



Influence Function Based Variance Estimation and Missing Data Issues in Case-Cohort Studies

STEVEN D. MARK

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA

sm7v@nih.gov

HORMUZD KATKI

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA

Received August 7, 2000; Revised February 13, 2001; Accepted February 22, 2001

Abstract. Recognizing that the efficiency in relative risk estimation for the Cox proportional hazards model is largely constrained by the total number of cases, Prentice (1986) proposed the case-cohort design in which covariates are measured on all cases and on a random sample of the cohort. Subsequent to Prentice, other methods of estimation and sampling have been proposed for these designs. We formalize an approach to variance estimation suggested by Barlow (1994), and derive a robust variance estimator based on the influence function. We consider the applicability of the variance estimator to all the proposed case-cohort estimators, and derive the influence function when known sampling probabilities in the estimators are replaced by observed sampling fractions. We discuss the modifications required when cases are missing covariate information. The missingness may occur by chance, and be completely at random; or may occur as part of the sampling design, and depend upon other observed covariates. We provide an adaptation of S-plus code that allows estimating influence function variances in the presence of such missing covariates. Using examples from our current case-cohort studies on esophageal and gastric cancer, we illustrate how our results are useful in solving design and analytic issues that arise in practice.

Keywords: case-cohort, confounding, Cox proportional hazards model, influence function, missing covariates, nested case-control, stratified sample, robust variance, weighted estimating equations

1. Introduction

Prentice (1986) proposed the case-cohort design as a means of estimating the relative risk in a Cox proportional hazards (CPH) model without collecting covariate information on all cohort members. In particular, he proposed measuring covariates on all the cases that occurred in the cohort and on a random sample of individuals drawn from the entire cohort (both cases and controls). Since this initial work there have been a number of papers using variations of the same basic sampling idea, but suggesting different estimating equations than those of Prentice. The primary motivations have been to further develop the properties of case-cohort estimators (Self and Prentice, 1988; Kalbfleisch and Lawless, 1988), to accommodate more complicated sampling schemes (Pugh, 1993; Robins, Rotnitzky, and Zhao, 1994; Barlow, 1994; Kim and De Gruttola, 1999; Borgan et al., 2000), to increase the efficiency of estimation of the relative risk (Robins, Rotnitzky, and Zhao, 1994; Kim and De Gruttola, 1999; Borgan et al., 2000), or to establish a computationally simpler estimator for the variance of the relative risk (Lin and Ying, 1993). In the past year this journal has

published several papers comparing features, proposing extensions, and examining through simulations the performance of a number of these estimating equations (Kim and De Gruttola, 1999; Borgan et al., 2000), as well as a paper reviewing aspects of variance estimation.

The latter work of Therneau and Li (1999) demonstrated that the variance estimator for the Self and Prentice (1988) estimating procedure, a procedure which leads to relative risk estimates asymptotically equivalent to the original estimates of Prentice (1986), can be explicitly expressed in terms of the variance one would have from a CPH model if covariates were measured on all the cohort members, plus the variance induced by the sampling procedure. In addition to providing a conceptually important insight into the Self and Prentice variance, the re-expression also gave rise to an algebraic formulation that is easy to implement in the S-plus software. Therneau and Li also reviewed two "robust variance estimators": the estimator of Lin and Ying (1993) and the estimator of Barlow (1994). Barlow (1994) proposed his variance estimator based on analogy with the infinitesimal jackknife estimator of the variance of the empirical influence function for the full cohort CPH model. He demonstrated by simulation that this estimator accommodated a more complicated sampling scheme and estimating equation than the one proposed by Prentice (1986). The goal of our paper is to derive the influence function variance estimator for all case-cohort designs, and discuss their applied uses and implementation. In section 2 we develop a notation for estimating equations for the CPH model relative risk that includes both the full cohort and the case-cohort estimating equations. In section 3 we utilize results from Reid and Crepeau (1985), and Lin and Wei (1989), and derive the influence function variance estimator for the relative risks that solve these general estimating equations. Having expressed the influence function variance estimator in a notation that covers all the proposed estimating equations, we then explicitly demonstrate the relation between the robust estimators proposed by Lin and Ying (1993) and by Barlow (1994), and show that they are both influence function variance estimators. We emphasize that our derivation of this influence function requires that we know the sampling weights. For some estimators we derive the correct influence function when known sampling weights are replaced with observed sampling fractions. In section 4, we discuss the impact on variance estimation that occurs when some cases are missing covariate information. Since all the proposals explicitly devised for case-cohort sampling assume that covariate information exists for all cases, the latter issue has not received attention. Our experience, based largely on covariate measurements obtained from biological specimens (Mark et al., 2000) is that even when the intention is to measure covariates on all the cases, there is always a proportion of cases for whom this is not possible. We also discuss designs and give analytic formula for the analysis of case-cohort studies in which deliberate fractional and/or differential sampling of cases occurred. We give examples from our work in which considerations of efficient allocation of resources and the control of confounding gave rise to such designs.

