B.2 Modern Survival Problem sheet 2: Nonparametric estimation of survival curves

To be turned in by noon on Friday, 30 October, 2015

(1) Consider a situation where the multiplicative intensity model holds, and there is unobserved right censoring. That is, for some individuals we observe the event time T_i and $\delta_i = 1$; for others, we observe $\delta_i = 0$ and no event time. Suppose the right censoring is independent of event times, all n individuals are independent, and the distribution of censoring times is known to have cdf G. (So $G(c)$ is the probability of being censored before time c.) Let $t_1 < t_2 < \cdots < t_k$ be the observed event times (assumed distinct).

Show that

$$\hat{A}(t) = \sum_{t_i \le t} \Big((n - i + 1)\big(1 - G(t_i)\big) \Big)^{-1}$$

is an unbiased estimator for the cumulative hazard, and derive an estimator for the variance.

(2) Nonparametric estimators are inevitably less efficient (that is, have larger errors, on average) than parametric estimators. Consider the case when n individuals are observed up to time t. Their event times are independent and exponentially distributed with unknown parameter λ, and we observe all event times. We wish to compare two different possible estimators for the cumulative hazard up to time t: First, taking advantage of the knowledge that the data come from an exponential distribution; and second, using the nonparametric Nelson–Aalen estimator.

(a) i. Show that the MLE for $\Lambda(t)$ under the exponential model is

$$\frac{nt}{\sum_{i=1}^{n} T_i}.$$

ii. Compute the (approximate) variance for this estimator.

(b) Using the inequality

$$\log n + \gamma \le \sum_{i=1}^{n} \frac{1}{i} \le \log n + \gamma + \frac{1}{2n},$$

show that the Nelson–Aalen estimator for $S(t)$ is approximately $Y(t+)/n$ (the empirical fraction surviving to time n), and find a bound for the error — that is, for the maximum difference between $\widetilde{S}(s)$ and $Y(s+)/n$ on $0 \le s \le t$.

(c) Use this to estimate the variance of $\hat{A}(t)$, the Nelson–Aalen estimator. Show that this variance is larger than the variance for the parametric estimator above.

(d) Plot the ratio of the variances for a range of values of t, for the case $\lambda = 1$ and $n = 1000$.

(e) Why might one prefer to use the nonparametric estimator, even when there is no censoring?

(3) You may find `R` code at http://steinsaltz.me.uk/survival/countingprocess.R that simulates and plots a survival process, that starts with $n = 50$ individuals, each with constant mortality rate $\alpha = 1$. What is the intensity λ of the survival process at time t?

   (a) `R` code runs faster if you replace loops with vector operations. Can you get rid of the loop in this code?
   (b) Modify the code to plot the martingale $N(t) - \int_0^t \lambda(s)ds$.
   (c) Add a routine that computes the optional variation process. Run a simulation and plot it.
   (d) Modify the program to apply to $\alpha(t) = 2t$.

(4) The data set `ovarian`, included in the `survival` package, presents data for 26 ovarian cancer patients, receiving one of two treatments, which we will refer to as the *single* and *double* treatments. (They appear in the data set as the `rx` variable, taking on values 1 and 2 respectively.)

   (a) Create a survival object for the times in this database.
   (b) Compute and plot the Kaplan–Meier estimator for the survival curves. (For a small extra challenge, plot the single-treatment survival curve black, and the double-treatment curve red.) You may use the `survfit` function.
   (c) Compute the Nelson–Aalen survival curve estimate. Make a table of the relevant data (time of events, number of events, number at risk).
   (d) Compute the standard error for the probability of survival past 400 days in each group, as estimated by the Nelson–Aalen and Kaplan–Meier estimators.