Biometrika (2000), **87**, 1, *pp.* 37–47

Printed in Great Britain

On fitting Cox's proportional hazards models to survey data

By D. Y. LIN

Department of Biostatistics, Box 357232, University of Washington, Seattle, Washington 98195, U.S.A.

danyu@biostat.washington.edu

Summary

Binder (1992) proposed a method of fitting Cox's proportional hazards models to survey data with complex sampling designs. He defined the regression parameter of interest as the solution to the partial likelihood score equation based on all the data values of the survey population under study, and developed heuristically a procedure to estimate the regression parameter and the corresponding variance. In this paper, we provide a formal justification of Binder's method. Furthermore, we present an alternative approach which regards the survey population as a random sample from an infinite universe and accounts for this randomness in the statistical inference. Under the alternative approach, the regression parameter retains its original interpretation as the log hazard ratio, and the statistical conclusion applies to other populations. The related problem of survival function estimation is also studied.

Some key words: Complex survey design; Estimating function; Failure time; Finite-population sampling; Model misspecification; Partial likelihood; Regression analysis; Superpopulation; Survey sampling; Survival data.

1. Introduction

The Cox (1972) proportional hazards model has been widely used to study the effects of covariates on a failure time. This model specifies that the hazard function of the failure time T associated with a vector of possibly time-varying covariates X satisfies

$$h(t \mid X) = h_0(t) e^{\beta_0' X(t)}, \tag{1·1}$$

where $h_0(.)$ is an unspecified baseline hazard function, and β_0 is a vector-valued unknown regression parameter pertaining to the log hazard ratio or log relative risk.

Typically, the failure time is subject to right censoring. Let C be the censoring time. Also, let $\tilde{T} = \min(T, C)$, $\Delta = I(T \leqslant C)$ and $Y(t) = I(\tilde{T} \geqslant t)$, where $I(.)$ is the indicator function. If $\{\tilde{T}_i, \Delta_i, X_i(.)\}$ $(i = 1, \ldots, N)$ is a random sample from the joint distribution of $\{\tilde{T}, \Delta, X(.)\}$, then β_0 can be estimated by the partial likelihood score function

$$U(\beta) := \sum_{i=1}^{N} \Delta_i \left\{ X_i(\tilde{T}_i) - \frac{S^{(1)}(\beta, \tilde{T}_i)}{S^{(0)}(\beta, \tilde{T}_i)} \right\}, \tag{1·2}$$

where

$$S^{(0)}(\beta, t) = N^{-1} \sum_{i=1}^{N} Y_i(t) e^{\beta' X_i(t)}, \quad S^{(1)}(\beta, t) = N^{-1} \sum_{i=1}^{N} Y_i(t) e^{\beta' X_i(t)} X_i(t).$$

Denote the solution to $U(\beta) = 0$ by B. Under mild conditions, $N^{\frac{1}{2}}(B - \beta_0)$ is asymptotically zero-mean normal with a covariance matrix that can be consistently estimated by $\mathscr{I}^{-1}(B)$, where $\mathscr{I}(\beta) = -N^{-1}\partial U(\beta)/\partial\beta$ (Andersen & Gill, 1982).

In population-based surveys, the sample is drawn from a finite survey population via a complex design, such as stratified multi-stage sampling. Binder (1992) described a procedure for fitting proportional hazards models to such data, which has been implemented in major software packages, such as SUDAAN, and commonly used by data analysts. To be specific, suppose that a sample of size n is drawn from a survey population of N units through a complex design, the sampling weights being denoted by the w_i's. Binder (1992) defined the finite-population parameter of interest as B, which is the root of (1·2) based on the fixed values $\{\tilde{T}_i, \Delta_i, X_i(.)\}$ $(i = 1, \ldots, N)$ of the survey population. He then proposed to estimate B by the estimating function

$$\hat{U}(\beta) := \sum_{i=1}^{n} w_i \Delta_i \left\{ X_i(\tilde{T}_i) - \frac{\hat{S}^{(1)}(\beta, \tilde{T}_i)}{\hat{S}^{(0)}(\beta, \tilde{T}_i)} \right\}, \tag{1·3}$$

where

$$\hat{S}^{(0)}(\beta, t) = \sum_{i=1}^{n} w_i Y_i(t) e^{\beta' X_i(t)}, \quad \hat{S}^{(1)}(\beta, t) = \sum_{i=1}^{n} w_i Y_i(t) e^{\beta' X_i(t)} X_i(t).$$

Denote the resulting estimator by $\hat{B}$. Binder (1992) derived heuristically a variance estimator for $\hat{B}$ by treating $\{\tilde{T}_i, \Delta_i, X_i(.)\}$ $(i = 1, \ldots, N)$ as fixed. Note that the target parameter B is an implicit function of the failure times, censoring times and covariate values of the survey population, and does not have the hazard ratio interpretation.

In this paper, we develop an alternative inference procedure which treats $\{\tilde{T}_i, \Delta_i, X_i(.)\}$ $(i = 1, \ldots, N)$ of the survey population as a random sample from the joint distribution of $\{\tilde{T}, \Delta, X(.)\}$ rather than as fixed quantities. We refer to this as the superpopulation inference as opposed to the finite-population inference of Binder (1992); see Särndal (1978), Binder (1983) and Godambe & Thompson (1986) for general discussion of superpopulation versus finite-population philosophies.

