Notes on Greenwood's Variance Estimator for the Kaplan-Meier Estimator

Jon A. Wellner

January 30, 2010

1. Introduction

Suppose that $X_1, \ldots, X_n$ are i.i.d. F and $Y_1, \ldots, Y_n$ are i.i.d. G independent of the X_i's. We observe $(Z_i, \Delta_i) = (X_i \wedge Y_i, 1_{[X_i \le Y_i]})$, $i = 1, \ldots, n$. We do not assume (here) that F or G is continuous.

Let $(Z, \Delta) \stackrel{d}{=} (Z_1, \Delta_1)$. We write

$$
\begin{aligned}
H^{uc}(t) &= P(Z \le t, \Delta = 1), \\
H^{c}(t) &= P(Z \le t, \Delta = 0), \\
H(t) &= P(Z \le t).
\end{aligned}
$$

Thus $1 - H(t) = P(Z > t) = (1 - F(t))(1 - G(t))$. The corresponding empirical distributions $\mathbb{H}_n^{uc}$, $\mathbb{H}_n^{c}$, and $\mathbb{H}_n$ are given by

$$
\begin{aligned}
\mathbb{H}_n^{uc}(t) &= n^{-1} \sum_{i=1}^{n} \Delta_i 1\{Z_i \le t\}, \\
\mathbb{H}_n^{c}(t) &= n^{-1} \sum_{i=1}^{n} (1 - \Delta_i) 1\{Z_i \le t\}, \\
\mathbb{H}_n(t) &= n^{-1} \sum_{i=1}^{n} 1\{Z_i \le t\}.
\end{aligned}
$$

2. Heuristics via the "Binomial likelihood"

Allowing ties, we let $0 < T_1 < \cdots < T_k < \infty$ denote the ordered distinct values of $Z_1, \ldots, Z_n$, and set $d_j = n \Delta \mathbb{H}_n^{uc}(T_j)$ for $j = 1, \ldots, k$. Thus d_j is the number of uncensored observations Z_i occurring at T_j. We also set $r_j = n(1 - \mathbb{H}_n(T_j-))$ for $j = 1, \ldots, k$. Thus r_j is the total number of observations Z_i with values at or beyond T_j.

As shown in problem 1 of problem set number 4 (Stat 582, Winter 2010), the natural nonparametric likelihood for the right-censoring problem as outlined above is

$$L_n(\underline{\lambda}) = \prod_{j=1}^{k} \lambda_j^{d_j} (1-\lambda_j)^{r_j - d_j} \tag{2.1}$$

where $\lambda_j \equiv \Delta\Lambda(T_j) = p_j / \sum_{i=j}^{k+1} p_i$ for $j = 1, \ldots, k$. The natural maximizer is, term by term,

$$\hat{\lambda}_j = \frac{d_j}{r_j} = \frac{n\Delta\mathbb{H}_n^{uc}(T_j)}{n(1-\mathbb{H}_n(T_j-))},$$

and the resulting estimator of F is

$$1 - \hat{F}_n(t) = \prod_{j:T_j<t} (1-\hat{\lambda}_j).$$

By the formal equivalence of (2.1) to the likelihood for k independent Binomial random variables, Binomial(r_j, λ_j), $j = 1, \ldots, k$, the information matrix for $\underline{\lambda} = (\lambda_1, \ldots, \lambda_k)^T$ is

$$I(\underline{\lambda}) = \text{diag}\left(\frac{r_j}{\lambda_j(1-\lambda_j)}\right),$$

and we expect that

$$\begin{aligned} Var(\log(1-\hat{F}_n(t))) &= Var(\sum_{j:T_j\le t} \log(1-\hat{\lambda}_j)) \\ &\doteq \sum_{j:T_j\le t} Var(\log(1-\hat{\lambda}_j)) \\ &\doteq \sum_{j:T_j\le t} \frac{1}{(1-\lambda_j)^2} \frac{\lambda_j(1-\lambda_j)}{r_j} \\ &= \sum_{j:T_j\le t} \frac{\lambda_j}{r_j(1-\lambda_j)}. \end{aligned}$$

Thus a natural estimator of this variance is

$$\begin{aligned} \widehat{Var}(\log(1-\hat{F}_n(t))) &= \sum_{j:T_j\le t} \frac{\hat{\lambda}_j}{r_j(1-\hat{\lambda}_j)} \\ &= \sum_{j:T_j\le t} \frac{d_j/r_j}{r_j(r_j-d_j)/r_j} \\ &= \sum_{j:T_j\le t} \frac{d_j}{r_j(r_j-d_j)}. \end{aligned}$$

