Causal effect of treatment on right-censored survival time

Let $O = (W, A, \Delta = I(T \leq C), \tilde{T} = \min(T, C))$, T is a survival time, C is a right-censoring time, A is a binary treatment, W are baseline covariates. Suppose that C is independent of T, given A, W. Consider the conditional survivor function:

$$S(t_0 \mid A, W) = P(T > t_0 \mid A, W).$$

Let

$$\lambda(t \mid A, W) = P(T = t \mid A, W, T \geq t)$$

be the conditional hazard of survival at time t. We have

$$S(t_0 \mid A, W) = \prod_{s \leq t_0} (1 - \lambda(s \mid A, W)).$$

If T is continuous, then this writes as

$$S(t_0 \mid A, W) = \prod_{s \leq t_0} (1 - d\Lambda(s \mid A, W)).$$

If C is independent of T, given A, W, then we can identify the conditional hazard as follows:

$$\lambda(t \mid A, W) = P(\tilde{T} = t, \Delta = 1 \mid \tilde{T} \geq t, A, W).$$

Let's denote the right-hand side with $\tilde{\lambda}(t \mid A, W)$. Define

$$\Psi_a(P) = E_P S(t_0 \mid A = a, W).$$

Under a causal model in which $T = T_A$, T_0, T_1 treatment specific counterfactual survival times, and the assumption that A is independent of T_0, T_1, given W, we have

$$\Psi_a(P) = P(T_a > t_0),$$

the counterfactual survival function at t_0 under intervention $A = a$.

Prediction of survival based on right-censored data

Suppose that we want to estimate $\psi_0(t_0, A, W) = S_0(t_0 \mid A, W)$ at a given point t_0, based on the right-censored data structure $O = (W, A, \Delta, \tilde{T})$. If there would not be right-censoring, then a valid loss function would be

$$L^F(\psi)(W, A, T) = (I(T > t_0) - \psi(t_0, A, W))^2.$$

Let $\bar{G}(t \mid A, W) = P(C \geq t \mid A, W)$. We can define the Inverse probability of censoring weighted loss function:

$$L_{G_0}(\psi)(O) = \frac{L^F(\psi)(W, A, \tilde{T})\Delta}{\bar{G}(\tilde{T} \mid A, W)}.$$

An improved IPCW-loss is given by:

$$L_{G_0}(\psi)(O) = \frac{L^F(\psi)(W, A, \tilde{T})\{I(\tilde{T} \leq t_0, \Delta = 1) + I(\tilde{T} > t_0)\}}{\bar{G}_0(\min(\tilde{T}, t_0) \mid A, W)}.$$

The cross-validation selector is now defined by:

$$k_n = \arg\min_k \frac{1}{V} \sum_v \sum_{i \in VAL(v)} L_{G_{n,v}}(\hat{\Psi}_k(P_{n,v})(O_i),$$

where $G_{n,v}$ is an estimator of G_0 based on the training sample $P_{n,v}$.

Candidate estimators of the conditional survival function

Direct estimators IPC-weighted logistic regression estimators regressing $I(T > t_0)$ on A, W using weights

$$\frac{I(\tilde{T} \leq t_0, \Delta = 1) + I(\tilde{T} > t_0)}{\bar{G}_0(\min(\tilde{T}, t_0) \mid A, W)}.$$

Through estimation of the conditional hazard: Consider the case that T is discrete. We can estimate the hazard $\tilde{\lambda}(t \mid A, W)$ with logistic regression of the binary outcome $I(\tilde{T} = t, \Delta = 1)$, given A, W, among observations with $\tilde{T} \geq t$. We can also run a pooled logistic regression by creating a row of data for each $s \leq \tilde{T}$, define by $(W, A, I(\tilde{T} = s, \Delta = 1))$, $s = 1, \ldots, \tilde{T}$, thereby simultaneously fitting the conditional hazard as a function of (t, A, W).

If T is continuous, we can discretize T using an interval partition of its range. We can now use the above logistic regression approach to estimate the corresponding discrete conditional hazard, and by dividing $\lambda_d(t \mid A, W)$ by the width $h(t)$ of the interval that contains t, we obtain an estimate of the actual continuous conditional hazard. The choice of interval partition, width of the interval, can now be a tuning parameter