Causal Inference - assignment 3

Amir Gavrieli

19.12.19

Question 1

Consider the DAG in Figure 1. For each of the following pairs of nodes, list: (1) all possible paths between them (2) what set of nodes is required (in a d-separation sense) to block all of their paths:
Note that I wrote in (1) paths that are also blocked by colliders, that is if the collider is not observed in the path then the information doesn't "flow" through that path.

0.1 $W \rightarrow S$

0.1.1 List of all possible paths between them:

- $W \rightarrow Y \rightarrow R \rightarrow S$
- $W \rightarrow Y \rightarrow R \leftarrow Z \rightarrow S$
- $W \rightarrow Y \leftarrow X \rightarrow Z \rightarrow S$
- $W \rightarrow Y \leftarrow X \rightarrow Z \rightarrow R \rightarrow S$

0.1.2 What set of nodes is required to block all of their paths:

We will call the set of nodes $\mathbf{Z}$ (not to be confused with the z in the graph). Let us consider what are the constraints on $\mathbf{Z}$ for each path respectively:

- $\mathbf{Z} \cap \{Y, R\} \neq \varnothing$
- $\mathbf{Z} \cap \{Y, R\} \neq \varnothing \vee R \notin \mathbf{Z}$
- $\mathbf{Z} \cap \{X, Z\} \neq \varnothing \vee Y \notin \mathbf{Z}$
- $\mathbf{Z} \cap \{X, Z, R\} \neq \varnothing \vee Y \notin \mathbf{Z}$

From the intersection of these constraints we get: $\mathbf{Z} = \{Y, X\}$.

0.2 $X \rightarrow T$

0.2.1 List of all possible paths between them:

- $X \rightarrow Y \rightarrow R \rightarrow T$
- $X \rightarrow Z \rightarrow R \rightarrow T$
- $X \rightarrow Z \rightarrow S \leftarrow R \rightarrow T$

0.2.2 What set of nodes is required to block all of their paths:

We will call the set of nodes $\mathbf{Z}$ (not to be confused with the z in the graph). Let us consider what are the constraints on $\mathbf{Z}$ for each path respectively:

- $\mathbf{Z} \cap \{Y, R\} \neq \varnothing$
- $\mathbf{Z} \cap \{Z, R\} \neq \varnothing$
- $\mathbf{Z} \cap \{Z, R\} \neq \varnothing \vee S \notin \mathbf{Z}$

From the intersection of these constraints we get: $\mathbf{Z} = \{R\}$. Note that we could have also chosen $\mathbf{Z} = \{Z, Y\}$, $\mathbf{Z} = \{R\}$ is simply the smallest set.

0.3 $Y \rightarrow H$

0.3.1 List of all possible paths between them:

- $Y \leftarrow X \rightarrow Z \leftarrow H$
- $Y \rightarrow R \leftarrow Z \leftarrow H$
- $Y \rightarrow R \rightarrow S \leftarrow Z \leftarrow H$

0.3.2 What set of nodes is required to block all of their paths:

Similar to the sections above the constraints on $\mathbf{Z}$ are as follows:

- $\mathbf{Z} \cap \{X\} \neq \varnothing \vee Z \notin \mathbf{Z}$
- $\mathbf{Z} \cap \{Z\} \neq \varnothing \vee R \notin \mathbf{Z}$
- $\mathbf{Z} \cap \{R, Z\} \neq \varnothing \vee S \notin \mathbf{Z}$

The easiest $\mathbf{Z}$ to consider is $\varnothing$ since every path has a collider.

Question 2

Consider the causal graph in Figure 2.

(i) List all the sets of variables that satisfy the backdoor criterion to determine the causal effect of T on Y.

The first part of the back-door criterion is that no decendants of T can be in any $\mathbf{Z}$, hence we know that for every set $\mathbf{Z}$, $W \notin \mathbf{Z}$. To satisfy the second backdoor criterion we wish to block all paths between T and Y that end with an arrow into T, these paths are:

- $T \leftarrow Z \rightarrow Y$
- $T \leftarrow Z \leftarrow C \rightarrow D \rightarrow Y$
- $T \leftarrow A \leftarrow B \rightarrow Z \rightarrow Y$
- $T \leftarrow A \leftarrow B \rightarrow Z \leftarrow C \rightarrow D \rightarrow Y$

The criteria for d-separating each path are as follows:

- $Z \in \mathbf{Z}$
- $\mathbf{Z} \cap \{Z, C, D\} \neq \varnothing$
- $\mathbf{Z} \cap \{A, B, Z\} \neq \varnothing$
- $Z \notin \mathbf{Z} \vee \mathbf{Z} \cap \{A, B, C, D\} \neq \varnothing$

Thus the sets are as follows:

- $\{Z, A\}$
- $\{Z, B\}$
- $\{Z, C\}$
- $\{Z, D\}$
- $\{Z, A, B\}$
- $\{Z, A, C\}$
- $\{Z, A, D\}$
- $\{Z, B, C\}$
- $\{Z, B, D\}$
- $\{Z, C, D\}$
- $\{Z, A, B, C\}$

- $\{Z, A, B, D\}$
- $\{Z, A, C, D\}$
- $\{Z, B, C, D\}$
- $\{Z, A, B, C, D\}$

(ii) List all of the minimal sets of variables that satisfy the backdoor criterion to determine the causal effect of the T on Y (i.e., any set of variables such that, if you remove any one of the variables from the set, it would no longer meet the criterion).

