Exercises for Causal Thinking (Math-352)

November 20, 2024

1 Exercise Sheet 10

Exercise 1 (*Challenging*: a comparison of variance)**.** (From Vock, Homework 2)

Consider two estimators for the average response: $\frac{1}{n}\sum_{i=1}^{n} Y_i^{a=1}$ and $\frac{1}{n}\sum_{i=1}^{n} \frac{A_i Y_i^{a=1}}{\pi(A_i|L_i)}$ and suppose $\pi(\cdot \mid \cdot)$ is known and the potential outcomes are known.[1]

(a) By assuming conditional exchangeability $Y_i^a \perp\!\!\!\perp A_i \mid L_i$, show that the first has lower variance than the second (that is, we pay some penalty for not observing all subjects in the data set being treated).

Hint: Show that the second estimator can be written as the first plus something else, and then demonstrate that the two terms are uncorrelated.

(b) Compute the difference in variance between the estimators in a if A is randomized with probability $P(A = 1) = \frac{1}{2}$ (i.e. $\pi = \frac{1}{2}$)

Exercise 2 (*Challenging*: Doubly Robustness)**.** Justify Theorem 1.

Theorem 1 (Doubly robust estimator of $\mathbb{E}(Y \mid L, A = a)$)**.** If either the propensity model $\pi(a \mid l; \gamma)$ or the outcome regression model $Q(l, a; \beta)$ is correctly specified, then

$$\mathbb{E}\left[\frac{I(A=a)Y}{\pi(a \mid L;\gamma)} + \left(1 - \frac{I(A=a)}{\pi(a \mid L;\gamma)}\right)Q(L,a;\beta)\right] = \mathbb{E}[\mathbb{E}(Y \mid L, A = a)].$$

Hints:

- Suppose first that $\pi(a \mid l; \gamma)$ is correctly specified, but the outcome model $Q(l, a; \beta)$ is mis-specified. Then, show that $\mathbb{E}\left\{\left(1 - \frac{I(A=a)}{\pi(a|L;\gamma)}\right)Q(L,a;\beta)\right\} = 0$ using the law of total expectation.

[1] The first estimator is an estimator that is typically impossible to compute because all the counterfactuals are not observed. However, in this exercise we have assumed that $Y_i^{a=1}$ is observed.

- Next, suppose that $\pi(a \mid l; \gamma)$ is mis-specified, but the outcome model $Q(l, a; \beta)$ is correctly specified. Then, show that $\mathbb{E}\left[\frac{I(A=a)}{\pi(a|L;\gamma)}\{Y - Q(L, a; \beta)\}\right] = 0$ using the law of total expectation.

Exercise 3 (More on doubly robustness)**.** Let A, L, Y be binary random variables. Consider a logistic regression model

$$\text{logit } E[Y \mid A, L] = \beta_1 + \beta_2 A + \beta_3 L \ .$$

The maximum likelihood estimator of $(\beta_1, \beta_2, \beta_3)$ is the solution of the score equations

$$\left(\frac{\partial \log \mathcal{L}(\beta)}{\partial \beta_1}, \frac{\partial \log \mathcal{L}(\beta)}{\partial \beta_2}, \frac{\partial \log \mathcal{L}(\beta)}{\partial \beta_3}\right)^T = 0 \ ,$$

where $\beta \equiv (\beta_1, \beta_2, \beta_3)^T$.

(a) Argue that the likelihood $\mathcal{L}(\beta)$ takes the form

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} p_i^{Y_i}(1 - p_i)^{1-Y_i} \tag{1}$$

where $p_i \equiv \text{expit}(\beta^T X_i) = \frac{e^{\beta^T X_i}}{1+e^{\beta^T X_i}}$ and $X_i \equiv (1, A_i, L_i)^T$.

(b) Argue that the score equations can be written as

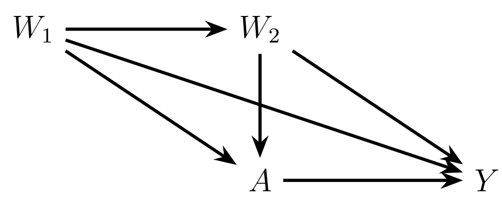
$$\sum_{i=1}^{n}\left(Y_i - \frac{e^{\beta^T X_i}}{1 + e^{\beta^T X_i}}\right) = 0 \ ,$$
$$\sum_{i=1}^{n} A_i\left(Y_i - \frac{e^{\beta^T X_i}}{1 + e^{\beta^T X_i}}\right) = 0 \ ,$$
$$\sum_{i=1}^{n} L_i\left(Y_i - \frac{e^{\beta^T X_i}}{1 + e^{\beta^T X_i}}\right) = 0 \ .$$

(c) Use the answer to part (b) to justify the following Lemma 1.

