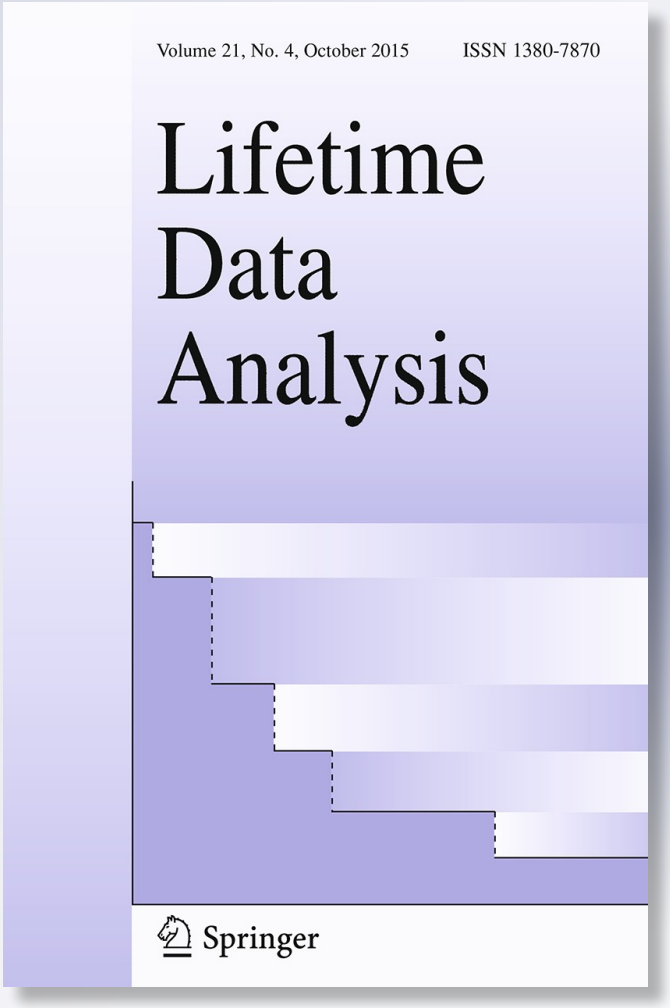
Z-estimation and stratified samples: application to survival models

Norman E. Breslow, Jie Hu & Jon A. Wellner

Lifetime Data Analysis
An International Journal Devoted to Statistical Methods and Applications for Time-to-Event Data

ISSN 1380-7870
Volume 21
Number 4

Lifetime Data Anal (2015) 21:493-516
DOI 10.1007/s10985-014-9317-5



Springer

Lifetime Data Anal (2015) 21:493–516
DOI 10.1007/s10985-014-9317-5

Z-estimation and stratified samples: application to survival models

Norman E. Breslow · Jie Hu · Jon A. Wellner

Received: 17 July 2014 / Accepted: 29 December 2014 / Published online: 15 January 2015

Abstract The infinite dimensional Z-estimation theorem offers a systematic approach to joint estimation of both Euclidean and non-Euclidean parameters in probability models for data. It is easily adapted for stratified sampling designs. This is important in applications to censored survival data because the inverse probability weights that modify the standard estimating equations often depend on the entire follow-up history. Since the weights are not predictable, they complicate the usual theory based on martingales. This paper considers joint estimation of regression coefficients and baseline hazard functions in the Cox proportional and Lin–Ying additive hazards models. Weighted likelihood equations are used for the former and weighted estimating equations for the latter. Regression coefficients and baseline hazards may be combined to estimate individual survival probabilities. Efficiency is improved by calibrating or estimating the weights using information available for all subjects. Although inefficient in comparison with likelihood inference for incomplete data, which is often difficult to implement, the approach provides consistent estimates of desired population parameters even under model misspecification.

Keywords Semiparametric models · Proportional hazards · Additive hazards · Calibration of sampling weights · Model misspecification · Survey sampling

N. E. Breslow (✉) · J. Hu
Department of Biostatistics, University of Washington, Seattle, WA, USA
e-mail: norm@uw.edu

J. Hu
e-mail: hujie0704@gmail.com

J. A. Wellner
Department of Statistics, University of Washington, Seattle, WA, USA
e-mail: jaw@stat.washington.edu

1 Introduction

Large cohort studies provide important evidence regarding causes of disease. The Atherosclerosis Risk in Communities study, for example, has followed a cohort of nearly 16,000 subjects to investigate environmental and genetic factors leading to cardiovascular disease (Williams 1989). Over 26,000 women randomized to hormone therapy or placebo as part of the Women's Health Initiative have been followed similarly to determine disease risks associated with exposure to exogenous estrogens and other risk factors (Anderson et al. 2003). Both studies used stratified random sampling to select limited numbers of stored serum samples for assay of biomarkers. Both also routinely ignored extensive data on standard risk factors available for subjects in the main cohort who were not selected as cases or controls for the biomarker studies. Our goal is to provide a general statistical framework (theory) that will facilitate incorporation of this additional information into the analysis.

In previous work (Breslow and Wellner 2007, 2008) we discussed inference in semiparametric models fitted to stratified samples using inverse probability weighted (IPW) versions of the likelihood equations. We demonstrated how calibration or estimation of sampling weights using information available for the entire cohort improved the precision of regression coefficients, particularly of main effect or interaction terms involving variables known for all (Breslow et al. 2009a, b). We also explored simultaneous inference on both finite and infinite dimensional parameters in semiparametric models, for example, for estimation of individual survival probabilities (Breslow and Lumley 2013).

Most of this earlier work assumed that model assumptions held. Here we focus instead on inference in the face of general misspecification. The basic tool is the infinite dimensional Z-estimation theorem, an extension of Huber's theorem for parametric models (Huber 1967; van der Vaart 1995). We start with a statement of this theorem and demonstrate how, once it has been used to develop properties of estimates based on complete cohort data, the analogous properties of calibrated IPW estimates based on stratified samples quickly follow. The general theory is then applied to asymptotic inference, both on and off the model, when fitting the Cox (1972) proportional and Lin and Ying (1994) additive hazards models to stratified samples. For the most part the mathematical exposition is informal, without close attention being paid to regularity conditions.

2 Huber's theorem and its extension

Huber's (1967) paper "The behavior of maximum likelihood estimates under nonstandard conditions" was a seminal contribution that opened up new research fields in both probability and statistics. Huber's own interest was primarily the development of estimators that had bounded influence functions and hence were less susceptible than usual to the effects of "outliers" (Huber 1980). Royall (1986) emphasized the utility of his results for construction of "robust" variances and confidence intervals for parameters of interest when the assumed parametric probability distribution was mis-

specified. The "sandwich" variance that Huber and others[1] derived ultimately became a key element in generalized estimating equation (GEE) methodology for the analysis of clustered data (Liang and Zeger 1986). Robust variances and confidence intervals are now commonplace in applied statistics and, perhaps predictably, this popularity has provoked some backlash (Freedman 2006).

Let $X_1, \ldots, X_N$ denote a series of independent and identically distributed (i.i.d.) random variables, each with distribution P. Denote expectations by $Pf = \int f dP$ so that $\mathbb{P}_N f = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$, where $\mathbb{P}_N$ is the empirical measure. Here is a version of Huber's theorem due to van der Vaart (1998, Theorem 5.21); see also Bickel et al (1993, Sect. 7.6 and Theorem A.10).

Theorem 1 *For each $\theta \in \Theta \subset \mathbb{R}^p$ let $x \mapsto \psi_\theta(x)$ be a measurable, vector valued estimating function. Define θ_0 by $P\psi_\theta = 0$ and assume that, for every θ_1 and θ_2 in a neighborhood of θ_0, $||\psi_{\theta_1}(x) - \psi_{\theta_2}(x)|| \leq \dot{\psi}(x)||\theta_1 - \theta_2||$ where $\dot{\psi}$ is a measurable function with $P\dot{\psi}^2 < \infty$. Suppose the map $\theta \mapsto P\psi_\theta$ is differentiable at θ_0 with a nonsingular derivative matrix $\dot{\Psi}$ and that $\widehat{\theta}_N$ satisfies $\mathbb{P}_N\left(\psi_{\widehat{\theta}_N}\right) = o_p\left(N^{-1/2}\right)$ and $\widehat{\theta}_N \xrightarrow{p} \theta_0$. Then*

$$\sqrt{N}\left(\widehat{\theta}_N - \theta_0\right) = -\dot{\Psi}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_{\theta_0}(X_i) + o_p(1)$$
$$\rightsquigarrow \mathbb{Z} \sim \mathcal{N}\left(0, \dot{\Psi}^{-1}P\psi_{\theta_0}\psi_{\theta_0}^{\mathrm{T}}\left(\dot{\Psi}^{-1}\right)^{\mathrm{T}}\right)$$

where $\rightsquigarrow$ denotes convergence in distribution, in this case to $\mathbb{Z}$ having a p-dimensional normal distribution with mean 0 and the "sandwich" form of variance.

Often the estimating functions are likelihood scores from a parametric model: $\psi_\theta(x) = \partial \log f_\theta(x)/\partial\theta, \theta \in \Theta$, and the map $\theta \mapsto P \log f_\theta$ has a second-order Taylor expansion with derivative matrix $\dot{\Psi}$. Then θ_0, assumed to be the unique solution to $P\psi_\theta = 0$, maximizes the function $\theta \mapsto P \log f_\theta$ and yields the model distribution closest to P in the sense of Kullback–Leibler distance.

The estimators in Theorem 1 were termed by Huber and others M-estimators since they often maximized the likelihood. We call them Z-estimators to emphasize that in some cases the zeros of estimating equations need not solve a maximization problem (van der Vaart 1998, p. 41). Parameters in the semiparametric models of interest have both finite and infinite dimensional components and a Z-estimation theorem involving infinitely many estimating equations is needed. The version we cite is again due to van der Vaart (1995, 1998, Theorem 19.26). See also van der Vaart and Wellner (1996, Theorem 3.3.1).