Looking at the previous section the minimal sets are:

- $\{Z, A\}$
- $\{Z, B\}$
- $\{Z, C\}$
- $\{Z, D\}$

(iii) Give a minimal set of variables that need to be measured in order to identify the effect of D on Y

The first part of the back-door criterion is that no decendants of D can be in any $\mathbf{Z}$, but since Y is the only decendant of D we don't need to worry about it. To satisfy the second backdoor criterion we wish to block all paths between D and Y that end with an arrow into D. Since all these paths go through C, we can simply define
$\mathbf{Z} = \{C\}$, note that a smaller set would have been the empty set, but that would only apply if we had no path between D and Y that end with an arrow into D, but we do. For example: $D \leftarrow C \rightarrow Z \rightarrow Y$

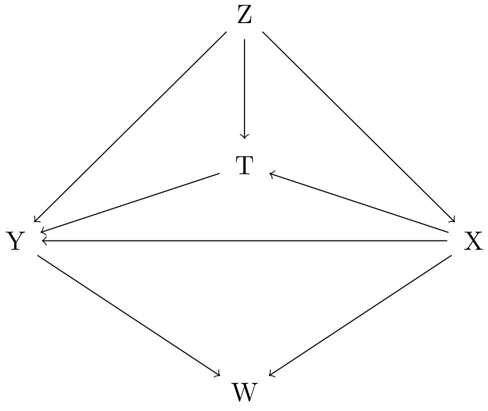

Figure 1: Causal graph describing the below experiment

Question 3

Consider the following scenario, which is loose adaptation of the LaLonde experiment (as described in the tutorial). You receive data collected from an experiments run in a 3 years (2024-2026) span in Gotham City, USA, trying to estimate the effect of job training on real income a year following the training. In the data (See Table 1), you have binary indicators for: prior education (Z), whether the annual real income in 2024 is higher than $50K\$$ (X), had job training (T), whether the annual real income in 2026 is higher than $100K\$$ (Y), and whether the citizen bought a house in 2026 (W). We know the following:

1. The income in 2024 depends solely on the prior education.
2. A person is selected to the job training program based on her prior education, and income in 2024.
3. The income in 2026 depends on the job training, prior education and income in 2024.
4. Whether a citizen purchased a house is directly based on her income in 2024 and 2026.

Your task is as follows:

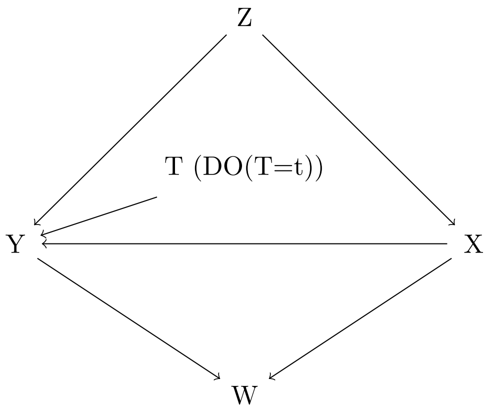

Figure 2: Causal graph that enables us to measure ATE.

0.4 Draw the causal graph that describe the above experiment.

0.5 Calculate the ATE of the experiment (derive the necessary probabilities from Table 1)

We want to understand the income in 2026 (Y) given the treatment (T). First we perform the DO operator on T (removing all arrows into T).

Formula development

Let us note a couple of things:

1. We recall that $ATE = \mathbb{E}[y|do(T=1)] - [y|do(T=0]$.
2. Since $Y \in \{0,1\}$, $\mathbb{E}[y|do(T=t)] = P(y=1|do(T=t))$.
3. From (1),(2): $ATE = P(y|do(T=1)) - P(y|do(T=0))$.
4. We know that for a set of variables $\mathbf{Z}$ that satisfies the back-door criterion for T, Y:
$P(y|do(T=t)) = \sum_{z \in \mathbf{Z}} P(y|T=t, Z=z) \cdot P(Z=z)$
5. To satisfy the second backdoor criterion we wish to block all paths between T and Y that end with an arrow into T, these paths are:
    - $T \leftarrow Z \rightarrow Y$
    - $T \leftarrow X \rightarrow Y$
    - $T \leftarrow Z \rightarrow X \rightarrow Y$

   The criteria for d-separating each path are as follows:

- $Z \in \mathbf{Z}$
- $X \in \mathbf{Z}$
- $\mathbf{Z} \cap \{Z, X\} \neq \emptyset$

So we get that $\mathbf{Z} = \{X, Z\}$, these corresponds to in some sense to what we did to the graph.