2. Case-Cohort Design: Background and Notation

We assume that the risk of failure for an individual in our cohort follows the Cox proportional hazards model (CPH), $\lambda(u|Z_i) = \lambda_o(u) \exp\beta Z_i(u)$ where $\lambda_o(u)$ is an unspecified

hazard at time u , and $Z_i(u)$ a vector of individual level covariates. Under the usual assumption that censoring time C_i is conditionally independent of the failure time, T_i (Anderson et al. 1991), the relative risk β is consistently estimated by the $\hat{\beta}$ that solves the partial likelihood score equations (1) (Andersen et al., 1991). Using standard counting process notation where $N_i(u)$ is the event counting process ($N_i(u) = 1$ iff $T_i \leq u$, and $T_i \leq C_i$), $Y_i(u)$ is the at risk process ($Y_i(u) = 1$, iff $(C_i \wedge T_i) \leq u$), and τ is the total time of the study, the estimating equation for the CPH model is

$$U(\beta) = \sum_{i=1}^n \int_0^\tau \{Z_i(s) - \bar{Z}(\beta, s)\} dN_i(s) = 0 \tag{1a}$$

where

$$\bar{Z}(\beta, s) = \frac{\sum_{j=1}^n Y_j(s) Z_j(s) \exp\{\beta Z_j(s)\}}{\sum_{j=1}^n Y_j(s) \exp\{\beta Z_j(s)\}} \tag{1b}$$

$\bar{Z}(\beta, s)$ is the weighted mean of the covariates for those in the risk set at time s , where the weights are proportional to the probability of failure at s (Anderson et al., 1991).

In epidemiologic research the $Z_i(u)$ often consists of two components, $Z_i(u) = \{V_i(u), J_i(u)\}$, where the $J_i(u)$ are easily obtainable observations, and the $V_i(u)$ require more elaborate measurement procedures. For instance, $J_i(u)$ could be age, sex, weight, or other measurements obtained by interview or a simple physical exam, while the $V_i(u)$ might be the results of costly laboratory tests requiring biologic specimens such as blood or other body tissue. Prentice (1986) proposed the case-cohort design because he recognized that the efficiency in estimating relative risk is largely constrained by the number of cases, and that measuring $V_i(u)$ on all n cohort members might be fiscally or logistically prohibitive. Though the sampling scheme he proposed accommodates both sampling while the cohort is under study and sampling at the end of the study, we will, for the sake of concreteness and without loss of generality, assume that the sampling occurs at the end of the observation period and that the covariates $Z_i(u)$ are time invariant. In this setting all individuals in the cohort have observable $(J_i, Y_i(u), N_i(u))$, while V_i is measured only on the cases and the members of the random sample, called the subcohort, \tilde{C} . To estimate β , Prentice proposed a set of estimating equations identical in form to (1), with the exception that now the average covariate, $\bar{Z}(\beta, s)$ was to be estimated from a specific subset of the individuals with the fully observed covariates (J_i, V_i) . The subsequent proposals for estimation in case-cohort designs follow the spirit of Prentice's proposal and differ only in the choice of what estimator, $\bar{\tilde{Z}}(\beta, s)$, is used to replace the unobservable $\bar{Z}(\beta, s)$. Assuming for now that V_i is measured on all cases we can use a single notation and write all these various proposals as

$$\tilde{U}(\beta) = \sum_{i=1}^n \int_0^\tau \{Z_i(s) - \bar{\tilde{Z}}(\beta, s)\} dN_i(s) = 0 \tag{2a}$$

$$\bar{\bar{Z}}(\beta, s) = \frac{\sum_{j=1}^n w_j(s) r_j(s) Y_j(s) Z_j \exp \{\beta Z_j\}}{\sum_{i=1}^n w_j(s) r_j(s) Y_j(s) \exp \{\beta Z_j\}} \quad (2b)$$

We call $r_i(s) \in \{0, 1\}$ the risk set selector, since it is an indicator variable that determines what members at risk ($Y_i(s) > 0$) to include in the estimation of $\bar{\bar{Z}}(\beta, s)$; $w_i(s) > 0$ is the weight that a selected individual contributes. The proposed approaches suggest three different choices for $r_i(s)$: $r_i(s) = 1$ if $i \in \tilde{C}$, or i is the failure at s ($dN_i(s) = 1$) (Prentice, 1986; Barlow, 1994); $r_i(s) = 1$ if $i \in \tilde{C}$ (Self and Prentice, 1988; Lin and Ying, 1993; Therneau and Li, 1999; Borgan et al., 2000); $r_i(s) = 1$ for all individuals (subcohort members and all cases) with completely measured covariates (Kalbfleisch and Lawless, 1988; Pugh, 1993; Robins, Rotnitzky, and Zhao, 1994). The suggestions for $w_i(s)$, though more varied, reduce to two themes: Prentice (1986) and Self and Prentice (1988) use $w_i(s) = 1$ for everyone. Thus, Self and Prentice estimate $\bar{\bar{Z}}(\beta, s)$ by a simple average of the random sample that composes \tilde{C} . The other proposals all use weights equal to the inverse of the sampling fraction and accommodate more complicated sampling schemes such as stratified sampling. For Kalbfleisch and Lawless (1988), Pugh (1993), and Robins, Rotnitzky, and Zhao (1994), these are time invariant weights with $w_i(s) = 1$ for all cases; for the non-cases, $w_i(s)$ is the inverse of the probability of selecting a non-cases into \tilde{C} . Barlow (1994), Therneau and Li (1999), and Borgan et al. (2000) allow the $w_i(s)$ to potentially vary with time. Specifically, they propose that for members of the subcohort, $w_i(s)$ be the inverse of the observed proportion of the cohort that is at risk and in the subcohort at s . Note that for $r_i(s) = 1$, $w_i(s) = 1$ for all n , equation (2) becomes the estimating equation for the full CPH model (1).

