Fixed vs. Random Censoring: Is Ignorance Bliss?

Stephen Portnoy

June 19, 2007

In many situations where censored observations are observed, it is not unreasonable to assume that the censoring values are known for all observations (even the uncensored ones). For example, one of the earliest approaches to censored regression quantiles was introduced by work of Powell in the mid 1980's. Powell assumed that the censoring values were constant, thus positing observations of the form $Y = \min(T, c)$ (where Y is observed and T is the possibly unobserved survival time that is assumed to obey some linear model). More generally, we may be willing to assume that we observe a sample of censoring times $\{c_i\}$ and a sample of censored responses $Y_i = \min(T_i, c_i)$, a model that could apply to a single sample. In this case, one could use the empirical distributions of the $\{Y_i\}$ and $\{c_i\}$ and take the ratio of empirical survival functions to estimate the survival function of T. This is asymptotically equivalent to applying the Powell method on a single sample.

Despite some optimality claims of Newey and Powell, it turns out that the Kaplan Meier estimate is better (asymptotically, and by simulations in finite samples) even though it does not use the full sample of $\{c_i\}$ values. More generally, even in multiple regression settings, the censored regression quantile estimators (Portnoy, JASA, 2003) are better in simulations than Powell's estimator (even for the constant censoring situation for which Powell's estimator was developed). Remarkably, in the one sample case, replacing the empirical function of $\{c_i\}$ by the true survival function (assuming it is known) yields an even less efficient estimator. Thus, it appears that discarding what appears to be pertinent information improves the estimators. The talk will try to quantify and explain this conundrum.