The term superpopulation has sometimes been used to imply that the sampling weights are not used in the analysis (Binder, 1983). This is not the position taken here. The proposed superpopulation inference takes into account the complex design of the survey sample while regarding the survey population under study as having been drawn from an infinite universe. The key difference between the proposed superpopulation inference and Binder's finite-population inference is that the former accounts for the random variation from one survey population to another. This superpopulation approach allows one to make analytical inference about the original regression parameter β_0 of model (1·1), and the statistical conclusions are generalisable to other similar populations. Although it is not widely recognised by traditional survey samplers, the superpopulation approach taken here is the prevailing school of thought in the current literature on population-based case-control studies; see Scott & Wild (1997) and the references therein.

In the next section, we provide a theoretical justification for Binder's work on finite-population inference of the regression parameter. Furthermore, we study the estimation of the cumulative hazard function and survival function, a problem not considered by Binder. Building on the results of § 2, we develop in § 3 a superpopulation theory for estimating β_0 as well as the cumulative hazard function and survival function under model (1·1) from survey data. In § 4, we report some simulation results. A few remarks are given in § 5. Most of the technical detail is relegated to the Appendix.

2. Finite-population inference

As mentioned in § 1, Binder (1992) defined the finite-population parameter of interest as B, the root of the score function $U(\beta)$ based on the survey population values $\{\tilde{T}_i, \Delta_i, X_i(.)\}$ $(i=1, \ldots, N)$. He proposed to estimate B by $\hat{B}$, the root of $\hat{U}(\beta)$. He derived heuristically a variance estimator for $\hat{B}$ conditional on the survey population values. He also suggested that $\hat{B}$ is consistent and asymptotically normal.

The key step in Binder's derivation is the approximation of $\hat{U}(B)$ by a weighted sum of $u_i(B)$ $(i=1, \ldots, n)$ given in (3·6) and (3·7) of his paper. This approximation was derived from the Taylor series expansion of the stochastic integral $\int \{\hat{S}^{(1)}(t, B)/\hat{S}^{(0)}(t, B)\}\, d\hat{G}(t)$ around $\hat{S}^{(0)} = S^{(0)}$, $\hat{S}^{(1)} = S^{(1)}$ and $\hat{G} = G$, where

$$G_i(t) = \Delta_i I(\tilde{T}_i \leqslant t), \quad G(t) = N^{-1} \sum_{i=1}^{N} G_i(t), \quad \hat{G}(t) = \sum_{i=1}^{n} w_i G_i(t).$$

Binder did not provide a formal justification for this expansion except to mention the consistency of $\hat{S}^{(0)}$, $\hat{S}^{(1)}$ and $\hat{G}$ and to cite the work of Lin & Wei (1989). As is evident from the Appendix of this paper, the techniques required for ascertaining the asymptotic properties of $\hat{U}(B)$ are extremely delicate and differ substantially from those of Andersen & Gill (1982) and Lin & Wei (1989) because the latter authors dealt with random sampling from an infinite population, whereas $\hat{U}$ involves correlated observations obtained from a finite population with unequal probability sampling.

To facilitate theoretical development, we rewrite (1·2) and (1·3) as

$$U(\beta) = \sum_{i=1}^{N} \int_0^\infty \left\{ X_i(t) - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right\} dG_i(t),$$

$$\hat{U}(\beta) = \sum_{i=1}^{N} \int_0^\infty \frac{\xi_i}{\pi_i} \left\{ X_i(t) - \frac{\hat{S}^{(1)}(\beta, t)}{\hat{S}^{(0)}(\beta, t)} \right\} dG_i(t),$$

where ξ_i indicates, by the values 1 versus 0, whether or not the ith unit of the survey population is selected into the sample, and π_i is the inclusion probability for the ith unit. It is assumed that $\pi_i > 0$ for all i. We also rewrite

$$\hat{G}(t) = N^{-1} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} G_i(t), \quad \hat{S}^{(0)}(\beta, t) = N^{-1} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} Y_i(t) e^{\beta' X_i(t)},$$

$$\hat{S}^{(1)}(\beta, t) = N^{-1} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} Y_i(t) e^{\beta' X_i(t)} X_i(t).$$

Note that the only randomness in $\hat{U}(\beta)$ is generated by the ξ_i's since the inference is conditional on $\{\tilde{T}_i, \Delta_i, X_i(.)\}$ $(i=1, \ldots, N)$. We are concerned with the asymptotic behaviour of $\hat{U}$, $\hat{B}$ etc. when both n and N tend to infinity. In the sequel, the summation is taken from 1 to N and the limit is taken as $N \to \infty$ if not explicitly indicated.

Define

$$s^{(0)}(\beta, t) = \lim_{N\to\infty} S^{(0)}(\beta, t), \quad s^{(1)}(\beta, t) = \lim_{N\to\infty} S^{(1)}(\beta, t);$$

$$a = \lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} \int_0^\infty X_i(t)\, dG_i(t), \quad g(t) = \lim_{N\to\infty} G(t).$$

Clearly, $N^{-1}U(\beta)$ converges to

$$u(\beta) := a - \int_0^\infty \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)}\, dg(t), \qquad (2{\cdot}1)$$

which is also the probability limit of $N^{-1}\hat{U}(\beta)$ since $\pi_i = \mathrm{pr}(\xi_i = 1)$. Thus, it follows from the arguments given in the proof of Lemma 3.1 of Andersen & Gill (1982) that $\hat{B}$ is a consistent estimator of B. More precisely, $\hat{B}$ and B converge to the same limit.

