Computing the Cox Model for Case Cohort Designs

Terry M. Therneau

Hongzhe Li

Technical Report Number 62

June 1998

Technical Report Series

Section of Biostatistics

Mayo Clinic, Rochester, Minnesota

Abstract

Prentice [9] proposed a case-cohort design as an efficient subsampling mechanism for survival studies. Several other authors have expanded on these ideas to create a family of related sampling plans, along with estimators for the covariate effects. We describe how to obtain the proposed parameter estimates and their variance estimates using standard software packages, with SAS and S-Plus as particular examples.

1 Introduction

Prentice [9] proposed a case-cohort design for large survey studies such as the Women's Health Study, where the population size makes it infeasible to collect data on all of the cases. If there is a concurrent registry which can be used to identify all of the subjects who experience an event, then it is possible to collect covariate data on only a subcohort of the subjects SC, randomly sampled from the population at large, and (perhaps at a later date) augment the sample with all those subjects who experience an event. Self and Prentice [12] derived an appropriate parameter estimate $\hat{\beta}$ for the design along with a variance estimator V, and proved results on their asymptotic distribution. The proposed variance estimate is algebraically complex, however, and other simpler estimates have been proposed. Lin and Ying [8] discussed Cox regression with incomplete covariate measurements and treated the case-cohort design as a special case of their general results. Barlow [1] gave a robust variance estimate based on the approximate jackknife; he also uses a survey sampling approach to allow for more complicated sampling

mechanisms. Binder [2] has given general results for Cox proportional hazards models and survey sampling designs.

In this paper, we describe how to obtain the proposed parameter estimates and their variance estimates using any proportional hazards regression program that support an *offset* statement, *dfbeta* residuals, and the *(start, stop]* notation to describe intervals of time at risk. Case weights, if supported by the package, can be used in place of the *offset* command. The SAS `phreg` procedure supports the first three of these (but only integer case weights), for instance, and the `coxph` function of S-Plus has all four options. For the Self and Prentice [12] estimates only the first two are needed. This translation to a common form, i.e. the computer code, reveals that the different variance estimates are actually closely related.

2 The case-cohort design

To fix notation, let n_c be the size of the full cohort C, n_{sc} be the size of the subcohort SC, and d the size of the set E of subjects with an event. By definition, all of the subjects in E are in C. In the original example of Prentice [9], the subcohort SC was chosen before all of the events were observed, and may contain a portion of the subjects in E.

As another example, consider the analysis of very large data sets. Deng, Quigley and Van Order [4] discuss a data set containing 1,489,372 observations on single family mortgage loans issued from 1976 to 1983 and purchased by the Federal Home Loan Mortgage Corporation. Data on the outcome of the loans was current through 1992, at which point somewhat less than 1% of the loans have defaulted. The authors point out that the

computational burden of a Cox model can be daunting, and suggest an alternative discrete time model for this reason. Instead, a case-cohort analysis could be done using a random sample of 10–20,000 of the loans, augmented with the set of defaults.

We will use the now-familiar counting process notation to describe a Cox model: $N_i(t)$ is the per subject counting process, which equals the number of observed events for that subject up to and including time t, with $\overline{N} \equiv \sum N_i$. The $Y_i(t)$ process for a subject is 1 when the subject is at risk and under observation, and 0 otherwise. $Z_i(t)$ is the vector of possibly time-dependent covariates for the subject, and $r_i(t) \equiv \exp\{\beta' Z_i(t)\}$ is the subject's risk score at time t.

If data from the full cohort were used in the analysis, then an estimate of β would be based on the usual Cox model score equation

$$U(\beta, t) = \sum_{i=1}^{n_c} \int_0^t \{Z_i(t) - \bar{Z}(\beta, t)\} dN_i(t)\,, \tag{1}$$

where $\bar{Z}$ is the weighted mean

$$\bar{Z}(\beta, t) = \frac{\sum_{i=1}^{n_c} Y_i(t) r_i(\beta, t) Z_i(t)}{\sum_{i=1}^{n_c} Y_i(t) r_i(\beta, t)}\,.$$

The kernel of the score equation is the term $Z_i(t) - \bar{Z}(\beta, t)$, which compares the covariate vector of the subject with an event at time t to the weighted average covariate vector at that time in the population, where the weights are the per-subject risks $r_i(t)$. For a case-cohort method of sampling, some modification of the equation is obviously required. If, as expected, certain covariate values are associated with a higher risk of an event, then an average over all subjects in the data sample $SC \cup E$ will be a biased estimate of $\bar{Z}$ for the cohort.

Two methods for correcting this immediately present themselves. The first is to compute $\bar{Z}$ only over the random subcohort SC, that is use

$$\bar{Z}_{SC}(\beta, t) = \frac{\sum_{i \in SC} Y_i(t) r_i(\beta, t) Z_i(t)}{\sum_{i \in SC} Y_i(t) r_i(\beta, t)}. \tag{2}$$

This estimate is proposed by Self and Prentice [12]. The original proposal of Prentice [9] was nearly identical; it included one more observation — the event occurring at time t — in the mean, based on a binomial argument. If the subcohort size n_{sc} is large this extra observation will have only a minimal impact on the estimate.

