BIS 537 Final Project Report

Can Meng and Waveley Qiu

2023-12-11

Background

Survival Analysis and Causal Inference

While the ideal setting to conduct causal inference on survival outcomes are randomized experiments, as is true for most areas of research, it is often the case that such an experimental setting is difficult or impossible to achieve. As such, when treatment and control groups are not exchangeable, it may be necessary to implement causal inference methods to address the absence of direct counterfactual proxies.

Time-to-event outcomes are of great interest to investigators across a variety of clinical settings, however present analytical difficulties even within the constrains of a randomized controlled trial. In particular, the non-observation of an event yet loss of a subject's data, or censoring, presents another missing data issue that the investigator has need to consider. This issue is compounded in the setting of a non-randomized or observational study, where exchangeability between treatment groups cannot be assumed of the collected data.

In this study, we will investigate two possible methods in which this dual-level missing data issue can be addressed. First, is a two-arm weighting method, in which weights based on the probability of treatment and the probability of censoring are calculated and used together in estimating the causal estimand of interest. The second is a single-arm sequential weighting method, in which weights based on the probability of treatment are used to construct a pseudo-population from which weights based on the probability of censoring indication are derived – the final compositely-weighted dataset will be used to estimate the causal estimand..

Causal Survival Estimands

The estimand of interest for this simulation study will be the restricted average survival causal effect, defined as follows (Mao 2018):

$$\begin{aligned}\Delta_{RACE} &= \frac{E[\omega(e_i)\min(T_{1i}, t^*)]}{E[\omega(e_i)]} - \frac{E[\omega(e_i)\min(T_{0i}, t^*)]}{E[\omega(e_i)]} \\ &= \int_0^{t^*} S_1(t)dt - \int_0^{t^*} S_0(t)dt\end{aligned}$$

This estimand is interpreted as the average difference in survival time between the treatment and control groups, if both potential outcomes are observed, under the upper bound time restriction of t^*.

Data Generation Settings

We will be using three covariates in this simulation study: X_1, X_2, X_3. These covariates have the following distributions: $X_1 \sim$ Bernoulli(0.6), $X_2 \sim N(0, 1)$, $X_3 \sim$ Gamma$(1, 1)$. Patterned after the simulation in Mao 2018, we will define our true models as follows:

Propensity Score Model

The propensity score model will be defined as the following logistic regression:

$$g^{-1}(E[Z=1|\mathbf{X}_i]) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i},$$
$$\text{where } g(\mathbf{X_i}) = \frac{\exp(\boldsymbol{\beta}\mathbf{X}_i^T)}{1+\exp(\boldsymbol{\beta}\mathbf{X}_i^T)} = E[Z=1|\mathbf{X}_i] = e(\mathbf{X}_i)$$

In our simulation, β_1, β_2, and β_3 will be varied to create settings of varying overlap (i.e., weak and strong overlaps), and β_0 will be varied to specify treatment proportions (i.e., low and high treatment proportions).

Survival Times Model

The outcomes model will be defined as the following Cox-Weibull model:

$$h(t|\mathbf{X}_i) = h_0(t)\exp(L_i),$$
$$\text{where } h_0(t) = \lambda\nu t^{\nu-1}, \text{ and } L_i = a_0 Z_i + a_1 X_{1i} + a_3 X_{3i}$$

$$\begin{aligned} S(T_i^S = t) &= \exp(-\lambda t^v \exp(L_i)) \\ &= 1 - F(t) \end{aligned}$$

Then, the survival time for subject i is drawn from:

$$T_i^S = \left(\frac{-\log(u_i^S)}{\lambda\exp(L_i)}\right)^{1/\nu}, \text{where } u_i^S \sim \text{Unif}(0,1)$$

Censoring Model

The censoring model will be defined as the following exponential model:

$$T_i^C \sim \text{Exponential}(\lambda\exp(K_i)),$$
$$\text{where } K_i = \gamma_0 + \gamma_2 X_{2i} + \gamma_3 X_{3i}$$

$$P(T_i^C = t) = \exp(-\lambda t\exp(K_i))$$

For simplicity, note that T_i^C is independent of T_i^S as well as treatment Z_i. This indicates that we are making the assuming that a subject's censoring time is dependent only on baseline covariates and is not influenced by their actual survival time or treatment assigned treatment.