We show in the Appendix that

$$N^{-\frac{1}{2}}\hat{U}(B) = N^{-\frac{1}{2}} \sum_{i=1}^{N} \frac{\xi_i - \pi_i}{\pi_i} U_i(B) + o_p(1), \qquad (2{\cdot}2)$$

where

$$U_i(\beta) = \int_0^\infty \left\{X_i(t) - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)}\right\} \left\{dG_i(t) - \frac{Y_i(t) e^{\beta' X_i(t)}\, dg(t)}{s^{(0)}(\beta, t)}\right\}. \qquad (2{\cdot}3)$$

Expression (2·3) is essentially the same as (3·7) of Binder (1992). Therefore, Binder's approximation for $\hat{U}(B)$ is valid. As argued in the Appendix, approximation (2·2) entails that $N^{-\frac{1}{2}}\hat{U}(B)$ is asymptotically zero-mean normal with covariance matrix

$$V(B) := \lim_{N\to\infty} N^{-1} \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} U_i(B)U_j'(B), \qquad (2{\cdot}4)$$

where $\pi_{ij} = \mathrm{pr}(\xi_i\xi_j = 1)$. It then follows from the Taylor series expansion that $N^{\frac{1}{2}}(\hat{B} - B)$ is asymptotically zero-mean normal with covariance matrix $D^{-1}(B)V(B)D^{-1}(B)$, where $D(B) = \lim \mathscr{I}(B)$. We can estimate $D(B)$ and $V(B)$ consistently by

$$\hat{D}(\hat{B}) := -N^{-1}\partial\hat{U}(\beta)/\partial\beta|_{\beta=\hat{B}},$$

$$\hat{V}(\hat{B}) := N^{-1} \sum_{i=1}^{N}\sum_{j=1}^{N} \xi_i\xi_j \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}\pi_i\pi_j} \hat{U}_i(\hat{B})\hat{U}_j'(\hat{B}), \qquad (2{\cdot}5)$$

where

$$\hat{U}_i(\beta) = \Delta_i \left\{X_i(\tilde{T}_i) - \frac{\hat{S}^{(1)}(\beta, \tilde{T}_i)}{\hat{S}^{(0)}(\beta, \tilde{T}_i)}\right\} - N^{-1} \sum_{j=1}^{N} \frac{\xi_j \Delta_j Y_i(\tilde{T}_j) e^{\beta' X_i(\tilde{T}_j)}}{\pi_j \hat{S}^{(0)}(\beta, \tilde{T}_j)} \left\{X_i(\tilde{T}_j) - \frac{\hat{S}^{(1)}(\beta, \tilde{T}_j)}{\hat{S}^{(0)}(\beta, \tilde{T}_j)}\right\}.$$

In practice, it is seldom necessary to evaluate (2·5) directly. Instead, one can express the variance estimator in a computationally simpler form for each specific survey design.

It is often of interest to estimate/predict the survival experience given specific covariate values. Let $H_0(t)$ and $F_0(t)$ denote the baseline cumulative hazard function and baseline survival function under model (1·1), i.e.

$$H_0(t) = \int_0^t h_0(u)\, du, \quad F_0(t) = e^{-H_0(t)},$$

which become the cumulative hazard and survival functions under $X = x_0$ when the X_i's are centred at x_0. Based on a random sample of size N, the Breslow (1972) estimator of $H_0(t)$ is

$$\tilde{H}_0(t; B) := \int_0^t \frac{dG(u)}{S^{(0)}(B, u)},$$

and the corresponding estimator of $F_0(t)$ is $\tilde{F}_0(t) := e^{-\tilde{H}_0(t;B)}$. By analogy with B, we take $\tilde{H}_0$ and $\tilde{F}_0$ with fixed $\{\tilde{T}_i, \Delta_i, X_i(.)\}$ $(i = 1, \ldots, N)$ as the parameters of interest.

Given the survey data, we can estimate $\tilde{H}_0(t; B)$ consistently by

$$\hat{H}_0(t; \hat{B}) := \int_0^t \frac{d\hat{G}(u)}{\hat{S}^{(0)}(\hat{B}, u)} \tag{2·6}$$

because of the consistency of $\hat{G}$, $S^{(0)}$ and $\hat{B}$. It is shown in the Appendix that $N^{\frac{1}{2}}\{\hat{H}_0(t; \hat{B}) - \tilde{H}_0(t; B)\}$ converges weakly to a zero-mean Gaussian process with covariance function

$$\sigma(t_1, t_2; B) := \lim_{N \to \infty} N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} h_i(t_1; B) h_j(t_2; B) \tag{2·7}$$

at (t_1, t_2), where $h_i(t; \beta) = v_i(t; \beta) + r'(t; \beta) D^{-1}(\beta) U_i(\beta)$,

$$v_i(t; \beta) = \int_0^t \frac{dG_i(u) - Y_i(u) e^{\beta' X_i(u)}\, d\tilde{H}_0(u; \beta)}{s^{(0)}(\beta, u)}, \quad r(t; \beta) = -\int_0^t \frac{s^{(1)}(\beta, u)\, dg(u)}{s^{(0)}(\beta, u)^2}.$$