Another option is to treat the data as the results of a weighted random sample, as in survey methods. Barlow [1] uses this approach for more general designs than the original case-cohort, these include among others the augmented case cohort design, where the subcohort SC is enriched part way through a study. Let $n_c(t)$ and $n_{sc}(t)$ be the numbers of cohort and subcohort subjects which are at risk at time t. The subject with an event is in the sampled risk set with probability 1, but each of the other subjects with probability $p(t) = n_{sc}(t)/n_c(t)$. Then with sampling weights $w_i(t) = 1/p(t)$ for the subcohort, 1 for the event at time t and 0 for the other (unsampled) subjects the weighted mean

$$\bar{Z}_w(\beta, t) = \frac{\sum_{i=1}^{n_c} Y_i(t) w_i(t) r_i(\beta, t) Z_i(t)}{\sum_{i=1}^{n_c} Y_i(t) w_i(t) r_i(\beta, t)} \tag{3}$$

will also be a consistent estimate of $\bar{Z}$ from the full cohort.

Clearly, both the Self and Prentice [12] and Barlow [1] estimators will converge to the true β in large samples, since both use a consistent estimate of $\bar{Z}$. If $p(t)$ is constant over time, then the two proposals are very similar, and differ only in how

much weight is given to the actual event at time t in computing the weighted mean at t. Below we will see that the variance estimates for these approaches, though quite different on the surface, are actually closely related as well.

2.1 Self and Prentice estimator

The Self and Prentice estimate of $\hat{\beta}$ can be computed fairly easily, using any Cox model program that allows for *offset* terms. In particular, let x be a constructed variable which is equal to 0 for subjects in the random subcohort SC and takes some large negative value, -100 say, for the subjects who have died. If there are subjects who are both in SC and have died, then enter them into the data set as two separate observations, one with $x = 0$ and status = censored and the other with $x = -100$ and status = event. Now fit the model with `offset(x)` as a term on the right hand side of the model. Observations which are not part of the subcohort SC, although formally a part of the estimation of $\bar{Z}$, do not in actuality affect the result since they have a relative weight of $\exp(-100) < 10^{-40}$ as compared to the SC subjects when computing the mean $\bar{Z}$. (Do not set x to too large a number, or the computer's exp function may fail.) Below is a sample fragment of SAS code that illustrates the method. The variable `subco` is assumed to be 1 for observations from the subcohort, `status` contains the outcome indicator, and `x1, x2` are the covariates of interest.

```
data temp; set mydata;
    if (status=1) then do;
        dummy = -100;
        output;
```

```
        end;
    if (subco=1) then do;
        dummy = 0;
        status= 0;
        output;
        end;

proc phreg data=temp;
    model time * status(0) = x1 x2;
    offset dummy;
```

Assume that we have computed the Self and Prentice [12] estimate by using a standard Cox model program with an offset, as per above. Because of the oversampling of cases with an event, the usual estimate of variance will overestimate the precision of $\hat{\beta}$. Self and Prentice proposed an asymptotically consistent estimate of var($\hat{\beta}$), which has been criticized as overly complex for practical use [8]. However, as shown in the appendix it also can be calculated by standard packages as

$$V = \mathcal{I}^{-1} + (1 - \alpha) D'_{SC} D_{SC}$$

where D_{SC} is a subset of the matrix of *dfbeta* residuals that contains only those rows for the subcohort SC, and $\alpha = n_{sc}/n_c$ is the proportion of cases sampled. For those computer packages which return *dfbeta* residuals, this represents a very simple calculation to correct the "standard" variance estimate $\mathcal{I}^{-1}$ returned by the Cox model program. Here is SAS code to compute and print the additional term; it adds two lines to the `phreg` call found above.

```
proc phreg data=temp;
    model time * status(0) = x1 x2;
```

```
    offset dummy;
    output out=temp2 dfbeta=dx1 dx2;
    id subco;   *retains "subco" in output data;

data temp3; set temp2;
    if (subco=1);

proc iml;
     use temp3;
     read all var dx1 dx2 into d;
     v = d' *d;
     print , v;
```

Below is the S-Plus code for the same hypothetical data set containing the variables time, status (1=event, 0=censored), subco (1=subcohort), and covariates of interest x1and x2. It makes use of the fact that a Cox model object can contain both a `var` and `naive.var` component, and both are displayed in the standard printout.

```
# build the data set
temp1 <- data.frame(mydata[status==1, ], dummy= -100,
                                         group=1)
temp2 <- data.frame(mydata[subco==1 , ], dummy= 0,
                                         group=0)
newdata <- rbind(temp1, temp2)

#fit the model
fit <- coxph(Surv(time, group) ~ x1 + x2 + offset(dummy),
               data=newdata)
```

```
# Fix the variance
dfbeta <- resid(fit, type='dfbeta')
d2 <- dfbeta[newdata$subco==1,]
fit$naive.var <- fit$var
fit$var <- fit$var + (1-alpha)* t(d2)%*% d2

print(fit)
```

