Biost 514: Applied Biostatistics I
Emerson, Fall 2009

Homework #5 Key
November 9, 2009

Questions for Biost 514 only:

It is sometimes said that all of statistics is founded on the Central Limit Theorem and a Taylor's Expansion. The *Delta Method* is the most common way that Taylor's Expansion is used in deriving asymptotic distribution of statistics. In this homework, you will derive Greenwood's formula for the standard error of the Kaplan-Meier estimator.

Censored survival data:

Suppose random variable T measures time to some event and that we are interested in estimating the survival distribution $S(t)= Pr\ (T > t) = 1 - F_T(t)$.

Suppose further that we cannot always directly observe T. Instead, there is some censoring variable $C \sim G(c)$ independent of T, and we can only observe the smaller of T and C: Define

$$Y= \min(T,C)$$
$$\delta = 1_{[Y=T]}$$

Under noninformative censoring, we can estimate $S(t)$ from the pairs (Y, δ) using the Kaplan-Meier estimates.

Kaplan-Meier estimator:

Suppose we have potentially censored observations $(Y_1,\delta_1), (Y_2, \delta_2), \ldots, (Y_n, \delta_n)$ defined as above.

Let $0 < \tau_1 < \tau_2 < \ldots < \tau_D$ be the D distinct times at which at least one failure was observed. Then for each τ_k define the number at risk N_k and the number of events d_k as

$$N_k = \sum_{i=1}^{n} 1_{[Y_i \geq \tau_k]} \qquad\qquad d_k = \sum_{i=1}^{n} \delta_i 1_{[Y_i = \tau_k]}$$

We want to estimate $S(t) = Pr(T > t)$. This can be effected by noting that for any ordered set of times $0= t_0 < t_1 < t_2 < \ldots t_k$, we can compute $S(t_k)$ as

$$S(t) = \prod_{i=1}^{k} \Pr(T > t_i \mid T > t_{i-1}).$$

Then, because an estimate of $\Pr(T > \tau_i \mid T > \tau_{i-1})$ can be computed from $1 - d_i / N_i$, the Kaplan-Meier estimator is given by

$$\hat{S}(t) = \prod_{k:\tau_k \leq t} \left(1 - \frac{d_k}{N_k}\right).$$

In deriving Greenwood's formula, as well as other methods of computing CI for estimated survival probabilities, we will find it useful to make use of the delta method:

Prop (delta method) : Suppose g is a differentiable function at θ and $a_n \to \infty$ as $n \to \infty$, then

$$a_n(Z_n - \theta) \quad \to_d \quad Z$$

implies

$$a_n(g(Z_n) - g(\theta)) \quad \to_d \quad g'(\theta)Z$$

where

$$g'(\theta) = \frac{\partial}{\partial \theta} g(\theta)$$

Proof: The proof of this proposition just makes use of a first order Taylor expansion.

5. Derive Greenwood's formula (the asymptotic variance in part c) and other appropriate methods for defining confidence intervals for the survival distribution.
   a. Under the assumption that the size N_k of the risk set at each time τ_i is large, find an approximate (asymptotic) distribution for the estimated hazards, where the hazard $\lambda(t)$ and its estimator are defined by

$$\lambda(t) = \lim_{h \to 0} \frac{\Pr(T < t + h \mid T \geq t)}{h}$$

$$\hat{\lambda}(t) = \begin{cases} \dfrac{d_k}{N_k} & t = \tau_k \\ 0 & t \notin \{\tau_1, \tau_2, \ldots, \tau_D\} \end{cases}.$$

   Express the asymptotic distribution in the form

$$\sqrt{n}\left(\hat{\lambda}(t) - \lambda(t)\right) \to_d N(0, V).$$

Ans: ***d_k*** **represents the number of events at time** ***τ_k*** **among the** ***N_k*** **subjects at risk. As each of the subjects are presumed to have the same hazard, then** ***d_k*** **has a binomial distribution with parameters** ***$n = N_k$*** **and** ***$p = \lambda_k = \lambda(\tau_k)$*****. The estimated hazard is thus just a sample mean of i.i.d., Bernoulli random variables having mean** ***λ_k*** **and variance** ***$\lambda_k (1 - \lambda_k)$*****. Hence, by the Levy central limit theorem**

$$\sqrt{N_k}\left(\hat{\lambda}_k - \lambda_k\right) = \sqrt{N_k}\left(\frac{d_k}{N_k} - \lambda_k\right) \to_d N(0, \lambda_k(1 - \lambda_k)).$$

   b. Use the delta method to find an asymptotic distribution for

$$\sqrt{n}\left(\log(1 - \hat{\lambda}(t)) - \log(1 - \lambda(t))\right) \to_d N(0, V^*).$$

Ans: **Let** ***$g(x) = \log(1 - x)$*****. Then** ***$g'(x) = -1 / (1 - x)$*****, and by the delta method**

$$\sqrt{N_k}\left(g\left(\hat{\lambda}_k\right)-g\left(\lambda_k\right)\right)=\sqrt{N_k}\left(\log\left(1-\frac{d_k}{N_k}\right)-\log\left(\lambda_k\right)\right)\rightarrow_d N\left(0,\frac{\lambda_k}{\left(1-\lambda_k\right)}\right).$$

c. Argue heuristically for the asymptotic distribution for

$$\sqrt{n}\left(\log\left(\hat{S}(t)\right)-\log\left(S(t)\right)\right)\rightarrow_d N\left(0,V^{**}\right).$$

Ans: **Now**

$$\hat{S}(t)=\prod_{k:\tau_k\le t}\left(1-\frac{d_k}{N_k}\right) \quad \Rightarrow \quad \log\left(\hat{S}(t)\right)=\sum_{k:\tau_k\le t}\log\left(1-\frac{d_k}{N_k}\right).$$

