

Cox Models for Subsampled Data

Terry M. Therneau
Mayo Clinic

February 1997

Women's Health Study

Goal – examine the effect of covariates on a population level, on the city of Seattle say.

Prentice proposed a case-cohort design:

- Assume that breast cancer is the endpoint.
- It is not cost-feasible to retrieve all of the records for every woman in Seattle.
- At the start, enroll a random subset of $p = n/N$ subjects.
- Via a public health registry, also abstract the records of every woman who gets breast cancer during the study period.
- Analyse using a modified Cox model.

This looks much like a classic case-control study, except

- Cases and controls are not matched.
- A random subset of subjects is used as the “control”.
The same subjects could be the control group for many questions.
- Abstraction of the controls would start before the cases were found.
- The primary statistical question is “incidence rate of breast cancer”, leading to a Cox model, leading to a relative hazard estimate rather than an odds ratio.
- It appears that we need a special computer program.

Outline

- Weighted models
 - frequency weights
 - sampling weights
- Case-cohort designs
 - basic strategy
 - alternative estimates
 - variance (?)
- An excursion
 - Leverage
 - Robust Variance
- Case-cohort designs, variance
- Summary

Weighted Cox Models

Case weights

- Frequency weights (unusual)
 - A weight of 3 means that 3 data points were actually observed, had the same values, and have been collapsed into a single observation to save space.
 - The normal `coxph` variance will be correct.
- Sampling weights
 - If 10% of the low-income mortgages and 1% of the others were sampled to create a data set, then weights of 1:10 will give the correct coefficient.
 - The final weighted coefficient $\hat{\beta}$ depends only on the ratios of the weights, not on their values.
 - The information matrix depends on the actual weights. Weights of 1 and 10 will give too small a variance; the computation assumes the data has more “information” than it actually contains.

Two applications

- Survey sampling
 - weights
 - clustering
 - strata

- Case-cohort designs
 - use *all* of the events
 - along with a *sample* of the subjects
 - a particular case of choice-based sampling

Weights, either frequency or sampling, can be added to S-Plus easily:

```
fit <- coxph(Surv(time, status) ~ x + z, weight=w, ...)  
rr <- resid(fit)
```

The `coxph` function returns unweighted residuals, i.e.

$$\sum_{i=1}^n w_i \widehat{M}_i = 0$$

Sometimes this is best, sometimes weighted residuals would be better.

SAS `phreg` does not accept fractional weights.

Survey Designs

Binder(1992) discusses fitting Cox models to survey data.

- His variance estimate is $D'WD$ where W “depends on the survey design”. For the simple case of sampling weights (i.e., no correlations due to clusters of households, etc), W is diagonal with $W_i = w_i^2$, the squared sampling weights.
- This estimate *is* invariant to multiplication of the weights by a constant.
- It is automatically computed when using case weights and a `cluster` term in the model (S-Plus 3.4).

Weighted models have been proposed as a mechanism to properly handle the effects of case deletion due to missing values. See Lin and Ying (1993).

Pugh (1994) uses a logistic regression to compute approximate case weights for the non-missing observations, proportional to $w_i = 1/\Pr\{\text{non-missing}\}$. The W matrix then depends on both the estimated weights and the leverage residuals of the logistic model.

The basic idea:

- If subject ‘Smith’ has a low probability of complete data (.2 say)
- and *did* have complete data
- he will be in the sample with a weight of $1/.2 = 5$, standing in for an “underrepresented minority” (the four others with similar covariate patterns who are deleted due to missing values).

Case-cohort designs

Why won't an ordinary Cox model program work for case-cohort designs?

The Cox model score equation is

$$\begin{aligned} U &= \sum_{deaths} [X_i(t) - \bar{X}(t)] \\ &= \sum_{i=1}^n \int [X_i(t) - \bar{X}(t)] dN_i(t) \end{aligned}$$

where \bar{X} is a weighted average value of the covariates *in the population* at the time of the death; with weights that depend on each subject's risk.

In a standard Cox model the data set is a random sample from the population, and \bar{X} is computed over everyone in the sample (everyone still alive).

In a survey sample we use a weighted \bar{X} .

In a case-cohort study, \bar{X} for the *sample* is not a good estimate of \bar{X} for the *population*.