Even more interesting than the computational simplification that was gained from rewriting the Self and Prentice [12] estimate in this new form, is the further insight that this form gives into the meaning of the estimate. Let β_p be the true coefficient for the (infinite) population at large, $\hat{\beta}_c$ the estimate for the cohort, if data were collected on all of the subjects therein, and $\hat{\beta}_{sc}$ the value for the actual study as conducted. The first term in the Self and Prentice variance, $\mathcal{I}^{-1}$, is an estimate of $\text{var}(\hat{\beta}_c)$, the variance we would have gotten if all of the subjects in C had been used in the computation. The second term is an estimate of the finite sample contribution $\text{var}(\hat{\beta}_{sc}|\text{cohort})$.

For the first term, note that both the score equation and the information matrix for a Cox model are sums, with one term per observed event. Each term is the estimated variance of the covariate vector X at that point in time and is not (other than accuracy) a function of the sample size at that time point. Since SC is a random sample, the subcohort computation based on SC is term by term a consistent estimate of the computation based on the full sample C. To see that the second term is an estimate of $\text{var}(\hat{\beta}_{sc}|\text{cohort})$, consider the matrix D of dfbeta residuals from a fit to the full cohort C. Assume that in computing this fit, subjects who experience an event are again represented as two rows of data with offset and status variables

as in the examples above. The matrix D can be divided into three sets of rows: the events, the influence for the subcohort SC (no deaths), and the influence for $\overline{SC}$. We make use of the following three algebraic properties of D.

1. The column sums of D are 0 at $\hat{\beta}$. (The column sums are the Newton-Raphson step for the next iteration of the computing algorithm [7]. Since the algorithm has converged, the next update step must be 0.)
2. The column sums of D_E, the rows corresponding to the events, are zero as well.
3. The ith row of D is an estimate of the change in $\hat{\beta}$ that would occur if observation i were removed (the motivating definition of D [3]). The approximate change in $\hat{\beta}$ from removing a group of observations is a sum of the appropriate rows of D.

Then

$$\begin{aligned} \hat{\beta}_{sc} &\approx \hat{\beta}_c + 1' D_{s\bar{c}} \\ &= \hat{\beta}_c - 1' D_{sc} \end{aligned} \tag{4}$$

where $\overline{SC}$ are the rows not in SC or E, which are the observations that are "removed" when moving from $\hat{\beta}_c$ to $\hat{\beta}_{sc}$. Thus $(\hat{\beta}_{sc} - \hat{\beta}_c)$ is a sum from a finite sample of rows with a known mean of 0, and the standard finite sample variance estimate is

$$\mathrm{var}(\hat{\beta}_{sc} - \hat{\beta}_c) \approx (1-\alpha) D'_{SC} D_{SC} \, .$$

This connection suggests ways in which the Self and Prentice estimate might be extended to more complex designs.

2.2 Lin and Ying's Estimate

Lin and Ying [8] give estimating equations for a Cox model with missing covariates. They treat the case-cohort design as a special case of their method, and show that their proposed estimates of $\bar{Z}$ and $\hat{\beta}$ are identical to those of Self and Prentice in this case.

As an estimate of the covariance matrix they propose

$$V = \mathcal{I}^{-1}(\sum_{i \in SC \cup E} \widehat{W}_i' \widehat{W}_i)\mathcal{I}^{-1}.$$

Noting that in their notation the case-cohort design corresponds to $H_i = 1$, $H_{0i} = I_{i \in SC}$, their definition of W_i (equation 6) can be rewritten as

$$\widehat{W}_i(\beta) = \int_0^{t_i} \{Z_i(s) - \bar{Z}(\beta, s)\} \{dN_i(s) - \frac{Y_i(s) r_i(s)}{\sum_{j \in SC} Y_j(s) r_j(s)} d\overline{N}(s)\},$$

from which we see that the W_i are the score residuals.

The variance estimate is thus $V = D'D$ where D is the matrix of *dfbeta* residuals. This is precisely the robust variance estimator for a Cox model whose properties are explored by Lin and Wei [5], and which was derived earlier as an infinitesimal jackknife estimate by Reid and Crépeau [10].

In computing the estimate, however, note that we have to "undo" part of the data setup. Subjects in the subcohort SC who later have an event were broken into two synthetic observations (rows) in the data set, and will have two rows in D as well, corresponding to the per *observation* leverage. The two rows in D for such a subject must be added together after the fit to get a single per *subject* dfbeta matrix $\widetilde{D}$, before the matrix product $\widetilde{D}'\widetilde{D}$ is formed. In the S-Plus package this is particularly easy. Assume that the data set of our prior example also

contained a variable `id` which is unique for each subject. Then, using the same data setup as the earlier example, the following will produce a model with the Lin and Ying variance:

```
fit <- coxph(Surv(time, status) ~  x1 + x2 + x3
          + offset(dummy) + cluster(id), data=newdata)
```

The `cluster` directive automatically performs the grouped jackknife variance, adds that variance to the standard printout, and uses it as the basis for the z-statistics $\beta/\text{se}(\beta)$.