Algorithmically, the censoring time for subject i will be drawn from:

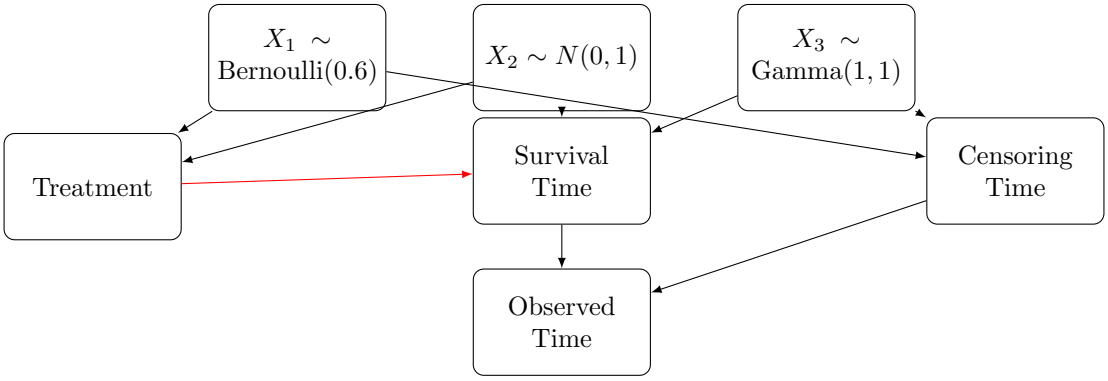
$$T_i^C = \frac{-\log(u^C)}{\lambda\exp(K_i)}, \text{where } u^C \sim \text{Unif}(0,1)$$

For subject i, the observed time, T^{obs} is the minimum of T_i^C and T_i^S ($T^{obs} = \min(T_i^C, T_i^S)$), the censoring indicator for subject i, given the survival time T_i and censoring time T^C will be assigned as follows:

$$C_i(T_i^S, T_i^C) = \begin{cases} 1, & \text{where } T_i^S > T_i^C \\ 0. & \text{where } T_i^S \leq T_i^C \end{cases}$$

Data Simulation

DAG



Treatment and Propensity Score

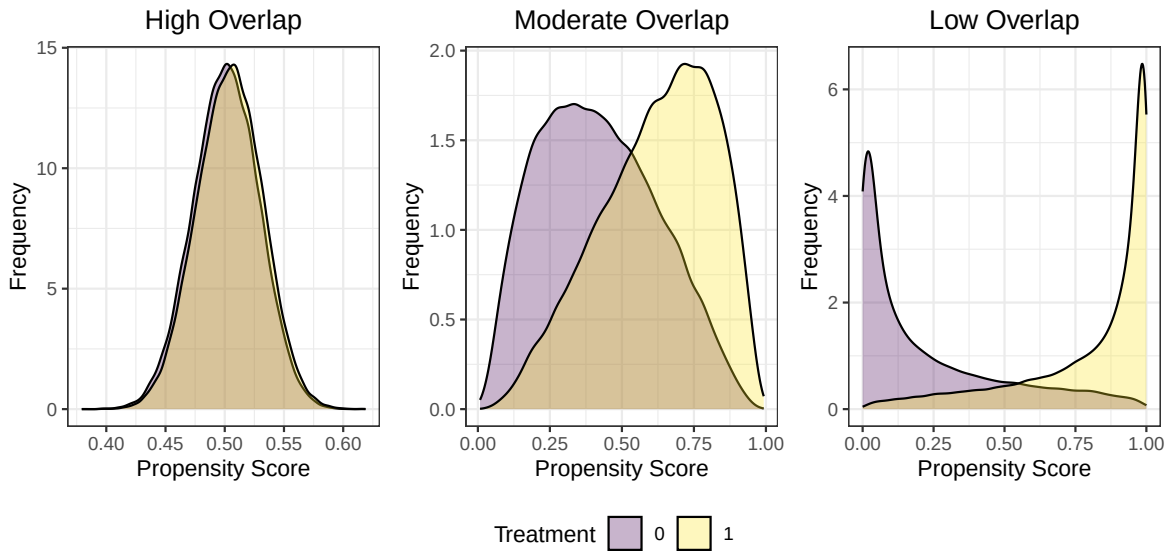
The propensity score is defined as follows:

$$
\begin{aligned}
e(\mathbf{X}_i) = P(Z = 1|\mathbf{X}_i) &= E[Z = 1|\mathbf{X}_i] \\
\boldsymbol{\beta}\mathbf{X}_i^T &= \text{logit}(E[Z = 1|\mathbf{X}_i]) \\
\implies E[Z = 1|\mathbf{X}_i] &= \frac{\exp(\boldsymbol{\beta}\mathbf{X}_i^T)}{1 + \exp(\boldsymbol{\beta}\mathbf{X}_i^T)} \\
E\Big[P(Z = 1)\Big] &= E_X\Big[E_Z\Big(Z = 1|\mathbf{X}_i\Big)\Big] \\
&= E_X\left[\frac{\exp(\boldsymbol{\beta}\mathbf{X}_i^T)}{1 + \exp(\boldsymbol{\beta}\mathbf{X}_i^T)}\right] = p
\end{aligned}
$$

As described previously, the covariates X_1 and X_2 in this model will be drawn from Bernoulli(0.6) and $N(0, 1)$ distributions, respectively. Then, we see that we will need to select three β coefficients to satisfy the form $g^{-1}(E[Z = 1|X]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, where $g(\mathbf{X}) = \frac{\exp(\boldsymbol{\beta}\boldsymbol{X}^T)}{1+\exp(\boldsymbol{\beta}\boldsymbol{X}^T)} = E[Z = 1|X]$.