A consistent estimator for $\sigma(t_1, t_2; B)$ is

$$\hat{\sigma}(t_1, t_2) := N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} \xi_i \xi_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} \hat{h}_i(t_1; \hat{B}) \hat{h}_j(t_2; \hat{B}),$$

where $\hat{h}_i(t; \beta) = \hat{v}_i(t; \beta) + R'(t; \beta) \hat{D}^{-1}(\beta) \hat{U}_i(\beta)$,

$$\hat{v}_i(t; \beta) = \frac{\Delta_i I(\tilde{T}_i \leqslant t)}{\hat{S}^{(0)}(\beta, \tilde{T}_i)} - N^{-1} \sum_{j=1}^{N} \frac{\xi_j \Delta_j I(\tilde{T}_j \leqslant t) Y_i(\tilde{T}_j) e^{\beta' X_i(\tilde{T}_j)}}{\pi_j \hat{S}^{(0)}(\beta, \tilde{T}_j)^2},$$

$$R(t; \beta) = -N^{-1} \sum_{i=1}^{N} \frac{\xi_i \Delta_i I(\tilde{T}_i \leqslant t) \hat{S}^{(1)}(\beta, \tilde{T}_i)}{\pi_i \hat{S}^{(0)}(\beta, \tilde{T}_i)^2}.$$

Given $\hat{H}_0$, we can estimate $\tilde{F}_0(t)$ by $\hat{F}_0(t) := e^{-\hat{H}_0(t;\hat{B})}$. It follows from the functional δ-method that $N^{\frac{1}{2}}\{\hat{F}_0(t) - \tilde{F}_0(t)\}$ converges weakly to a zero-mean Gaussian process with estimated covariance function $\hat{F}_0(t_1)\hat{F}_0(t_2)\hat{\sigma}(t_1, t_2)$.

3. Superpopulation inference

In § 2, the targets of inference are restricted to the summary statistics of the survey population data, B, $\tilde{H}_0$ and $\tilde{F}_0$, which do not have any probabilistic interpretation. In this section, we regard the survey population as a random sample from an infinite superpopulation and make analytic inference about β_0, H_0 and F_0 of model (1·1) by taking into account the sampling of the survey population from the superpopulation as well as that of the survey sample from the survey population. Under this superpopulation approach, the parameters of interest have simple probabilistic interpretation, and the inference pertains to the probability distribution of the failure time in the superpopulation.

For the superpopulation inference, $\{\tilde{T}_i, \Delta_i, X_i(.)\}$ $(i = 1, \ldots, N)$ of the survey population are not treated as fixed quantities, but rather as a random sample from the joint distribution of $\{\tilde{T}, \Delta, X(.)\}$. Let $\mathscr{F}$ be the sigma-field generated by $\{\tilde{T}_i, \Delta_i, X_i(.)\}$ $(i = 1, \ldots, N)$. The inclusion probabilities are allowed to depend on $\mathscr{F}$ and are expressed as

$$\pi_i = \mathrm{pr}(\xi_i = 1 \mid \mathscr{F}) \quad (i = 1, \ldots, N). \tag{3·1}$$

Since $U(\beta)$ is the partial likelihood score function for β_0 calculated from a random sample of size N and, as argued in § 2, $N^{-1}\hat{U}(\beta)$ converges to the same limit as $N^{-1}U(\beta)$, it follows from the consistency of the maximum partial likelihood estimator that $\hat{B}$, the solution to $\hat{U}(\beta)=0$, is consistent for the hazard ratio parameter β_0. Likewise, $\hat{H}_0(t;\hat{B})$ is a consistent estimator of the baseline cumulative hazard function $H_0(t)$ because of the consistency of the Breslow estimator $\tilde{H}_0(t;B)$ and the convergence of $\hat{H}_0(t;\hat{B})$ and $\tilde{H}_0(t;B)$ to the same limit. Thus, the estimators of the finite-population parameters discussed in § 2 are also consistent estimators of the model parameters for the superpopulation inference. The variances of the estimators will be different because the superpopulation inference takes into account the variation from one survey population to another.

To obtain the asymptotic distribution of $\hat{B}$ for the superpopulation inference, we need to derive the corresponding distribution of $\hat{U}(\beta_0)$. Clearly,

$$N^{-\frac{1}{2}}\hat{U}(\beta_0)=N^{-\frac{1}{2}}U(\beta_0)+N^{-\frac{1}{2}}\{\hat{U}(\beta_0)-U(\beta_0)\}. \qquad (3\cdot2)$$

Under mild conditions, $N^{-\frac{1}{2}}U(\beta_0)$ is asymptotically zero-mean normal with covariance matrix $D(\beta_0)$ (Andersen & Gill, 1982). The second term on the right-hand side of (3·2) can be written as

$$N^{-\frac{1}{2}}\sum_{i=1}^{N}\int_0^\infty \frac{\xi_i-\pi_i}{\pi_i}X_i(t)\,dG_i(t)-N^{\frac{1}{2}}\int_0^\infty\left\{\frac{\hat{S}^{(1)}(\beta_0,t)}{\hat{S}^{(0)}(\beta_0,t)}d\hat{G}(t)-\frac{S^{(1)}(\beta_0,t)}{S^{(0)}(\beta_0,t)}dG(t)\right\},$$

which is in the same form as the right-hand side of (A·1) in the Appendix. It then follows from the arguments of the Appendix that, conditional on $\mathscr{F}$,

$$N^{-\frac{1}{2}}\{\hat{U}(\beta_0)-U(\beta_0)\}=N^{-\frac{1}{2}}\sum_{i=1}^{N}\frac{\xi_i-\pi_i}{\pi_i}U_i(\beta_0)+o_p(1), \qquad (3\cdot3)$$