The computation in SAS is longer but also straightforward. Example 23.8 in the SAS documentation of the `phreg` procedure [11] shows the necessary code for computing the collapsed D matrix and subsequent robust variance $\widetilde{D}'\widetilde{D}$, for a different case (multiple events per subject) where a given subject may be represented by multiple rows.

2.3 Barlow's Estimate

Barlow [1] explicitly proposes the use of $D'D$ as a variance estimate for case-cohort studies, derives the estimator in a natural way, and notes its relationship to the work of Reid and Crépeau [10] and of Lin and Wei [5]. More importantly, he shows that the jackknife motivation for the estimate allows it to be used in more complex sampling schemes for which the calculation of an asymptotic information matrix estimate would be daunting.

For a general design he proposed a weighted mean with weights

$$w_i(t) = \begin{cases} n_c(t)/n_{sc}(t) & \text{if } i \in SC(t) \\ 1 & \text{if } i \in E(t) \end{cases}$$

where $n_c(t)$ and $n_{sc}(t)$ are the number of subjects in the cohort and subcohort at time t, respectively. A subject who is in both

$SC(t)$ and $E(t)$ is given a weight of 1. We then use the weighted estimator $\bar{Z}_s$.

For the simple case-cohort design discussed thus far, $E(t)$ contains the subject with an event at t and no one else, but it may be more general. Barlow gives for an example a study which included a random cohort of females along with all subjects who had a diagnosis of breast cancer, and whose endpoint was death due to breast cancer. In this case a subject would enter $E(t)$ at diagnosis and leave it at death.

Both SAS and S-Plus return unweighted residuals, i.e. the weighted sum of the residuals $w'D$ will equal zero. With case weights added to the model, the jackknife estimate of variance is $D'W^2D$ where W is a diagonal matrix of the weights. (Giving every observation a weight of 2, for instance, will not change the estimated variance.) For the time-varying weights suggested above, each subject would be represented by multiple rows of data, one row for each event time for which they were at risk, the per-subject leverage matrix $\widetilde{D}$ is formed by summing rows of WD and the variance is $\widetilde{D}'\widetilde{D}$.

We would suggest one change from the weights proposed by Barlow [1], i.e., a subject in $SC(t)$ should retain their weight of $n_c(t)/n_{sc}(t)$ when (and if) they become a member of $E(t)$, instead of being converted to a weight of 1. The primary reason for this is that weights as proposed by Barlow are not a predictable process, i.e., they are not a function of the covariates and risk set just *before* an event occurs. Thus the data set used to fit his model is somewhat contrived. A subject in SC with an event must have a change of weight at the time of his/her event, without prior warning. Such a subject's last interval of time must be set to $(t_i - \epsilon,\, t_i]$, t_i being their event time and ϵ some

small number, with status=event and case weight of 1. (Neither SAS nor S-Plus accept intervals of length zero, so it is necessary to make the interval of some small width ϵ, where ϵ is small enough so that this interval does not overlap any other event times. If time is measured in days, for instance, then $\epsilon = .5$ would suffice.)

3 Nested case-control designs

The decomposition $\beta_c + (\beta_{sc} - \beta_c)$ used to justify the Self and Prentice [12] formula might also be applied to the situation of a nested case-control design. It is well known, however, that a correction term is *not* required in this situation — the ordinary Cox model variance estimate can be used. It turns out that our decomposition does not disagree with this.

First, note that in the prior justification for the Self and Prentice estimate that a second order effect was ignored. The Cox information matrix is a sum of terms, one per death, each of which is an empirical variance of the covariate vector Z over the subjects at risk at that time. If we let $\sigma^2(t)$ be the true variance matrix at each death time and $\mathcal{I}(t)$ be the computed term, then simple algebra shows that

$$E[\mathcal{I}(t)] = \sigma^2(t) \left(1 - \frac{\sum Y_i(t) r_i^2(t)}{[\sum Y_i(t) r_i(t)]^2} \right) .$$

For an unweighted mean this reduces to the usual $(n-1)/n$ correction. If the number of subjects at each risk set were constant for both the cohort and case-cohort design, then at $\beta = 0$ ($r_i \equiv 1$) the information matrix for the latter will be smaller by a factor of $[1 - 1/n_{sc}]/[1 - 1/n_c]$. The first term in the Self and Prentice formula, which in our decomposition should estimate

the variance from a fit to the full cohort, is in expectation a small amount too large.

