

# Survival curve estimation under complex sampling

Thomas Lumley

Under iid sampling, formulas for the standard errors for the survival curve  $\hat{S}(t)$  in a one-sample problem or the predicted curve  $\hat{S}(t; z, \hat{\beta})$  in a proportional hazards model simplify because the cumulative hazard  $\Lambda(t)$  is the compensator of the event counting process  $N(t)$ , so that  $N - \Lambda$  is a (local) martingale.

These simplifications are not present under more general sampling schemes. This note gives computational formulas for the one-sample and regression estimators of the survival curve under arbitrary finite-population sampling, and under two-phase subsampling, and with calibration of weights. The sample design enters only in the estimation of population totals, so the formulas could readily be adapted to other probability sampling designs. The basic approach follows that of Williams (1995) for the one-sample problem under independent and identically distributed sampling of clusters from an infinite population or superpopulation.

## 1 Estimation of totals in finite population sampling.

In single-phase finite-population sampling each individual in the population is sampled with non-zero probability  $\pi_i = E[R_i]$ , and these probabilities are known for individuals in the sample. The pairwise sampling probabilities  $\pi_{ij} = E[R_i R_j]$  are also non-zero, and are known for individuals in the sample. The Horvitz–Thompson unbiased estimator of the population total of a variable  $X$  is

$$\hat{T}_X = \sum_{i:R_i=1} \frac{1}{\pi_i} X_i \equiv \sum_{i:R_i=1} \check{X}_i$$

and an unbiased estimator of its variance is

$$\widehat{\text{var}} \left[ \hat{T}_X \right] = \sum_{i,j:R_i R_j=1} \frac{X_i X_j}{\pi_{ij}} - \frac{X_i}{\pi_i} \frac{X_j}{\pi_j}.$$

In two-phase sampling a phase-one sample is taken with sampling probabilities  $\pi_{1,i}$  and pairwise probabilities  $\pi_{1,ij}$ . Some variables are measured on this sample, then a phase-two subsample is taken from this sample with probabilities  $\pi_{2|1,i}$  and  $\pi_{2|1,ij}$  that may depend on measured values in the entire phase-one sample. The values  $\pi_i^* = \pi_{1,i} \pi_{2|1,i}$  are not

the marginal sampling probabilities  $\pi_i$  for individuals;  $\pi_i$  would be the average of  $\pi_i^*$  over all possible phase-one samples that include element  $i$  and so is typically not available. However,  $\pi_i^*$  and the pairwise values  $\pi_{ij}^* = \pi_{1,ij}\pi_{2|1,ij}$  can be used for estimation in the same way as the marginal sampling probabilities (Särndal et al, 1992), so that

$$\hat{T}_X = \sum_{i:R_i=1} \frac{1}{\pi_i^*} X_i \equiv \sum_{i:R_i=1} \check{X}_i$$

is unbiased for the population total of  $X$  and its variance can be estimated by

$$\widehat{\text{var}} \left[ \hat{T}_X \right] = \sum_{i,j:R_i R_j=1} \frac{X_i X_j}{\pi_{ij}^*} - \frac{X_i}{\pi_i^*} \frac{X_j}{\pi_j^*}.$$

If the population total is known for some variable(s)  $Z$ , calibration of weights replaces the sampling weights  $1/\pi_i$  by calibrated weights  $g_i/\pi_i$  chosen so that the estimated total matches the known value:

$$T_Z = \sum_{i=1}^M Z_i = \sum_{i:R_i=1} \frac{g_i}{\pi_i} Z_i = \hat{T}_Z.$$

The standard error of an estimated total after calibration is

$$\widehat{\text{var}} \left[ \hat{T}_X \right] = \sum_{i,j:R_i R_j=1} \frac{r_i r_j g_i g_j}{\pi_{ij}} - \frac{r_i g_i}{\pi_i} \frac{r_j g_j}{\pi_j}.$$

where  $r_i$  is the residual from projecting  $X_i$  on to the space spanned by the calibration variables  $Z$ . A two-phase sample can be calibrated based on known population totals or on known phase-one totals, giving analogous formulas for the standard error (Särndal et al, 1992).

## 2 One-sample survival curve estimation

$dN_i(t)$  counts the events for person  $i$ ,  $Y_i(t)$  is the at-risk process for person  $i$ . The population totals of these are

$$\bar{d}N(t) = \sum_{i=1}^M dN_i(t)$$

and

$$\bar{Y}(t) = \sum_{i=1}^M Y_i(t).$$

Using the appropriate estimators from the previous section we can compute unbiased estimators  $d\hat{N}(t)$  and  $\hat{Y}(t)$  at every time  $t$  where  $d\hat{N}(t) > 0$ , and estimate the variances, and the covariances of  $\hat{N}$  and  $\hat{Y}$  at any finite set of times.

