EXERCISES FOR RANDOMIZATION AND CAUSATION (MATH-336)

Exercise Sheet 8

Exercise 1 (IPW and M-estimation)**.** This is the continuation of Exercise 1 of last week and we will consider the same setup.

(a) Find the asymptotic variance of the IPW estimator of the ATE when the propensity score is known.

(b) Suppose the propensity score is unknown. Write down the expression for the IPW estimator, $\hat{\text{ATE}}_{\text{IPW},u}$, of $E[Y^1 - Y^0]$;

(c) Suppose we do not want to posit a parametric model on the propensity score. Find sufficient conditions for an estimator $\hat{\pi}(L)$ of $\pi(L)$ that guarantee $\hat{\text{ATE}}_{\text{IPW},u}$ to be a consistent estimator of $E[Y^1 - Y^0]$.
Hint: use the properties of the IPW estimator when the propensity score is known; use the triangular inequality.

(d) Can we use the result of point c) to build confidence intervals?

Solution:

(a) define

$$C(\gamma) = E\left[-\dot{M}(A, L, Y; \gamma)\right]$$
$$B(\gamma) = E\left[M(A, L, Y; \gamma) M(A, L, Y; \gamma)^T\right]$$

all of which exists under exchangeability, positivity, and consistency. Remember that we defined M as $M(A, L, Y; \gamma) = \frac{AY}{\pi(L)} - \frac{(1-A)Y}{1-\pi(L)} - \gamma$ If these three quantities exist and regularity conditions we have

$$\sqrt{n}(\hat{\gamma} - \gamma_0) \xrightarrow{d} \mathcal{N}\left(0, \underbrace{C(\gamma_0)^{-1} B(\gamma_0) C(\gamma_0)^{-1^T}}_{:=V(\gamma_0)}\right)$$

Let us compute these quantities.

$$\dot{M}(a, l, y; \gamma) = -1$$

so that

$$C(\gamma_0) = 1$$

and thus

$$C(\gamma_0)^{-1} = 1$$

Next,

$$B(\gamma_0) = E\left[\left(\frac{AY}{\pi(L)} - \frac{(1-A)Y}{1-\pi(L)} - \gamma_0\right)^2\right]$$

Since

$$\gamma_0 = E\left[\frac{AY}{\pi(L)} - \frac{(1-A)Y}{1-\pi(L)}\right],$$

we deduce that

$$B(\gamma_0) = Var\left(\frac{AY}{\pi(L)} - \frac{(1-A)Y}{1-\pi(L)}\right)$$

Finally, the asymptotic variance of the IPW estimator of the ATE when the propensity score is known is :

$$V(\gamma_0) = Var\left(\frac{AY}{\pi(L)} - \frac{(1-A)Y}{1-\pi(L)}\right)$$

(b)

$$\hat{\text{ATE}}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i 1(A_i=1)}{\hat{\pi}(L_i)} - \frac{Y_i 1(A_i=0)}{1-\hat{\pi}(L_i)}$$

where $\hat{\pi}(\cdot)$ is an estimator of $\pi\left(\cdot\right)$.

(c) let us denote with $\hat{\text{ATE}}_{\text{IPW},k}$ the IPW estimator of the ATE when the propensity score is known. By the triangular inequality for real numbers it follows

$$\begin{aligned}\mid \hat{\text{ATE}}_{\text{IPW},u} - E\left[Y^1 - Y^0\right] \mid &= \mid \hat{\text{ATE}}_{\text{IPW},u} - \hat{\text{ATE}}_{\text{IPW},k} + \hat{\text{ATE}}_{\text{IPW},k} - E\left[Y^1 - Y^0\right] \mid \\ &\leq \underbrace{\mid \hat{\text{ATE}}_{\text{IPW},u} - \hat{\text{ATE}}_{\text{IPW},k}}_{:=\, a_n} \mid + \mid \underbrace{\hat{\text{ATE}}_{\text{IPW},k} - E\left[Y^1 - Y^0\right] \mid}_{:=\, b_n} \text{ pointwise.}\end{aligned}$$

