Statistical Science
2008, Vol. 23, No. 3, 313–317
DOI: 10.1214/08-STS244B
Main article DOI: 10.1214/07-STS244

Comment: Quantifying Information Loss in Survival Studies

Hani Doss

In their paper, Nicolae, Meng and Kong (henceforth NMK) propose several very interesting methods for quantifying the fraction of missing information in a sample, and focus their attention on genetic studies. Survival analysis is another area in statistics where missing information plays an important role. Here, censoring complicates study design, for example when we want to determine how big a clinical trial should be in order to have a good chance of detecting a treatment effect in a Cox model. Most current methods for dealing with this difficult problem involve two stages, where in the first stage we make a projection of what the variance of the coefficient of the treatment effect would be if there was no censoring, and in the second stage we make a correction to adjust for the censoring. Often this is done under restrictive parametric (e.g., exponential) assumptions for the underlying distributions. It would be desirable to use the methods proposed by NMK in the survival analysis setting. I tried to carry over their methods to the Cox model, and encountered some problems. The difficulties I discovered led me to consider modifications of their proposals, which I believe work well. Below I discuss the setup I consider, my experiences, the issues, and some approaches I think are promising.

1. SURVIVAL STUDIES FOR ASSESSING THE EFFICACY OF A NEW TREATMENT

A typical clinical trial with a survival outcome involves a fixed time frame, say five years. Patients enter the trial continuously during the first four years, are randomly assigned to treatment or control, and the last year is a followup year, during which no patients enter the study. Some patients die during the study, in which case their survival time is observed. But some patients die from other causes or are lost to followup, and some are still alive at the time the trial is ended; so in these cases the survival time is censored: for each individual in this group, there is a time t and we know only that the individual's survival is greater than t.

Clearly the censoring reduces information regarding the efficacy of the new treatment. When designing a subsequent study in the hope of getting stronger evidence against the null hypothesis of no treatment effect, we now have two choices: increase the number of patients in the study, which can be expensive, or try to reduce the censoring. We can reduce the censoring either by putting more resources into followup, or by extending the length of the period of time after the end of the accrual period. These result in costs which are financial and also ethical because increasing the length of the final followup period postpones publication of results that are of potential benefit to other patients. The decision of whether to increase the number of patients or to reduce the censoring depends crucially on the amount of information loss due to censoring, so being able to measure this is extremely important in the design of future studies. This situation is very similar to the one discussed by NMK.

By far the most commonly used model for regression with censored survival data is the Cox proportional hazards model. Suppose that individual i has covariate vector $Z_i = (Z_{i1}, \ldots, Z_{ip})$, where Z_{i1} is the indicator that the individual receives the treatment. Let X_i be the death time of individual i if there was no censoring, and let Y_i be the censoring time. For each individual, we observe the minimum $T_i = \min(X_i, Y_i)$ and also the indicator δ_i that X_i was not censored, that is, $\delta_i = I(X_i \le Y_i)$. So the data for individual i is the triple (T_i, δ_i, Z_i).

The proportional hazards model stipulates that the hazard rate for an individual with covariate vector Z is given by

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta' Z), \tag{1}$$

where β is a p-dimensional vector of coefficients, and λ_0 is the hazard function for an individual with covariate vector 0. For our purposes (as will be clear later), it is preferable to define the model in terms of cumulative hazard functions, and so by integrating (1), the

Hani Doss is Professor, Department of Statistics, University of Florida, Gainesville, Florida 32611, USA (e-mail: doss@stat.ufl.edu).