O., Goldstein, L. & Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Anal.* **6**, 39–58.

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

Breslow, N. E. & Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J. Roy. Statist. Soc. Ser. B* **59**, 447–461.

Breslow, N., McNeney, B. & Wellner, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Ann. Statist.* **31**, 1110–1139.

Chen, K. (2001). Generalized case-cohort sampling. *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* **63**, 791–809.

Cox, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34**, 187–220.

Henmi, M. & Eguchi, S. (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika* **91**, 929–941.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**, 663–685.

Kalbfleisch, J. D. & Lawless, J. F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statist. Med.* **7**, 149–160.

Kulich, M. & Lin, D. Y. (2000). Additive hazards regression for case-cohort studies. *Biometrika* **87**, 73–87.

Kulich, M. & Lin, D. Y. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *J. Amer. Statist. Assoc.* **99**, 832–844.

Lawless, J. F., Kalbfleisch, J. D. & Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* **61**, 413–438.

Lin, D. Y. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika* **87**, 37–47.

Lin, D. Y. & Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61–71.

Manski, C. F. & Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica* **45**, 1977–1988.

Nan, B. (2004). Efficient estimation for case-cohort studies. *Can. J. Statist.* **32**, 403–419.

Nan, B., Emond, M. J. & Wellner, J. A. (2004). Information bounds for Cox regression models with missing data. *Ann. Statist.* **32** 723–753.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *J. Amer. Statist. Assoc.* **33**, 101–116.
Overton, S. W. & Stehman, S. V. (1995). The Horvitz–Thompson theorem as a unifying perspective for probability sampling: With examples from natural resource sampling. *Amer. Statist.* **49**, 261–268.
Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Ann. Statist.* **10**, 475–478.
Præstgaard, J. & Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21**, 2053–2086.
Pugh, M., Robins, J., Lipsitz, S. & Harrington, D. (1994). *Inference in the Cox proportional hazards model with missing covariates*. Technical Report 758Z. Harvard School of Public Health, Boston, MA.
Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846–866.
Scheike, T. H. & Martinussen, T. (2004). Maximum likelihood estimation for Cox's regression model under case-cohort sampling. *Scand. J. Statist.* **31**, 283–293.
Scott, A. & Wild, C. (2002). On the robustness of weighted methods for fitting models to case-control data. *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* **64**, 207–219.
Scott, A. J. & Wild, C. J. (1986). Fitting logistic models under case-control or choice based sampling. *J. Roy. Statist. Soc. Ser. B* **48**, 170–182.
Scott, A. J. & Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57–71.
Self, S. G. & Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.* **16**, 64–81.
Skinner, C. J., Holt, D. & Smith, T. M. F. (eds) (1989). *Analysis of complex surveys.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Ltd, Chichester.
Therneau, T. M. & Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model.* Statistics for Biology and Health. Springer–Verlag, New York, NY.
van der Vaart, A. W. (1998). *Asymptotic statistics, vol. 3 of Cambridge series in statistical and probabilistic mathematics*. Cambridge University Press, Cambridge.
van der Vaart, A. W. & Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, NY.
Xie, Y. & Manski, C. F. (1989). The logit model and response-based samples. *Sociol. Methods Res.* **17**, 283–302.

Received November 2005, in final form March 2006

Norman E. Breslow, Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195-7232, USA.
E-mail: norm@u.washington.edu

Appendix

In appendices A and B we establish two results slightly more general than needed for the development in section 4. (See the end of appendix B for the special case required.) Although the 'i.i.d. within strata' representation given by proposition A.1 is very simple, it is crucial for a rigorous application of the exchangeable bootstrap limit theorem of Præstgaard & Wellner (1993) as sketched in section 4. The notation in these two appendices should be understood to be independent of the notation in the body of the paper.