We proved in the previous exercise that $\hat{\text{ATE}}_{\text{IPW},k}$ converges in probability to $E\left[Y^1 - Y^0\right]$, i.e., $b_n = o_P(1)$. So what's left to prove is that $a_n = o_P(1)$ too. Indeed

$$\hat{\text{ATE}}_{\text{IPW},u} - \hat{\text{ATE}}_{\text{IPW},k} = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i Y_i\left(\pi(L_i) - \hat{\pi}(L_i)\right)}{\hat{\pi}(L_i)\pi(L_i)} - \frac{(1-A_i)Y_i\left(\hat{\pi}(L_i) - \pi(L_i)\right)}{\left(1-\hat{\pi}(L_i)\right)\left(1-\pi(L_i)\right)}.$$

Furthermore, suppose *(C_I)* there exists M such that $M \geq\mid Y \mid$ w.p.1 and *(C_{II})* $\exists b > 0 :$ $b \leq \pi(L) \leq 1-b$, w.p.1 then, by the triangle inequality

$$\begin{aligned}\mid \hat{\text{ATE}}_{\text{IPW},u} - \hat{\text{ATE}}_{\text{IPW},k} \mid &\leq \frac{1}{n}\sum_{i=1}^{n} \frac{M}{b}\left(\mid \frac{\pi(L) - \hat{\pi}(L)}{\hat{\pi}(L)} \mid + \mid \frac{\hat{\pi}(L) - \pi(L)}{1-\hat{\pi}(L)} \mid\right) \text{ w.p.1} \\ &\leq \underbrace{\frac{2M}{b} \underbrace{\sup \mid \frac{\pi(L) - \hat{\pi}(L)}{\inf(\hat{\pi}(L), 1-\hat{\pi}(L))} \mid}_{:=\, c_n}}_{:=c_n'} \text{ w.p.1}\end{aligned}$$

If *(C_{III})* $c_n = o_P(1)$, then $c_n' = o_P(1)$, since $\frac{2M}{b}$ is a constant. Now, $\forall \epsilon > 0, P(a_n \geq \epsilon) \leq P(\{a_n \geq \epsilon\} \cap \{c_n' \geq a_n\}) + P(\{a_n \geq \epsilon\} \cap \{c_n' < a_n\}) \leq P(c_n' \geq \epsilon)$, since $P(\{c_n' < a_n\}) = 0$ and $P(\{a_n \geq \epsilon\} \cap \{c_n' \geq a_n\}) \leq P(c_n' \geq \epsilon)$ and therefore $a_n = o_P(1)$. But $a_n + b_n = o_P(1)$, by the Slutsky's theorem. We conclude by the definition of convergence in probability.

(d) If an estimator is consistent, we know that it will converge in probability to the true parameter as $n \to \infty$. This is important in practice, because it ensures that as we get more data the estimator will converge closer to the truth. However, consistency[1] is insufficient to make inference about the uncertainty of our estimates, because we would also like to know the rate of convergence. In particular, it does not give us information about the (asymptotic) variance of the estimator from which we can construct confidence intervals.

Exercise 2 (Pivot intervals)**.** In this exercise we show that bootstrap pivot confidence intervals are valid. Let $\hat{\theta}_n = g(X_1, \dots, X_n)$, where X_i are i.i.d. random variables, be an estimator for a true parameter θ and let $(\hat{\theta}^*_{n,1}, \dots, \hat{\theta}^*_{n,B})$ be bootstrap replications of $\hat{\theta}_n$, where $B \to \infty$.

We denote $R_n = \hat{\theta}_n - \theta$ the pivot. The cumulative distribution function (CDF) of R_n is

$$H(r) = \mathbb{P}(R_n \leq r).$$

Consider the interval $C^*_n(c_1, c_2)$ where

$$c_1 = \hat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right), \text{ and } c_2 = \hat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right),$$

for $\alpha \in (0, 1)$.

(a) Show that $\mathbb{P}(c_1 \leq \theta \leq c_2) = 1 - \alpha$.

(b) Can we conclude that C^*_n is an exact $1 - \alpha$ confidence interval for θ?

(c) A fellow student says that the result is interesting, but it's not directly useful because we don't know the function $H(r)$. Suggest a bootstrap estimator $\hat{H}(r)$ of $H(r)$.

