Info 6751. Fall 2022. Problem Set 7. Due on Canvas by 5pm on 17 Oct.

Intro to the problem set

This problem set uses the data file `pset7.csv`, which is a simulated setting with one confounder L,

$$L \longrightarrow A \longrightarrow Y$$
$$L \longrightarrow\!\!\!\!\!\!\;\curvearrowright\; Y$$

where data are generated by the following data generating process.

For $i = 1, \ldots, n$

$$L_i \sim \text{Bernoulli}(0.5) \qquad \text{binary confounder} \qquad (1)$$

$$\vec{\lambda}_i = \begin{cases} \left(\frac{2}{3} \quad \frac{1}{6} \quad \frac{1}{6}\right)^T & \text{if } A_i = 0 \\ \left(\frac{1}{6} \quad \frac{1}{6} \quad \frac{2}{3}\right)^T & \text{if } A_i = 1 \end{cases} \qquad \text{probability of treatment values } \{1,2,3\} \qquad (2)$$

$$A_i \sim \text{Multinomial}\left(\vec{\lambda}_i\right) \qquad \text{categorical treatment} \qquad (3)$$

$$\mu_i = L_i + A_i + L_i A_i \qquad \text{outcome mean} \qquad (4)$$

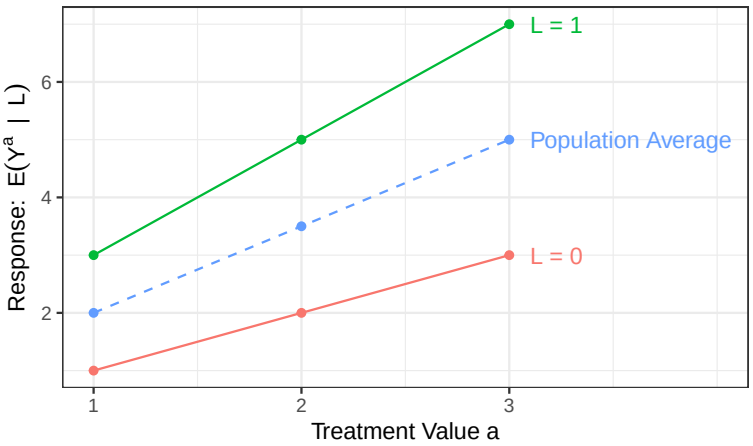
$$Y_i \sim \text{Normal}(\text{Mean} = \mu_i, \text{SD} = 5) \qquad \text{continuous outcome} \qquad (5)$$

Eq 2 is designed so that L is imbalanced across the treatment.

- Much higher probability mass on $A = 1$ when $L = 0$
- Much higher probability mass on $A = 3$ when $L - 1$

Eq 4 is designed so that L directly affects the outcome, and so that the response surface is interactive.

The figure below visualizes the response surface.



Although a lines are shown, the data are discrete with only treatment values $A \in \{1, 2, 3\}$.

You will submit

- A PDF with your answers
- A file with your code, or have code embedded within the PDF above

1 (25 points) Material covered Tuesday

Part 1 is about the **inverse probability weighting**.

1.1. (5 points) Nonparametrically estimate each unit's generalized propensity score: $\hat{\pi}_i = \hat{\mathrm{P}}(A = a_i \mid \vec{L} = \vec{\ell}_i)$. To facilitate grading, report the estimated propensity score for the first unit in the dataset.

1.2. (5 points) Create the inverse probability weight $\hat{w}_i = \frac{1}{\hat{\pi}_i}$ for each unit. To facilitate grading, report the estimated weight for the first unit in the dataset.

1.3. (10 points) For each treatment value $a = \{1, 2, 3\}$, estimate the population-average response $\mathrm{E}(Y^a)$ by inverse probability weighting. Report the three estimates.

1.4. (5 points) Summarize the population-average response curve estimate in a graph.

2 (25 points) Material covered Thursday

Part 2 is about the **marginal structural models**.

In Part (1) we estimated entirely by treatment modeling with no assumptions about the shape of the response surface. In the parametric g-formula of prior weeks, we assumed a functional form for the entire response surface $\mathrm{E}(Y \mid A, \vec{L})$. A marginal structural model is between these extremes: we make assumptions about the form of the population-average treatment response $\mathrm{E}(Y^a)$ marginalized over the population distribution of confounders. In this part, we will assume a linear marginal structural model,

$$\mathrm{E}(Y^a) = \alpha + \beta a \tag{6}$$

Note that the marginal structural model (Eq 6) is a model for the mean of potential outcomes rather than factual outcomes ($\mathrm{E}(Y^a)$ instead of $\mathrm{E}(Y \mid A = a)$), so it is not a standard OLS equation and must be estimated with inverse probability weighting.

2.1. (15 points) Estimate the marginal structural model using OLS weighted by the inverse probability of treatment weights from 1.2. Report the coefficient estimate $\hat{\beta}$.

2.2. (5 points) Produce a graph summarizing the population-average response curve as estimated by both methods (IPW and MSM-IPW).

2.3. (5 points) Comment on the differences between the two estimates. Why might we prefer one or the other?