Inverse Propensity Score Weighting

Tianyu Du

June 10, 2020

Let X, Y, Z denote the cause, the effect, and confounding variables. Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ denote domains of above-mentioned random variables. Let $f(Y)$ denote the random variable of interest. Then,

$$\begin{aligned}
\mathbb{E}[f(Y)|do(X=x)] &= \sum_{y\in\mathcal{Y}, z\in\mathcal{Z}} f(y)P(X=x, Y=y, Z=z|do(X=x)) && (1)\\
&= \sum_{y\in\mathcal{Y}, z\in\mathcal{Z}} f(y)P(Y=y|X=x, Z=z)P(Z=z) && (2)\\
&= \sum_{y\in\mathcal{Y}, z\in\mathcal{Z}} f(y)\frac{P(X=x, Y=y|Z=z)}{P(X=x|Z=z)}P(Z=z) && (3)\\
&= \sum_{y\in\mathcal{Y}, z\in\mathcal{Z}} f(y)\frac{P(X=x, Y=y, Z=z)}{P(X=x|Z=z)} && (4)
\end{aligned}$$

where $P(X=x|Z=z)$ is the **propensity score**, which denotes the probability of receiving treatment $X=x$ given characteristics $Z=z$.

Assume there is a finite dataset of size N, $(x_i, y_i, z_i)_{i=1}^N$, and we wish to infer the interventional distribution of $f(Y)$ using this dataset.

One method used for discrete variables is to simply count the occurrence of each (X, Y, Z) in the dataset.

$$\hat{P}(X=x, Y=y, Z=z) = \frac{\sum_{i=1}^N \mathbb{1}\{x_i=x\}\mathbb{1}\{y_i=y\}\mathbb{1}\{z_i=z\}}{N} \tag{5}$$

Therefore,

$$\begin{aligned}
\hat{\mathbb{E}}[f(Y)|do(X=x)] &= \sum_{y\in\mathcal{Y}, z\in\mathcal{Z}} \frac{f(y)}{P(X=x|Z=z)}\frac{\sum_{i=1}^N \mathbb{1}\{x_i=x\}\mathbb{1}\{y_i=y\}\mathbb{1}\{z_i=z\}}{N} && (6)\\
&= \frac{1}{N}\sum_{i=1}^N \frac{f(y_i)}{P(X=x_i|Z=z_i)} && (7)
\end{aligned}$$

The estimation (7) is the inverse-propensity-score weighted mean of $f(y_i)$.

Figure 1: The Causal Graph