(d) Use the bootstrap estimator to argue that $C_n = (\hat{c}_1, \hat{c}_2)$ is an approximate $1-\alpha$ confidence interval of θ, where

$$\hat{c}_1 = 2\hat{\theta}_n - \hat{\theta}^*_{1-\alpha/2},$$
$$\hat{c}_2 = 2\hat{\theta}_n - \hat{\theta}^*_{\alpha/2},$$

and $\hat{\theta}^*_\alpha$ is the α sample quantile of $(\hat{\theta}^*_{n,1}, \dots, \hat{\theta}^*_{n,B})$.

(e) Explain how to find Bootstrap pivot confidence interval for the parameter of the logistic model for the effect of pesticide in the previous exercises sheet.

Solution:

(a)

$$\begin{aligned}\mathbb{P}(c_1 \leq \theta \leq c_2) &= \mathbb{P}(\hat{\theta}_n - c_2 \leq \hat{\theta}_n - \theta \leq \hat{\theta}_n - c_1)\\ &= \mathbb{P}(\hat{\theta}_n - c_2 \leq R_n \leq \hat{\theta}_n - c_1)\\ &= H(\hat{\theta}_n - c_1) - H(\hat{\theta}_n - c_2).\end{aligned}$$

From the definition of c_1 and c_2

$$\begin{aligned}&= H\left(H^{-1}\left(1 - \frac{\alpha}{2}\right)\right) - H\left(H^{-1}\left(\frac{\alpha}{2}\right)\right)\\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2}.\\ &= 1 - \alpha\end{aligned}$$

[1] In this context with consistency we refer to asymptotic consistency, which is a property of an estimator. This is totally different from the hypothesis of causal consistency (e.g., $A = a \implies Y^a = Y$), which is a causal assumption and does not involve the definition of any estimator.

(b) Because $\mathbb{P}(c_1 \leq \theta \leq c_2) = 1 - \alpha$, C_n^* is an exact $1 - \alpha$ confidence interval for θ.

(c) Although H is unknown, we can form a bootstrap estimate $\hat{H}$ as

$$\hat{H}(r) = \frac{1}{B} \sum_{b=1}^{B} I(R_{n,b}^* \leq r),$$

where $R_{n,b}^* = \hat{\theta}_{n,b}^* - \hat{\theta}_n$.

(d) Let r_β^* denote the β sample quantile of $(R_{n,1}^*, \ldots, R_{n,B}^*)$ and let θ_β^* denote the β sample quantile of $(\theta_{n,1}^*, \ldots, \theta_{n,B}^*)$. Note that $r_\beta = \theta_\beta^* - \hat{\theta}_n$. It follows that an approximate $1 - \alpha$ confidence interval is $C_n = (\hat{c}_1, \hat{c}_2)$ where

$$\begin{aligned} \hat{c}_1 &= \hat{\theta}_n - \hat{H}^{-1}\left(1 - \frac{\alpha}{2}\right) = \hat{\theta}_n - r_{1-\alpha/2}^* = 2\hat{\theta} - \theta_{1-\alpha/2}^* \\ \hat{c}_2 &= \hat{\theta}_n - \hat{H}^{-1}\left(\frac{\alpha}{2}\right) \qquad = \hat{\theta}_n - r_{\alpha/2}^* = 2\hat{\theta} - \theta_{\alpha/2}^*. \end{aligned}$$

(e) Suppose $S = \{x_i, y_i\}_{i=1}^n$ are the samples from the logistic model. We draw B set of Bootstrap samples of size n from S. Call the resulting Bootstrap samples $S_j^* = \{x_{j,i}^*, y_{j,i}^*\}, j = 1, \ldots, B$. Now for each $j = 1, \ldots, B$, find the maximum likelihood estimate $\beta_j^* = (\beta_{0,j}^*, \beta_{1,j}^*)$. Denote by q_α^* and $q_{1-\alpha}^*$ the α and $1 - \alpha$ sample quantiles of $(\beta_1^*, \ldots, \beta_B^*)$ respectively. The resulting Bootstrap confidence interval for β is $(q_\alpha^*, q_{1-\alpha}^*)$.

Exercise 3 (Sandwich estimator of the variance of logistic model)**.** Consider again the logistic model in the previous exercise sheet, and assume $\beta_0 = 0$. In other words, consider the following logistic model:

$$\mathbb{E}(Z_i) = g^{-1}(x_i\beta), \quad g(\mu) = \log(\frac{\mu}{1-\mu}),$$

where Z_i are independent binary random variables and $x_i > 0$.