Furthermore, the asymptotic results in part b suggest the approximate distribution

$$\log\left(1-\frac{d_k}{N_k}\right) \;\dot{\sim}\; N\left(\log\left(1-\lambda_k\right),\frac{\lambda_k}{N_k\left(1-\lambda_k\right)}\right).$$

If we knew that the individual terms were totally independent, then we would be home free, because the sum of independent normals is normal, and we should be able to argue that if each of the independent terms that are asymptotically normal, then their sum will also be asymptotically normal under some reasonable conditions. But in the typical survival analysis setting, the same individual contributes to many different risk sets. However, under the assumption of noninformative censoring, each risk set has to look like a random sample from a population of subjects in the at-risk population. And as the sample sizes get large, the individual terms are asymptotically uncorrelated, and we have

$$\log\left(\hat{S}(t)\right)=\sum_{k:\tau_k\le t}\log\left(1-\frac{d_k}{N_k}\right)\dot{\sim}\,N\left(\log\left(S(t)\right),\sum_{k:\tau_k\le t}\frac{\lambda_k}{N_k\left(1-\lambda_k\right)}\right)$$

(More rigorously, we would define the probability of being at risk at time τ_k as the probability of neither failing nor being censored: Pr$(T \ge \tau_k, C \ge \tau_k) = \pi_k$ and note that N_k / n should be consistent for π_k, and we could rigorously derive that

$$\sqrt{n}\left(\log\left(\hat{S}(t)\right)-\log\left(S(t)\right)\right)\rightarrow_d N\left(0,\sum_{k:\tau_k\le t}\frac{\lambda_k}{\pi_k\left(1-\lambda_k\right)}\right).)$$

d. Use the delta method to find an asymptotic distribution for

$$\sqrt{n}\left(\hat{S}(t)-S(t)\right)\rightarrow_d N\left(0,V^{***}\right).$$

Ans: **Let $g(x) = \exp(x)$. Then $g'(x) = \exp(x)$, and by the delta method**

$$\sqrt{n}\left(\hat{S}(t)-S(t)\right)\rightarrow_d N\left(0,\left[S(t)\right]^2\sum_{k:\tau_k\le t}\frac{\lambda_k}{\pi_k\left(1-\lambda_k\right)}\right).$$

To be able to use this inferentially, we would need to estimate the standard error in the approximate distribution

$$\hat{S}(t) \;\dot{\sim}\; N\left(S(t), [S(t)]^2 \sum_{k:\tau_k \le t} \frac{\lambda_k}{n\pi_k(1-\lambda_k)} \right).$$

By Slutsky's theorem, we are allowed to use a consistent estimate of the standard error. So making the appropriate substitutions for *S(t)*, π_k , and λ_k , we would end up basing our inference on

$$\hat{S}(t) \;\dot{\sim}\; N\left(S(t), [\hat{S}(t)]^2 \sum_{k:\tau_k \le t} \frac{d_k}{N_k(N_k-d_k)} \right).$$

e. Show that confidence intervals derived using the asymptotic distributions in part d (or even part c) could have limits outside the interval (0,1). Show that this problem is avoided if the confidence intervals are defined first for log(-log(S(t))), and derive the method whereby this could be done.

<u>Ans:</u> Using the results of part d, we would obtain approximate 100(1-α)% CI as

$$\hat{S}(t) \;\pm\; z_{1-\alpha/2} \times \hat{S}(t) \sqrt{\sum_{k:\tau_k \le t} \frac{d_k}{N_k(N_k-d_k)}}.$$

Suppose that *n* = 1,000 and the first event is observed at time 1 prior to any censoring. Then the estimated survival is 0.999 and the estimated standard error is .999 ·√(1/999000) = 0.001, and the 95% CI for S(1) would be .997 to 1.001. Similarly, we can obtain a lower bound of a CI less than 0 on this additive scale which could conceivably result in any real number.

Similarly, if we use the results of part c, we would obtain the upper bound of an approximate 100(1-α)% CI as

$$\exp\left(\log(\hat{S}(t)) - z_{1-\alpha/2} \times \sqrt{\sum_{k:\tau_k \le t} \frac{d_k}{N_k(N_k-d_k)}} \right).$$

Supposing again that *n* = 1,000 and the first event is observed at time 1 prior to any censoring. Then the upper bound for the 95% CI for S(1) 1.001. In this case, we need not worry about obtaining a lower bound less than 0, because as the CI for log(*S(t)*) ranges from negative infinity to infinity, the exponentiation of that CI to be a CI for *S(t)* will range from 0 to infinity. But that includes values over 1.

However, if we apply the delta method with *g(x)* = log*(- x)* to the results of part c, we obtain

$$\log(-\log(\hat{S}(t))) \;\dot{\sim}\; N\left(\log(-\log(S(t))), \frac{1}{[\log(S(t))]^2} \sum_{k:\tau_k \le t} \frac{\lambda_k}{N_k(1-\lambda_k)} \right)$$

and we can use an estimated standard error and base CI on the approximate distribution

$$\log\left(-\log\left(\hat{S}(t)\right)\right) \dot{\sim} N\left(\log\left(-\log\left(S(t)\right)\right), \frac{1}{\left[\log\left(\hat{S}(t)\right)\right]^2}\sum_{k:\tau_k \le t}\frac{d_k}{N_k\left(N_k - d_k\right)}\right)$$

A CI for log(-log(*S(t)*) will range between negative infinity and infinity. When we exponentiate that, we will get a number between 0 and infinity. The additive inverse of that first exponentiation will therefore be negative, and exponentiating that last result will be between 0 and 1.