Theorem 2 *For each θ in a normed space Θ and every h in an arbitrary set H, let $x \mapsto \psi_{\theta,h}(x)$ be a measurable function such that the class $\left\{\psi_{\theta,h} : ||\theta - \theta_0|| < \delta\right\}$, for*

[1] Godambe (1960) had earlier studied variances based on the "information sandwich", but was concerned with inefficient estimators on the model rather than with misspecification. Cox (1961) derived the sandwich in an informal treatment of tests of separate families of hypotheses, later crediting Huber for a rigorous discussion of the distributional result.

some $\delta > 0$, is contained in a P-Donsker class $\mathscr{F}$ with finite envelope function. Assume that the map $\Psi : \Theta \mapsto \ell^\infty(H)$ given by $\Psi(\theta) = P\psi_\theta$ is Fréchet-differentiable at a zero θ_0, with a derivative $\dot{\Psi} : \text{lin}\Theta \mapsto \ell^\infty(H)$ that has a continuous inverse on its range. Furthermore, assume that $||P\left(\psi_{\theta,h} - \psi_{\theta_0,h}\right)^2||_H \longrightarrow 0$ as $\theta \longrightarrow \theta_0$. If $\|\mathbb{P}_N \psi_{\widehat{\theta}_N}\|_H = o_p(1/\sqrt{N})$ and $\widehat{\theta}_N \xrightarrow{p} \theta_0$, then, uniformly for $h \in H$,

$$\dot{\Psi}\sqrt{N}\left(\widehat{\theta}_N - \theta_0\right)h \;=\; -\mathbb{G}_N \psi_{\theta_0,h} + o_p(1)$$

where

$$\mathbb{G}_N = \sqrt{N}\left(\mathbb{P}_N - P\right) \;\rightsquigarrow\; \mathbb{G} \;\; in \;\; \ell^\infty(\mathscr{F})$$

is the empirical process indexed by the Donsker class $\mathscr{F}$ that converges in distribution to the Brownian bridge process $\mathbb{G}$.

Once again we think of θ_0 as being *defined* as the solution, assumed unique, of the population version of the estimating equations, to which the estimator converges in large samples. Here are two examples that illustrate the content of the theorem.

Example 1 (Empirical DF) Take for Θ the class of distribution functions (DFs) F for a real valued random variable T. Let $H = \mathbb{R}$, the real line, and set $\psi_{F,t} = \mathbf{1}[T \le t] - F(t)$, $t \in \mathbb{R}$. The solution to $\sup_t |\mathbb{P}_N \psi_{\hat{F}_N,t}| = 0$ is the empirical DF $\hat{F}_N(t) = \mathbb{P}_N \mathbf{1}[T \le t] = N^{-1}\sum_{i=1}^N \mathbf{1}[T_i \le t]$. The map $\Psi : \Theta \mapsto \ell^\infty(H)$ is given by $\Psi(F)(t) = (F_0 - F)(t)$, where $F_0(t) = P(T \le t)$ is the true DF. Since $\Psi(F) - \Psi(F^*) = F^* - F$ is already a linear map, its Fréchet derivative $\dot{\Psi}(F - F^*) = \Psi(F) - \Psi(F^*) = F^* - F$ is the negative of the identity map, as is the inverse derivative. The conclusion of the theorem is thus

$$-\dot{\Psi}\sqrt{N}\left(\hat{F}_N - F_0\right) \;=\; \sqrt{N}\left(\hat{F}_N - F_0\right) \;\rightsquigarrow\; \mathbb{G},$$

where $\mathbb{G}$ is the mean zero Gaussian process indexed by $t \in R$ having

$$\text{Cov}(\mathbb{G}(t), \mathbb{G}(s)) = P(\mathbf{1}[T \le t] - F_0(t))(\mathbf{1}[T \le s] - F_0(s)) = F_0(t \wedge s) - F_0(t)F_0(s).$$

This is the classical Donsker theorem (van der Vaart 1998, Theorem 19.3) and the conclusion of Theorem 2, that the class of functions $\mathscr{F} = \{\mathbf{1}[-\infty, t] : t \in \mathbb{R}\}$ is Donsker, is effectively just a restatement of the hypotheses.

Example 2 (Nelson-Aalen estimator) This is the standard nonparametric estimator of the cumulative hazard from censored observations $X = (T, \Delta)$ where T is the survival time, here assumed absolutely continuous, and Δ is the 0/1 indicator of whether T is censored or fully observed. We take for Θ a collection of functions of the form $\Lambda(t) = \int_0^t \lambda(s)ds$, where λ denotes the hazard, that are uniformly bounded over a finite interval $[0, \tau]$. Let H denote the unit ball in the space $\text{BV}[0, \tau]$ of bounded functions of bounded variation on $[0, \tau]$. Set $\mathbb{N}(t) = \Delta \cdot \mathbf{1}[T \le t]$ the usual counting process, $Y(t) = \mathbf{1}[T \ge t]$ the "at risk" process, $M_\Lambda(t) = \mathbb{N}(t) - \int_0^t Y(s)d\Lambda(s)$ and

suppose $PY(\tau) > 0$. M_Λ is a martingale when $\Lambda = \Lambda_0$, the true cumulative hazard under P. Let

$$\psi_{\Lambda,h}(X) \;=\; \int_0^\tau h dM_\Lambda \;=\; \int_0^\tau h\,(d\mathbb{N} - Yd\Lambda) \;=\; \Delta h(T) - \int_0^\tau hYd\Lambda.$$

These functions form a Donsker class since h and $\int hYd\Lambda$ are of uniformly bounded variation (van der Vaart 1998, Example 19.11). Setting $h_t(s) = \mathbf{1}[s \le t]/\mathbb{P}_N Y(s)$ for arbitrary $t \in [0, \tau]$,[2] the solution $\hat{\Lambda}_N$ to $||\mathbb{P}_N \psi_{\hat{\Lambda}_N}||_H = 0$ satisfies

$$\mathbb{P}_N \psi_{\hat{\Lambda}_N, h_t} = \mathbb{P}_N \left\{ \frac{\Delta \cdot \mathbf{1}[T \le t]}{(\mathbb{P}_N Y)(T)} \right\} - \int_0^\tau \frac{\mathbf{1}[s \le t]\mathbb{P}_N Y(s)}{\mathbb{P}_N Y(s)} d\hat{\Lambda}_N(s) \;=\; 0,$$

$$i.e., \quad \hat{\Lambda}_N(t) = \mathbb{P}_N \left\{ \frac{\Delta \cdot \mathbf{1}[T \le t]}{(\mathbb{P}_N Y)(T)} \right\},$$

the estimator proposed by Nelson (1972) and Aalen (1976). We calculate

$$\Psi(\Lambda)h \;=\; P\psi_{\Lambda,h} \;=\; \int_0^\tau hPY\, d\Lambda_0 - \int_0^\tau hPY\, d\Lambda \;=\; -\int_0^\tau hPY\, d(\Lambda - \Lambda_0).$$

Since for Λ, $\Lambda^* \in \Theta$, $\Psi(\Lambda) - \Psi(\Lambda^*)$ is linear in $\Lambda - \Lambda^*$, this difference equals the Fréchet derivative:

$$\dot{\Psi}(\Lambda - \Lambda^*)h \;=\; \Psi(\Lambda)h - \Psi(\Lambda^*)h \;=\; -\int_0^\tau hP(Y)d(\Lambda - \Lambda^*).$$

For η in the range of $\dot{\Psi} \subset \ell^\infty(H)$ given by $\eta h = \int_0^\tau h d\eta$, the inverse map is thus

$$\dot{\Psi}^{-1}(\eta)h \;=\; -\int_0^\tau \frac{h}{PY} d\eta.$$

We conclude from Theorem 2 that, for random processes indexed by $h \in H$,

$$\sqrt{N}\left(\hat{\Lambda}_N - \Lambda_0\right)h \;=\; -\dot{\Psi}^{-1}\mathbb{G}_N\left(\psi_{\Lambda_0,h}\right) + o_p(1) \;\rightsquigarrow\; \mathbb{G}\left(\dot{\Psi}^{-1}\psi_{\Lambda_0,h}\right).$$

Thus, with h_t now given by $h_t(s) = \mathbf{1}[s \le t]$, $\sqrt{N}\left(\hat{\Lambda}_N - \Lambda_0\right)(t) = \sqrt{N}\left(\hat{\Lambda}_N - \Lambda_0\right)h_t$, $t \in [0, \tau]$, is an asymptotically Gaussian stochastic process with covariance function

$$P\left(\dot{\Psi}^{-1}\psi_{\Lambda_0,h_t}\dot{\Psi}^{-1}\psi_{\Lambda_0,h_s}\right) = P\left(\int_0^\tau \frac{h_t}{PY} dM_0 \cdot \int_0^\tau \frac{h_s}{PY} dM_0\right)$$
$$= \int_0^\tau \frac{h_t h_s}{(PY)^2} PY d\Lambda_0 \;=\; \int_0^{t\wedge s} \frac{d\Lambda_0(u)}{P(T \ge u)},$$

[2] Although h_t is not itself in H, it is of bounded variation and hence may be renormalized to be in H, which is all that is needed in the sequel since the estimating equations are linear in h.

where $M_0 = \mathbb{N} - \int_0^\tau Y d\Lambda_0$ is a martingale and we have used standard results for the covariance of martingale integrals (Aalen et al. 2008, Eqs. 2.31, 2.43). This asymptotic distribution for $\hat{\Lambda}_N$ was established by Breslow and Crowley (1974). The example illustrates several features that arise in the application of Theorem 2 to survival models under both simple random and stratified sampling.

We now turn to the modifications of Theorem 2 needed when $\widehat{\theta}_N$ solves, at least to order $o_p(1/\sqrt{N})$, an IPW version of the estimating equations.

3 Two-phase stratified sampling

As mentioned in the Introduction, the studies that motivated this work involved stratified random sampling from a large cohort, the "Phase I sample", to select subjects for a smaller Phase II sample for whom additional covariates were ascertained. Thus X is not fully observed for all N subjects. It is commonplace, however, that some *auxiliary* variables U are observed for all subjects. These are most useful if correlated with the portions of X not observed for everyone. They are used for selection of the Phase II subjects, to attempt to maximize their informativeness vis-à-vis study hypotheses, and for calibration or estimation of the weights, so as to bring into the analysis more of the Phase I information. We denote by $V \in \mathcal{V}$ the auxiliary variables U plus the portions of X observed for all subjects. $R_1, \ldots, R_N$ denote sampling indicators that indicate whether ($R_i = 1$) or not ($R_i = 0$) the i^{th} subject is selected at Phase II.

3.1 Finite population stratified sampling

The Phase II sample is typically selected using stratified sampling. We partition $\mathcal{V}$ into J strata: $\mathcal{V} = \mathcal{V}_1 \cup \cdots \cup \mathcal{V}_J$, count the number N_j of Phase I subjects in stratum j, and sample n_j of them at random without replacement. The resulting sample sizes are as shown in Table 1.