3. Influence Function Based Variance Estimation

The properties of the relative risk estimator $\hat{\beta}$ from the full CPH, and the simplicity in form and calculation of its variance, derives from the fact that the score for the partial likelihood has a martingale structure (Anderson et al., 1991). That is, conditional on events prior to s , the score increments $\{Z_i(s) - \bar{\bar{Z}}(\beta, s)\} dN_i(s)$ have mean 0 and are independent. In case-cohort estimation either some cases are excluded from the weighted mean (2b) or, the risk set selectors, $r_j(s)$, and / or weights, $w_j(s)$, are not functions of information prior to time s . Hence, it is not possible to construct a nested sequence of conditioning events that lead to mean 0 independent increments of the score. Prentice (1986) showed that a non-zero covariance occurs between all increments $t < s$ for which the case at s is not in \tilde{C} , and explicitly calculated the covariance under the CPH model and the sampling scheme. Self and Prentice (1988), and Borgan et al. (2000) for the weighted version of a stratified Self and Prentice estimator, calculated the variance by decomposing the estimating equations into the martingale process of the full cohort and the difference between the average covariate estimator for the full CPH model, $\bar{\bar{Z}}(\beta, s)$, and the case-cohort model, $\bar{\bar{Z}}(\beta, s)$. They proved that the two parts are asymptotically uncorrelated, and used standard martingale theory for the variance of the first part, and finite sampling arguments to

obtain the variance of the second. All these derivations were model based since the requirement that $\{Z_i(s) - \bar{Z}(\beta, s)\}$ have conditional mean 0 depends on the CPH being a correct specification of the hazard.

Influence functions are a general approach for obtaining non-model based variance estimates. In particular, for an estimator $\hat{\beta}$ that can be written as $\hat{\beta} = t(\hat{F}_n)$, where $t(\cdot)$ is a function independent of sample size n , and \hat{F}_n is the empirical distribution function of the observed data z_i , the variance of $\hat{\beta}$ can be found through a first order differential analysis of $t(\cdot)$ in the neighborhood of the true distribution function F_o (Huber, 1977).

$$t(\hat{F}) \approx t(F_o) + \frac{1}{n} \sum_{i=1}^n U(z_i, F_o) \tag{3}$$

The influence functions $U(z_i, F_o)$ are independent with expectation 0. Under regularity conditions (Huber, 1977; Anderson et al., 1991) which assure that the remainder term in the approximation is of the order $o_p(n^{-1})$, the variance of $t(\hat{F})$ is $\frac{1}{n} E_{F_o}(U^2(z_i, F_o))$. Since the influence function variance only depends on the expression of β as a functional of the empirical cumulative distribution, and makes no assumptions regarding the relationship of β to parameters that characterize the underlying distribution function, it is robust to misspecification of those parameters. An estimator of the variance is obtained by averaging the sums of squares of the empirical influence function $\hat{U}(z_i, \hat{F}_n)$.

Reid and Crepeau (1985) derived the influence function for the $\hat{\beta}$ defined by the CPH estimating functions (1) by expressing $\hat{\beta}$ as a functional of the joint, $F(t, z, \delta)$, and marginal, $F(t, z)$, cumulative distribution functions of the failure time T_i , the covariate vector Z_i , and the observed event indicator $\Delta_i = I(T_i < C_i)$. Taking all expectations with respect to $F(t, z)$, and dropping the distribution function from the argument, the influence function for $\hat{\beta}$ is

$$U(T_i, Z_i, \Delta_i) = I^{-1} W_i(T_i, Z_i, \Delta_i) \tag{4}$$

where

$$I = \int_0^\tau \delta \left\{ \frac{E[I(T_i \geq t)Z_i Z_i' \exp\{\beta Z_i\}]}{E[I(T_i \geq t) \exp\{\beta Z_i\}]} - \left(\frac{E[I(T_i \geq t)Z_i \exp\{\beta Z_i\}]}{E[I(T_i \geq t) \exp\{\beta Z_i\}]} \right) \times \left(\frac{E[I(T_i \geq t)Z_i \exp\{\beta Z_i\}]}{E[I(T_i \geq t) \exp\{\beta Z_i\}]} \right)' \right\} dF(t, z, \delta) \tag{5}$$

and

$$w_i(T_i, Z_i, \Delta_i) = \Delta_i \left\{ Z_i - \frac{E[I(T_i \geq t)Z_i \exp\{\beta Z_i\}]}{E[I(T_i \geq t) \exp\{\beta Z_i\}]} \right\} - \int_0^\tau \left\{ Z_i - \frac{E[I(T_i \geq t)Z_i \exp\{\beta Z_i\}]}{E[I(T_i \geq t) \exp\{\beta Z_i\}]} \right\} \frac{\exp Z_i \beta \delta I(T_i \geq t)}{E[I(T_i \geq t) \exp\{\beta Z_i\}]} \times dF(t, z, \delta) \tag{6}$$

As noted in section 2, the $\hat{\beta}$ for the full CPH is not only a solution to the classic CPH estimating equations 1, but also a solution to equation 2. Provided that the risk set selectors and weights in equation 2 are non-stochastic functions of T_i , Z_i , and Δ_i , and are finite and bounded away from 0, Reid and Crepeau's derivation (1985) of the influence function of the full CPH applies to the $\hat{\beta}$ that solves (2). For the full CPH model $r_i(s)$ and $w_i(s)$ equal 1, and can be regarded as constant functions of the random variables.

