EXERCISES FOR RANDOMIZATION AND CAUSATION (MATH-336)

Exercise Sheet 12

Exercise 1. In this exercise, you will study partial identification (bounds) of the average treatment effect. Suppose that Z, A, Y, U satisfy the single-world causal model corresponding to the graph below. Suppose that the measured variables $Z, A, Y \in \{0, 1\}$ are binary.

(a) Show that, without using Z, the average treatment effect of A and Y satisfies the following inequalities

$$-P(Y = 0, A = 1) - P(Y = 1, A = 0) \leq \mathbb{E}(Y^{a=1} - Y^{a=0}) \leq P(Y = 1, A = 1) + P(Y = 0, A = 0).$$

What is the difference between the upper and the lower bounds $(UB - LB)$?

(b) Suppose $A = 1$ if an individual elects to get the annual influenza vaccine and $A = 0$ otherwise. Let $Y^a = 1$ if an individual subsequently does develop flu-like symptoms when $A = a$, and $Y^a = 0$ otherwise. Suppose that the investigator is comfortable with assuming that each individual is more or as likely to develop flu-like symptoms if they are unvaccinated versus if they are vaccinated.[1]

(i) Formalize the investigator's assumption as a counterfactual inequality.

(ii) What is the upper bound on $\mathbb{E}(Y^{a=1} - Y^{a=0})$ under this assumption?

(iii) Can we derive a tighter lower bound without adding additional assumptions?

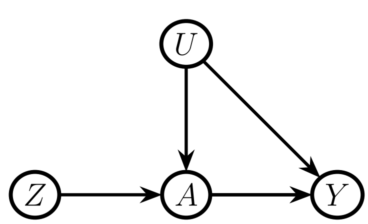
(c) Now you will show some famous bounds using the instrumental variable Z. Suppose that necessary consistency and positivity assumptions hold. Let $p(y, a \mid z)$ denote $P(Y = y, A = a \mid Z = z)$ and $p(y \mid z)$ denote $P(Y = y \mid Z = z)$. Show that

$$LB \leq \mathbb{E}(Y^{a=1} - Y^{a=0}) \leq UB,$$

where

$$\begin{aligned} LB = \max\{ & -p(0,1 \mid 0) - p(1,0 \mid 0), \\ & -p(0,1 \mid 1) - p(1,0 \mid 1), \\ & p(1 \mid 0) - p(1 \mid 1) - p(1,0 \mid 0) - p(0,1 \mid 1), \\ & p(1 \mid 1) - p(1 \mid 0) - p(1,0 \mid 1) - p(0,1 \mid 0)\}, \end{aligned}$$

[1] In this exercise we ignore interference, and suppose that individuals are iid and that positivity and consistency hold.



and

$$\begin{aligned}UB = \min\{&p(1,1 \mid 0) + p(0,0 \mid 0),\\ &p(1,1 \mid 1) + p(0,0 \mid 1),\\ &p(1 \mid 0) - p(1 \mid 1) + p(0,0 \mid 0) - p(1,1 \mid 1),\\ &p(1 \mid 1) - p(1 \mid 0) + p(0,0 \mid 1) + p(1,1 \mid 0)\},\end{aligned}$$

Conclude that

$$UB - LB \leq \min\{P(A=0 \mid Z=0) + P(A=1 \mid Z=1), P(A=0 \mid Z=1) + P(A=1 \mid Z=0)\} \leq 1.$$

and that $UB - LB = 1$ if and only if $A \perp\!\!\!\perp Z$.

Exercise 2 (Efficiency of linear adjustment)**.** (Inspired by [1]) Consider 3 different linear models defined by population least squares,

$$\beta^* = \arg\min_\beta \mathbb{E}\big[(Y - \beta_1 - \beta_2 A)^2\big]$$

$$\beta' = \arg\min_\beta \mathbb{E}\big[(Y - \beta_1 - \beta_2 A - \beta_3^T L)^2\big] \text{ (ANCOVA model)}$$

$$\beta^\dagger = \arg\min_\beta \mathbb{E}\big[(Y - \beta_1 - \beta_2 A - \beta_3^T L - \beta_4^T AL)^2\big]$$

Suppose (L, A, Y) are i.i.d., $A \perp\!\!\!\perp L$, $\mathbb{E}(L) = 0$.

(a) Show that[2] $\beta_1^* = \beta_1' = \beta_1^\dagger$ and $\beta_2^* = \beta_2' = \beta_2^\dagger$.

(b) A classical result from M-estimation theory implies that

$$\sqrt{n}(\widehat{\beta}_1^m - \beta_1) \xrightarrow{d} N(0, V^m),$$

where $m \in \{*, ', \dagger\}$, $\pi = P(A = a \mid L)$, $V^m = \frac{E[(A-\pi)^2 \epsilon_m^2]}{\pi^2(1-\pi)^2}$ and $\epsilon_{i*}, \epsilon_{i'}, \epsilon_{i\dagger}$ are the error terms in the regression estimators, for example,

$$\epsilon_{i,\dagger} = Y_i - (\beta_1^\dagger + \beta_2^\dagger A_i + \beta_3^{\dagger T} L_i + \beta_4^{\dagger T} A_i L_i)\ .$$

Use this result to show that

$$V^\dagger \leq \min\{V', V^*\}\ .$$

In other words, asymptotically it is more efficient to use covariates L in the model indicated by $\dagger$.[3]

References

[1] Qingyuan Zhao. Lecture Notes on Causal Inference. page 109.

[2] We have not said anything about the linear model being correctly specified. We have not given an argument why $\mathbb{E}(L) = 0$. However, we could center L_i by using $L_i - \frac{1}{n}\sum_{i=1}^n L_i$, which will give the same point estimates of the β's but $\beta^\dagger$ has larger variance.

[3] Careful consideration is required to decide whether or not it is more efficient to use L in a finite sample.