Appendix A: a representation of stratified sampling

Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space and $W : (\Omega, \mathcal{A}) \to (\mathcal{W}, \mathcal{B})$. Write P^W for the measure induced by W on $(\mathcal{W}, \mathcal{B})$; in the notation of section 2, $P^W = \tilde{P}_0$. Suppose that $\mathcal{W}_1, \ldots, \mathcal{W}_J$ is a (measurable) partition of $\mathcal{W}$, i.e.: (i) $\mathcal{W}_j \in \mathcal{B}$, $j = 1, \ldots, J$; (ii) $\mathcal{W}_j \cap \mathcal{W}_{j'} =$ for $j \neq j'$; and (iii) $\cup_{j=1}^{J} \mathcal{W}_j = \mathcal{W}$. We will assume that $P(W \in \mathcal{W}_j) \equiv p_j > 0$ for $j = 1, \ldots, J$. Now consider a new probability space $(\Omega^\dagger, \mathcal{A}^\dagger, P^\dagger)$, where $\Omega^\dagger = \Omega_0^\dagger \times \Omega_1^\dagger \times \cdots \times \Omega_J^\dagger$,

$\mathcal{A}^\dagger = \mathcal{A}_0^\dagger \times \mathcal{A}_1^\dagger \times \cdots \times \mathcal{A}_J^\dagger$ and $P^\dagger = P_0^\dagger \cdot P_1^\dagger \cdots P_J^\dagger$. Let random variables $\Delta = (\Delta_1, \ldots, \Delta_J)$ and $W_1^\dagger, \ldots, W_J^\dagger$ be defined thereon as follows: for $\omega^\dagger = (\omega_0^\dagger, \omega_1^\dagger, \ldots, \omega_J^\dagger) \in \Omega^\dagger$,

$$\Delta(\omega^\dagger) = \Delta(\omega_0^\dagger) \sim \text{multinomial}_J(1, (p_1, \ldots, p_J))$$
$$W_j^\dagger(\omega^\dagger) = W_j^\dagger(\omega_j^\dagger) \sim P_j^\dagger$$

for $j = 1, \ldots, J$ where $p_j = P(W \in \mathcal{W}_j)$, $j = 1, \ldots, J$, and $P_j^\dagger$ is defined by

$$P_j^\dagger(W_j \in B) = \frac{P(W \in B \cap \mathcal{W}_j)}{P(W \in \mathcal{W}_j)} = \frac{P^W(B \cap \mathcal{W}_j)}{P^W(\mathcal{W}_j)}, \quad B \in \mathcal{B}. \tag{36}$$

Now define a random variable $W^\dagger : (\Omega^\dagger, \mathcal{A}^\dagger) \to (\mathcal{W}, \mathcal{B})$ by

$$W^\dagger(\omega^\dagger) = \Delta_1(\omega_0^\dagger) W_1^\dagger(\omega_1^\dagger) + \cdots + \Delta_J(\omega_0^\dagger) X_J^\dagger(\omega_J^\dagger).$$

Note that Δ, $W_1^\dagger, \ldots, W_J^\dagger$ are independent by construction.

Proposition A.1
$W^\dagger \stackrel{d}{=} W$ on $(\mathcal{W}, \mathcal{B})$. That is, $P^{W^\dagger} = P^W$ as measures on $(\mathcal{W}, \mathcal{B})$.

Proof. First note that

$$\begin{aligned} P^\dagger(W^\dagger \in \mathcal{W}_j) &= P^\dagger(W_j^\dagger \in \mathcal{W}_j, \Delta_j = 1) \\ &= P^\dagger(W_j^\dagger \in \mathcal{W}_j) P^\dagger(\Delta_j = 1) = 1 \cdot p_j = p_j \end{aligned} \tag{37}$$

using independence of Δ and $W_j^\dagger$, the fact that $W_j^\dagger$ takes values in $\mathcal{W}_j$ with $P^\dagger$-probability 1, and $P^\dagger(\Delta_j = 1) = p_j$ by the definition of $P^\dagger$.

