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?

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

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:

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

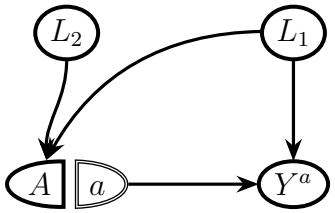
where Z_i are independent binary random variables and $x_i > 0$.

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:

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

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[1] $A, L_1, L_2, Y, Y^{a=0}, Y^{a=1}$ 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 $W_{L_1,L_2} = \frac{1}{\pi(A|L_1,L_2)}$.

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

(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.

References

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