This leads to

$$\widehat{Var}(1-\hat{F}_n(t)) = (1-\widehat{F}_n(t))^2 \sum_{j:T_j\le t} \frac{d_j}{r_j(r_j-d_j)} \tag{2.2}$$

$$= n^{-1}(1-\widehat{F}_n(t))^2 \int_{[0,t]} \frac{d\mathbb{H}_n^{uc}(s)}{(1-\mathbb{H}_n(s-))(1-\mathbb{H}_n(s-)-\Delta\mathbb{H}_n^{uc}(s))}. \tag{2.3}$$

This is Greenwood's (1926) estimate of the variance of life-table estimators, and the above derivation is based on the treatment in Cox and Oakes (1984), pages 50-51.

3. An approach via martingale theory

From Shorack and Wellner (1986), Proposition 7.2.1, page 301,

$$\frac{1-\hat{F}_n(t)}{1-F(t)} - 1 = -\int_{[0,t]} \frac{1-\hat{F}_{n-}}{1-F} d(\hat{\Lambda}_n - \Lambda);$$

equivalently,

$$\mathbb{Z}_n(t) \equiv \frac{\sqrt{n}(\hat{F}_n(t)-F(t))}{1-F(t)} = \int_{[0,t]} \frac{1-\hat{F}_{n-}}{1-F} d\mathbb{B}_n$$

$$= \int_{[0,t]} \frac{1-\hat{F}_{n-}}{1-F} \frac{1}{1-\mathbb{H}_n(s-)} d\mathbb{M}_n(s).$$

where

$$\mathbb{B}_n(t) = \int_{[0,t]} \frac{1}{1-\mathbb{H}_{n-}} d\mathbb{M}_n, \quad \mathbb{M}_n(t) = \sqrt{n}\left(\mathbb{H}_n^{uc}(t) - \int_{[0,t]} (1-\mathbb{H}_n(s-))d\Lambda(s)\right).$$

Here $\mathbb{M}_n$ is a (normalized) counting process martingale with predictable variation process

$$\langle \mathbb{M}_n \rangle(t) = \int_{[0,t]} (1-\mathbb{H}_n(s-))(1-\Delta\Lambda(s))d\Lambda(s).$$

It follows that the predictable variation process of $\mathbb{Z}_n$ is, using $1-\Delta\Lambda = (1-F)/(1-F_-)$,

$$\langle \mathbb{Z}_n \rangle(t) = \int_{[0,t]} \left(\frac{1-\hat{F}_{n-}}{1-F}\right)^2 \frac{1}{(1-\mathbb{H}_n(s-))^2}(1-\mathbb{H}_n(s-))(1-\Delta\Lambda(s))d\Lambda(s)$$

$$\to_p \int_{[0,t]} \left(\frac{1-F_-}{1-F}\right)^2 \frac{1}{(1-H(s-))}(1-\Delta\Lambda(s))d\Lambda(s)$$

$$= \int_{[0,t]} \frac{1}{(1-H(s-))(1-\Delta\Lambda(s))} d\Lambda(s)$$

$$
\begin{aligned}
&= \int_{[0,t]} \frac{1-G(s-)}{(1-F(s))(1-F(s-))(1-G(s-))^2} dF(s) \\
&= \int_{[0,t]} \frac{1-G(s-)}{(1-F(s))(1-F(s-))(1-G(s-))^2} dF(s) \\
&= \int_{[0,t]} \frac{1}{(1-F(s))(1-G(s-))(1-H(s-))} dH^{uc}(s) \\
&= \int_{[0,t]} \frac{1}{(1-H(s-)) \cdot [1-F(s-)-\Delta F(s)](1-G(s-))} dH^{uc}(s) \\
&= \int_{[0,t]} \frac{1}{(1-H(s-)) \cdot [1-H(s-)-\Delta H^{uc}(s)]} dH^{uc}(s).
\end{aligned}
$$

Thus the natural estimator of $\hat{F}_n(t)$ is just (2.3).

REFERENCES

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman and Hall, London.

Greenwood, M. (1926). The errors of sampling of the survivorship tables. In *Reports on Public Health and Statistical Subjects*, no. 33. London: HMSO. Appendix 1.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York. (Reprinted by SIAM, 2009.)

Department of Statistics
University of Washington
P.O. Box 354322
Seattle, Washington 98195-4322
U.S.A.
e-mail: jaw@stat.washington.edu