Incorporating the $r_i(s)$ and $w_i(s)$; re-expressing Reid and Crepeau's results in terms of counting process notation; replacing expectations by their sample averages; and replacing $F(t, z, \delta = 1)$ by the scaled aggregated counting process

$$n^{-1}\bar{N}(t) = n^{-1} \sum_{i=1}^n N_i(t) \quad (7)$$

the empirical influence function for (2) is

$$\hat{U}_i = \hat{I}^{-1} \hat{W}_i. \quad (8)$$

\hat{I} , though no longer the observed information of a partial likelihood, has the same functional form as \hat{I} in the full cohort model (Anderson et al., 1991) and

$$\begin{aligned} \hat{W}_i = & \int_0^{\tau} Y_i(s) \{Z_i - \bar{Z}(\hat{\beta}, s)\} dN_i(s) - \int_0^{\tau} \left(w_i(s) r_i(s) Y_i(s) \right. \\ & \left. \times \{Z_i - \bar{Z}(\hat{\beta}, s)\} \exp\{\hat{\beta} Z_i\} \frac{d\bar{N}(s)}{\sum_{j=1}^n w_j(s) r_j(s) Y_j(s) \exp\{\hat{\beta} Z_j\}} \right) \end{aligned} \quad (9)$$

The influence function estimator for the variance of $\hat{\beta}$ is

$$\text{var}(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^n \hat{U}_i \hat{U}_i' \quad (10)$$

Without invoking influence function arguments Lin and Wei (1989) derived a robust variance estimator for the CPH model identical to (10) with $w_i(s) = r_i(s) = 1$. Assuming (assumption VII 2a in Anderson et al., 1991) that the numerator and denominator of $\bar{Z}(\beta, s)$ in equation 2 converge to their expectations, the arguments for consistency given by Lin and Wei apply to equation (10). In her thesis, Pugh (1993) gives a detailed proof of the consistency of (10) based on weaker assumptions.

As show by (2), the variation in case-cohort estimating equations is entirely due to variation in $r_i(s)$ and $w_i(s)$. To gain insight regarding the algebraic effect these differences produce on the influence function variance, it is useful to expand (8) into two terms. As noted by Reid and Crepeau (1985) the first term of the empirical influence function

$$\hat{I}^{-1} \int_0^{\tau} Y_i(s) \{Z_i - \bar{Z}(\hat{\beta}, s)\} dN_i(s)$$

is exactly the expression one would have if the estimating equation were the sum of independent random variables. This term does not reflect the risk set aspect of the full CPH or case-cohort models. The second term,

$$\widehat{I}^{-1} \int_0^\tau w_i(s)r_i(s)Y_i(s)\{Z_i - \bar{Z}(\widehat{\beta}, s)\} \exp\{\widehat{\beta}Z_i\} \frac{d\bar{N}(s)}{\sum_{j=1}^n w_j(s)r_j(s)Y_j(s)\exp\{\widehat{\beta}Z_j\}}$$

accounts for the impact of the *i*'th observation on the estimation of all the risk sets to which it contributes (risk sets at times *s* where $w_i(s)r_i(s)Y_i(s) > 0$). This decomposition is algebraically more transparent in Cain and Lange's (1984) informal derivation of the empirical influence function which only requires taking partial derivatives of the estimating equations.

The influence variance estimator (10) is a general expression for all the previously proposed robust variances for case-cohort estimators. Without reference to influence function arguments Lin and Ying (1993) derived a robust variance estimator for the Self and Prentice equations. The Lin and Ying estimator is produced when one substitutes in equation 9 the Self and Prentice values of $w_i(s)=1$ for all *i*, and $r_i(s)=1$ iff $i \in \tilde{C}$. Similarly, using the appropriate definition for $w_i(s)$ and $r_i(s)$ given in section 2, produces Barlow's suggested variance for Prentice type estimating equations, and the case-cohort version of the estimator of Pugh (1993) and Robins, Rotnitzky, and Zhao (1994). The finite sample performance of each of these estimators has been examined in simulations (Pugh, 1993; Barlow 1994; Therneau and Li, 1999; Kim and De Gruttola, 1999).

In a number of the procedures for estimating $\widehat{\beta}$ described in section 2, the $w_i(s)$ are the inverse of the probability that person *i* is sampled. If one replaces the true probability of sampling with an estimated probability, then (8) will not in general be a consistent estimate of the influence function, and (10) will not consistently estimate the variance. For instance, one could modify the Pugh (1993) and Robins, Rotnitzky, and Zhao (1994) equations, and substitute the observed proportion of controls sampled, for the true sampling probability. For missing at random data, Robins, Rotnitzky, and Zhao (1994) derived a general expression relating the influence function with known weights, to the influence function with estimated weights. They proved that the variance of the $\widehat{\beta}$ using these estimated weights, is less than or equal to the variance when the true sampling probability is used. Pugh's simulations confirm these results. Applying proposition 6.1 of Robins, Rotnitzky, and Zhao (1994), it is simple to derive and estimate the correct influence function for this estimator where the true probability of including a control in the subcohort, is replaced by the observed proportion of sampled controls. For cases the influence function remains unchanged and is estimated by (8). For the controls in the subcohort, instead of using (8) the correct influence function is estimated by

$$\widehat{U}_i = \widehat{I}^{-1} \{ \widehat{W}_i - \widehat{E}[\widehat{W}_i | \tilde{C}, \delta_i = 0] \}$$

where $\widehat{E}[\widehat{W}_i | \tilde{C}, \delta_i = 0]$ is the observed average of \widehat{W}_i among the controls in the subcohort. The unsampled controls now also have a non-zero influence which is estimated by

$$\widehat{U}_i = \widehat{I}^{-1} \{ -\widehat{E}[\widehat{W}_i | \tilde{C}, \delta_i = 0] \}.$$