Now let $B \in \mathcal{B}$. Then since $p_j > 0$ for $j = 1, \ldots, J$,

$$\begin{aligned} P^\dagger(W^\dagger \in B) &= \sum_{j=1}^J P^\dagger(W^\dagger \in B \cap \mathcal{W}_j) = \sum_{j=1}^J \frac{P^\dagger(W^\dagger \in B \cap \mathcal{W}_j)}{P^\dagger(W^\dagger \in \mathcal{W}_j)} P^\dagger(W^\dagger \in \mathcal{W}_j) \\ &= \sum_{j=1}^J \frac{P^\dagger(W_j^\dagger \in B)}{P^\dagger(W_j^\dagger \in \mathcal{W}_j)} p_j \quad \text{by (37)} \\ &= \sum_{j=1}^J \frac{P^W(B \cap \mathcal{W}_j)/P^W(\mathcal{W}_j)}{1} \cdot p_j \quad \text{by (36)} \\ &= \sum_{j=1}^J P^W(B \cap \mathcal{W}_j) = P^W(B) = P(W \in B). \end{aligned}$$

□

If $W_1, \ldots, W_N$ are i.i.d. P^W, then we can represent the W_i's in terms of $(\Delta_i, W_{1,i}^\dagger, \ldots, W_{J,i}^\dagger)$, $i = 1, \ldots, N$, i.i.d. as $(\Delta, W_1^\dagger, \ldots, W_J^\dagger)$ as described in proposition A.1. It follows that

$$\begin{aligned} \mathbb{P}_{j,N_j} &= \frac{1}{N_j} \sum_{i=1}^N \delta_{W_i} 1_{\mathcal{W}_j}(W_i) \qquad (38) \\ &= \frac{1}{N_j} \sum_{j'=1}^J \sum_{i=1}^N \Delta_{j',i} \delta_{W_{j',i}^\dagger} 1_{\mathcal{W}_j}(W_{j,i}^\dagger) \\ &= \frac{1}{N_j} \sum_{i=1}^{N_j} \delta_{W_{j,i}} \qquad (39) \end{aligned}$$

by relabelling the $W^{\dagger}_{j,i}$'s and where $N_j = \sum_{i=1}^{N} \Delta_{j,i}$ on the right side is independent of the $W^{\dagger}_{j,i}$'s. This yields the promised doubly indexed form of the stratum-specific empirical measure in terms of independent $W_{j,i}$'s distributed according to $P_{0|j}$, where $P_{0|j}(B) = P_0(B1_{\mathcal{W}_j})/P_0(1_{\mathcal{W}_j})$ for $B \in \mathcal{B}$. □

Appendix B: proof of weak convergence of the stratum-specific empirical process

Let $\mathbb{P}_{j,N_j}$ be as defined in (38), where $N^{-1}N_j = \mathbb{P}_N(1_{\mathcal{W}_j}) \to_{a.s.} P_0(\mathcal{W}_j) \equiv v_j > 0$.

Proposition B.1
If $\mathcal{F}$ is P_0-Donsker and $v_j > 0$, then $\mathcal{F}$ is $P_{0|j}$-Donsker on stratum $\mathcal{W}_j$ in the sense that

$$\mathbb{G}_{j,Nj} \equiv \sqrt{N_j}(\mathbb{P}_{j,N_j} - P_{0|j}) \rightsquigarrow \mathbb{G}_j \quad \text{in } \ell^{\infty}(\mathcal{F}) \tag{40}$$

where $\mathbb{G}_j$, defined by

$$\mathbb{G}_j(f) = v_j^{-1/2}\mathbb{G}_{P_0}((f - P_{0|j}(f))1_{\mathcal{W}_j}), \quad f \in \mathcal{F}, \tag{41}$$

is a $P_{0|j}$-Brownian bridge process.

Remark 1. Note that

$$\mathrm{var}(\mathbb{G}_j(f)) = v_j^{-1}P_0[(f - P_{0|j}(f))^2 1_{\mathcal{W}_j}] = \mathrm{var}_j(f) \equiv \mathrm{var}(f(W) \mid W \in \mathcal{W}_j).$$

Remark 2. The proposition implies that the process $\sqrt{N_j}(\mathbb{P}_{j,N_j} - P_{0|j})$ behaves asymptotically the same as that of a sample of fixed size drawn from the conditional distribution $P_{0|j}$.