In this exercise we use M-estimation theory to derive sandwich estimator of the variance for the above logistic model where the M-estimator is the MLE.

(a) Write down the form of $M(z, \beta)$, and let $\hat{\beta}$ be the resulting M-estimator (MLE in this case).

(b) Use $\hat{\beta}$ to derive an empirical estimator $\hat{C}$ for $C(\beta) = \mathbb{E}[-\dot{M}(Z_i, \beta)]$.

(c) Use $\hat{\beta}$ to derive an empirical estimator $\hat{B}$ for $B(\beta) = \mathbb{E}[M(Z_i, \beta)^2]$.

(d) Derive the sandwich estimator of the variance of $\hat{\theta}$ by:

$$\hat{\Sigma} = \hat{C}^{-1}\hat{B}\hat{C}^{-1}/n.$$

Solution:

(a) Define

$$M(X, Y; \beta) : \ \beta \longmapsto X\left(Y - \frac{e^{\beta X}}{1 + e^{\beta X}}\right), \ \beta \in \mathbb{B} \subset \mathbb{R}.$$

When we compute the MLE estimator of a logistic model we implicitly consider it conditional on $(x_1, \dots, x_n)$. Hence, the solution $\hat{\beta}$ of

$$\sum_{i=1}^{n} x_i \left(Y_i - \frac{e^{\beta x_i}}{1 + e^{\beta x_i}} \right) = 0$$

is the same solution of

$$\sum_{i=1}^{n} M(x_i, Y_i; \hat{\beta}) = 0.$$

(b) In the exercise about logistic regression model, we derived the second-derivative of log-likelihood function of this model (derivative of M with respect to β in this case) . Thus we use the corresponding empirical function to get

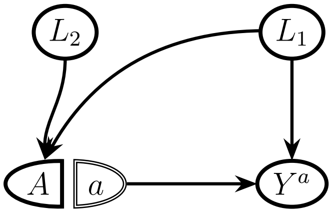
$$\hat{C} = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \frac{e^{\hat{\beta} x_i}}{(1 + e^{\hat{\beta} x_i})^2}.$$

(c) We can estimate B by

$$\hat{B} = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \left(Y_i - \frac{e^{\hat{\beta} x_i}}{1 + e^{\hat{\beta} x_i}} \right)^2 .$$

(d) Derive the estimator using the given formula in the question.

Exercise 4 (Predicting the propensity scores)**.** (Based on [1], p. 191) Consider a scenario under which L_1 is a common cause of treatment A and outcome Y, whereas L_2 only causes the outcome via the treatment (as illustrated by the SWIG below). This could for example describe a situation where $A \in \{0, 1\}$ denotes two possible treatments and $L_2 \in \{0, 1\}$ indicates the hospital at which an individual is treated, where hospital 1 provides treatment $A = 1$ more often than hospital $A = 0$.



Suppose A, L_1, L_2, Y, Y^a are drawn from the following data-generating mechanism:

$$\begin{aligned}
L_1, L_2 &\sim Ber\left(p = \frac{1}{2}\right) \\
A &\sim Ber\left(p = \text{logit}^{-1}(-4 + L_1 + 8L_2)\right) \\
Y &\sim Ber\left(p = \text{logit}^{-1}(-2L_1 + 2A)\right) \\
Y^{a=1} &\sim Ber\left(p = \text{logit}^{-1}(-2L_1 + 2 \cdot 1)\right) \\
Y^{a=0} &\sim Ber\left(p = \text{logit}^{-1}(-2L_1 + 2 \cdot 0)\right)
\end{aligned}$$

subject to the constraint

$$Y = Y^{a=1} I(A = 1) + Y^{a=0} I(A = 0) \ .$$

(a) Simulate[2] $A, L_1, L_2, Y, Y^{a=0}, Y^{a=1}$ for a population of 1 000 individuals using R.

(b) Plot the CDF of the weights $W_{L_1} = \frac{1}{\pi(A|L_1)}$ and $W_{L_1,L_2} = \frac{1}{\pi(A|L_1,L_2)}$.