In a nested case-control design, a set of controls is chosen separately at each risk set. The controls are randomly chosen from those available without reference to other risk sets, and are used in the computation of $\bar{Z}$ only for this risk set. As in the case-cohort design, the variance of the estimate can be written as the variance of $\hat{\beta}_c$ plus a correction term, but because the risk sets are independent this correction term can be written as a sum of separate terms, one for each death time. At each risk set the leverage of a subject can be written as equation 5, without the integration. The contribution to the finite sample ("extra") portion of the variance is the sum of squares of these over the risk set or

$$\frac{\sum Y_i(s)[Z_i(s) - \bar{Z}(s)][Z_i(s) - \bar{Z}(s)]' r_i^2(s)}{[\sum Y_i(s) r_i(s)]^2}.$$

At $\beta = 0$ this reduces to $\mathcal{I}(t)/n_{sc}$.

Thus, the variance for the nested case-control design from this perspective is the sum of two terms. The first is approximately $(1 - 1/nsc)$ times the usual Cox model variance estimate and the second approximately $1/n_{sc}$ times the usual estimate.

A much more complete discussion of this second term and its anticipated size for various designs is found in Langholz and Thomas [6].

4 Comparing the estimates

As α, the proportion of cohort contained in our sample, goes to 1, the Self and Prentice [12] variance converges to the usual Cox model variance $\mathcal{I}^{-1}$, whereas the Lin and Ying [8] estimate

converges to the infinitesimal jackknife estimate $D'D$. Others have suggested that the jackknife estimate may have a larger mean-squared error than information matrix calculations, which in turn suggests a potential superiority for the Self and Prentice approach.

A small simulation study was conducted to examine the performance of different variance estimates, and to compare them with the variance estimate of full cohort. A large cohort size of 5000 was selected, but the total number of events was purposely kept small to represent studies with 80 and 90% power, a common range for clinical research. Failure times were exponential with a hazard function of $\lambda(t) = \exp(Z\beta)$ with $\beta = 0.5$; the covariate Z was Uniform(0,2). Censoring was independent of survival and uniform on (a, b) where a and b were chosen to give approximately 100 or 150 events for the case of .8 and .9 power, respectively. For each of the two powers we generated two subcohorts: (a) small cohort, where the cohort size was taken as the expected number of events; (b) bigger cohort of size where the cohort size was taken as three times the expected number of events. The second subcohort size was motivated by the common wisdom in epidemiologic case-control studies, that the efficiency of a study is not much improved by abstraction of more than three controls per case.

One thousand such samples were generated. The first four rows of table 1 show that the average estimated treatment effect was essentially identical for fits using the full cohort ($n = 5000$) and using the much smaller case-cohort sample. The actual variance of the estimate, over the 1000 simulations, was over twice as large for the case-cohort design with a small cohort, however, and approximately 30% larger for the 3:1 design.

Table 1: *Simulation Results. All variance estimates in the table are multiplied by 100.*

Estimate of β ($\hat{\beta}$)	.8,a	.8,b	.9,a	.9,b
full cohort	.50	.50	.50	.49
case-cohort	.51	0.51	0.51	0.49
var($\hat{\beta}$), full cohort	3.5	3.0	2.6	2.6
var($\hat{\beta}$), case-cohort	7.9	4.4	5.3	3.5

Estimate of var($\hat{\beta}$)	.8,a	.8,b	.9,a	.9,b
full cohort	3.3(.37)	3.3(.37)	2.5(.25)	2.5(.25)
case-cohort(Naive)	3.3(.51)	3.3(.43)	2.6(.34)	2.5(.28)
case-cohort(Self)	7.5(.99)	4.6(.50)	5.2(.59)	3.3(.32)
case-cohort($D^{'}D$)	7.6(1.38)	4.6(.71)	5.3(.85)	3.3(.46)

The next four lines compare the estimators of variance. The first of these shows that the usual Cox variance estimator based on the full data is approximately unbiased for var($\hat{\beta}_c$), with a standard error (in parenthesis) of 10 to 11% of the estimate. The naive estimate for the case-cohort data set, i.e., the estimated variance printed out by the programs when no correction for the sampling method is applied, also estimates the full-cohort accuracy of $\hat{\beta}$. That is, it is an accurate estimate of the wrong quantity. Both the Self and Prentice and the Barlow estimates are approximately unbiased estimators of the actual variation in $\hat{\beta}_{SC}$; the former estimate appears to have a smaller variance, especially for the smallest data sets.

5 Example

Related to their long periods of immobility and the consequent pooling of blood in the lower extremities, patients who have sustained acute spinal cord injury (SCI) have a high incidence of deep vein thrombosis (DVT) and its possibly fatal complication, pulmonary embolism (PE), should the clot dislodge and travel to the lungs. The high prevalence of DVT in the SCI population has stimulated the investigation of several forms of DVT prophylaxis, but the best methods of surveillance and prevention remain unclear.

