BIS 537 Final Project Report

Waveley Qiu (wq46)

2023-11-28

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.

In this study, we will investigate two possible methods in which this dual-level missing data issue can be addressed. First, is an iterated 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. We then will investigate a matching and weighting combination method, in which the probability of censoring based on baseline covariates will be used to match censored records with complete records and the matched records will be weighted based on their propensity scores.

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: $X_1 \sim N(0, 1)$, $X_2 \sim$ Bernoulli(0.6), $X_3 \sim$ 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:

$$g^{-1}(E[Z=1|\mathbf{X}_i]) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i},$$
$$\text{where } g(\mathbf{X_i}) = \frac{\exp(\boldsymbol{\beta}\mathbf{X}_i^T)}{1+\exp(\boldsymbol{\beta}\mathbf{X}_i^T)} = E[Z=1|\mathbf{X}_i] = e(\mathbf{X}_i)$$

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 Models

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

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

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

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

Data Simulation

Treatment and Propensity Score

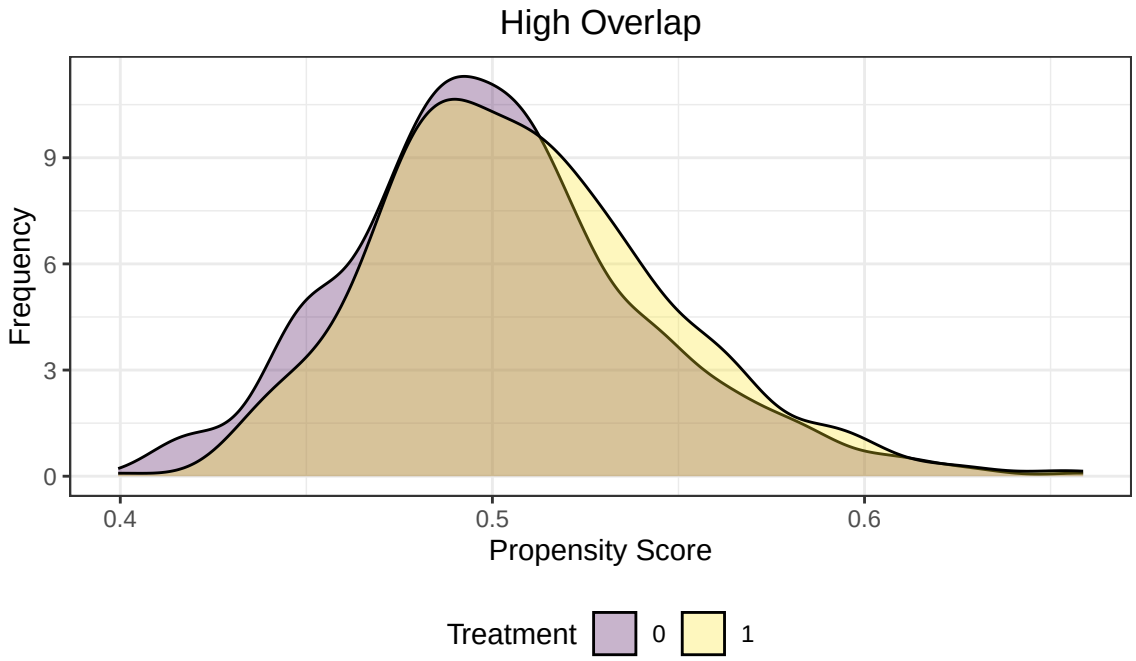
The propensity score is defined as follows:

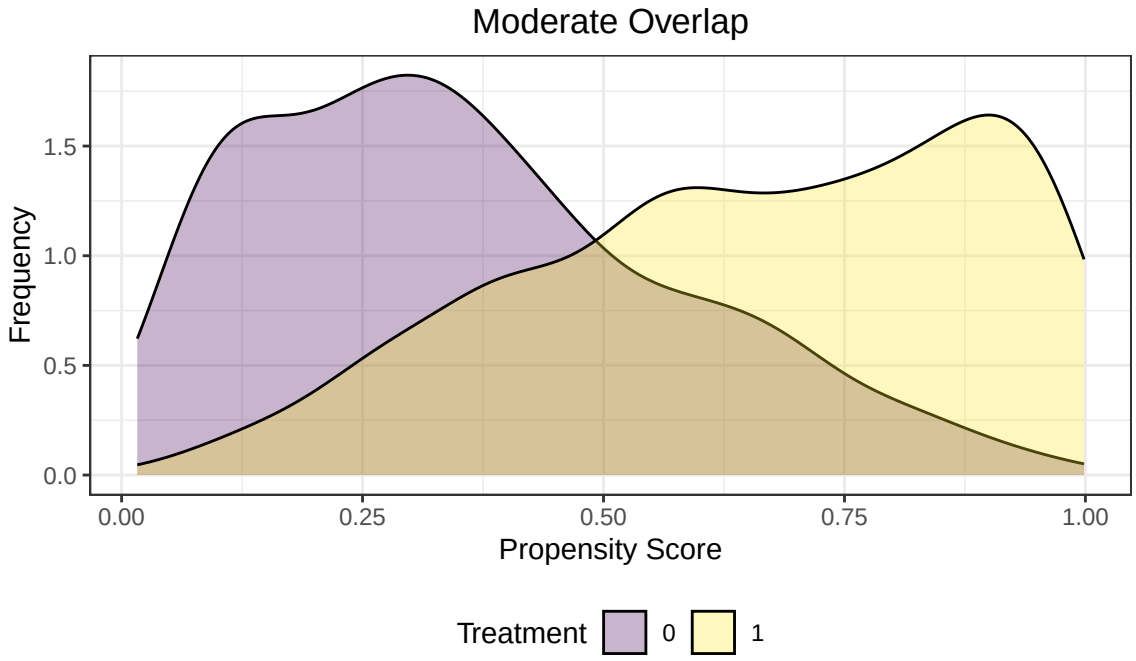
$$\begin{aligned} e(\mathbf{X}_i) = P(Z=1|\mathbf{X}_i) &= E[Z=1|\mathbf{X}_i] \\ \boldsymbol{\beta}\mathbf{X}_i^T &= \text{logit}(E[Z=1|\mathbf{X}_i]) \\ &\implies E[Z=1|\mathbf{X}_i] = \frac{\exp(\boldsymbol{\beta}\mathbf{X}_i^T)}{1+\exp(\boldsymbol{\beta}\mathbf{X}_i^T)} \\ E\Big[P(Z=1)\Big] &= E_X\Big[E_Z\Big(Z=1|\mathbf{X}_i\Big)\Big]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

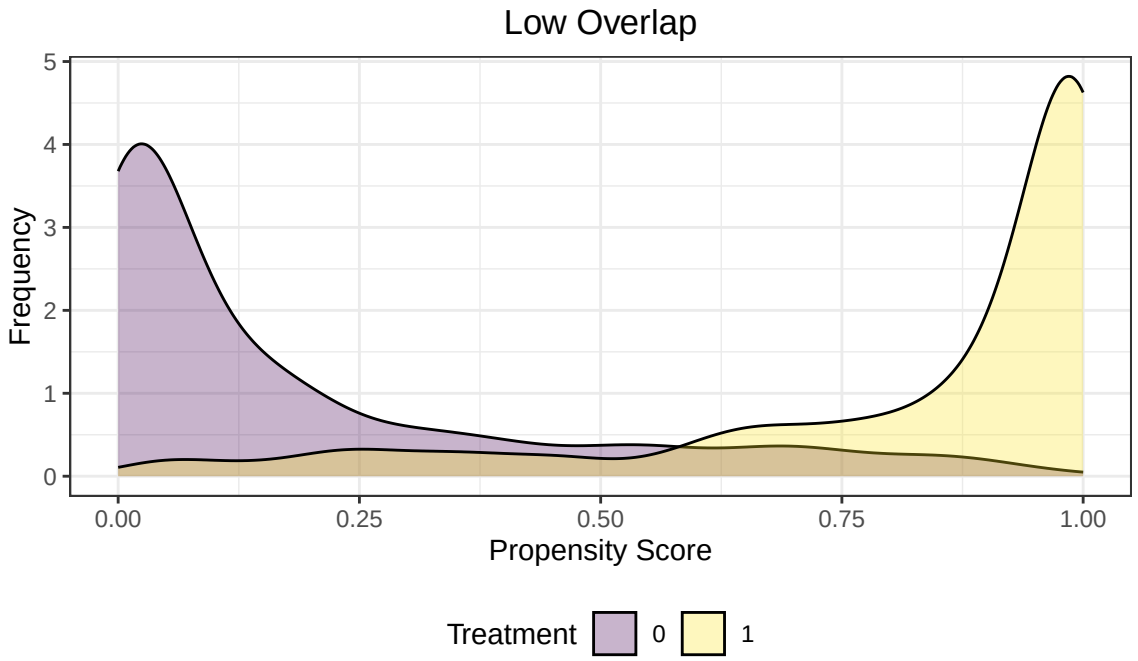
Let $X_1 \sim N(0,1)$, $X_2 \sim \text{Bernoulli}(0.6)$, representing one continuous and one binary covariate upon which treatment is determined for each subject. Then, we see that we will need to select three β coefficients to satisfy the form $g^{-1}(E[Z=1|X]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, 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:


High Overlap
Frequency
9
6
3
0
0.4
0.5
0.6
Propensity Score
Treatment
0
1


Moderate Overlap
Frequency
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
Propensity Score
Treatment
0
1



Outcomes

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