(c) Because the causal model satisfies exchangeability conditions $Y^a \perp\!\!\!\perp A \mid L_1$ and $Y^a \perp\!\!\!\perp A \mid L_1, L_2$, we may use either $\pi(A \mid L_1)$ or $\pi(A \mid L_1, L_2)$ to identify $E[Y^a]$ from the observed data using inverse probability weighting. Compute the mean and variance of $\hat{\mu}_{IPW}$ for $\pi(A \mid L_1)$ and $\pi(A \mid L_1, L_2)$ by simulating 5 000 instances of the above population, using logistic models to estimate the propensity scores. Deduce that $\hat{\mu}_{IPW}$ has larger variance when we use L_2 to estimate the propensity scores, even though we might get more accurate predictions by including L_2 in the conditioning set.

Solution: Generating data:

```
library(ipw)
n <- 1000
set.seed(259)
# Generating data (Y.0 and Y.1 are counterfactuals)
genData<- function(n){
  L1 <- rbinom(n, size=1, prob=.5)
  L2 <- rbinom(n, size=1, prob=.5)
  A <- rbinom(n, size =1, prob=plogis(-4+L1+8*L2))
  Y.0 <- rbinom(n, size=1, prob= plogis(-2 +  L1 + 2*0 ))
  Y.1 <- rbinom(n, size=1, prob= plogis(-2 +  L1 + 2*1 ))
  Y<- as.numeric(A==0)*Y.0 + as.numeric(A==1)*Y.1
  wt.L1 <-ipwpoint(A, family='binomial', link='logit',
                   denominator = ~L1, data=ObsData)$ipw.weights
  wt.L12 <- ipwpoint(A, family='binomial', link='logit',
                     denominator = ~L1+L2, data=ObsData)$ipw.weights
  data.frame(L1,L2,A,Y,Y.0,Y.1,wt.L1,wt.L12) }

ObsData<- genData(n)
write.csv(ObsData, file="exercise_sheets_week_by_week/R/prediction_propensity.csv", row.
```

Here, we have used the function `ipwpoint()` to estimate the IPW weights (the estimates are identical to those we would obtain using the procedure outlined in the solutions of Exercise 3 from Exercise Sheet 7). We plot the CDF of the weights (displayed in Fig. 1 and Fig. 2):

```
library(latex2exp)
# Plotting the CDF of the IPW weights
plot(ecdf(ObsData$wt.L1), main=TeX('$W = \\frac{1}{\\pi(A| L_1)}$'),
     ylab=TeX('$P(W\\leq w)$'), xlab=TeX('$w$'))
```

and

[2] The implementation in R proceeds analogously to Exercise 3 of Exercise Sheet 7.

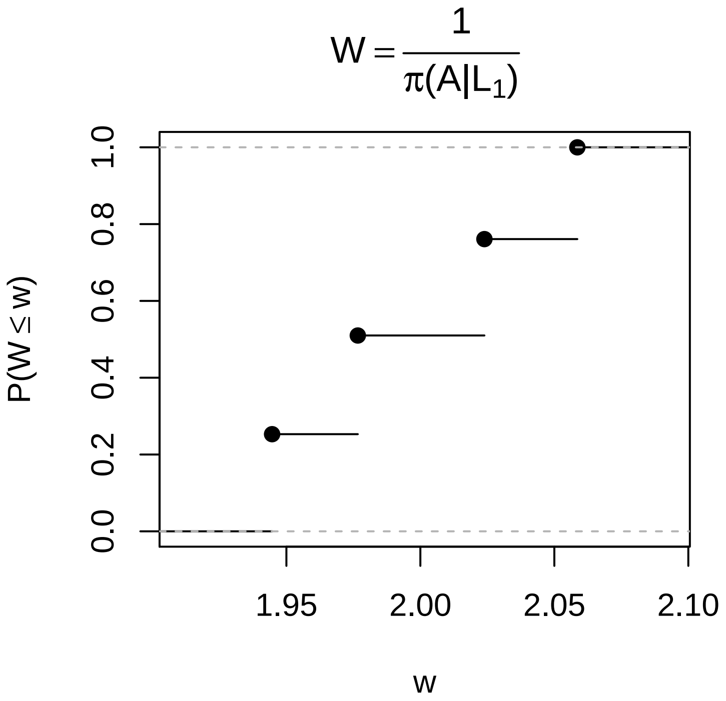


FIGURE 1. CDF of IPW weights

```r
# Plotting the CDF of the IPW weights
plot(ecdf(ObsData$wt.L12), main=TeX('$W = \\frac{1}{\\pi(A| L_1,L_2)}$'),
     ylab=TeX('$P(W\\leq w)$'), xlab=TeX('$w$'))
```