Proof. First proof. By the discussion at the beginning of section 2.10.4, p. 200, van der Vaart and Wellner (1996), $\mathcal{F}_j \equiv \{f1_{\mathcal{W}_j} : f \in \mathcal{F}\}$ is P_0-Donsker,and hence the collection $\tilde{\mathcal{F}}_j \equiv \{f1_{\mathcal{W}_j} : f \in \mathcal{F} \cup \{1\}\}$ is also P_0-Donsker. Now we write

$$\begin{aligned}
\sqrt{N_j}(\mathbb{P}_{j,N_j}f - P_{0|j}f) &= \sqrt{N_j}\left(\frac{(1/N)\sum_{i=1}^{N} f(W_i)1_{\mathcal{W}_j}(W_i)}{(1/N)\sum_{i=1}^{N} 1_{\mathcal{W}_j}(W_i)} - \frac{P_0(f1_{\mathcal{W}_j})}{P_0(1_{\mathcal{W}_j})}\right) \\
&= \sqrt{\frac{N_j}{N}}\left\{\frac{\mathbb{G}_N(f1_{\mathcal{W}_j})}{N_j/N} - \frac{\mathbb{G}_N(1_{\mathcal{W}_j})P_0(f1_{\mathcal{W}_j})}{(N_j/N)P_0(\mathcal{W}_j)}\right\} \\
&= \frac{1}{\sqrt{N_j/N}}\left\{\mathbb{G}_N(f1_{\mathcal{W}_j}) - \mathbb{G}_N(1_{\mathcal{W}_j})P_{0|j}(f)\right\} \\
&= \frac{1}{\sqrt{N_j/N}}\mathbb{G}_N((f - P_{0|j}(f))1_{\mathcal{W}_j}) \\
&\Rightarrow \frac{1}{\sqrt{v_j}}\mathbb{G}_{P_0}((f - P_{0|j}(f))1_{\mathcal{W}_j}) \equiv \mathbb{G}_{P_{0|j}}(f),
\end{aligned}$$

and, in fact,

$$\left\{\frac{1}{\sqrt{v_j}}\mathbb{G}_{P_0}((f - P_{0|j}(f))1_{\mathcal{W}_j}) : f \in \mathcal{F}\right\} \stackrel{d}{=} \{\mathbb{G}_{P_{0|j}}(f) : f \in \mathcal{F}\}.$$

Second proof. By the second representation of the stratum-specific empirical measure $\mathbb{P}_{j,N_j}$ as $\mathbb{P}_{j,N_j} = N_j^{-1}\sum_{i=1}^{N_j}\delta_{W_{j,i}}$ where the $W_{j,i}$'s are i.i.d. $P_{0|j}$, it follows that the empirical process $\mathbb{G}_{j,N_j} = \sqrt{N_j}(\mathbb{P}_{j,N_j} - P_{0|j})$ is just the empirical process of i.i.d. $W_{j,i}$'s, but with a random

sample size N_j independent of the $W_{j,i}$'s. As $N_j/N \to v_j > 0$, it follows from theorem 3.5.1, p. 339, van der Vaart and Wellner (1996) that $\mathbb{G}_{j,N_j} \rightsquigarrow \mathbb{G}_j$ in $\ell^\infty(\mathcal{F})$, where $\mathbb{G}_j$ is a $P_{0|j}$-Brownian bridge process as before. □

In the application of the results of appendices A and B in section 4 we take $\mathcal{W}_1, \ldots, \mathcal{W}_J$ to be the measurable partition of $\mathcal{W}$ induced by the partition $\mathcal{V}_1, \ldots, \mathcal{V}_J$ of $\mathcal{V}$ (i.e. $\mathcal{W}_j = V^{-1}(\mathcal{V}_j)$ for $j = 1, \ldots, J$ where $V(W) \equiv (\tilde{X}(X), U)$). Moreover, the Donsker class $\mathcal{F}$ in proposition B.1 is taken to be a Donsker class of functions of X only rather than functions of $W = (X, U)$. This is exactly what is needed for the development in section 4.