which converges weakly to a zero-mean normal random vector with covariance matrix $V(\beta_0)$. Since as a limit $V(\beta_0)$ is a deterministic matrix which does not depend on $\mathscr{F}$, the weak convergence also holds unconditionally. In view of (3·1) and (3·3), we can use the law of conditional expectation to show that the two terms on the right-hand side of (3·2) are asymptotically independent. Hence, $N^{-\frac{1}{2}}\hat{U}(\beta_0)$ is asymptotically zero-mean normal with covariance matrix $D(\beta_0)+V(\beta_0)$. The Taylor series expansion then implies that $N^{\frac{1}{2}}(\hat{B}-\beta_0)$ is asymptotically zero-mean normal with covariance matrix

$$\Omega := D^{-1}(\beta_0)+D^{-1}(\beta_0)V(\beta_0)D^{-1}(\beta_0),$$

which can be estimated consistently by $\hat{\Omega}:=\hat{D}^{-1}(\hat{B})+\hat{D}^{-1}(\hat{B})\hat{V}(\hat{B})\hat{D}^{-1}(\hat{B})$. The covariance matrix comprises two components: $D^{-1}(\beta_0)$ is the variation due to the sampling of the survey population from the superpopulation and $D^{-1}(\beta_0)V(\beta_0)D^{-1}(\beta_0)$ is the variation due to the sampling of the survey sample from the survey population.

To establish the weak convergence of $\hat{H}_0(t;\hat{B})$, we make the decomposition

$$N^{\frac{1}{2}}\{\hat{H}_0(t;\hat{B})-H_0(t)\}=N^{\frac{1}{2}}\{\tilde{H}_0(t;B)-H_0(t)\}+N^{\frac{1}{2}}\{\hat{H}_0(t;\hat{B})-\tilde{H}_0(t;B)\}.$$

The two terms on the right-hand side are asymptotically independent. It follows from the existing theory for the Breslow estimator (Andersen & Gill, 1982) that $N^{\frac{1}{2}}\{\tilde{H}_0(t;B)-H_0(t)\}$ converges weakly to a zero-mean Gaussian process with covariance function

$$\phi(t_1,t_2):=\int_0^{\min(t_1,t_2)}\frac{dH_0(u)}{s^{(0)}(\beta_0,u)}+r'(t_1;\beta_0)D^{-1}(\beta_0)r(t_2;\beta_0).$$

By the arguments of the Appendix, $N^{\frac{1}{2}}\{\hat{H}_0(t;\hat{B})-\tilde{H}_0(t;B)\}$ converges weakly to a zero-mean Gaussian process with covariance function $\sigma(t_1,t_2;\beta_0)$. Hence, $N^{\frac{1}{2}}\{\hat{H}_0(t;\hat{B})-H_0(t)\}$ converges weakly to a zero-mean Gaussian process with covariance function $\phi(t_1,t_2)+\sigma(t_1,t_2;\beta_0)$, which can be estimated consistently by $\hat{\phi}(t_1,t_2)+\hat{\sigma}(t_1,t_2)$, where

$$\hat{\phi}(t_1,t_2)=N^{-1}\sum_{i=1}^{N}\frac{\xi_i\Delta_i I\{\tilde{T}_i\leqslant\min(t_1,t_2)\}}{\pi_i\hat{S}^{(0)}(\hat{B},\tilde{T}_i)^2}+R'(t_1;\hat{B})\hat{D}^{-1}(\hat{B})R(t_2;\hat{B}).$$

Consequently, $N^{\frac{1}{2}}\{\hat{F}_0(t)-F_0(t)\}$ converges weakly to a zero-mean Gaussian process with estimated covariance function $\hat{F}_0(t_1)\hat{F}_0(t_2)\{\hat{\phi}(t_1,t_2)+\hat{\sigma}(t_1,t_2)\}$. These results enable one to predict survival experience for people inside and outside the survey population.

The variances for $\hat{B}$, $\hat{H}_0$ and $\hat{F}_0$ are always larger under the superpopulation inference than under the finite-population inference. The extra variance for $N^{\frac{1}{2}}\hat{B}$ is $D^{-1}(\beta_0)$ and that for $N^{\frac{1}{2}}\hat{H}_0(t;\hat{B})$ is $\phi(t,t)$. The extra variances, which can be estimated from standard software, are numerically negligible if the π_i's are small because the relative order of $D^{-1}(\beta_0)$ and $\phi(t_1,t_2)$ to $D^{-1}(\beta_0)V(\beta_0)D^{-1}(\beta_0)$ and $\sigma(t_1,t_2;\beta_0)$ is n/N. Thus, the price to pay for making the superpopulation inference is minimal if the inclusion probabilities are low. However, the variance estimators given in § 2 will be too small for making the superpopulation inference if the inclusion probabilities are high.

4. Numerical studies

We conducted a series of simulation experiments to evaluate the performance of the proposed superpopulation method. Failure times were generated from model (1·1) in which X_1 is a Bernoulli variable with 0·5 success probability and X_2 is a unit-variance normal variable with mean 0 if $X_1=1$ and mean $-0{\cdot}5$ if $X_1=0$; we set $\lambda_0=1$, $\beta_{01}=\beta_{02}=0{\cdot}5$. Failure times were subject to independent censoring by a Un[0, 1] variable, creating approximately 40% observed failures and 60% non-failures. We considered $N=1000$ and 10 000 together with $n=200$, 500 and 1000, and applied stratified simple random sampling. The population was stratified by failures versus non-failures and by $X_2+Z\leqslant 0$ versus $X_2+Z>0$, where Z is an independent standard normal variable. Equal numbers of units were drawn from each of the four strata in the survey population. For each combination of N and n, we generated 10 000 simulation samples.