Previously (Breslow and Wellner 2007) we considered IPW estimation in semiparametric models under finite population stratified sampling, assuming the truth of the model. The sampling indicators satisfy $\sum_{i=1}^N R_i \mathbf{1}(V_i \in \mathcal{V}_j) = n_j$, where the n_j are fixed by the experimenter. Hence they are dependent random variables, albeit exchangeable within strata, which rules out application of Theorem 2. Many researchers prefer to develop their theory under Bernoulli sampling, which preserves the i.i.d. structure of the problem, and this is the path we now follow. In the sequel we show how results obtained under the two sampling designs are related.

Table 1 Two-phase stratified sampling

	Stratum				
	1	2	$\cdots$	J	Total
Phase I	N_1	N_2	$\cdots$	N_J	N
Phase II	n_1	n_2	$\cdots$	n_J	n
Sampling fractions	$\frac{n_1}{N_1}$	$\frac{n_2}{N_2}$	$\cdots$	$\frac{n_J}{N_J}$	$\frac{n}{N}$

3.2 Bernoulli sampling

Instead of observing the N values of V all at once for the Phase I subjects, suppose instead they are examined one by one. As subjects enter the study we generate, independently for each one, a Phase II sampling indicator R_i with

$$\Pr(R = 1|U, X) = \Pr(R = 1|V) = \pi_0(V)$$

where π_0 is a known function such that $\pi_0(v) \geq \delta > 0$ for all $v \in \mathcal{V}$. Since π_0 depends only on V, the portions of X unobserved for subjects not sampled at Phase II are "missing at random" (Little and Rubin 2002). For stratified Bernoulli sampling, $\pi_0(v) = p_j$ for $v \in \mathcal{V}_j$ where the p_j are known probabilities. This setup preserves the i.i.d. structure of the observations $\{(V_i, R_i, R_i X_i), i = 1, \ldots, N\}$.

4 Inverse probability weighted Z-estimation

We extend P from a distribution for X alone to a distribution for (R, U, X). Define the IPW empirical measure $\mathbb{P}_N^\pi$ via

$$\mathbb{P}_N^\pi(f) \;=\; \mathbb{P}_N\left(\frac{R}{\pi_0} f\right) \;=\; \frac{1}{N}\sum_{i=1}^{N} \frac{R_i}{\pi_0(V_i)} f\left(X_i\right).$$

4.1 Application of Theorem 2

Suppose that the conditions of Theorem 2 have been verified for an i.i.d. random sample $X_1, \ldots, X_N$ but that $\widehat{\theta}_N$ now satisfies

$$\left\|\mathbb{P}_N^\pi\left(\psi_{\widehat{\theta}_N}\right)\right\|_H \;=\; \sup_{h\in H}\left|\frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\pi_0(V_i)}\psi_{\widehat{\theta}_N}\left(X_i\right)h\right| \;=\; o_p\left(N^{-\frac{1}{2}}\right).$$

In our applications this equation is satisfied exactly, *i.e.*, with 0 replacing $o_p(N^{-\frac{1}{2}})$. Consistency of the IPW version of $\widehat{\theta}_N$ generally follows from consistency of the ordinary version and the fact that the sampling probabilities are bounded away from zero. Since $P\left(\frac{R}{\pi_0}\psi_\theta\right) = P\left(\psi_\theta\right)$ for all θ, the parameter θ_0 estimated by this approach is the same as would be estimated by fitting the model to complete Phase I data. This is especially important when the model has been misspecified. The derivative $\dot{\Psi}$ is also unchanged, which means that most of the work in determining the asymptotic distribution of $\widehat{\theta}_N$ has already been accomplished.

The conclusion of Theorem 2 for the IPW estimator is thus, uniformly for $h \in H$,

$$\begin{aligned}\dot{\Psi}\sqrt{N}\left(\widehat{\theta}_N - \theta_0\right)h \;&=\; -\mathbb{G}_N\left(\frac{R}{\pi_0}\psi_{\theta_0,h}\right) + o_p(1).\\ &=\; -\mathbb{G}_N\left(\psi_{\theta_0,h} + \frac{R-\pi_0}{\pi_0}\psi_{\theta_0,h}\right) + o_p(1). \qquad (1)\end{aligned}$$

The two terms in parentheses on the RHS are *uncorrelated*. Hence the asymptotic variance is

$$\mathrm{Var}_{\mathrm{A}}\left[\dot{\psi}\sqrt{N}\left(\widehat{\theta}_N-\theta_0\right)h\right] \;=\; P\left(\psi^2_{\theta_0,h}\right)+P\left(\frac{1-\pi_0}{\pi_0}\psi^2_{\theta_0,h}\right).$$

This equals the variance obtained by fitting the model to complete data (the Phase I variance) plus a term that reflects the loss of information due to the Phase II sampling. Efficiency is improved by reducing this second term, the Phase II variance.

4.2 Calibration of the sampling weights

Survey samplers (Deville and Särndal 1992) improve their estimates of finite population totals by "calibration" of the sampling weights using known population totals of auxiliary variables associated with the variable of interest. The same technique may be applied to improve the efficiency of IPW estimates in semiparametric models. Let $C=(C_1,C_2,\ldots,C_K)^{\mathrm{T}}$, where $C_k=C_k(V)$, denote a vector of variables known for all Phase I subjects and suppose PCC^{T} is nonsingular. The basic idea is to select a new set of weights w_i that are as close as possible to the sampling or "design" weights, $d_i=\pi_0^{-1}(V_i)$, according to a specified distance measure G. The new weights must also satisfy the *calibration equations*, whereby the population totals of the C_k are exactly estimated: $\sum_{i=1}^N C_k(V_i)=\sum_{i=1}^N R_i w_i C_k(V_i)$. When the distance measure is the Poisson deviance, $G(w,d)=w\log(w)-w+d$, this is accomplished by choosing new weights

$$w_i \;=\; \frac{\exp\left[-\widehat{\lambda}_N^{\mathrm{T}}C\left(V_i\right)\right]}{\pi_0\left(V_i\right)}$$

where $\widehat{\lambda}_N$ is a K-vector of Lagrange multipliers for the constrained (by the calibration equations) optimization problem. Known in the survey literature as "raking", the procedure yields weights that are always positive; other choices for G may not.

To derive the asymptotic behavior of the estimator $\widehat{\theta}_N(\widehat{\lambda}_N)$ obtained by IPW using calibrated weights, we use a variant of the Z-theorem with estimated nuisance parameters (Breslow and Wellner 2008). Consider λ as a parameter in the calibrated IPW equations

$$\mathbb{P}_N\left(\frac{R}{\pi_\lambda(V)}\psi_\theta\right) \;=\; 0,\quad \text{where}\quad \pi_\lambda(V) \;=\; \exp\left[\lambda^{\mathrm{T}}C(V)\right]\pi_0(V),$$

and suppose that $\widehat{\lambda}_N$ is the estimator that solves the calibration equations

$$\sum_{i=1}^N \frac{R_i}{\pi_{\widehat{\lambda}_N}(V_i)}C\left(V_i\right) \;=\; \sum_{i=1}^N C\left(V_i\right).$$

Under standard regularity conditions for design based inference, which hold for our two-phase sampling setup, or arguing directly from the preceding equation, one may show (Deville and Särndal 1992; Breslow et al. 2009a)

$$\sqrt{N}\widehat{\lambda}_N \;=\; \left(PCC^{\mathrm{T}}\right)^{-1}\mathbb{G}_N\left(\frac{R-\pi_0}{\pi_0}C\right)+o_p(1). \tag{2}$$

Thus $\widehat{\lambda}_N$ converges in probability to zero as $N \uparrow \infty$ and the calibrated weights converge to the design weights. Suppose also that the map

$$\lambda \mapsto P\left(\frac{R}{\pi_\lambda(V)}\psi_{\theta_0,h}(X)\right)$$

is differentiable at $\lambda = 0$, uniformly in h, with derivative $\dot{\Psi}_{\lambda,h} = -P\psi_{\theta_0,h}C^{\mathrm{T}}$. Combining Eqs. (1) and (2) with Theorem 1(iv) of Breslow and Wellner (2008) we then have, uniformly for $h \in H$,

$$\begin{aligned}
&\dot{\psi}\sqrt{N}\left(\widehat{\theta}_N(\widehat{\lambda}_N)-\theta_0\right)h = \dot{\psi}\sqrt{N}\left(\widehat{\theta}_N(0)-\theta_0\right)h - \dot{\Psi}_{\lambda,h}\sqrt{N}\widehat{\lambda}_N + o_p(1)\\
&= -\mathbb{G}_N\left\{\psi_{\theta_0,h}+\frac{R-\pi_0}{\pi_0}\left[\psi_{\theta_0,h}-P\psi_{\theta_0,h}C^{\mathrm{T}}\left(PCC^{\mathrm{T}}\right)^{-1}C\right]\right\}\\
&\quad + o_p(1)\\
&= -\mathbb{G}_N\left[\psi_{\theta_0,h}+\frac{R-\pi_0}{\pi_0}\left(\psi_{\theta_0,h}-\Pi_C\psi_{\theta_0,h}\right)\right]+o_p(1)
\end{aligned} \tag{3}$$

where Π_C denotes population least squares projection on $[C_1,\ldots,C_K]$. Comparing (3) with (1), the advantage of calibration is that we replace the scores in the Phase II variance term with the residuals after their projection on $[C_1,\ldots,C_K]$. The improvement in precision would be greatest if we could select the calibration variables to be highly correlated with the scores (Lumley 2012, Sect. 8.5.1).[3]

To avoid having to write out expressions like (3) in the sequel, we define

$$\mathbb{G}_N^{\widehat{\pi}}(f) \;=\; \mathbb{G}_N\left\{f+\frac{R-\pi_0}{\pi_0}\left[f-\Pi_C(f)\right]\right\},\quad f\in\mathscr{F}, \tag{4}$$

and refer to it as the (calibrated) IPW empirical process.