With these modifications (10) is a consistent estimate of the variance of $\hat{\beta}$. Our simulations (data not shown) confirm the performance of the variance estimator based on these modified influence functions. The situation is less tractable for Barlow's (1994) and Therneau and Li's (1999) proposal for time varying weights. They sample the subcohort with a fixed time invariant probability, but for $w_i(s)$ suggest using the inverse of the fraction of the cohort at risk that is in the subcohort at time s . Provided that the subcohort members are all sampled with the same probability, then the weights in (2b), though varying from one time to the next, are invariant at each time s . Hence, they factor out and produce the usual estimating equations of Prentice (1986), and Self and Prentice (1988). Then the variance given by (10) is consistent. However, in the nontrivial case, such as the stratified random sample proposed by Borgan et al. (2000), (8) has not been shown to be the correct influence function. Furthermore, the approach we used above for correcting the influence function for weight estimation, implicitly requires the parameters in the model for the estimated weights to be ancillary to the relative risk parameters, β , of the CPH model. Hence it is not applicable to the time varying weights of Barlow (1994) and Therneau and Li (1999) which involve the probability distribution of $Y_i(s)$, and thus are implicitly a function of β . It is of interest to note that estimating equations (2) do include the estimating equations for the nested case-control design in which random samples of controls are actually drawn at each failure time s , and in which time varying weights $w(s)$ can be used (Borgan, Goldstein, and Langholz, 1995). If controls are chosen by binomial sampling at each time s , (10) is a consistent robust variance estimate for these studies. In the usual nested case-control design in which a fixed number of controls are drawn at each time s , the influence function(8) requires modifications analogous to those given above for time invariant weights.

Borgan et al. (2000) proposed a Self and Prentice type estimator for stratified case-cohort studies where the data is analyzed by a non-stratified CPH model. Their estimator requires case weights that depend on the stratifying variables. They propose both time invariant and time dependent weights, where the latter are estimated as described above. They acknowledge, however, that their variance estimator is also not valid for their time dependent version. They state that they could not derive a consistent estimator, but that they could demonstrate that the difference between using the true and the estimated weights was asymptotically not negligible. Through simulation they found the empirical variance was lower in the time dependent weight version than in the time invariant version.

Rather than using a non-stratified model as Borgan et al. (2000), an alternative approach to analyzing a stratified sampling design is to use a stratified CPH model with separate baseline hazards for each stratum. This allows one to control for the effect of the stratifying variables without modeling their effects on the hazard. The only modification in relative risk estimation is to replace for each individual i in stratum k , the overall covariate average $\bar{Z}(\beta, s)$ in (2b), with stratum specific covariate averages $\bar{Z}_k(\beta, s)$. To estimate the influence function variance we must define aggregated counting process $\bar{N}_k(t)$ separately for each stratum. \bar{W}_i in equation 9 is then calculated by substituting $d\bar{N}_k(s)$ for $d\bar{N}(s)$, $\bar{Z}_k(\beta, s)$ for $\bar{Z}(\beta, s)$, and restricting the summation in the denominator of the second term of (9) to the individuals in stratum k . With those changes, (10) is the influence function variance estimator for the stratified Cox proportional hazards model. In the discussion

section we describe the design and analytic utility of stratified sampling and analysis for the control of confounding.

4. Variance Estimation with Cases Missing Covariate Measurements with Examples from Esophageal and Gastric Cancer Research

With the exception of Pugh (1993) and Robins, Rotnitzky, and Zhao (1994) who propose a general scheme for estimating the CPH model with missing covariates, approaches specifically proposed for the case-cohort study assume that all cases have complete covariate measurements, $\{V_i, J_i\}$. Since obtaining V_i is usually either expensive, intrusive, or logistically difficult to obtain, V_i may not be assessable on all individuals even if that were the investigator's intent. We have recently implemented nine case-cohort studies of esophageal and gastric cancers, all of which had unanticipated missing case data. In our studies V_i consisted of measurements made on serum samples collected at the beginning of a chemoprevention trial. For approximately 10% of the cases (Mark et al., 2000) these serum measurement were unavailable. Some of this missingness was known early in the design stage; some did not become apparent until analysis. In the four studies in which we intended to measure V_i on all the cases, missingness was due to completely random processes and occurred with equal frequency among cases and non-cases. Causes of missingness included the inability to obtain adequate blood at the start of the study; mishaps in blood processing, transport or storage; or mishaps during the laboratory measurement procedures. In five of our studies we deliberately sampled only a fraction of available cases. In contrast to the larger studies where we wished to obtain as precise an estimate of the disease-serologic marker relationship as possible, the goal of these smaller studies was to obtain more accurate knowledge of the relative risks and exposure distributions before committing irreplaceable resources on a larger study. In one instance we examined the relationship of a serum measurement of fumonosins (a fungal toxin) to esophageal cancer. Though animal studies and ecological association had suggested a link between fumonosins and cancer, no data existed for humans. We sampled 100 of the possible 1179 cancer cases. On analyses (Abnet et al., 2001) we found that the measurement error associated with the procedure for quantifying fumonosins in human serum was too great to allow detection of moderately elevated relative risks. Therefore we have decided not to augment the study with the full complement of cases until measurement issues can be resolved. For a study examining serologic markers of H. Pylori infection and gastric cancer we sampled 200 cases (Limburg et al., 2001). In contrast to the fumonosin study, measurement of H. Pylori infection in serum is a well established test, and the association with gastric cancer is sufficiently documented that in 1994 the International Agency for Research on Cancer classified H. Pylori as a class I human carcinogen (Moller et al., 1994). Though not supported by much data, the general consensus has been that this risk extends only to cancers arising in the body of the stomach and not those arising in the cardia of the stomach. Our study has an unusually larger number of gastric cancers from the cardia (Mark et al., 2000), so to test this hypothesis we sampled 100 cases from each of the two sites. We found that H. pylori increased the risk at both sites (Limburg et al., 2001). Despite a recent pooled analysis of all prospective studies of

H. pylori and gastric cancer in which we participated (Helicobacter and Cancer Collaborative Group, 2001), there is still considerable disagreement as to whether the *H. pylori*-gastric cardia association is real. We are now extending our study to include an additional 800 cases of the gastric cardia cases and will have a definitive answer at completion. Since in cancer research the number of exposures measurable in biological specimens continues to grow at a fast rate, and the possibilities for allocation are increasing, we anticipate that case-cohort studies with fractional case sampling will become more common.