Table 1. *Summary statistics for the simulation studies*

					Superpopulation		Finite-population	
N	n	n/N	Mean($\hat{B}_1$)	SE($\hat{B}_1$)	SEE	Cov. pr.	SEE	Cov. pr.
1000	200	20%	0·506	0·229	0·228	0·947	0·202	0·913
1000	500	50%	0·502	0·144	0·144	0·953	0·100	0·822
10 000	200	2%	0·509	0·231	0·227	0·945	0·224	0·942
10 000	500	5%	0·503	0·145	0·144	0·948	0·140	0·941
10 000	1000	10%	0·503	0·102	0·102	0·948	0·096	0·932

SEE, mean of standard error estimates. Cov. pr., coverage probability of 95% confidence interval.

Table 1 displays the Monte Carlo estimates for the sampling mean and sampling standard error of $\hat{B}_1$ as well as for the means of the standard error estimates and the empirical coverage probabilities of the 95% confidence intervals for β_{01} based on the superpopulation and finite-population variance estimators. Apparently, $\hat{B}_1$ has little, if any, bias. The superpopulation variance estimator reflects very well the true variance, and the corresponding confidence intervals have accurate coverage probabilities. When the inclusion probabilities are low, less than 5%, say, the finite-population variance estimator is close to the superpopulation variance estimator, and the corresponding confidence intervals have reasonable coverage probabilities. If the inclusion probabilities are high, however, the use of the finite-population variance estimator would result in severe underestimation of the superpopulation variance and poor coverage probabilities of the confidence intervals for the log hazard ratio.

The results for estimating β_{02} are similar to those for β_{01} and are thus omitted from the table. Additional studies revealed that the standard partial likelihood method which ignores the complex design of the survey could yield seriously misleading results; the sampling means of the maximum partial likelihood estimators of β_{01} and β_{02} were found to be about 0·47 and 0·36, respectively.

5. REMARKS

There is a long-standing debate in survey sampling on whether survey data should be analysed within the finite-population or superpopulation framework. The former approach is a valuable descriptive tool, but not well suited for regression analysis. When the survey population is fixed, there is no probability model governing the relationship between the response variable and covariates. Consequently, the interpretation of the regression parameter is awkward and the prediction of the response based on covariate values is difficult.

Numerically, the superpopulation approach advocated here may not be a dramatic departure from the traditional finite-population approach. In fact, the same parameter estimators are used, and the variance estimators are similar when the inclusion probabilities are low. However, by treating the survey population as a random sample from the superpopulation and by adjusting for this extra randomness in the variance estimation, one can make inference about parameters which have clear probabilistic interpretations, and the statistical conclusion extends beyond the survey population under study. Conceptually, this is a more natural and appealing approach to regression analysis.

A major motivation for Binder's finite-population inference is that it is estimating a well-defined quantity even when model (1·1) fails. This is also true for the superpopulation inference after a variance adjustment. Under misspecified models, B converges to a well-defined limit β^*, which is the root of $u(\beta)$ given in (2·1) (Lin & Wei, 1989). Thus, $\hat{B}$ converges to β^*, which is also the limit of $\hat{B}$ and B in the finite-population inference. Under the superpopulation approach, β^* can be interpreted as the value of β minimising a generalised Kullback–Leibler distance between the assumed and true conditional hazard functions (Hjort, 1992). By the arguments of the Appendix and Lin & Wei (1989), $N^{\frac{1}{2}}(\hat{B}-\beta^*)$ is asymptotically zero-mean normal with a covariance matrix consistently estimated by

$$\hat{D}^{-1}(\hat{B})\{\hat{Q}(\hat{B})+\hat{V}(\hat{B})\}\hat{D}^{-1}(\hat{B}),$$

where $\hat{Q}(\beta)=N^{-1}\sum(\xi_i/\pi_i)\hat{U}_i(\beta)\hat{U}_i'(\beta)$. Furthermore, $\hat{H}_0(t;\hat{B})$ converges to

$$H_0^*(t):=\int_0^t\{s^{(0)}(\beta^*,u)\}^{-1}\,dg(u),$$

and $N^{\frac{1}{2}}\{\hat{H}_0(t;\hat{B})-H_0^*(t)\}$ and $N^{\frac{1}{2}}\{\hat{F}_0(t)-e^{-H_0^*(t)}\}$ are asymmetrically zero-mean Gaussian with approximate covariance functions

$$\{\hat{\phi}^*(t_1, t_2) + \hat{\sigma}(t_1, t_2)\}, \quad \hat{F}_0(t_1)\hat{F}_0(t_2)\{\hat{\phi}^*(t_1, t_2) + \hat{\sigma}(t_1, t_2)\},$$

respectively, where $\hat{\phi}^*(t_1, t_2) = N^{-1} \sum (\xi_i/\pi_i)\hat{h}_i(t_1; \hat{B})\hat{h}_i(t_2; \hat{B})$.

Acknowledgement

This research was supported by the U.S. National Institutes of Health. The author thanks the reviewers and Zhiliang Ying for helpful comments.