4.3 Calibrating to stratum totals

When the calibration variables are simply the stratum indicators, $C_j = \mathbf{1}[V \in \mathscr{V}_J]$, the projection onto $[C_1,\ldots,C_J]$ yields the stratum specific mean:

[3] Indeed, the term $\mathbb{G}_N[(R-\pi_0)/\pi_0]\psi_{\theta_0,h}$ in (1), which has the same limiting distribution whether the $\psi_{\theta_0,h}$ are regarded as random or fixed by conditioning (van der Vaart and Wellner 1996, Sect. 2.9), is the normalized error arising from IPW estimation of the Phase I total of the scores. The solution to the sample survey problem, to estimate this unknown total using two phase stratified sampling, is best achieved when the calibration variables used to adjust the sampling weights are highly correlated with the scores.

$$\Pi_C(f) = \sum_{j=1}^{J} P_j(f) C_j, \text{ where } P_j(A) = P(A \cap \mathscr{V}_j)/P(\mathscr{V}_j).$$

Whereas the design weights are inverses of the *a priori* sampling probabilities p_j, the calibrated weights are inverses of the actual sampling fractions n_j/N_j. In view of (3), furthermore, with Var_j denoting the stratum-specific variance,

$$\text{Var}_A\left[\dot{\psi}\sqrt{N}\left(\widehat{\theta}_N(\widehat{\lambda}_N) - \theta_0\right)h\right] = P\left(\psi^2_{\theta_0,h}\right) + \sum_{j=1}^{J} P\left(\mathscr{V}_j\right)\frac{1-p_j}{p_j}\text{Var}_j\left(\psi_{\theta_0,h}\right).$$

This is precisely the asymptotic variance that Breslow and Wellner (2007) derived for IPW estimation under finite population stratified sampling, where the design weights were the observed N_j/n_j.[4] Hence calibration to stratum frequencies enables one to reconcile the apparent difference between the two sampling schemes. Further calibration would in principle improve the efficiency of estimation under either scheme (Saegusa and Wellner 2013).

5 Semiparametric models

The preceding discussion considered a single infinite dimensional parameter θ. In semiparametric models one works with a parameter $\theta = (\beta, \Lambda)$ that is partitioned into a parametric part, $\beta \in \Gamma \subset \mathbb{R}^p$, and a nonparametric part, $\Lambda \in \mathscr{H} \subset \mathscr{B}$ where $\mathscr{B}$ is a normed space. We consider the special case where Λ is a finite measure. We also assume that the conditions of Theorem 2 hold so that $\sqrt{N}\left(\widehat{\beta}_N - \beta_0, \widehat{\Lambda}_N - \Lambda_0\right)$ is asymptotically Gaussian under i.i.d. random sampling.

For the survival models, $X = (T, \Delta, Z) : \quad 0 \le T \le \tau < \infty, \quad \Delta \in \{0, 1\}, \quad Z \in \mathbb{R}^p$, where T is the right-censored survival time, Δ the censoring indicator, and Z a p-dimensional vector of covariates. The β are regression coefficients, which may be interpreted as log hazard ratios or as excess hazards depending on the model. Λ denotes the baseline hazard function, which we interpret as a measure on $[0, \tau]$. We further impose the "partly unnecessary" assumptions made by van der Vaart (1998, Sect. 25.12.1) to guarantee applicability of Theorem 2 to fitting the Cox model with complete Phase I data. These include that Z is bounded, that the survival and censoring distributions are continuous and that $P[T \ge \tau] > 0$, where τ is a time at which a non-zero proportion of the cohort is still "at risk" of "death". See Breslow and Wellner (2007, Sect. 7) for a more complete statement.

[4] This result would be of no surprise to a survey sampler. For estimation of a population total using stratified Bernoulli sampling, it is well known that conditioning on the Phase II stratum totals $\{n_1, \ldots, n_J\}$ (see Table 1) is equivalent to finite population stratified sampling (Särndal et al. 1992, Sect. 9.8, Example 9.14).

5.1 Semiparametric likelihood equations

We follow closely the development in van der Vaart (1998, Sect. 25.12) for the model $P_{\beta,\Lambda}(X)$ with density $p_{\beta,\Lambda}(x)$. Let $\dot{\ell}_{\beta,\Lambda}$ denote the usual likelihood scores for β

$$\dot{\ell}_{\beta,\Lambda} = \frac{\partial \log p_{\beta,\Lambda}}{\partial \beta}$$

and let $B_{\beta,\Lambda}$ denote the score operator that maps directions $h \in H$, from which Λ is approached by paths $\Lambda_{t,h}$ in one dimensional submodels given by $d\Lambda_{t,h} = (1+ht)d\Lambda$, into the corresponding likelihood scores:

$$B_{\beta,\Lambda} h \;=\; \left.\frac{\partial \log p_{\beta,\Lambda_{t,h}}}{\partial t}\right|_{t=0}.$$

Then the IPW version of the likelihood equations with calibrated weights is

$$\begin{aligned}
\mathbb{P}_N^{\widehat{\pi}} \dot{\ell}_{\beta,\Lambda} &= \frac{1}{N}\sum_{i=1}^{N} \frac{R_i}{\pi_{\widehat{\lambda}_N}(V_i)} \dot{\ell}_{\beta,\Lambda}(X_i) = 0 \\
\mathbb{P}_N^{\widehat{\pi}} B_{\beta,\Lambda} h &= \frac{1}{N}\sum_{i=1}^{N} \frac{R_i}{\pi_{\widehat{\lambda}_N}(V_i)} B_{\beta,\Lambda} h(X_i) = 0, \quad h \in H,
\end{aligned}$$

where by $\mathbb{P}_N^{\widehat{\pi}}$ we mean the IPW empirical distribution using calibrated weights:

$$\mathbb{P}_N^{\widehat{\pi}} f \;=\; \mathbb{P}_N\left(\frac{R}{\pi_{\widehat{\lambda}_N}} f\right) \;=\; \frac{1}{N}\sum_{i=1}^{N} \frac{R_i}{\pi_{\widehat{\lambda}_N}(V_i)} f(X_i).$$

5.2 Semiparametric inference under the model

When the model holds, *i.e.*, $P = P_0 = P_{\beta_0,\Lambda_0}$, the Fréchet derivative $\dot{\Psi}$ that figures in Theorem 2 may often be expressed in terms of $\dot{\ell}_0 = \dot{\ell}_{\beta_0,\Lambda_0}$, $B_0 = B_{\beta_0,\Lambda_0}$ and B_0^*, the adjoint of B_0, all evaluated at the "true value" (van der Vaart 1998, Eq. 25.91). For β estimation we then find (van der Vaart 1998, Sect. 25.5.1), see also (Bickel et al. 1993, Chap. 5),

$$\begin{aligned}
\ell_0^* &= \left[I - B_0\left(B_0^* B_0\right)^{-1} B_0^*\right]\dot{\ell}_0, \text{ the efficient score,} \\
\tilde{I}_0 &= P_0 \ell_0^* \left(\ell_0^*\right)^{\mathrm{T}}, \text{ the efficient information, and} \\
\tilde{\ell}_0 &= \tilde{I}_0^{-1} \ell_0^*, \text{ the efficient influence function.}
\end{aligned}$$

Invertibility of the information operator $B_0^* B_0$ follows from the assumed invertibility of $\dot{\Psi}$. Further defining the operator $A : H \mapsto L_2(P_0)$ by

$$Ah \;=\; B_0\left(B_0^* B_0\right)^{-1} h - P_0\left[B_0\left(B_0^* B_0\right)^{-1} h \dot{\ell}_0^{\mathrm{T}}\right]\tilde{\ell}_0,$$

and completing the arguments in van der Vaart (1998, p. 424) as detailed in Breslow and Lumley (2013), the conclusion of the Z-theorem shown in (3) for IPW estimation with calibrated weights becomes

$$\sqrt{N}\left(\widehat{\beta}_N(\widehat{\lambda}_N) - \beta_0\right) = \mathbb{G}_N^{\widehat{\pi}}(\tilde{\ell}_0) + o_p(1) \tag{5}$$

$$\sqrt{N}\left(\widehat{\Lambda}_N(\widehat{\lambda}_N) - \Lambda_0\right) h = \mathbb{G}_N^{\widehat{\pi}}(Ah) + o_p(1). \tag{6}$$

These expansions provide the basics needed for asymptotic inference.

6 Applications to survival models

Cox model based inferences follow directly from Sect. 5.2. These are considered first, followed by separate applications of Theorem 2 for "robust" inferences under general misspecification for estimating parameters in both the Cox and the Lin–Ying models.

6.1 Cox regression on the model

Recall that $X = (T, \Delta, Z)$ for the survival models. Let $\mathbb{N}(t)$ and $Y(t)$ denote the counting and "at risk" processes as in Example 2. Define

$$M_{\beta,\Lambda}(t) = \mathbb{N}(t) - \int_0^t e^{Z^{\mathrm{T}}\beta} Y(s) d\Lambda(s). \tag{7}$$

$M_0 = M_{\beta_0,\Lambda_0}$ is a martingale under the model P_0. The likelihood scores may then be expressed

$$\dot{\ell}_{\beta,\Lambda}(X) = \int_0^\tau Z dM_{\beta,\Lambda}$$

$$B_{\beta,\Lambda} h(X) = \int_0^\tau h dM_{\beta,\Lambda}, \quad h \in H,$$

where H is again taken to be the unit ball in $\mathrm{BV}[0, \tau]$. Then van der Vaart (1998, Sect. 25.12.1) showed

$$B_0^* \dot{\ell}_0 = P_0 Z e^{Z^{\mathrm{T}}\beta_0} Y$$

$$B_0^* B_0 h = h P_0 e^{Z^{\mathrm{T}}\beta_0} Y$$

$$\left(B_0^* B_0\right)^{-1} h = h / P_0 e^{Z^{\mathrm{T}}\beta_0} Y.$$

Setting $m(t) = P_0 Z e^{Z^{\mathrm{T}}\beta_0} Y(t) / P_0 e^{Z^{\mathrm{T}}\beta_0} Y(t) = P_0(Z|T = t, \Delta = 1)$, the conditional expectation of Z for a subject who dies at t, it follows that

$$\ell_0^* = \int_0^\tau [Z - m]\, dM_0 \,,$$

$$\tilde{I}_0 = P_0 e^{Z^{\mathrm{T}}\beta_0} \int_0^\tau [Z - m]^{\otimes 2}\, Y d\Lambda_0 \qquad \text{and}$$

$$Ah = \int_0^\tau \frac{h}{P_0 e^{Z^{\mathrm{T}}\beta_0} Y} dM_0 - P_0 \left(\int_0^\tau \frac{h}{P_0 e^{Z^{\mathrm{T}}\beta_0} Y} dM_0 \dot{\ell}_0^{\mathrm{T}} \right) \tilde{\ell}_0. \tag{8}$$

The joint asymptotic distribution of $(\widehat{\beta}_N, \widehat{\Lambda}_N)$ is obtained by plugging these expressions into (5) and (6). For the special case that all Phase I subjects have complete data, *i.e.*, under simple random sampling with $\widehat{\beta}_N$ and $\widehat{\Lambda}_N$ denoting the ordinary, unweighted estimates, we have

$$\sqrt{N}\left(\widehat{\beta}_N - \beta_0\right) = \mathbb{G}_N\left(\tilde{\ell}_0\right) + o_p(1)$$

$$\sqrt{N}\left(\widehat{\Lambda}_N - \Lambda_0\right) h = \mathbb{G}_N \left[\int_0^\tau \frac{h}{P_0 e^{Z^{\mathrm{T}}\beta_0} Y} dM_0 - P_0 \left(\int_0^\tau \frac{h}{P_0 e^{Z^{\mathrm{T}}\beta_0} Y} dM_0 \dot{\ell}_0^{\mathrm{T}} \right) \tilde{\ell}_0 \right] + o_p(1)$$

with $\tilde{\ell}_0 = \tilde{I}_0^{-1} \ell_0^*$ as shown above. The asymptotic expansions for calibrated IPW estimators are found by replacing $\mathbb{G}_N$ in these equations by the $\mathbb{G}_N^{\widehat{\pi}}$ in (4).