Appendix C: proof of equation (26)

Besides the consistency and asymptotic linearity (24) for $\hat{\alpha}$ assumed in section 6, we further assume that $0 < \sigma \le \pi_\alpha(v)$ as in (3) and that

$$\left| \frac{1}{\pi_\alpha(v)} - \frac{1}{\pi_{\alpha_0}(v)} - \frac{-\dot{\pi}_0^T(v)}{\pi_0^2(v)}(\alpha - \alpha_0) \right| \le \psi(v)|\alpha - \alpha_0|^{1+\zeta} \tag{42}$$

for α in a neighbourhood of α_0, where $\zeta > 0$ and ψ satisfies $E\psi^2(V) < \infty$. The second assumption will typically follow from the first provided that π_α has a continuous second derivative. For example, suppose that π_α is given by a logistic regression model with linear predictor $\tilde{v}^T\alpha$, where $\tilde{v} = \tilde{v}(v) \in \mathbb{R}^q$. Then Taylor's formula with remainder shows that the left-hand side of (42) equals $\left|\frac{1}{2}\exp(-\tilde{v}^T\alpha^*)(\alpha - \alpha_0)^T\tilde{v}\tilde{v}^T(\alpha - \alpha_0)\right|$ with α^* on the line segment between α and α_0. Thus the condition holds with $\zeta = 1$ provided $\exp(\tilde{v}^T\alpha) = \pi_\alpha(v)/[1 - \pi_\alpha(v)]$ is bounded away from 0 and $\tilde{V}$ has finite fourth moment. It follows that

$$\begin{aligned}
\left(\mathbb{P}_N^{\hat{\pi}} - \mathbb{P}_N^{\pi_0}\right)\tilde{\ell}_0 &= \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\mathcal{V}_0^c}(V_i)\left(\frac{\xi_i}{\hat{\pi}_i} - \frac{\xi_i}{\pi_0}\right)\tilde{\ell}_0(X_i) \\
&= \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\mathcal{V}_0^c}(V_i)\xi_i\tilde{\ell}_0(X_i)\left[\frac{1}{\pi_{\hat{\alpha}}(V_i)} - \frac{1}{\pi_{\alpha_0}(V_i)} - \frac{-\dot{\pi}_0^T(V_i)}{\pi_0^2(V_i)}(\hat{\alpha} - \alpha_0)\right] \\
&\quad + \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\mathcal{V}_0^c}(V_i)\xi_i\tilde{\ell}_0(X_i)\left[\frac{-\dot{\pi}_0^T(V_i)}{\pi_0^2(V_i)}\right](\hat{\alpha} - \alpha_0) \\
&\equiv R_N - \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\mathcal{V}_0^c}(V_i)\frac{\xi_i}{\pi_0(V_i)}\tilde{\ell}_0(X_i)\left[\frac{\dot{\pi}_0^T(V_i)}{\pi_0(V_i)}\right](\hat{\alpha} - \alpha_0),
\end{aligned} \tag{43}$$

where by (3), the similar assumption for π_α and (42),

$$\begin{aligned}
|R_N| &\le \left| \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\mathcal{V}_0^c}(V_i)\xi_i\tilde{\ell}_0(X_i)\left[\frac{1}{\pi_{\hat{\alpha}}(V_i)} - \frac{1}{\pi_{\alpha_0}(V_i)} - \frac{-\dot{\pi}_0^T(V_i)}{\pi_0^2(V_i)}(\hat{\alpha} - \alpha_0)\right] \right| \\
&\le \frac{1}{\sigma^2}\frac{1}{N}\sum_{i=1}^N \psi(V_i)|\tilde{\ell}_0(X_i)| \cdot |\hat{\alpha} - \alpha_0|^{1+\zeta} \\
&= O_p(1)|\hat{\alpha} - \alpha_0||\hat{\alpha} - \alpha_0|^\zeta = O_p(1)O_p(N^{-1/2})o_p(1).
\end{aligned}$$

Multiplying through (43) by $\sqrt{N}$, we conclude that (26) holds by virtue of $\sqrt{N}R_N = o_p(1)$ and the strong law of large numbers.