Appendix

Finite-population sampling distributions of $\hat{U}(B)$ and $\hat{H}_0(t; \hat{B})$

In order for $\hat{U}(B)$, $\hat{H}_0(t; \hat{B})$ or any other statistics calculated from the survey data to be asymptotically normal, the survey design needs to permit the application of the central limit theorem to the normalised Horvitz–Thompson estimator; that is, for possibly vector-valued variates $(Z_1, \ldots, Z_N)$,

$$N^{-\frac{1}{2}} \sum_{i=1}^{N} \frac{\xi_i - \pi_i}{\pi_i} Z_i$$

is asymptotically zero-mean normal. The conditions required for the central limit theorem to hold are given by Hájek (1960, 1964) and Rosén (1972). By the central limit theorem, the finite-dimensional distributions of the processes

$$N^{\frac{1}{2}}\{\hat{G}(t) - G(t)\}, \quad N^{\frac{1}{2}}\{\hat{S}^{(0)}(\beta, t) - S^{(0)}(\beta, t)\}, \quad N^{\frac{1}{2}}\{\hat{S}^{(1)}(\beta, t) - S^{(1)}(\beta, t)\}$$

are asymptotically multivariate zero-mean normal since these three processes are normalised Horvitz–Thompson estimators at every t. We refer to such processes as normalised Horvitz–Thompson processes and assume that they are tight under the survey design of interest. The tightness, together with the finite-dimensional normality, implies that normalised Horvitz–Thompson processes converge weakly to zero-mean Gaussian processes. 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, §§ 2.9, 3.6, 3.7) can be applied to possibly stratified simple random sampling and can potentially be extended to other survey designs.

In addition to the above results, we will also appeal to the following lemma in our derivation of the asymptotic distributions for $\hat{U}(B)$ and $\hat{H}_0(t; \hat{B})$.

Lemma 1. *Let $W(t)$ and $Z(t)$ be two sequences of bounded processes. Suppose that $W(t)$ is monotone and converges to $w(t)$ uniformly in t in probability and that $Z(t)$ converges weakly to a zero-mean process with continuous sample paths. Then*

$$\int_0^t \{W(u) - w(u)\}\, dZ(u) \to 0, \quad \int_0^t Z(u)\, d\{W(u) - w(u)\} \to 0$$

uniformly in t in probability.

This lemma follows from the strong embedding theorem (Shorack & Wellner, 1986, p. 47) and Helly's Theorem (Serfling, 1980, p. 352); see the Appendix of Lin, Wei & Ying (1998) for the proof of a similar result.

We are now in position to study the asymptotic properties for $\hat{U}(B)$. By definition, $U(B) = 0$. Thus, $\hat{U}(B) = \hat{U}(B) - U(B)$, that is

$$N^{-\frac{1}{2}}\hat{U}(B) = N^{-\frac{1}{2}} \sum_{i=1}^{N} \int_0^\infty \frac{\xi_i - \pi_i}{\pi_i} X_i(t)\, dG_i(t) - N^{\frac{1}{2}} \int_0^\infty \left\{ \frac{\hat{S}^{(1)}(B, t)}{\hat{S}^{(0)}(B, t)}\, d\hat{G}(t) - \frac{S^{(1)}(B, t)}{S^{(0)}(B, t)}\, dG(t) \right\}. \tag{A·1}$$

The second term on the right-hand side of (A·1) equals

$$N^{\frac{1}{2}} \int_0^\infty \frac{\hat{S}^{(1)}(B,t)}{\hat{S}^{(0)}(B,t)}\, d\{\hat{G}(t)-G(t)\} + N^{\frac{1}{2}} \int_0^\infty \left\{\frac{\hat{S}^{(1)}(B,t)}{\hat{S}^{(0)}(B,t)} - \frac{S^{(1)}(B,t)}{S^{(0)}(B,t)}\right\} dG(t). \quad (A{\cdot}2)$$

As stated above, $N^{\frac{1}{2}}\{\hat{G}(t)-G(t)\}$ converges weakly to a zero-mean Gaussian process. The limiting process can be shown to have continuous sample paths via the Kolmogorov–Centsov Theorem (Karatzas & Shreve, 1988, p. 53). Clearly, $\hat{S}^{(0)}(B,t)$ is a monotone function in t. In addition, $\hat{S}^{(1)}(B,t)$ is a sum of two monotone functions since $X_i(t)=\max\{X_i(t),0\}-\max\{-X_i(t),0\}$. It then follows from Lemma 1 that

$$N^{\frac{1}{2}} \int_0^\infty \frac{\hat{S}^{(1)}(B,t)}{\hat{S}^{(0)}(B,t)}\, d\{\hat{G}(t)-G(t)\} = N^{\frac{1}{2}} \int_0^\infty \frac{s^{(1)}(B,t)}{s^{(0)}(B,t)}\, d\{\hat{G}(t)-G(t)\} + o_p(1). \quad (A{\cdot}3)$$

The second term of (A·2) equals

$$\int_0^\infty \left[\frac{N^{\frac{1}{2}}\{\hat{S}^{(1)}(B,t)-S^{(1)}(B,t)\}}{\hat{S}^{(0)}(B,t)} - \frac{N^{\frac{1}{2}}\{\hat{S}^{(0)}(B,t)-S^{(0)}(B,t)\}S^{(1)}(B,t)}{\hat{S}^{(0)}(B,t)S^{(0)}(B,t)}\right] dG(t),$$

which can be shown to be

$$\int_0^\infty \left[\frac{N^{\frac{1}{2}}\{\hat{S}^{(1)}(B,t)-S^{(1)}(B,t)\}}{s^{(0)}(B,t)} - \frac{N^{\frac{1}{2}}\{\hat{S}^{(0)}(B,t)-S^{(0)}(B,t)\}s^{(1)}(B,t)}{s^{(0)}(B,t)^2}\right] dg(t) + o_p(1) \quad (A{\cdot}4)$$