Breslow and Lumley (2013) used Taylor's formula with the preceding equations to derive a well known result for the estimated cumulative hazard for a subject with $Z = z_0$. For simple random sampling they showed

$$\sqrt{N}\left[e^{z_0^{\mathrm{T}}\widehat{\beta}_N} \widehat{\Lambda}_N(t) - e^{z_0^{\mathrm{T}}\beta_0} \Lambda_0(t) \right] = e^{z_0^{\mathrm{T}}\beta_0} \mathbb{G}_N \left\{ \int_0^t \frac{dM_0}{P_0 e^{Z^{\mathrm{T}}\beta_0} Y} + \left[\int_0^t (z_0 - m)^{\mathrm{T}} d\Lambda_0 \right] \tilde{\ell}_0 \right\} + o_p(1) \tag{9}$$

uniformly for $0 \le t \le \tau$. Since the efficient influence function $\tilde{\ell}_0$ is constructed to be orthogonal in $L_2(P_0)$ to the closure of the range space of B_0, *i.e.*, to the "nuisance tangent space", the two terms inside the curly brackets are uncorrelated. Hence the process (9) converges in distribution to the mean zero process

$$e^{z_0^{\mathrm{T}}\beta_0} \left[\mathbb{Z}(t) + \int_0^t (z_0 - m)^{\mathrm{T}} d\Lambda_0 \cdot \mathbb{Z}_* \right] \tag{10}$$

where $\mathbb{Z}$ and $\mathbb{Z}_*$ are independently Gaussian. Using the martingale calculus, the covariance function of $\mathbb{Z}$ is

$$\mathrm{Cov}\left(\mathbb{Z}(t), \mathbb{Z}(s)\right) = \int_0^{t \wedge s} \frac{1}{P_0 e^{Z^{\mathrm{T}}\beta_0} Y} d\Lambda_0,$$

while the covariance matrix of Z_* is the inverse of the efficient information $\tilde{I}_0$. This result agrees with that of Begun et al. (1983), who had an additional term

$\exp(-e^{z_0^{\mathrm{T}}\beta_0}\Lambda_0)$ multiplying (9) and (10) since they worked with the survival instead of the cumulative hazard function. See also Tsiatis (1981) and Andersen and Gill (1982).

For calibrated IPW estimation one replaces $\mathbb{G}_N$ by $\mathbb{G}_N^{\widehat{\pi}}$. The first term in the asymptotic variance function is that for complete Phase I data as just described. Calculation of the Phase II variance is more involved. For stratified Bernoulli sampling, this involves terms of the form

$$\sum_{j=1}^{J} P_0(\mathscr{Y}_j)\frac{1-p_j}{p_j}\mathrm{Cov}_{0,j}\left[f_t-\Pi_C(f_t),\, f_s-\Pi_C(f_s)\right]$$

where

$$f_t = e^{z_0^{\mathrm{T}}\beta}\left\{\int_0^t \frac{dM_0}{P_0 e^{Z^{\mathrm{T}}\beta_0}Y} + \left[\int_0^t (z_0-m)^{\mathrm{T}} d\Lambda_0\right]\tilde{\ell}_0\right\}$$

and where $\mathrm{Cov}_{0,j}$ denotes covariance under the distribution $P_{0,j}$. Unfortunately, the two terms between curly brackets in f_t are not orthogonal in $L_2(P_{0,j})$.

6.2 Cox regression off the model

Suppose now $(\widehat{\beta}_N, \widehat{\Lambda}_N)$ solves the estimating equations shown in Sect. 5.1, but that P does not necessarily satisfy the Cox model. One must then apply Theorem 2 directly rather than rely on the general results of Sect. 5.2. Note that $M_{\beta,\Lambda}(t)$ as defined in (7) generally is not a martingale under P, not even at (β_0, Λ_0). The calibrated IPW estimating equations may be written

$$\mathbb{P}_N^{\widehat{\pi}}\int Z dM_{\beta,\Lambda} = \mathbb{P}_N^{\widehat{\pi}}\left[\Delta Z - \Lambda\left(Ze^{Z^{\mathrm{T}}\beta}Y\right)\right] = 0 \tag{11}$$

$$\mathbb{P}_N^{\widehat{\pi}}\int h dM_{\beta,\Lambda} = \mathbb{P}_N^{\widehat{\pi}}\left[\Delta h(T) - \Lambda\left(he^{Z^{\mathrm{T}}\beta}Y\right)\right] = 0, \quad h \in H. \tag{12}$$

Recall that $\Lambda(h) = \int_0^\tau h d\Lambda$ for Λ a measure.

These equations are easily solved. Substituting $\mathbf{1}[s \le t]/P_N^{\widehat{\pi}} e^{Z^{\mathrm{T}}\beta}Y(s)$ for $h(s)$ in (12) and arguing as in Example 2 leads to

$$\widehat{\Lambda}_N(\beta)(t) = \mathbb{P}_N^{\widehat{\pi}}\left(\frac{\Delta \mathbf{1}[T \le t]}{\left(P_N^{\widehat{\pi}} e^{Z^{\mathrm{T}}\beta}Y\right)(T)}\right),$$

an IPW version of the Breslow estimator. Inserting this $\widehat{\Lambda}_N(\beta)$ in (11), the resulting $\widehat{\beta}_N$ solves

$$\mathbb{P}_N^{\widehat{\pi}}\Delta\left(Z - \frac{\mathbb{P}_N^{\widehat{\pi}} Ze^{Z^{\mathrm{T}}\beta}Y}{\mathbb{P}_N^{\widehat{\pi}} e^{Z^{\mathrm{T}}\beta}Y}(T)\right) = 0,$$

an IPW version of the Cox partial likelihood equations. Similarly, the "true values" of the parameters $(\beta_0, \Lambda_0) = (\beta_0(P), \Lambda_0(P))$ are now *defined* as the solutions to the population version of the estimating equations $\Psi_{\beta,\Lambda} = 0$, which we write as

$$\begin{aligned} \Psi_{1;\beta,\Lambda} &= P \int Z dM_{\beta,\Lambda} = P\Delta Z - \Lambda\left(PZe^{Z^{\mathrm{T}}\beta}Y\right) = 0 \\ \Psi_{2;\beta,\Lambda} h &= P \int h dM_{\beta,\Lambda} = P\Delta h(T) - \Lambda\left(hPe^{Z^{\mathrm{T}}\beta}Y\right) = 0, \quad h \in H. \end{aligned} \tag{13}$$

Taking $h = PZe^{Z^{\mathrm{T}}\beta}Y / Pe^{Z^{\mathrm{T}}\beta}Y$, and subtracting, β_0 solves

$$P\Delta\left[Z - \frac{PZe^{Z^{\mathrm{T}}\beta}Y}{Pe^{Z^{\mathrm{T}}\beta}Y}(T)\right] = 0.$$

The arguments of Struthers and Kalbfleisch (1986) may be adapted to demonstrate that, under standard regularity conditions, $\widehat{\beta}_N \xrightarrow{p} \beta_0$. From this it follows that

$$\widehat{\Lambda}_N(t) = \mathbb{P}_N^{\widehat{\pi}}\left(\frac{\Delta \mathbf{1}[T \le t]}{\left(\mathbb{P}_N^{\widehat{\pi}} e^{Z^{\mathrm{T}}\widehat{\beta}_N}Y\right)(T)}\right) \xrightarrow{p} \Lambda_0(t) = P\left(\frac{\Delta \mathbf{1}[T \le t]}{\left(Pe^{Z^{\mathrm{T}}\beta_0}Y\right)(T)}\right)$$

uniformly for $0 \le t \le \tau$.

Partitioning $\dot{\Psi}$ in the same fashion as Ψ in (13), the conclusion of Theorem 2 is

$$\dot{\Psi}_{11}\sqrt{N}\left(\hat{\beta}_N - \beta_0\right) + \dot{\Psi}_{12}\sqrt{N}\left(\hat{\Lambda}_N - \Lambda_0\right) = -\mathbb{G}_N^{\widehat{\pi}}\left(\int Z dM_0\right) + o_p(1) \tag{14}$$

$$\dot{\Psi}_{21}\sqrt{N}\left(\hat{\beta}_N - \beta_0\right) h + \dot{\Psi}_{22}\sqrt{N}\left(\hat{\Lambda}_N - \Lambda_0\right) h = -\mathbb{G}_N^{\widehat{\pi}}\left(\int h dM_0\right) + o_p(1), \tag{15}$$

which result may be compared with van der Vaart (1998, Theorem 25.90). Here $M_0 = M_{\beta_0,\Lambda_0}$ has mean zero in view of (13) but, as a reminder, in general is not a martingale. The components of $\dot{\Psi}$ are readily found from (13); in fact, $\dot{\Psi}_{12}$ and $\dot{\Psi}_{22}$ follow immediately from the linearity of these equations in Λ:

$$\begin{aligned} \dot{\Psi}_{11}(\beta - \beta_0) &= -\Lambda_0\left(PZ^{\otimes 2}e^{Z^{\mathrm{T}}\beta_0}Y\right)(\beta - \beta_0) \\ \dot{\Psi}_{12}(\Lambda - \Lambda_0) &= -(\Lambda - \Lambda_0)\left(PZe^{Z^{\mathrm{T}}\beta_0}Y\right) \\ \dot{\Psi}_{21}(\beta - \beta_0)h &= -\Lambda_0\left(hPZ^{\mathrm{T}}e^{Z^{\mathrm{T}}\beta_0}Y\right)(\beta - \beta_0) \\ \dot{\Psi}_{22}(\Lambda - \Lambda_0)h &= -(\Lambda - \Lambda_0)\left(hPe^{Z^{\mathrm{T}}\beta_0}Y\right). \end{aligned}$$