Offset terms

An offset term is a variable entered into the model for which β is already known. Offset variables are not necessary in an ordinary linear model, since the term βX can simply be subtracted from both sides of the equation and analysis done with the adjusted y values. When there is a transformation involved, such as glm models, offset terms can be quite useful.

One use is to include a known covariate such as a risk score. This technique has been successfully used in later studies of PBC, for instance. When the latter study is quite small (< 20 deaths say), we may well feel that a risk score adjustment based on the coefficients of the natural history model (n=418) will be more accurate than a re-determination using the data at hand.

Case-cohort designs

Assume that a fraction p of the population is selected for inclusion in the study, and these patients are followed forward in time. As well, all of the deaths in the population are identified as they occur (through some public registry perhaps), and these subjects's histories are then retrospectively abstracted and added to the data.

The key computational issues are

- Since all of the deaths are present in the sample, the individual terms of the score equation should not be weighted.
- An ordinary Cox model will not be correct, however, since \overline{X} for the *sample* is not an unbiased estimate of the \overline{X} of the *population*.
- One correct method is to compute \overline{X} only over the randomly selected subjects.

The weight for each subject in \overline{X} is

$$Y_i(t)e^{\beta_1 X_1 + \beta_2 X_2 + \dots} = Y_t(t)r_i,$$

where Y is 1 for those still at risk and 0 otherwise. Adding an offset changes this to

$$Y_i(t)r_i e_i^o$$

.

Let

- **time** = time from entry to death or last follow-up.
- **group** = 0 for the original subjects, 1 for the deaths. A subject who was chosen in the original cohort and dies should be entered into the data set twice: as a cohort subject and as a death.
- **dummy** = 0 for the original subjects, -200 for the deaths. The precise value of this constant is not important; we want $\exp(\text{dummy})$ to be 1 for the cohort and ≈ 0 for the deaths, without causing a numeric exception in the computer's *exp* function.
- **x1, x2** = covariates of interest.

Then the following code will give give the coefficients suggested by Self and Prentice.

```
fit <- coxph(Surv(time, group) ~ x1 + x2 + offset(dummy),  
            data=mydata)
```

(With a relative weight of $\exp(-200) < 1/10^{86}$, the group=1 observations have no actual effect on \bar{X} .)

Variations

Any method that gives an unbiased \bar{X} will give the correct coefficients.

- Self and Prentice use the mean only over the subsample, as has been done here.
- The earlier paper of Prentice also included the subject who died. Thus his mean includes 1 more observation. To accomplish this we use, instead of an offset, a risk interval via the (start, stop] notation of `coxph`. Let ϵ be some small number, such that $(t - \epsilon, t]$ includes only the death at t . Subjects who are in SC are coded as 1 or 2 observations, with offset, as before. Subjects only in E are coded as a short interval.
- Barlow suggests the weighted estimator, where $w = 1/p$ for SC and 1 for E . Subjects in E enter the risk set at $t_i - \epsilon$, as above.

How does this selection effect the variance?

Leverage

Let D be the matrix of approximate leverage residuals, the *dfbeta* residuals of SAS and Splus.

D is an $n \times p$ matrix, $-D_{ij}$ is the approximate change in β_j if observation i were removed from the sample.

- $D'D$ is the approximate jackknife estimate of variance. This is recommended as a 'robust' variance estimate by Lin & Wei.
- $D'D$ is robust to some failures of the Cox model, it is also somewhat less stable than the usual variance.
- If a single subject is represented by multiple rows (observations) in the data set (multiple events/subject, time-dependent covariates, ...), then it is important to use the *per subject* rather than the *per observation* leverage.

Case-cohort designs

The Self and Prentice variance estimate can then be computed using the `dfbeta` residuals. Let D be the matrix of `dfbeta` residuals, and D_{SC} be the submatrix containing only the rows for which `dummy=0`. Then

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

is the Self and Prentice variance estimate, where the first term is the usual Cox model variance, and p is the proportion of the original population included in the sample.

```
rr <- resid(fit, type='dfbeta')
r2 <- rr[dummy==0,]
fit$naive.var <- fit$var
fit$var <- fit$var + (1-p) * t(r2) %*% r2
print(fit)
```

As an alternative, one can use the approximate jackknife variance $D'D$. This is explored in papers by Barlow (1994) and Ying and Lin (1993). It is easier to compute but somewhat more variable. Assume that the variable `id` is the subject identifier.

```
fit <- coxph(Surv(time, group) ~ x1 + x2 + offset(dummy) +
             cluster(id), data=mydata)
print(fit)
```

The SAS code for the variance is considerably more tedious, but straightforward.