by Lemma 1, together with the uniform consistency of $\hat{S}^{(0)}$, $\hat{S}^{(1)}$, $S^{(0)}$, $S^{(1)}$ and G and the weak convergence of $N^{\frac{1}{2}}\{\hat{S}^{(0)}(B,t)-S^{(0)}(B,t)\}$ and $N^{\frac{1}{2}}\{\hat{S}^{(1)}(B,t)-S^{(1)}(B,t)\}$. The combination of (A·1)–(A·4) yields the linear approximation for $N^{-\frac{1}{2}}\hat{U}(B)$ given in (2·2) of § 2. Since the right-hand side of (2·2) is a normalised Horvitz–Thompson estimator, the central limit theorem implies that $N^{-\frac{1}{2}}\hat{U}(B)$ is asymptotically zero-mean normal with covariance matrix $V(B)$ given in (2·4).

To establish the weak convergence of $\hat{H}_0(t;\hat{B})$, we make the decomposition

$$\hat{H}_0(t;\hat{B})-\tilde{H}_0(t;B)=\{\hat{H}_0(t;B)-\tilde{H}_0(t;B)\}+\{\hat{H}_0(t;\hat{B})-\hat{H}_0(t;B)\}.$$

Clearly,

$$N^{\frac{1}{2}}\{\hat{H}_0(t;B)-\tilde{H}_0(t;B)\} = N^{\frac{1}{2}} \int_0^t \frac{d\{\hat{G}(u)-G(u)\}}{\hat{S}^{(0)}(B,u)} - N^{\frac{1}{2}} \int_0^t \frac{\{\hat{S}^{(0)}(B,u)-S^{(0)}(B,u)\}\, dG(u)}{\hat{S}^{(0)}(B,u)S^{(0)}(B,u)}.$$

It then follows from the arguments used to establish the asymptotic approximation for (A·2) that

$$N^{\frac{1}{2}}\{\hat{H}_0(t;B)-\tilde{H}_0(t;B)\} = N^{-\frac{1}{2}} \sum_{i=1}^N \frac{\xi_i-\pi_i}{\pi_i}\, v_i(t;B). \quad (A{\cdot}5)$$

By Taylor expansion, $N^{\frac{1}{2}}\{\hat{H}_0(t;\hat{B})-\hat{H}_0(t;B)\}=R'(t;B^*)N^{\frac{1}{2}}(\hat{B}-B)$, where B^* is on the line segment between $\hat{B}$ and B. It is easy to show that $R(t;B^*)$ converges to $r(t;B)$ uniformly in t. Thus,

$$N^{\frac{1}{2}}\{\hat{H}_0(t;\hat{B})-\hat{H}_0(t;B)\} = r'(t;B)D^{-1}(B)N^{-\frac{1}{2}} \sum_{i=1}^N \frac{\xi_i-\pi_i}{\pi_i}\, U_i(B)+o_p(1). \quad (A{\cdot}6)$$

Combining (A·5) and (A·6), we have

$$N^{\frac{1}{2}}\{\hat{H}_0(t;\hat{B})-\tilde{H}_0(t;B)\} = N^{-\frac{1}{2}} \sum_{i=1}^N \frac{\xi_i-\pi_i}{\pi_i}\{v_i(t;B)+r'(t;B)D^{-1}(B)U_i(B)\}+o_p(1),$$

which is a normalised Horvitz–Thompson process. Therefore, $N^{\frac{1}{2}}\{\hat{H}_0(t;\hat{B})-\tilde{H}_0(t;B)\}$ converges weakly to a zero-mean Gaussian process with covariance function (2·7).

REFERENCES

ANDERSEN, P. K. & GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100–20.

BINDER, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int. Statist. Rev.* **51**, 279–92.

BINDER, D. A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* **79**, 139–47.

BRESLOW, N. (1972). Contribution to the discussion of the paper by D. R. Cox. *J. R. Statist. Soc.* B **34**, 216–7.

COX, D. R. (1972). Regression models and life-tables (with Discussion). *J. R. Statist. Soc.* B **34**, 187–220.

GODAMBE, V. P. & THOMPSON, M. E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *Int. Statist. Rev.* **54**, 127–38.

HÁJEK, J. (1960). Limiting distributions in simple random sampling from a finite population. *Pub. Math. Inst. Hungar. Acad. Sci.* **5**, 361–74.

HÁJEK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Statist.* **35**, 1491–523.

HJORT, N. L. (1992). On inference in parametric survival data models. *Int. Statist. Rev.* **60**, 355–87.

KARATZAS, I. & SHREVE, S. E. (1988). *Brownian Motion and Stochastic Calculus.* New York: Springer-Verlag.

LIN, D. Y. & WEI, L. J. (1989). The robust inference for the Cox proportional hazards model. *J. Am. Statist. Assoc.* **84**, 1074–8.

LIN, D. Y., WEI, L. J. & YING, Z. (1998). Accelerated failure time models for counting processes. *Biometrika* **85**, 605–18.

ROSÉN, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement. I and II. *Ann. Math. Statist.* **43**, 373–97, 748–76.

SÄRNDAL, C.-E. (1978). Design-based and model-based inference in survey sampling. *Scand. J. Statist.* **5**, 27–52.

SCOTT, A. J. & WILD, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57–71.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

SHORACK, G. R. & WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics.* New York: Wiley.

VAN DER VAART, A. W. & WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes, with Applications to Statistics.* New York: Springer.

[*Received August* 1998. *Revised June* 1999]