Substituting $h = PZe^{Z^{\mathrm{T}}\beta_0}Y / Pe^{Z^{\mathrm{T}}\beta_0}Y$ in (15) and subtracting from (14), one finds

$$\sqrt{N}\left(\widehat{\beta}_N - \beta_0\right) = \mathbb{G}_N^{\widehat{\pi}}(D^{-1}G) + o_p(1), \quad \text{where}$$

$$D = \int_0^\tau \left[\frac{PZ^{\otimes 2}e^{Z^{\mathrm{T}}\beta_0 Y}}{Pe^{Z^{\mathrm{T}}\beta_0 Y}} - \left(\frac{PZe^{Z^{\mathrm{T}}\beta_0 Y}}{Pe^{Z^{\mathrm{T}}\beta_0 Y}}\right)^{\otimes 2}\right] Pe^{Z^{\mathrm{T}}\beta_0 Y} d\Lambda_0$$

$$= \int_0^\tau \left[\frac{PZ^{\otimes 2}e^{Z^{\mathrm{T}}\beta_0 Y}}{Pe^{Z^{\mathrm{T}}\beta_0 Y}} - \left(\frac{PZe^{Z^{\mathrm{T}}\beta_0 Y}}{Pe^{Z^{\mathrm{T}}\beta_0 Y}}\right)^{\otimes 2}\right] (t) dP(T \le t, \Delta = 1) \quad (16)$$

equals the efficient information $\tilde{I}_0$ under the model and

$$G = \int_0^\tau \left(Z - \frac{PZe^{Z^{\mathrm{T}}\beta_0 Y}}{Pe^{Z^{\mathrm{T}}\beta_0 Y}}\right) dM_0.$$

When P is the model distribution P_0, G is the efficient score, M_0 is a martingale and $\mathrm{Var}(G) = D = \tilde{I}_0$. For complete data, where $\mathbb{G}_N$ replaces $\mathbb{G}_N^{\widehat{\pi}}$, the asymptotic variance of the normalized $\widehat{\beta}_N$ is then the standard $\tilde{I}_0^{-1}$. Off the model one has the sandwich variance $D^{-1}\mathrm{Var}(G)D^{-1}$. The second equation (16) for D makes it easier to check that this is indeed the sandwich variance derived by Lin and Wei (1989). See Therneau and Grambsch (2000, pp. 159-60). This is also the Phase I variance of the calibrated IPW estimator off the model, with the total variance being found using (4) applied to $f = D^{-1}G$.

A similar argument leads to an expansion for $\sqrt{N}(\widehat{\Lambda}_N - \Lambda_0)$. In (15) we substitute $h/Pe^{Z^{\mathrm{T}}\beta_0 Y}$ for h and subtract (14) from (15) to find

$$\sqrt{N}\left(\widehat{\Lambda}_N - \Lambda_0\right) h = \mathbb{G}_N^{\widehat{\pi}}\left[\int_0^\tau \frac{h}{Pe^{Z^{\mathrm{T}}\beta_0 Y}} dM_0 - \Lambda_0\left(\frac{hPZ^{\mathrm{T}}e^{Z^{\mathrm{T}}\beta_0 Y}}{Pe^{Z^{\mathrm{T}}\beta_0 Y}}\right) D^{-1}G\right] + o_p(1).$$

This is the same formula as derived earlier under the model. To see this, note that the second term in brackets on the RHS of the analogous expansion in Sect. 6.1 may be rewritten using the martingale calculus as

$$\begin{aligned} P_0\left(\int_0^\tau \frac{h}{P_0 e^{Z^{\mathrm{T}}\beta_0 Y}} dM_0 \dot{\ell}_0^{\mathrm{T}}\right)\tilde{\ell}_0 &= P_0\left(\int_0^\tau \frac{h}{P_0 e^{Z^{\mathrm{T}}\beta_0 Y}} dM_0 \int_0^\tau Z^{\mathrm{T}} dM_0\right)\tilde{\ell}_0 \\ &= P_0\left(\int_0^\tau \frac{hZ^{\mathrm{T}}e^{Z^{\mathrm{T}}\beta_0 Y}}{P_0 e^{Z^{\mathrm{T}}\beta_0 Y}} d\Lambda_0\right)\tilde{\ell}_0 \\ &= \Lambda_0\left(\frac{hP_0 Z^{\mathrm{T}}e^{Z^{\mathrm{T}}\beta_0 Y}}{P_0 e^{Z^{\mathrm{T}}\beta_0 Y}}\right) D^{-1}G. \end{aligned}$$

Under the model, in $L_2(P_0)$, the first and second terms in brackets in the expansion for $\sqrt{N}(\widehat{\Lambda}_N - \Lambda_0)$ are orthogonal. Off the model, in $L_2(P)$, they may not be. Similarly, the asymptotic expansion (9) for the estimated cumulative hazard holds off the model, but the two terms in curly brackets, and hence $\mathbb{Z}(t)$ and $\mathbb{Z}_*$ in (10), may be correlated.

The discussion in 6.1 regarding calculation of the Phase II variance for calibrated IPW estimates applies both on and off the model.

6.3 Additive hazards regression

Asymptotic properties of $(\widehat{\beta}_N, \widehat{\Lambda}_N)$ for the additive hazards model, whether for simple random or two phase stratified sampling, are similar to those for the Cox model. Only a brief outline is given here. Since no likelihood calculations are involved, the discussion is focussed primarily on properties under general misspecification. For further details, see the 2014 University of Washington PhD thesis by one of us (JH).

The cumulative hazard function conditional on covariates Z is assumed, under the model, to satisfy $\Lambda(t|Z) = \Lambda(t) + Z^{\mathrm{T}}\beta \cdot t$. Whereas the Cox model may be fit easily even in situations where it clearly does not hold, the additive hazards model imposes rather severe restrictions on P to ensure that estimated baseline and conditional hazards are non-negative, at least in the limit. Let $\mathbb{N}(t)$ and $Y(t)$ denote the standard counting and at risk processes. Set

$$M_{\beta,\Lambda}(t) \;=\; \mathbb{N}(t) - \int_0^t Y(s)d\Lambda(s) - \int_0^t Y(s)Z^{\mathrm{T}}\beta ds, \tag{17}$$

and, with $m(t) = PZY(t)/PY(t)$, define

$$D \;=\; \int_0^\tau P[Z - m(s)]^{\otimes 2} Y(s) ds. \tag{18}$$

We assume that D is non-singular and that, almost surely in Z,

$$P\int_0^t \frac{d\mathbb{N}(s)}{PY(s)} - \int_0^t m^{\mathrm{T}}(s)ds\; D^{-1}\; P\int_0^\tau [Z-m]d\mathbb{N} \quad \text{and} \tag{19}$$

$$P\int_0^t \frac{d\mathbb{N}(s)}{PY(s)} + \int_0^t [Z-m(s)]^{\mathrm{T}}ds\; D^{-1}\; P\int_0^\tau [Z-m]d\mathbb{N} \tag{20}$$

are non-decreasing in t on the interval $[0, \tau]$. These assumptions ensure that the limiting values of the baseline hazard, and of conditional hazards estimated under the model, are non-negative whether or not the model actually holds. When in fact $\Lambda(t|Z) = \Lambda_0(t) + Z^{\mathrm{T}}\beta_0 \cdot t$, (17) is for $(\beta, \Lambda) = (\beta_0, \Lambda_0)$ a martingale under $P = P_0 = P_{\beta_0,\Lambda_0}$. Using martingale arguments, (19) may be shown to equal $\Lambda_0(t)$ and (20) to equal $\Lambda_0(t) + Z^{\mathrm{T}}\beta_0 \cdot t$, both assumed non-decreasing in t. The same is true off the model (see below) for (β_0, Λ_0) defined in the usual manner as functions of P.

We start with equations motivated by work of McKeague and Sasieni (1994) (see below) that lead to the estimators proposed by Lin and Ying (1994) for simple random sampling and by Kulich and Lin (2000) for the stratified case-cohort design, but modify them using calibrated weights for more general two-phase sampling designs.

In partitioned form, the equations have the same form as under the Cox model:

$$\mathbb{P}_N^{\widehat{\pi}} \psi_{1;\beta,\Lambda}(X) = \mathbb{P}_N^{\widehat{\pi}} \int_0^\tau Z d M_{\beta,\Lambda} = 0 \tag{21}$$

$$\mathbb{P}_N^{\widehat{\pi}} \psi_{2;\beta,\Lambda}(X) h = \mathbb{P}_N^{\widehat{\pi}} \int_0^\tau h d M_{\beta,\Lambda} = 0, \quad h \in H. \tag{22}$$

Inserting $h = \mathbb{P}_N^{\widehat{\pi}} ZY / \mathbb{P}_N^{\widehat{\pi}} Y$ in (22) and combining with (17) lead to

$$\mathbb{P}_N^{\widehat{\pi}} \int_0^\tau ZY d\Lambda = \mathbb{P}_N^{\widehat{\pi}} \int_0^\tau \frac{\mathbb{P}_N^{\widehat{\pi}} ZY}{\mathbb{P}_N^{\widehat{\pi}} Y} d\mathbb{N} - \mathbb{P}_N^{\widehat{\pi}} \int_0^\tau \frac{\mathbb{P}_N^{\widehat{\pi}} ZY}{\mathbb{P}_N^{\widehat{\pi}} Y}(t) Y(t) Z^{\mathrm{T}} \beta dt.$$

Substituting this expression in turn for $\int_0^\tau \mathbb{P}_N^{\widehat{\pi}} ZY d\Lambda$ in (21), and solving for β, yields explicit formulas for the estimators, namely

$$\widehat{\beta}_N = \left\{ \mathbb{P}_N^{\widehat{\pi}} \int_0^\tau \left[Z - \frac{\mathbb{P}_N^{\widehat{\pi}} ZY(t)}{\mathbb{P}_N^{\widehat{\pi}} Y(t)} \right]^{\otimes 2} Y(t) dt \right\}^{-1} \mathbb{P}_N^{\widehat{\pi}} \int_0^\tau \left[Z - \frac{\mathbb{P}_N^{\widehat{\pi}} ZY}{\mathbb{P}_N^{\widehat{\pi}} Y} \right] d\mathbb{N}$$

$$\widehat{\Lambda}_N(t) = \mathbb{P}_N^{\widehat{\pi}} \left(\frac{\Delta \mathbf{1}[T \le t]}{(\mathbb{P}_N^{\widehat{\pi}} Y)(T)} \right) - \int_0^t \frac{\mathbb{P}_N^{\widehat{\pi}} Z^{\mathrm{T}} \widehat{\beta}_N Y(s)}{\mathbb{P}_N^{\widehat{\pi}} Y(s)} ds,$$

where the last equation follows from (22) with $h(s) = \mathbf{1}[s \le t] / \mathbb{P}_N^{\widehat{\pi}} Y(s)$. The estimated baseline cumulative hazard is the IPW Nelson-Aalen estimate of the cumulative hazard for the entire population, minus the IPW estimate of the time-weighted average excess risk.