Winemiller et al. [14] took advantage of a registry of SCI patients available at the the Mayo Clinic to examine the impact of several aspects of patient care on the risk of DVT; the registry has recorded all hospitalizations for SCI and their major complications since 1976. Detailed data on the treatment of each patient, however, is not present in the registry and must be abstracted from the permanent (paper) medical record. Two variables of particular interest, the day-by-day usage of heparin and/or elastic stockings (TEDs) require one to peruse the entire sequence of daily nursing notes, a tedious process. The presence of DVT/PE was, however, available from the data base. Because of the pressures of time all 84 records of the subjects who experienced an event were abstracted, but only 201 of the 344 remaining admissions from 1976 to 1995. Table 2 shows the counts grouped by years.

For the non-event subjects, a random subset of records from the calendar year was abstracted. The original goal was to select two controls for each event, with at least two as well for the years with no DVT cases, i.e., 1992 and 1995. However, the process

	76-80	81-85	86-90	91-95	Total
abstracted: event	27	36	14	7	84
abstracted: non-event	60	77	41	23	201
not abstracted	32	20	38	53	143

Table 2: *Number of subjects for the DVT study*

went better than anticipated (a rare event for a clinical study!), and a further 29 records were randomly sampled from across all years. The final proportion of non-cases abstracted for each individual calender year ranged from 12% to 100%.

For analysis, let α_j be the proportion of non-event records abstracted for year j, and let the weight for a subject i who was admitted in year j be $w_i = 1$ for a DVT subject and $w_i = 1/\alpha_j$ for an abstracted non-DVT subject. Time dependent covariates, e.g. use of heparin, were coded by breaking each subject up into multiple observations, each over an interval (start, stop]. Each observation contains the values of the covariates that apply over that interval, along with a status variable that indicates whether the interval was terminated with an event (1-yes, 0-no). Analysis was done using the S-Plus package since it supports non-integer case weights. Here is the code and results for one of the models.

```
> fit <- coxph(Surv(start, stop, status) ~ age10 + male +
                surveill + teds + cluster(id), weights=w)
> summary(fit)
          coef exp(coef) se(coef) robust se    z    p
   age10  0.06       1.1     0.06      0.07  0.9 0.35
    male  0.93       2.5     0.38      0.39  2.4 0.02
surveill  1.08       2.9     0.28      0.29  3.7 <.01
     ted -0.73       0.5     0.26      0.29 -2.6 0.01
```

```
Rsquare= 0.03   (max possible= 0.534 )
Likelihood ratio test= 31.5  on 4 df,   p=2.4e-06
Wald test            = 23.5  on 4 df,   p=1.0e-04
Score (logrank) test = 29.8  on 4 df,   p=5.4e-06
              Robust = 20.7  p=0.00037

  (Note: the likelihood ratio and score tests assume
     independence of observations within a cluster, the
     Wald and robust score tests do not).
```

The variable `age10` is age in decades, we see that the risk goes up only slightly (6%) for each ten years of age and in fact that age is not a significant predictor. Male gender carries a 2.5 fold risk in comparison to females. Active surveillance using doppler echo carries a 3-fold risk, presumably it increases the chance of finding a thrombus and not the chance of forming one. TEDs, as was expected, decrease the risk of DVT by approximately one half.

The `cluster` statement in the call directs the `coxph` function to create the robust variance estimate $\widetilde{D}'\widetilde{D}$ based on leverages for each unique value of `id`. The column labeled `se(coef)` contains the uncorrected variance $\mathcal{I}^{-1}$; z and p are based on the robust estimate. In this particular study they are not very different, probobly because proportion of unsampled data is low. In the overall analysis of the study, however, we found that the correction was somewhat more important for those fits that contained a larger number of covariates.

To do the analysis in the manner of Self and Prentice [12] requires a thought experiment. In their setup, a fraction α_j of the total admissions for the year would have been chosen, that

fraction of the total cases abstracted, and then the remaining subjects with an event added. In this study events were identified first and sampling done later. If the study had been done in the Self and Prentice manner, each of the abstracted events would have had probability α_j of being in the 'random' subset SC, and these events would have weight 1 in the computation of $\bar{Z}$ while the other events had weight 0. Instead, we might give each event weight α_j in the mean. Note that this is the same as giving events a weight of 1 and non-event observations a weight of $1/\alpha_j$, that is, this modified Self and Prentice approach will yield the same coefficients as above.

The modified Self and Prentice variance can be written as

$$\mathcal{I}^{-1} + D'W^*D$$

where W^* is a diagonal matrix of weights which are $(1-\alpha_j)$ for the random sample of non-cases, and $\alpha_j(1-\alpha_j)$ for each of the subjects with an event.

```
> dfbeta <- resid(fit, type='dfbeta', collapse=id)
> newvar <- fit$naive.var + t(dfbeta) %*% diag(Wstar) %*% dfbeta
> fit$var <- newvar
> print(fit)
          coef exp(coef) se(coef) robust se  z      p
   age10  0.06       1.1     0.06      0.06  0.9  0.34
    male  0.93       2.5     0.38      0.40  2.4  0.02
surveill  1.08       2.9     0.28      0.29  3.7  <.01
     ted -0.73       0.5     0.26      0.28 -2.6  0.01
```

Here `Wstar` contains the vector W^* of modified weights. The `collapse` argument to the residuals function causes a summation over subject to be performed so that the result has one row per unique value of `id`. The Self and Prentice variance is then

inserted into the fit object in place of the approximate jackknife variance, and the result printed out. The differences between this and the Barlow [1] estimate are quite small, for this data set.