Lemma 1 (Consistent RCT estimator, even if mis-specified)**.** The estimator $\frac{1}{n}\sum_{i=1}^{n} \text{expit}(\hat{\beta}_1 + \hat{\beta}_2 a + \hat{\beta}_3 L_i)$ based on MLE estimates from a logistic regression model

$$\text{logit}\{Q(l, a; \beta)\} = \beta_1 + \beta_2 a + \beta_3 l.$$

unbiasedly estimates $Q(l, a)$ if A is randomly assigned, even if the logistic regression model is mis-specified.



Exercise 4 (Exploring the IPW estimator)**.** (Based on Lab 4 of Maya L. Petersen and Laura B. Balzer)

In this exercise we will implement the IPW and Hajek (or stabilized IP) estimators numerically in R in order to explore their efficiency in cases with near violations of positivity. Consider treatment A and outcome Y with baseline covariates W_1, W_2 in the dataset `stabilized_weights.csv`, and suppose these satisfy the causal model below: The data was generated by drawing $n = 5000$ i.i.d. samples from the distributions

$$\begin{aligned} W_1, W_2 &\sim Ber\left(p = \frac{1}{2}\right) \\ A &\sim Ber\left(p = \text{logit}^{-1}(-1.3 - 3W_1 + 3W_2)\right) \\ Y &\sim Ber\left(p = \text{logit}^{-1}(-2 - 2W_1 + 3W_2 + 3A + 2AW_2)\right) \\ Y^{a=1} &\sim Ber\left(p = \text{logit}^{-1}(-2 - 2W_1 + 3W_2 + 3 \cdot 1 + 2 \cdot 1 \cdot W_2)\right) \\ Y^{a=0} &\sim Ber\left(p = \text{logit}^{-1}(-2 - 2W_1 + 3W_2 + 3 \cdot 0 + 2 \cdot 0 \cdot W_2)\right) , \end{aligned}$$

subject to the constraint

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

The true effect is given by $E[Y^{a=1} - Y^{a=0}] \approx 0.26$ (computed by evaluating $\frac{1}{n'} \sum_{i=1}^{n'} (Y_i^1 - Y_i^0)$ in a larger realization of the data with $n' = 100000$) .

(a) Import the dataset `stabilized_weights.csv` into R and use the `glm` command to perform the following logistic regression for the treatment mechanism $\pi(A \mid L)$:

$$\text{logit } \pi(A \mid L; \gamma) = \gamma_0 + \gamma_1 W_1 + \gamma_2 W_2 \ .$$

Plot the empirical cumulative distribution function of the IPW weights $\frac{1}{\pi(A_i \mid W_{1,i}, W_{2,i})}$ and use the weights to evaluate the IPW estimator

$$\hat{\mu}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_1 = a) Y_i}{\pi(A_i \mid W_{1,i}, W_{2,i}; \gamma)} \ .$$

(b) Compute $\hat{\mu}_{IPW}$ with truncated weights $\frac{I(\pi \leq 10)}{\pi} + 10 \cdot I(\pi > 10)$ instead of the weights $\frac{1}{\pi}$ in part (a).

(c) Evaluate the stabilized IPW estimator given by Eq. 2 using the weights as in part (a).

$$\hat{\mu}_{STIPW}(a) = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=a)Y_i}{\pi(A_i|L_i;\gamma)}}{\frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=a)}{\pi(A_i|L_i;\gamma)}}\ . \tag{2}$$

(d) Estimate the variance of the estimators in parts (a)-(d) by drawing $R = 5000$ different realizations of a population with $n = 5000$ i.i.d. individuals from the data generating mechanism outlined above.

References

Maya L. Petersen and Laura B. Balzer. Labs & Assignments. URL `https://www.ucbbiostat.com/labs`.

David M. Vock. PubH 7485 & 8485: Methods for Causal Inference (University of Minnesota School of Public Health). URL `https://sites.google.com/site/dmvock/courses-1/pubh-7485-8485-methods-for-causal-inference`.