Whether missingness is by happenstance or by design, provided we assume or know that the lack of measurement is unrelated to the value of V_i , and, for the time being J_i , the case-cohort estimating equations (2) remain asymptotically unbiased when limited to those cases with observed V_i . If one simply ignores the missing cases, the usual model based variance estimation procedures of Prentice (1986), or Self and Prentice (1988) continue to be consistent. Though the influence function variance (8) also remains correct as written, caution must be taken in implementation. In (9), \widehat{W}_i requires that the distribution function $F(t, \delta, z)$ be approximated by the aggregated counting process (7), where the sum is over all n members of the cohort. With missing cases this requires the estimating procedure use data from the full cohort, and not just the data from the sampled individuals with complete observations. This also implies that if, contrary to our assumption, $V_i(s)$ were time varying, we need to obtain measurements on all subcohort members at all times s where $dN(s) = 1$, regardless of whether the $V_i(s)$ were missing for the case at s . In the discussion section we provide code with which to estimate influence function variances when cases are missing covariate data.

A more general missing by design scenario would allow us to sample different fractions of the cases depending upon other baseline characteristics, J_i , such as age and sex. Because of the strong covariance of disease with age, and our desire to estimate effects within age groups, in our studies we did sampling (Mark et al., 2000) stratified on age, sex, and outcome. In the studies where we sampled only a portion of the cases, the case sampling fraction was greater in the younger age strata since we wanted to be able to estimate the within age-stratum risk with nearly equal precision. For such designs with differential case sampling, estimating equation (2) is no longer asymptotically unbiased. If the sampling fractions dependence on J_i were discrete, for example if the sampling fraction for cases were determined by three different categories of age, then one could estimate with any of the above sampling schemes provided one used a CPH model stratified on these categories. This was how we analyzed the data on the association of serum selenium and esophageal and gastric cancers (Mark et al., 2000). Pugh (1993), and Robins, Rotnitzsky, and Zhao (1994), offer an approach that accommodates sampling fractions as continuous functions of J_i , and that permits an unstratified analysis as in Borgan et al. (2000) even for a stratified sampling scheme. Rather than (2) their estimating equation can be written

$$\sum_{i=1}^n \int_0^{\tau} w_i \{Z_i(s) - \bar{Z}(\beta, s)\} dN_i(s) = 0 \quad (11)$$

where now the weights, w_i (the inverse of the sampling fraction for individual i), are used not only in the calculation of $\bar{Z}(\beta, s)$, but also to account for the incomplete measurement

of cases. With (9) modified so that the first term is multiplied by w_i , the variance of $\hat{\beta}$ that solves (11) can be estimated by (10). We are currently using these equations to analyze a case-cohort study where case sampling depended on J_i .

5. Discussion

Barlow (1994) was the first to explicitly use an empirical influence function to estimate the variance of the relative risk in samples from a case-cohort setting. The estimating functions he proposed were a weighted version of the Prentice estimating equations. He demonstrated the ability of this approach to accommodate a more complex sampling scheme than the simple design of Prentice, and showed that the suggested variance performed well in simulation.

In addition to Barlow's modification, a number of other estimating equations have been proposed to accommodate other sampling designs and to increase the ease and efficiency of estimation. By defining risk set selecting indicators, $r_i(s)$, and weights, $w_i(s)$, we have written all case-cohort estimating equations in the form of a single estimating equation (2). Additionally, the usual full cohort CPH estimating equations (1), are the special case of (2) where $r_i(s) = w_i(s) = 1$ for all i . Writing the estimating equations in a form that subsumes both the full cohort and the case-cohort models, has allowed us to use results on variance estimation for the relative risk in a full-cohort CPH model, (Reid and Crepeau, 1985, Lin and Wei, 1989) and derive influence function variances for the relative risks from case-cohort designs. When using the appropriate risk set selectors, $r_i(s)$ and weights $w_i(s)$ detailed in section 2, the influence function variance estimator we propose reproduces the specific estimators of Barlow (1994), Lin and Ying (1993), and the weighted version of the Lin and Wei estimate suggested by Pugh (1993).

Influence function variances for the "relative risk" parameters, β , that solve estimating equations (2) are based on an asymptotic approximation of the β 's expressed as a functional of the cumulative distribution functions (CDF). The CDF's (survival and censoring distributions) are estimated by the empirical CDF's. The property of the empirical CDF as a non-parametric maximum likelihood estimator does not depend on whether the underlying hazards of the survival times are correctly specified by a Cox proportional hazards model. Thus the influence function variances represented by (10) are robust to the modeling assumption about the hazards. From our viewpoint, however, the primary utility of the influence function variance is not its robustness to model misspecification. Rather the practical advantage of (10) is that it provides variance estimators that accommodate a wide variety of complex sampling schemes and estimating equations. Having a single algebraic form for the estimating equations (2) and their variance (10) also computationally facilitates the use of one data format and one statistical program to analyze different case-cohort studies.