6 Discussion

Binder [2] discusses the fitting of Cox's proportional hazards model to survey data. Case weights are applied to each observation consistent with the survey design, with a suggested variance estimate of $D'WD$ where W is formed "using design based methods." The score equation for $\hat{\beta}$ is a weighted one

$$U(\beta, t) = \sum_{i=1}^{n} w_i \int_0^t [Z_i(t) - \bar{Z}(\beta, t)] dN_i(t) \,.$$

In case-cohort sampling case weights for all of the events are fixed at 1, nevertheless the similarity to survey sampling ideas is obvious.

This connection of all of the methods to simple sampling ideas, unfortunately, has been obscured by the fact that all of the papers cited use a different notation for the quantities of interest. Given that several of the current statistical packages include dfbeta residuals for the Cox model as an option, both solutions and variance estimates for a wide variety of models should now be readily available to the statistical practictioner.

References

[1] Barlow, W.E. (1994). Robust variance estimation for the case-cohort design. *Biometrics* **50**, 1064-1072.

[2] Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* **79**, 139-47.

[3] Cain, K.C. and Lange, N.T. (1984). Approximate case influence for the proportional hazards regression model with censored data. *Biometrics* **40**, 493-9.

[4] Deng, Y., Quigley, J.M. and Van Order, R. (1995). Mortgage Terminations. *Institute of Business and Economic Research, University of California at Berkeley.* Working Paper 95-230.

[5] Lin, D.Y. and Wei, L.J. (1989). The robust inference for the Cox Proportional Hazards Model. *J. American Statistical Association* **84**, 1074-79.

[6] Langholz, B. and Thomas, D.C. (1991). Efficiency of cohort sampling designs: some surprising resluts. *Biometrics* **47**, 1563-72.

[7] Lipsitz, S.R., Dear, K.B.G. and Zhao, L. (1994). Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics* **50**, 842-46.

[8] Lin, D.Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *J. American Statistical Association*, **88**, 1341-9.

[9] Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73** 1-11.

[10] Reid, N. and Crépeau, H. (1985). Influence functions for proportional hazards regression. *Biometrika* **72**: 1-9.

[11] SAS Institute Inc. (1996). The PHREG Procedure. In *SAS/STAT Software: Changes and Enhancements through Release 6.11*, Cary, NC: SAS Institute Inc.

[12] Self, S.G. and Prentice, R.L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Annals Statistics* **16**, 64-81.

[13] Therneau, T.M., Grambsch P.M. and Fleming, T.R. (1990). Martingale based residuals for survival models. *Biometrika* **77**, 147-60.

[14] Winemiller, M.H., Stolp-Smith, K.A, Silverstein, M.D. and Therneau, T.M. (1998). Sequential pneumatic compression or heparin is effective in preventing venous thromboembolism in spinal cord injury patients. *in press.*

A Appendix

Let $\alpha = n_{sc}/n_c$ be the number of subjects in the chosen subcohort divided by the size of the total population. Self and Prentice [12] show that $n_c^{1/2}(\hat{\beta} - \beta)$ is an asymptotically normal with mean 0 and covariance matrix $\Sigma^{-1}(\Sigma+\Delta)\Sigma^{-1}$. The matrix Σ is consistently estimated by ${n_c}^{-1}\mathcal{I}$, where $\mathcal{I}$ is the information matrix from the Cox model above, and Δ is consistently estimated by

$$\hat{\Delta} = \frac{1}{{n_c}^2} \int \int \tilde{G}(\beta, x, t) d\bar{N}(x) d\bar{N}(t)$$

with

$$\begin{aligned} \tilde{G}(\beta, s, t) = (1-\alpha)/\alpha \qquad & [\{\tilde{S}^{(0)}(s)\tilde{S}^{(0)}(t)\}^{-1} \, \widetilde{H}^{(1)}(\beta, s, t) \\ + \quad & \{\tilde{S}^{(0)}(s)\tilde{S}^{(0)}(t)\}^{-2} \, \tilde{S}^{(1)}(s)\tilde{S}^{(1)}(t)\widetilde{H}^{(0)}(\beta, s, t) \end{aligned}$$

$$
\begin{aligned}
&+ \quad \tilde{S}^{(0)}(s)^{-1}\tilde{S}^{(0)}(t)^{-2}\,\tilde{S}^{(1)}(t)\widetilde{H}^{(2)}(\beta,s,t) \\
&+ \quad \tilde{S}^{(0)}(s)^{-2}\tilde{S}^{(0)}(t)^{-1}\,\tilde{S}^{(1)}(s)\widetilde{H}^{(2)}(\beta,s,t)]\,,
\end{aligned}
$$