The population cumulative hazard function is

$$\Lambda(t) = \int_0^t \frac{d\bar{N}(t)}{\bar{Y}(t)}$$

The cumulative hazard can be estimated by plugging in the Horvitz–Thompson estimators for  $\bar{N}$  and  $\bar{Y}$ :

$$\hat{\Lambda}(t) = \int_0^t \frac{d\hat{N}(t)}{\hat{Y}(t)}.$$

The estimated cumulative hazard is a step function with jumps at the observed event times. The variance and covariance of the steps can be estimated by the delta method:

$$\begin{aligned} \widehat{\text{cov}} \left[ d\hat{\Lambda}(t), d\hat{\Lambda}(s) \right] &= \frac{\widehat{\text{cov}} \left[ d\hat{N}(t), d\hat{N}(s) \right]}{\hat{Y}(t)\hat{Y}(s)} + \frac{d\hat{N}(t)d\hat{N}(s)\widehat{\text{cov}} \left[ \hat{Y}(t), \hat{Y}(s) \right]}{\hat{Y}^2(t)\hat{Y}^2(s)} \\ &\quad - \frac{d\hat{N}(s)\widehat{\text{cov}} \left[ d\hat{N}(t), \hat{Y}(s) \right]}{\hat{Y}(t)\hat{Y}(s)^2} - \frac{d\hat{N}(t)\widehat{\text{cov}} \left[ d\hat{N}(s), \hat{Y}(t) \right]}{\hat{Y}(s)\hat{Y}(t)^2}. \end{aligned}$$

An estimator of the variance of the cumulative hazard estimator is then

$$\widehat{\text{var}} \left[ \hat{\Lambda}(t) \right] = \sum_{s, s' \leq t} \widehat{\text{cov}} \left[ d\hat{\Lambda}(s), d\hat{\Lambda}(s') \right]$$

By separating the finite-population estimation of  $\bar{N}(t)$  and  $\bar{Y}(t)$  from the delta-method computations it is straightforward to obtain estimates under a wide range of one-phase and two-phase sampling designs.

Under iid sampling the standard errors from this formula agree closely with the martingale-based formula that includes only the terms in  $\text{var}[d\hat{N}(t)]$ , and under two-phase subsampling they agree with the `nested.km()` function of Mark & Katki.

### 3 The Cox model

The model is

$$d\Lambda(t; x, \beta) = e^{x\beta} d\Lambda_0(t)$$

and it is fitted by solving the sampling-weighted partial score equations of Binder (1991). For computational convenience, the covariates are centered at the sampling-weighted mean, so  $\Lambda_0$  is the cumulative hazard at the mean covariate value.

The estimated baseline survival curve follows Breslow’s proposal for the Cox model

$$d\hat{\Lambda}_0(t) = \frac{\sum_{i=1}^n dN_i(t)/\pi_i}{\sum_{j=1}^n e^{x_j\hat{\beta}} Y_j(t)/\pi_j}.$$

We write  $d\hat{N}$  for the numerator and  $\hat{Y}$  for the denominator of this expression and compute them as estimated population totals as for the one-sample estimator. With these definitions we can write the variance estimator as the sum of four terms. The first term  $V_1(t)$  in the variance is the same expression as the one-sample variance (but using the new definition of  $\hat{Y}$ ).

The second term in the variance depends on the variance  $V_\beta$  of  $\hat{\beta}$ . Its increment at time  $t$  is

$$dV_2(t) = d\hat{\Lambda}_0(t)^2 E(t) \hat{V}_\beta E(t)^T$$

where  $E(t)$  is the weighted average covariate over the risk set

$$E(t) = \frac{\sum_{i=1}^n Y_i x_i e^{x_i \hat{\beta}} / \pi_i}{\sum_{i=1}^n Y_i e^{x_i \hat{\beta}} / \pi_i}$$

The sum of the first two terms  $v_1(t) + v_2(t)$  is the variance of  $d\hat{\Lambda}_0$  conditional on the centering value for the covariates.

The remaining two terms depend on the value  $x_0$  for predicting survival.

$$dV_3(t) = \hat{\Lambda}_0(t) x_0 e^{x_0 \hat{\beta}} \hat{V}_\beta x_0^T e^{x_0 \hat{\beta}} \hat{\Lambda}_0(t)$$

and

$$dV_4(t) = -2\hat{\Lambda}_0(t) x_0 e^{x_0 \hat{\beta}} \hat{V}_\beta E(t)^T d\hat{\Lambda}_0(t) e^{x_0 \hat{\beta}}$$

The estimated variance of  $d\hat{\Lambda}(t; x)$  is given by integrating the variances of the increments

$$\widehat{\text{var}} \left[ \hat{\Lambda}(t; x) \right] = \sum_{s \leq t} V_1(s) + V_2(s) + V_3(s) + V_4(s)$$

and the variance of the estimated survival function is then

$$\widehat{\text{var}} \left[ \hat{S}(t; x) \right] = \widehat{\text{var}} \left[ \hat{\Lambda}(t; x) \right] \hat{S}(t; x)^2$$

It would be possible to combine  $v_2$ ,  $v_3$ ,  $v_4$  into a single term in  $E(t) - x_0$ , as Tsiatis did, but keeping them separate allows more computations to be shared across different values of  $x_0$ .

## 4 Notes

The Cox model estimator does not assume that the model is true, and so even under iid sampling differs from the standard martingale-based estimator of Tsiatis (1981). When data are simulated from a proportional hazards model the agreement seems to be very good.

Duplicating a data set to give clusters of two identical observations should give the same standard error estimate, and does.

The standard errors for the  $k$ -sample estimator in a two-phase sample from a cohort are very close to those computed by `nested.km` in Mark & Katki's `NestedCohort` package, for their zinc dataset.

## 5 References

- Binder DA. (1992) Fitting Cox's proportional hazards models from survey data. *Biometrika* 79: 139-147
- Särndal CE, Swensson B, Wretman J (1992) *Model Assisted Survey Sampling*. Springer.
- Tsiatis AA (1981) A Large Sample Study of Cox's Regression Model. *Annals of Statistics* 9(1) 93-108
- Williams RL (1995) Product-Limit Survival Functions with Correlated Survival Times. *Lifetime Data Analysis* 1: 171-186