Understanding the Self and Prentice formula.

The original Annals paper gives formal mathematical underpinnings:

$$V = \mathcal{I}^{-1}[\mathcal{I} + \Delta]\mathcal{I}^{-1}$$

where 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

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

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

$$\tilde{S}^{(0)}(\beta, s) = \frac{1}{n_{sc}} \sum_{i \in SC} Y_i(s) r_i(\beta, s)$$

$$\tilde{S}^{(1)}(\beta, s) = \frac{1}{n_{sc}} \sum_{i \in SC} Y_i(s) r_i(\beta, s) X_i(s)$$

$$\tilde{S}^{(2)}(\beta, s) = \frac{1}{n_{sc}} \sum_{i \in SC} Y_i(s) r_i(\beta, s) X_i(s) X_i'(s).$$

\tilde{H} is defined by

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

and \tilde{Q} as

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

This formula has been considered overly complex.

Consider a more heuristic justification.

Let β_p be 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. Then

$$\hat{\beta}_{sc} - \hat{\beta}_p = (\hat{\beta}_c - \hat{\beta}_p) + (\hat{\beta}_{sc} - \hat{\beta}_c)$$

- the first term in the variance, \mathcal{I}^{-1} , is an estimate of $\text{var}(\hat{\beta}_c)$.
- the second term is an estimate of the finite sample contribution $\text{var}(\hat{\beta}_{sc}|\text{cohort})$.

- Consider the matrix D of dfbeta residuals from a fit to the full cohort of data.
- In computing this fit, subjects who experience an event are again represented as 2 rows.
- The matrix D can be divided into 3 sets of rows
 - the events (E),
 - the other-than-event-time influence for SC ,
 - the other-than-event-time influence for \overline{SC} .

- At $\hat{\beta}_c$ the column sums of D are 0.
- Due to an identity of the score equation, the column sums of the submatrix D_E (the first set of rows) are also 0.
- Thus, the column sums of the non- E rows sum to 0.

Since D is a matrix of leverage residuals

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

where \overline{SC} are the rows not in SC or E .

Then $\text{var}[(\hat{\beta}_{sc} - \hat{\beta}_c) \mid \hat{\beta}_c] = (1 - p)\text{var}[D_{\overline{E}}]$, by standard finite sampling results, leading directly to the second term of the Self and Prentice variance.

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

Simulation

Number of cohorts	200			
Number of subcohorts/cohort	4			
Cohort size	5000	500	5000	5000
Subcohort size	300	100	150	150
Percent censored	97	70	92	99.2
E(#events)	150	150	400	40

Variations

Several suggestions have been made for how to choose SC

- Randomly
- Augmented sampling: if some of the later risk sets have few controls, add a few more from the set of subjects at risk at that time.
- Nested case control: at each death time, use a random sample of controls.
- Complex

A variance can be derived for any of them, using the above approach.

Summary

- Any method that gives an unbiased \bar{X} will give the correct coefficients.
- The two variance estimators appear to have similar performance.
- The efficiency of the method depends on how SC is chosen. Standard finite-sampling ideas should apply.

Why do we care?

Case-control

- For a given number of controls/case, is 0-30% more efficient than a case-cohort study.
- We are familiar with the software.
- It gives a familiar number, the odds-ratio, which we know how to explain.
- Every study and endpoint needs its own set of controls

Case-cohort

- Recognizes the time aspect, which is important for some studies.
- Can give absolute estimates of risk (Kaplan-Meier).
- The same cohort can be used for multiple studies.
- We would need to sample a weighted cohort (very young people are essentially uninformative on the risk of dementia, say).

Winemiller Study

Risk of DVT/PE in patients with paralysis due to traumatic spinal cord injury.

- Because of a departmental registry, he knew all of the DVT cases up front.
- The key question was efficacy of SCD, which is found in the nurses notes. There was not enough time to abstract all of the charts.
- For each case, he took the next 2 sequential admissions, plus a few more charts for the sparser years.