The joint asymptotic distribution of $(\widehat{\beta}_N, \widehat{\Lambda}_N)$ is easily determined from the explicit expressions given for these estimators and the limiting distribution of the calibrated IPW empirical process $\mathbb{G}_N^{\widehat{\pi}}$. Here we show that it may be obtained also by applying the Z-estimation theorem, the conclusions of which are precisely as shown in Eqs. (14) and (15) for the Cox model. Now, however, $M_0 = M_{\beta_0,\Lambda_0}$ is given by (17) and (β_0, Λ_0) are the limits in probability of $(\widehat{\beta}_N, \widehat{\Lambda}_N)$, obtained by substituting P for $P_N^{\widehat{\pi}}$ in the formulas just derived. In other words,

$$\beta_0 = D^{-1} P \int_0^\tau (Z - m)\, d\mathbb{N},$$

with D defined in (18), and

$$\Lambda_0(t) = P \int_0^t \frac{d\mathbb{N}}{PY} - \int_0^t m^{\mathrm{T}}(s) ds\; D^{-1}\; P \int_0^\tau (Z - m)\, d\mathbb{N}.$$

The assumption that (19) and (20) are non-decreasing ensures that, even off the model, $\Lambda_0(t)$ and $\Lambda_0(t) + Z^{\mathrm{T}} \beta_0 \cdot t$ are non-decreasing.

The Fréchet derivative of $\Psi_{\beta,\Lambda} = P \psi_{\beta,\Lambda}$, where the components of $\psi_{\beta,\Lambda}$ are shown in (21) and (22), follows trivially since $M_{\beta,\Lambda}$ and hence $\Psi_{\beta,\Lambda}$ itself are linear in both β and Λ:

$$\begin{aligned}
\dot{\Psi}_{11}(\beta-\beta_0) &= -\int_0^\tau P\left[ZZ^{\mathrm{T}}Y(t)\right]dt(\beta-\beta_0)\\
\dot{\Psi}_{12}(\Lambda-\Lambda_0) &= -(\Lambda-\Lambda_0)P(ZY)\\
\dot{\Psi}_{21}(\beta-\beta_0)h &= -\int_0^\tau h(t)P\left[Z^{\mathrm{T}}Y(t)\right]dt(\beta-\beta_0)\\
\dot{\Psi}_{22}(\Lambda-\Lambda_0)h &= -(\Lambda-\Lambda_0)hP(Y).
\end{aligned}$$

Inserting $h = m = P(ZY)/P(Y)$ in the new version of Eq. (15), where M_0 is defined by (17) rather than (7) and the components of $\dot{\Psi}$ are as shown above, and subtracting (15) from (14), we find

$$\sqrt{N}\left(\widehat{\beta}_N-\beta_0\right) = \mathbb{G}_N^{\widehat{\pi}}(D^{-1}G) + o_p(1) \tag{23}$$

where D is shown in (18) and now

$$G = \int_0^\tau (Z-m)\, dM_0.$$

Similar calculations lead to the expansion

$$\sqrt{N}\left(\widehat{\Lambda}_N-\Lambda_0\right)h = \mathbb{G}_N^{\widehat{\pi}}\left[\int_0^\tau \frac{h}{PY}dM_0 - \int_0^\tau h(t)\,m^{\mathrm{T}}(t)dt\,D^{-1}G\right] + o_p(1). \tag{24}$$

The estimators $(\widehat{\beta}_N, \widehat{\Lambda}_N)$ are the estimators used in practice for the additive risk model under simple random sampling (Lin and Ying 1994). Then $\mathbb{G}_N$ replaces $\mathbb{G}_N^{\widehat{\pi}}$ in the expansions above. Since the estimators are not efficient, one still has the sandwich form of variance for $\widehat{\beta}$:

$$\mathrm{Var}_{\mathrm{A}}\sqrt{N}\left(\widehat{\beta}_N-\beta_0\right) = D^{-1}\mathrm{Var}(G)D^{-1}.$$

The only difference between the asymptotic theory on and off the model is that, on the model, G is a martingale integral and its variance may be found via the martingale calculus to equal

$$\mathrm{Var}(G) = \int_0^\tau P\left\{[Z-m(t)]^{\otimes 2}\,Y(t)\left(d\Lambda_0(t) + Z^{\mathrm{T}}\beta_0 dt\right)\right\}.$$

The cumulative hazard at time t for a subject with covariates z_0 is estimated to be $\widehat{\Lambda}_N(t) + z_0^{\mathrm{T}}\widehat{\beta}_N \cdot t$ which converges to $\Lambda_0(t) + z_0^{\mathrm{T}}\beta_0 \cdot t$. Using (23) and (24), the normalized difference between the estimator and its limit satisfies

$$\begin{aligned}
&\sqrt{N}\left(\widehat{\Lambda}_N(t)-\Lambda_0(t)\right) + z_0^{\mathrm{T}}\sqrt{N}\left(\widehat{\beta}_N-\beta_0\right)t\\
&= \mathbb{G}_N^{\widehat{\pi}}\left\{\int_0^t \frac{dM_0}{PY} + \int_0^t [z_0-m(s)]^{\mathrm{T}}\, ds\; D^{-1}G\right\} + o_p(1),
\end{aligned}$$

which expression may be compared with that derived for the Cox model under simple random sampling in (9). The discussions at the end of Sects. 6.1 and 6.2 regarding calculation of Phase I and Phase II terms in the asymptotic variance of estimated cumulative hazards apply here as well as there.

Under the additive hazards model, the estimator $\widehat{\Lambda}_N$ of the baseline cumulative hazard may not be monotone increasing, though it is assumed to be so in the limit. While this does not affect the asymptotic theory, it is awkward in practice. There are two possible remedies. One, suggested by Lin and Ying (1994), is to use the modified estimator

$$\widehat{\Lambda}_N^{\dagger}(t) \equiv \sup_{0 \le s \le t} \widehat{\Lambda}_N(s).$$

Alternatively, much as suggested by Li and Tseng (2008), we could replace $\widehat{\Lambda}_N(t)$ by the isotonized estimator $\widehat{\Lambda}_N^{\text{iso}}(t)$ obtained by forming the greatest convex minorant $\hat{H}_N(t)$ of the cumulative sum process $H_N(t) \equiv \int_0^t \widehat{\Lambda}_N(s)ds$ and setting $\widehat{\Lambda}_N^{\text{iso}}(t) = \hat{H}_N'(t)$, where the derivative is the right derivative at each t. See, for example, Barlow et al (1972, pp. 9-17). Lin and Ying sketched an argument for the asymptotic equivalence of $\widehat{\Lambda}_N$ and their monotone modified version for complete Phase I data, assuming that the additive model held and that the baseline hazard was strictly positive on $[0, \tau]$. It remains to study these modified estimators for two-phase sampling and under model misspecification.

A referee asked about the extension of these results to the more general additive model of McKeague and Sasieni (1994). Here each subject has an additional q-vector of covariates W, Λ is redefined to be a q-vector of functions in BV$[0, \tau]$ and the cumulative hazard at t conditional on (W, Z) is assumed to equal

$$\text{cumhaz}(t|W, Z) \;=\; W^{\mathrm{T}}\Lambda(t) + Z^{\mathrm{T}}\beta t.$$

If we now define

$$M_{\beta,\Lambda} \;=\; \mathbb{N}(t) - \int_0^t Y(s)W^{\mathrm{T}}d\Lambda(s) - \int_0^t Y(s)Z^{\mathrm{T}}\beta ds,$$

and take for h a q-vector of functions in the unit ball of BV$[0, \tau]$, the estimating equations used by McKeague and Sasieni (1994) at the first iteration of their iterative procedure may be written

$$\begin{aligned} \mathbb{P}_N \int_0^\tau Z dM_{\beta,\Lambda} &= 0 \\ \mathbb{P}_N \int_0^\tau h^{\mathrm{T}} W dM_{\beta,\Lambda} &= 0, \;\; h \in H^q. \end{aligned}$$

Our estimating Eqs. (21) and (22) for the Lin–Ying model are a special case where $q = W = 1$ and $\mathbb{P}_N$ is replaced by $\mathbb{P}_N^{\hat{\pi}}$. Explicit solutions to the new equations generalize those shown above for $\hat{\beta}_N$, $\hat{\Lambda}_N$. Using smoothing to estimate the derivatives

of Λ, McKeague and Sasieni (1994) iterate their procedure to convergence to find the maximum likelihood estimates. They remark that often little is gained in practice over the simpler estimator, even when the model holds.

7 Discussion

We have applied the infinite dimensional Z-estimation theorem to derive asymptotic properties of both Euclidean and non-Euclidean parameters in semiparametric models fitted to data from two-phase stratified sampling designs. The approach works well both "on the model" and under general misspecification. When applied to simple random (i.i.d.) samples, it leads quickly to well known results for the Cox proportional and Lin–Ying additive hazards models for survival data. We have developed a general theory that extends those results, and undoubtedly many other well known results for semiparametric models where the non-Euclidean parameter is estimable at a $\sqrt{N}$ rate, to accomodate IPW estimators using calibrated weights for two-phase stratified sampling designs.