The mean and variance of the estimators can be estimated using the following code:

```r
# number of iterations
set.seed(259)
n<-1000
R<- 5000
# matrix for estimates from IPW
estimates<- matrix(, nrow = R, ncol = 2)
for(r in 1:R){
  # Redraw the data
  NewData<- genData(n)

  # IPW estimators for the two models for \pi
  IPW.L1<- mean( NewData$wt.L1*as.numeric(NewData$A==1)*NewData$Y) -
    +  mean( NewData$wt.L1*as.numeric(NewData$A==0)*NewData$Y)
  IPW.L12<- mean( NewData$wt.L12*as.numeric(NewData$A==1)*NewData$Y) -
    +  mean( NewData$wt.L12*as.numeric(NewData$A==0)*NewData$Y)
```

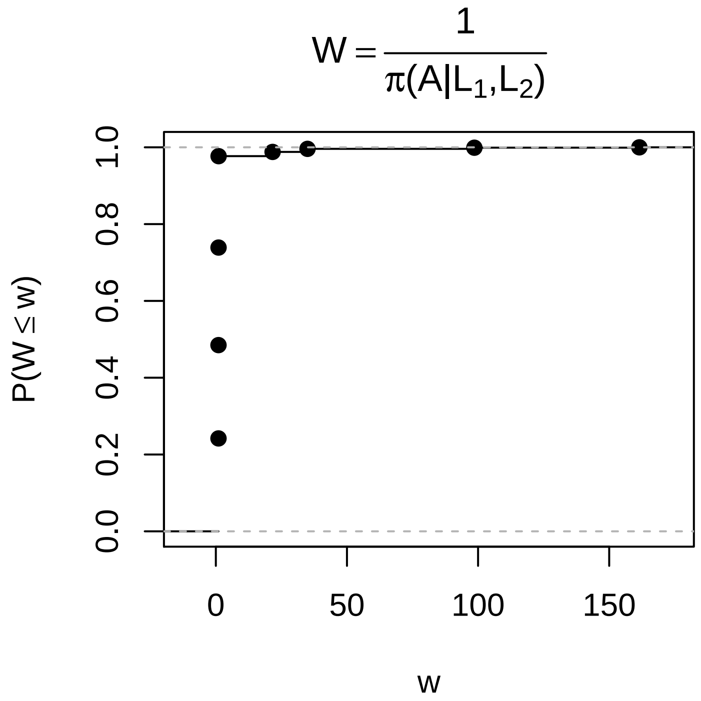


FIGURE 2. CDF of IPW weights

```
  estimates[r,]<-c(IPW.L1, IPW.L12)
}

# Average value of the estimates over R repetitions
colMeans(estimates)

# Variance of the estimators
diag(var(estimates))
```

which yields the estimates shown in Table. 1.

TABLE 1. Estimated means and variances

	$1/\pi(A \mid L_1)$	$1/\pi(A \mid L_1, L_2)$
Mean	0.4415398	0.4350833
Variance	0.001417936	0.025774764

The average causal effect, $\mu = E[Y^{a=1} - Y^{a=0}] = 0.4353$, can be estimated by simulating a population with 10 000 individuals:

```
n<-10000
set.seed(259)
big.population<-genData(n)
```

```
mu<- mean(big.population$Y.1-big.population$Y.0)
mu
```

As expected, $\hat{\mu}_{IPW}$ is unbiased regardless of whether we use $\pi(A \mid L_1)$ or $\pi(A \mid L_1, L_2)$ to predict the treatment propensities. The fact that the variance of $\hat{\mu}_{IPW}$ is larger for the latter case is not surprising in light of the larger variance of the weights in the CDF plot. Therefore, it is not necessarily true that we obtain better estimates by using more covariates to predict the treatment propensity.

References

[1] Miguel Hernan and James M. Robins. *Causal Inference*. Chapman & Hall, 2018.