$\widetilde{H}$ is defined by

$$
\begin{aligned}
\widetilde{H}^{(0)}(\beta,s,t) &= \widetilde{Q}^{(0)}(\beta,s,t) - \tilde{S}^{(0)}(\beta,s)\tilde{S}^{(0)}(\beta,t) \\
\widetilde{H}^{(1)}(\beta,s,t) &= \widetilde{Q}^{(1)}(\beta,s,t) - \tilde{S}^{(1)}(\beta,s)\tilde{S}^{(1)}(\beta,t) \\
\widetilde{H}^{(2)}(\beta,s,t) &= \widetilde{Q}^{(2)}(\beta,s,t) - \tilde{S}^{(0)}(\beta,s)\tilde{S}^{(1)}(\beta,t)\,,
\end{aligned}
$$

and $\widetilde{Q}$ as

$$
\begin{aligned}
\widetilde{Q}^{(0)}(\beta,s,t) &= \frac{1}{n_{sc}}\sum_{i\in SC} Y_i(s)Y_i(t)r_i(\beta,s)r_i(\beta,t) \\
\widetilde{Q}^{(1)}(\beta,s,t) &= \frac{1}{n_{sc}}\sum_{i\in SC} Y_i(s)Y_i(t)r_i(\beta,s)r_i(\beta,t)Z_i(s)Z_i'(t) \\
\widetilde{Q}^{(2)}(\beta,s,t) &= \frac{1}{n_{sc}}\sum_{i\in SC} Y_i(s)Y_i(t)r_i(\beta,s)r_i(\beta,t)Z_i'(t)\,.
\end{aligned}
$$

The $\tilde{S}^{(i)}$ are the weighted moments of Z:

$$
\begin{aligned}
\tilde{S}^{(0)}(\beta,s) &= \frac{1}{n_{sc}}\sum_{i\in SC} Y_i(s)r_i(\beta,s) \\
\tilde{S}^{(1)}(\beta,s) &= \frac{1}{n_{sc}}\sum_{i\in SC} Y_i(s)r_i(\beta,s)Z_i(s) \\
\tilde{S}^{(2)}(\beta,s) &= \frac{1}{n_{sc}}\sum_{i\in SC} Y_i(s)r_i(\beta,s)Z_i(s)Z_i'(s)\,.
\end{aligned}
$$

This definition appears quite formidable. However, if we introduce the rescaled weights

$$
w_i(s) \equiv \frac{Y_i(s)\,r_i(\beta,s)}{\sum_{j\in SC} Y_j(s)\,r_j(\beta,s)}\,,
$$

then the terms can be rearranged into the following simple form

$$
\tilde{G}(\beta,s,t) = n_{sc}(1-\alpha)/\alpha \sum_{i\in SC} w_i(s)w_i(t)\{Z_i(s)-\bar{Z}(\beta,s)\}\{Z_i(t)-\bar{Z}(\beta,t)\}'\,.
$$

Consider the set of score residuals from the data set used in the SAS example above, defined as

$$L_i(\beta) = \int \{Z_i(s) - \bar{Z}(\beta, s)\}\, dM_i(s)$$

where M is the martingale residual [13]. For the subjects in SC, there are no events and the martingale residual is then $M_i(t) = \int^t Y_i(s) r_i(\beta, s) d\Lambda_0(\beta, s)$. (Again making use of the trick that a subject in the subcohort SC who does experience an event will be represented as two lines of data, one in SC and one 'outside' SC.) If we substitute in the Aalen estimate of $\hat{\Lambda}$, which has a jump at each observed event, the observed residual is

$$\begin{aligned} L_i(\beta) &= \int \{Z_i(s) - \bar{Z}(\beta, s)\} \frac{Y_i(s) r_i(s)}{\sum_{j \in SC} Y_j(s) r_j(s)} d\overline{N}(s) \quad (5) \\ &= \int \{Z_i(s) - \bar{Z}(\beta, s)\} w_i(s) d\overline{N}(s)\,. \end{aligned}$$

If we let L_{SC} be a matrix of score residuals, containing one row for each of the observations in the subcohort SC, then straightforward algebra shows that

$$\hat{\Delta} = \frac{n_c - n_{sc}}{{n_c}^2} L'_{SC} L_{SC}\,.$$

The *dfbeta* residuals for a Cox model, as found in SAS and S-Plus for example, are defined as $L\mathcal{I}^{-1}$, based on the derivation of Cain and Lange [3]. They are returned as a matrix or data set with one row per observation and one column per variable. The Self and Prentice estimate of variance is then

$$V = \mathcal{I}^{-1} + (1 - \alpha) D'_{SC} D_{SC}$$

where D_{SC} is a subset of the *dfbeta* matrix that contains only those rows for the subcohort SC.