Our approach uses IPW versions of the standard estimating equations applied to data from the Phase II sample. When the model holds, the resulting estimates can be seriously inefficient, even after calibration. Likelihood equations for two phase samples, however, involve integration over the "missing data", here the portions of X not observed at Phase I, with respect to unknown distributions. This can pose a substantial challenge to implementation of an efficient methodology. In applications to the Cox model, for example, it involves consideration of three infinite dimensional parameters, namely, the distributions of the survival times, the censoring times and the covariates (Nan et al. 2004). To our knowledge, concrete proposals for implementation involve assumptions that may limit applicability, for example, to discrete covariates (Nan 2004), to complete independence (not just conditional on covariates) of survival and censoring times (Scheike and Martinussen 2004) and to situations where all the auxiliary variables are included in the model (Zeng and Lin 2014). Multiple imputation of the "missing data" is another option, but requires correct specification of the imputation model for consistency (Marti and Chavance 2011; Keogh and White 2013). Most of these proposals, furthermore, have been restricted to (stratified) versions of the case-cohort design in which all cases ("deaths") are sampled at Phase II (Prentice 1986; Borgan et al. 2000). In practice, including for the two studies mentioned in the introduction, Phase II data for many cases may be missing due to loss or degradation of stored tissue samples. Our approach handles Phase II sampling of both cases and controls.

The advantages of "efficient" methods are less clear when the model has been misspecified. Epidemiologists and survey samplers generally agree that the goal of stratified sampling designs is to estimate the same parameter as would have been estimated had complete data been available for the entire cohort (Phase I sample). This goal is achieved by the IPW methods, even those that involve calibrated weights to improve efficiency, but not by those based on likelihoods. Of course, careful model checking is important for any statistical analysis to detect departures from the assumed model. Lumley (2009) has argued that even when the model is "nearly correct", to the

extent that the departure could not be reliably detected, the bias of the semiparametric efficient estimate may be sufficient to outweigh its advantages in terms of lower variance. This suggests that, certainly for large samples, the IPW approach is preferred. For small or even moderately sized samples, provided that an efficient method is available, the choice is less clear. Further work is needed on this issue.

A major limitation of our approach is the restriction to covariates that are fixed in time. By contrast, the martingale theory for survival analyses of simple random samples, and the approaches by Borgan et al. (2000) and others to the analysis of case-cohort data, some of which which also involve IPW versions of estimating equations, easily accomodate time-dependent covariates. Close inspection of the equations we use to express our results reveals that they all "make sense" when the covariates $Z(t)$ or $z_0(t)$ are time dependent. One could envisage a generalization of our theory to the situation where the underlying data took the form $X = (T, \Delta, Z(\cdot))$. Substantial work would be needed, however, to clearly delineate the appropriate boundaries of application. In particular, results we borrowed from van der Vaart (1998, Sect. 25.12.1) regarding membership of the estimating equations in a Donsker class would need careful extension. The interpretation of cumulative hazards of the form $\Lambda[t|Z(t)]$ is unambiguous when the covariates are "external" in the sense of Kalbfleisch and Prentice (2002, Sect. 6.3). Correct interpretation of results based on "internal" time-dependent covariates requires much more care.

An alternative to calibration of the weights is their estimation using a parametric model, usually logistic regression, for $P(R = 1|V)$ (Robins et al. 1994). Incorporation of the sampling strata indicators into the regression equation is essential for consistent estimation; adding further Phase I variables, such as those used for calibration, increases asymptotic efficiency. Calibration and estimation yield the same weights when V is discrete, the calibration variables identify the separate strata defined by V and the estimation model is saturated (Lumley et al. 2011). They often yield similar estimates in other settings. A general theory for estimated weights could be developed along the lines presented here, again using the Z-estimation theorem with nuisance parameters (Breslow and Wellner 2008); see Breslow et al (2009a, Eq. 14) for some preliminary steps in this direction. Whether calibration or estimation is adopted, the question remains as to how best select the Phase I variables used for adjustment of the weights. For β estimation, the expansion (5) in terms of the efficient influence function, or its analog off the model where $D^{-1}G$ replaces $\tilde{\ell}_0$, suggests that the optimal calibration variables are $C^{\text{opt}}(V) = \text{E}(\tilde{\ell}_0|V)$. This choice of calibration variables yields the optimal estimator in the class of augmented inverse probability weighted (AIPW) estimators considered by Robins et al. (1994); see Lumley et al. (2011). Of course, since the distribution $[X|V]$ is unknown, some method of approximating $C^{\text{opt}}(V)$ is needed. Robins et al. (1994) suggest a regression approach and Kulich and Lin (2004) a "plug in" approach based on a (not necessarily correct) imputation model. Further study of this issue would be desirable.

In summary, using the Z-estimation theorem, we have outlined a systematic methodology for inference in semiparametric models using data from two-phase stratified samples. The approach incorporates calibration of the weights to improve efficiency and applies both on and off the model.

Acknowledgments Wellner's research was supported in part by National Science Foundation Grant DMS-1104832 and National Institute of Allegery and Infectious Diseases Grant 2R01 AI291968-04. Dedicated to Niels Keiding on the occasion of his 70th birthday.

References

Aalen O (1976) Nonparametric inference in connection with multiple decrement models. Scand J Stat 3:15–27

Aalen OO, Borgan O, Gjessing HK (2008) Survival and event history analysis. Springer, New York

Andersen PK, Gill RD (1982) Cox's regression model for counting processes: a large sample study. Ann Stat 10:1100–1120

Anderson GL, Manson J, Wallace R, Lund B, Hall D, Davis S, Shumaker S, Wang CY, Stein E, Prentice RL (2003) Implementation of the Women's Health Initiative study design. Ann Epidemiol 13:S5–S17

Barlow R, Bartholomew D, Bremner J, Brunk H (1972) Statistical inference under order restrictions. Wiley, New York

Begun JM, Hall WJ, Huang WM, Wellner JA (1983) Information and asymptotic efficiency in parametric–nonparametric models. Ann Stat 11:432–452

Bickel P, Klaassen C, Ritov Y, Wellner J (1993) Efficient and adaptive estimation for semiparametric models. The Johns Hopkins University Press, Baltimore

Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J (2000) Exposure stratified case–cohort designs. Lifetime Data Anal 6:39–58

Breslow N, Crowley J (1974) A large sample study of the life table and product limit estimates under random censorship. Ann Stat 2:437–453

Breslow NE, Lumley T (2013) Semiparametric models and two-phase samples: applications to Cox regression. In: IMS collections, vol. 9, Institute of Mathematical Statistics, Beachwood, OH, pp 65–77

Breslow NE, Wellner JA (2007) Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. Scand J Stat 34:86–102

Breslow NE, Wellner JA (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'. Scand J Stat 35:186–192

Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M (2009a) Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. Statist Biosci 1:32–49

Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M (2009b) Using the whole cohort in the analysis of case–cohort data. Am J Epidemiol 169:1398–1405

Cox DR (1961) Tests of separate families of hypotheses. In: Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, vol. 1, University of California Press, Berkeley, CA, pp 105–123

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

Deville JC, Särndal CE (1992) Calibration estimators in survey sampling. J Am Stat Assoc 87:376–382

Freedman DA (2006) On the so-called "Huber sandwich estimator" and "robust standard errors". Am Stat 60:299–302

Godambe VP (1960) An optimum property of regular maximum-likelihood estimation. Ann Math Stat 31:1208–1211

Huber PJ (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, University of California Press, Berkeley, CA, pp 221–233

Huber PJ (1980) Robust statistics. Wiley, New York

Kalbfleisch JD, Prentice R (2002) The statistical analysis of failure time data, 2nd edn. Wiley, Hoboken, NJ

Keogh RH, White IR (2013) Using full-cohort data in nested case–control and case–cohort studies by multiple imputation. Stat Med 32:4021–4043

Kulich M, Lin DY (2000) Additive hazards regression for case–cohort studies. Biometrika 87:73–87

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

Li G, Tseng CH (2008) Non-parametric estimation of a survival function with two-stage design studies. Scand J Stat 35:193–211

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22
Lin DY, Wei LJ (1989) The robust inference for the Cox proportional hazards model. J Am Stat Assoc 84:1074–1078
Lin DY, Ying Z (1994) Semiparametric analysis of the additive risk model. Biometrika 81:61–71
Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York
Lumley T (2009) Robustness of semiparametric efficiency in nearly-correct models for two-phase samples. UW Biostatistics Working Paper Series. http://biostats.bepress.com/uwbiostat/paper351, Accessed 22 November 2014
Lumley T (2012) Complex surveys: a guide to analysis using R. Wiley, Hoboken, NJ
Lumley T, Shaw PA, Dai JY (2011) Connections between survey calibration estimators and semiparametric models for incomplete data. Int Stat Rev 79:200–220
Marti H, Chavance M (2011) Multiple imputation analysis of case–cohort studies. Stat Med 30:1595–1607
McKeague IW, Sasieni PD (1994) A partly parametric additive risk model. Biometrika 81:501–514
Nan B (2004) Efficient estimation for case–cohort studies. Can J Stat 32:403–419
Nan B, Emond M, Wellner JA (2004) Information bounds for Cox regression models with missing data. Ann Stat 32:723–753
Nelson W (1972) Theory and applications of hazard plotting for censored failure data. Technometrics 14:945–966
Prentice RL (1986) A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73:1–11
Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression-coefficients when some regressors are not always observed. J Am Stat Assoc 89:846–866
Royall RM (1986) Model robust confidence-intervals using maximum-likelihood estimators. Int Stat Rev 54:221–226
Saegusa T, Wellner JA (2013) Weighted likelihood estimation under two-phase sampling. Ann Stat 41:269–295
Särndal C, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer, New York
Scheike TH, Martinussen T (2004) Maximum likelihood estimation for Cox's regression model under case-cohort sampling. Scand J Stat 31:283–293
Struthers CA, Kalbfleisch JD (1986) Misspecified proportional hazard models. Biometrika 73:363–369
Therneau TM, Grambsch PM (2000) Modeling survival data: extending the Cox model. Springer, New York
Tsiatis AA (1981) A large sample study of Cox's regression model. Ann Stat 9:93–108
van der Vaart AW (1995) Efficiency of infinite dimensional M-estimators. Stat Neerl 49:9–30
van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge, UK
van der Vaart AW, Wellner JA (1996) Weak convergence and empirical processes with applications in statistics. Springer, New York
Williams OD (1989) The Atherosclerosis Risk in Communities (ARIC) study—design and objectives. Am J Epidemiol 129:687–702
Zeng DL, Lin DY (2014) Efficient estimation of semiparametric transformation models for two-phase cohort studies. J Am Stat Assoc 109:371–383