For instance, as discussed in section 4, we have initiated a number of case-cohort studies of cancer using stratified sampling. Borgan et al. (2000) showed that such stratified sampling can lead to increases in the efficiency of estimating relative risks. Though efficiency was one concern in the design of our studies, the control of confounding was a more important consideration. It is well established that age is a prime determinant of

cancer rates, with rates increasing exponentially with age. Consequently, the most common means of cohort sampling in cancer research is nested case-control studies. For such studies, controls for individual risk sets for each case can be chosen using small age intervals for matching. This age matching allows the dependency of cancer rates on age to be conditioned out of the risk estimates. Though these nested case-control designs have considerable utility, there are situations, particularly if one is studying multiple cancer sites as we are, where the case-cohort design is preferable. Stratified sampling in case-cohort designs is one means of implementing confounding control by matching. For all our case-cohort studies we sampled from six stratum defined by sex and three ten-year age groups. For our large study on the relationship of serum selenium to esophageal and gastric cancers, we analyzed the data using a CPH model stratified on all six sampling strata (Mark et al., 2000). We controlled for the residual within stratum age variation by stratum specific covariates. We estimated the variance using the influence function variance estimator for this stratified CPH analysis given at the end of section 3. Since in this study the only missing cases were those missing by chance, we could have used a single baseline hazard and the version of (2) where the weights $w_i(s)$ account for the sampling dependence on covariates (Borgan et al., 2000). That approach, however, would have required us to control for the entire thirty year variation in age by modeling the effect on the hazard rather than by matching. For a smaller study that we are currently analyzing on the relationship of Epstein-Barr virus and gastric cancers, we are using a model stratified on age, but which accounts for the sex stratification with sampling weights. We chose this approach because sex is not a confounder in this data set and, unlike the selenium study where we had 1079 cases, here we have only 200 cases and are more concerned about efficiency of estimation. In this study, since the case sampling fractions were dependent upon sex, we must use equation 11 that accounts for this differential sampling, rather than equation 2.

Our derivation of the influence function estimated by (8), requires that the weights $w_i(s)$ be known functions of the observable random variables. If one estimates $w_i(s)$, for instance, replacing true binomial sampling probabilities by observed sampling proportions, then the variance estimator (10) is no longer guaranteed to be consistent. In section 3 we give the correct influence function when time invariant $w_i(s)$ are estimated. As we state, the analogous correction can be used for to generate an influence function variance for the usual time varying weights proposed for nested case-control sampling (Borgan, Goldstein, and Langholz, 1995). However, we have not derived an influence function for the estimated time varying $w_i(s)$ proposed by Barlow (1994) and by Therneau and Li (1999) for case-cohort sampling. The contention that (10) is a consistent estimator for those weights is unsupported, and we suspect that it is incorrect.

Though case-cohort designs are commonly used to estimate relative risks, we have developed estimators for the absolute risk and population attributable risk in stratified samples with missing data. We estimated that variation in serum selenium levels may account for one-fourth of the cancer epidemic in the region of Linxian, China (Mark et al., 2000). We are currently preparing a manuscript showing the derivation of these estimators, documenting their finite sample properties, and providing computer code for implementation. The same algebraic components and missing case adjustments required for the influence function estimators, are necessary for the implementation of these absolute risk estimators.

Previous case-cohort estimating procedures have assumed the availability of full covariate information on all of the cases. Our experience indicates that by chance or design, this is frequently not the situation. As long as the time of death, s , of each uncensored case is observed, and the covariate values for the other individuals at risk and with $r_i(s) = 1$ at s are observed, a consistent influence function variance can be calculated. Thus, in the case-cohort setting the variance is estimable. However, the procedure for estimating this variance through standard software packages requires modification. Therneau and Li (1999) provide S-plus code that can be used to estimate the Lin and Ying (1993) variance for the Self and Prentice (1988) equations when no cases are missing covariate data. As mentioned above, this is the influence function variance for those estimating equations. If missing cases are present, their procedure needs to be modified in the following manner: 1) estimate $\hat{\beta}$ excluding the case(s) j at s with missing covariates; 2) assign the missing case(s) the covariate value $Z_j = \bar{Z}(\hat{\beta}, s)$ calculated from the observed data; 3) treat the missing case(s) as if it occurred outside the subcohort and assign it a large negative offset (Therneau and Li suggest an offset of -100); 4) re-estimate and obtain the influence function variance using the code they give in section 2.2. Since the Borgan et al. (2000) equations are a stratified version of the Self and Prentice estimator, the same approach applies.

As discussed in section 4, when the missingness occurs by design, or if for some other reason the missingness is thought to depend on other observed covariates, then the contribution of each individual case to the estimating equation must be weighted by the inverse probability of sampling the cases as in (11). The variance for these equations is obtained as above with the exception that in steps one and four the S-plus weighted CPH procedure is used, and in step 4 no offset or cluster command is required.