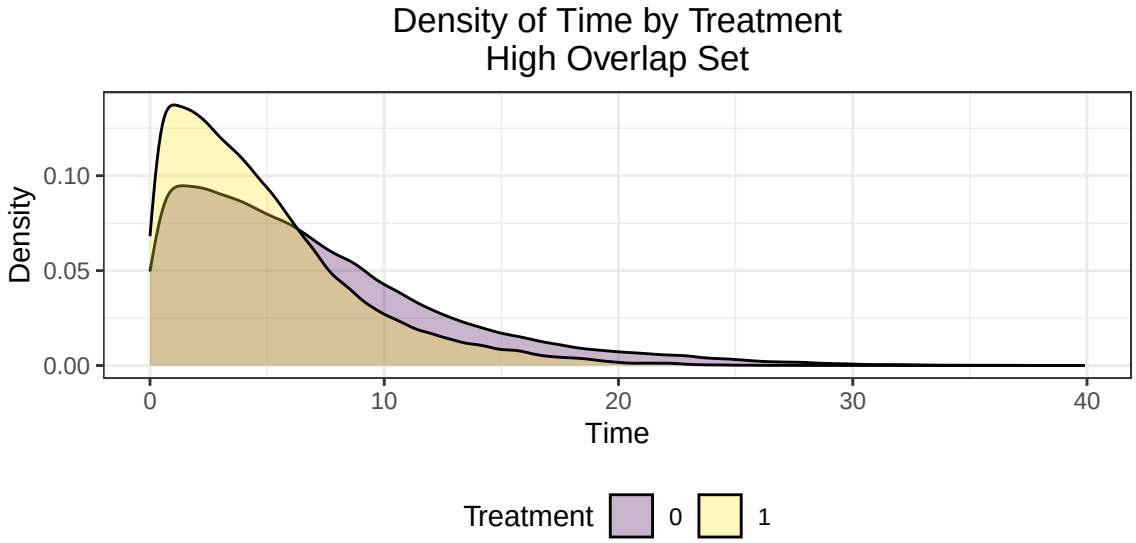
For weak overlap, let $\beta_1 = \beta_2 = \beta_3 = 3$. For moderate overlap, $\beta_1 = \beta_2 = \beta_3 = 1$. For strong overlap, $\beta_1 = \beta_2 = \beta_3 = 0.1$.

The following plots show the overlap of these three simulations:



Outcomes

The following generates the outcomes, drawn from a Cox-Weibull model.



Censoring

True Estimand

As we have previously established, $T(0)\Big|(X_1, X_3)$ and $T(1)\Big|(X_1, X_3)$ are drawn from the following models:

$$T(1)\Big|(X_1, X_3) \sim h_{Z=1}(t|X_1, X_3) = \lambda v t^{v-1} \exp(\alpha_0 + \alpha_1 X_1 + \alpha_3 X_3)$$
$$= \lambda^* \nu t^{\nu-1}, \text{ where } \lambda^* = \lambda \exp(\alpha_0 + \alpha_1 X_1 + \alpha_3 X_3)$$

$$T(0)\Big|(X_1, X_3) \sim h_{Z=0}(t|X_1, X_3) = \lambda v t^{v-1} \exp(\alpha_1 X_1 + \alpha_3 X_3)$$
$$= \lambda^* \nu t^{\nu-1}, \text{ where } \lambda^* = \lambda \exp(\alpha_1 X_1 + \alpha_3 X_3)$$

Our target estimand is the average causal effect, or $E[T(1) - T(0)]$, which is unconditional on (X_1, X_3). By the law of total expectations and iterated expectations, we see the following:

$$E[T_i] = E\Bigg[E\Big[T_i|X_1, X_3)\Big]\Bigg]$$
$$= \int_{X_1, X_3} E\Big[T_i|X_1 = x_1, X_3 = x_3]P(X_1 = x_1, X_3 = x_3)d\mu(x_1, x_3)$$
$$= \int_{X_1, X_3} E\Big[T_i|X_1 = x_1, X_3 = x_3]P(X_1 = x_1)P(X_3 = x_3)d\mu(x_1, x_3)$$

Let $X_{11}, ..., X_{1m} \sim P(X_1)$ and $X_{31}, ..., X_{3m} \sim P(X_3)$. Then, for m sufficiently large, we note:

$$E[T_i] = \int_{X_1, X_3} E\Big[T_i|X_1 = x_1, X_3 = x_3]P(X_1 = x_1)P(X_3 = x_3)d\mu(x_1, x_3)$$
$$\approx \sum_{j=1}^{m} E[T_i|X_1 = x_{1j}, X_3 = x_{3j}]$$

To obtain the true value estimates of $E[T(1)]$ and $E[T(0)]$, we will draw 1000000 samples each from the distributions of X_1 and X_3 to simulate the distributional behavior of each of these random variables, and sum the computed conditional expectations of $T(1)|(X_{1j}, X_{3j})$ and $T(0)|(X_{1k}, X_{3k})$ for each sample j and k, where $j = k = 1000000$.