Computationally efficient estimation of the influence function for Kaplan-Meier censoring

Johan Sebastian Ohlendorff

2022-12-20

We assume that the event times are sorted and possibly tied such that $\tilde{T}_1 < \ldots < \tilde{T}_i = \ldots = \tilde{T}_{i+k} < \tilde{T}_{i+k+1} < \ldots < \tilde{T}_n$. We use the following algorithm to preserve memory and the number of iterations for say $\mu^{(i)} = \int 1_{\{t \leqslant \tau, d=1\}} \frac{f_i(t-)}{G(t-)} P(dz)$. The idea is to split the sum into two terms:

$$\frac{1}{n} \sum_{j=1}^{n} \frac{\hat{f}_i(\tilde{T}_j-) 1_{\{\tilde{T}_j \leqslant \tau, \Delta_j = 1\}}}{\hat{G}(\tilde{T}_j-)} =$$
$$\frac{1}{n} \left(\sum_{j=2}^{i+k} \frac{g(j) 1_{\{\tilde{T}_j \leqslant \tau, \Delta_j = 1\}}}{\hat{G}(\tilde{T}_j-)} + h(i) \sum_{j=i+k+1}^{n} \frac{1_{\{\tilde{T}_j \leqslant \tau, \Delta_j = 1\}}}{\hat{G}(\tilde{T}_j-)} \right)$$

since $\hat{f}_i(\tilde{T}_j-)$ only depends on i for $i+k > j$ and only depends on j for $i+k \leqslant j$, so these values are calculated a priori. Also the first term will always be zero, since we are looking at the value of the integral before any observed event (hence the sum starts at $j = 2$). One can check in the estimation of the Influence Curve for the censoring, which does not depend on the covariates that we need to calculate $2n$ values (i.e. n values for $g(i)$ and n for $h(j)$). This is how we can avoid memory issues. The algorithm is:

$t := 1$
$\hat{\mu}_2 := \sum_{j=1}^{n} \frac{1_{\{\tilde{T}_j \leqslant \tau, \Delta_j = 1\}}}{\hat{G}(\tilde{T}_j -)}$
while $\tilde{T}_1 = \tilde{T}_t$ *and* $t \leqslant n$ **do**
  if $\tilde{T}_t \leqslant \tau$ *and* $\Delta_t = 1$ **then**
    $\hat{\mu}_2 = \hat{\mu}_2 - \frac{1}{G(\tilde{T}_t -)}$
  end
  $t = t + 1$
end
$tieEnd := t - 1$
$\hat{\mu}_1 := 0$
for $i = 1$ *to* n **do**
  $\hat{\mu}^{(i)} = \frac{1}{n}\left(\hat{\mu}_1 + h(i)\hat{\mu}_2\right)$
  if $tieEnd \leqslant i$ **then**
    $t = i + 1$
    while $\tilde{T}_1 = \tilde{T}_t$ *and* $t \leqslant n$ **do**
      if $\tilde{T}_t \leqslant \tau$ *and* $\Delta_t = 1$ **then**
        $\hat{\mu}_2 = \hat{\mu}_2 - \frac{1}{G(\tilde{T}_t -)}$
        $\hat{\mu}_1 = \hat{\mu}_1 + \frac{g(t) 1_{\{\tilde{T}_t \leqslant \tau, \Delta_t = 1\}}}{\hat{G}(\tilde{T}_t -)}$
      end
      Let $t = t + 1$
    end
  end
  Let $tieEnd = t - 1$
end
return $\hat{\mu}^{(i)}$ for each $i = 1, \ldots, n$

The idea is that but we keep on adding and subtracting the terms with tied values in the event times. Then we do not need to calculate a sum for each i.