We have focused on providing influence function variance estimators for all of the case-cohort estimating equations encompassed by (2), and have not extensively addressed which of the varying options to use. In some cases, for example stratified sampling, if a non-stratified Cox model is used for estimation, certain choices such as the Prentice (1986) or Self and Prentice (1988) estimators are eliminated. Several papers have used simulations (Kim and De Gruttola (1999); Borgan et al. (2000)) to investigate efficiency and to suggest which weights to use in equation 2. Robins, Rotnitzky, and Zhao (1994) have proposed a class of estimators which contains the most efficient estimator: that is, the estimator which reaches the semiparametric efficiency bound. What we have presented in equation 11 is just one member of this class. Their general class is defined by subtracting off a function, $\phi((T_i \wedge C_i), \Delta_i, J_i, w_i)$, from each individual increment of (11), and pre-multiplying by a function $h(s, Z_i)$. For the specific equations we have presented here $h(s, Z_i) = 1$, and $\phi((T_i \wedge C_i), \Delta_i, J_i, w_i) = 0$. The efficiency of the relative risk estimators in this class depends upon the choice of $\phi(\cdot)$ and $h(\cdot)$.

One simple means of increasing efficiency is to use the estimated rather than true weights in their model. The corrections to the influence functions we give in section 3 are, in fact, one function in the class of functions $\phi(\cdot)$. The $h(\cdot)$ and $\phi(\cdot)$ which produce the most efficient estimator are non-closed form solutions to integral equations which are functions of the sampling fractions and the unknown joint distribution of $\{Z_i, T_i, C_i\}$. We are currently working on adaptive estimation of $h(\cdot)$ and $\phi(\cdot)$ to see whether we can

use their theoretical results to obtain important increases in efficiency in an applied setting.

References

- C. C. Abnet, C. B. Borkowf, Y. L. Qiao, P. S. Albert, E. Wang, A. H. Merrill, S. D. Mark, Z. W. Dong, P. R. Taylor S. M. Dawsey, "Sphingolipids as biomarkers of fumonisin exposure and risk of esophageal squamous cell carcinoma," To appear *Cancer Epidemiology Biomarkers and Prevention*, 2001.
- P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding, *Statistical Models Based on Counting Processes*, Springer-Verlag: New York, NY, 1991.
- W. E. Barlow, "Robust variance estimation for the case-cohort design," *Biometrics*, vol 50 pp. 1064–1072, 1994.
- O. Borgan, B. Langholz, S. O. Samuelsen, L. Goldstein and J. Pogoda, "Exposure stratified case-cohort designs," *Lifetime Data Analysis*, vol 6 pp. 39–58, 2000.
- O. Borgan, L. Goldstein and B. Langholz, "Methods for the analysis of sampled cohort data in the cox proportional hazards model," *The Annals of Statistics*, vol 23 pp. 1749–1778, 1995.
- K. C. Cain and N. T. Lange, "Approximate case influence for the proportional hazards regression model in censored data," *Biometrics*, vol 40 pp. 493–499, 1984.
- Helicobacter and Cancer Collaborative Group, "Gastric cancer and Helicobacter Pylori: a combined analysis of eleven case-control studies nested within prospective cohorts," *Gut*, vol 3 pp. 347–353, 2001.
- P. J. Huber, *Robust Statistical Procedures*, Society for Industrial and Applied Mathematics: Philadelphia, PA, 1977.
- J. D. Kalbfleisch and J. F. Lawless, "Likelihood analysis of multi-state models for disease incidence and mortality," *Statistics in Medicine*, vol 7 pp. 149–160, 1988.
- S. Kim and V. De Gruttola, "Strategies for cohort sampling under the Cox proportional hazards model, application to an AIDS clinical trial," *Lifetime Data Analysis*, vol. 5 pp. 149–172, 1999.
- P. J. Limburg, C. Q. Wang, S. D. Mark, Y. L. Qiao, G. I. Perez-Perez, M. J. Blaser, P. R. Taylor, Z. W. Dong, S. M. Dawsey, "Helicobacter pylori seropositivity: Association with increased gastric cardia and non-cardia cancer risks in Linxian, China." *Journal of the National Cancer Institute*, 93, pp. 226–233, 2001.
- D. Y. Lin and L. J. Wei, "The robust inference for the Cox proportional hazards model," *Journal of the American Statistical Association*, vol 84 pp. 1074–1078, 1989.
- Y. Lin and Z. Ying, "Cox regression with incomplete covariate measurements," *Journal of the American Statistical Association*, vol 88 pp. 1341–1349, 1993.
- S. D. Mark, Y. L., S. M. Dawsey, H. Katki, E. W. Gunter, W. Yan-Ping, J. F. Fraumeni, W. J. Blot, Z. W. Dong, P. R. Taylor, "Higher serum selenium is associated with lower esophageal and gastric cardia cancer rates." *Journal of the National Cancer Institute*, vol 92 pp. 1753–1763, 2000.
- H. Moller, E. Heseltine, H. Vainio. "Working group report on schistosomes, liver flukes and Helicobacter pylori." *International Journal of Cancer*, vol 60 pp. 587–589, 1994.
- R. L. Prentice, "A case-cohort design for epidemiologic cohort studies and disease prevention trials," *Biometrika*, vol. 73 pp. 1–11, 1986.
- M. G. Pugh, *Inference in the Cox Proportional Hazards Model with Missing Covariate Data*, thesis, Harvard School of Public Health: Boston, MA, 1993.
- N. Reid and H. Crepeau, "Influence functions for proportional hazards regression," *Biometrika*, vol 72 pp. 1–9, 1985.
- J. M. Robins, A. Rotnitsky, and L. P. Zhao, "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, vol 89 pp. 846–866, 1994.
- S. G. Self and R. L. Prentice, "Asymptotic distribution theory and efficiency results for case-cohort studies," *The Annals of Statistics*, vol. 16 pp. 64–81, 1988.
- T. M. Therneau and H. Li, "Computing the Cox Model for Case Cohort Designs," *Lifetime Data Analysis*, vol 5 pp. 99–112, 1999.