Now Let us Calculate $\mathbb{E}[y|do(T = t)]$:

$$\begin{aligned}
&\mathbb{E}[y|do(T = t)] = P(y|do(T = t)) \\
&= \sum_{z \in \mathbf{Z}} P(y|T = t, Z = z) \cdot P(Z = z) \\
&= \sum_{z \in \{0,1\}} \sum_{x \in \{0,1\}} P(y|T = t, X = x, Z = z) \cdot P(X = x, Z = z) \\
&= \sum_{z \in \{0,1\}} \sum_{x \in \{0,1\}} P(y|T = t, X = x, Z = z) \cdot P(X = x|Z = z) \cdot P(Z = z) \\
&= P(y|T = t, X = 0, Z = 0) \cdot P(X = 0|Z = 0) \cdot P(Z = 0) \\
&+ P(y|T = t, X = 1, Z = 0) \cdot P(X = 1|Z = 0) \cdot P(Z = 0) \\
&+ P(y|T = t, X = 1, Z = 1) \cdot P(X = 1|Z = 1) \cdot P(Z = 1) \\
&+ P(y|T = t, X = 0, Z = 1) \cdot P(X = 0|Z = 1) \cdot P(Z = 1)
\end{aligned}$$

Since we don't have the actual probabilities we will estimate them from the data, we are in fact using estimates of the probabilities, $\hat{P}$, but for convenience we shall continue to use P.

- $\hat{P}(Z = 1) = \frac{12}{20} = \frac{3}{5}$, $\hat{P}(Z = 0) = \frac{8}{20} = \frac{2}{5}$
- $\hat{P}(X = 1|Z = 0) = \frac{2}{8} = \frac{1}{4}$, $\hat{P}(X = 0|Z = 0) = \frac{6}{8} = \frac{3}{4}$
- $\hat{P}(X = 1|Z = 1) = \frac{5}{12}$, $\hat{P}(X = 0|Z = 1) = \frac{7}{12}$
- $\hat{P}(Y = 1|T = 1, Z = 0, X = 0) = \frac{1}{2}$
- $\hat{P}(Y = 1|T = 1, Z = 1, X = 0) = \frac{1}{2}$
- $\hat{P}(Y = 1|T = 1, Z = 0, X = 1) = 1$
- $\hat{P}(Y = 1|T = 1, Z = 1, X = 1) = 1$
- $\hat{P}(Y = 1|T = 0, Z = 0, X = 0) = 0$
- $\hat{P}(Y = 1|T = 0, Z = 1, X = 0) = \frac{2}{3}$
- $\hat{P}(Y = 1|T = 0, Z = 0, X = 1) = 1$
- $\hat{P}(Y = 1|T = 0, Z = 1, X = 1) = \frac{2}{3}$

Let us calculate $Do(T = 1)$:

$$\begin{aligned}
&\mathbb{E}[y|do(T = 1)] = P(y|do(T = 1))\\
&= P(y = 1|T = 1, X = 0, Z = 0) \cdot P(X = 0|Z = 0) \cdot P(Z = 0)\\
&+ P(y = 1|T = 1, X = 1, Z = 0) \cdot P(X = 1|Z = 0) \cdot P(Z = 0)\\
&+ P(y = 1|T = 1, X = 1, Z = 1) \cdot P(X = 1|Z = 1) \cdot P(Z = 1)\\
&+ P(y = 1|T = 1, X = 0, Z = 1) \cdot P(X = 0|Z = 1) \cdot P(Z = 1)\\
&= \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{2}{5}\\
&+ 1 \cdot \frac{1}{4} \cdot \frac{2}{5}\\
&+ 1 \cdot \frac{5}{12} \cdot \frac{3}{5}\\
&+ \frac{1}{2} \cdot \frac{7}{12} \cdot \frac{3}{5}\\
&= \frac{27}{40}
\end{aligned}$$

Let us calculate $Do(T = 0)$:

$$\begin{aligned}
&\mathbb{E}[y|do(T = 1)] = P(y|do(T = 1))\\
&= P(y = 1|T = 0, X = 0, Z = 0) \cdot P(X = 0|Z = 0) \cdot P(Z = 0)\\
&+ P(y = 1|T = 0, X = 1, Z = 0) \cdot P(X = 1|Z = 0) \cdot P(Z = 0)\\
&+ P(y = 1|T = 0, X = 1, Z = 1) \cdot P(X = 1|Z = 1) \cdot P(Z = 1)\\
&+ P(y = 1|T = 0, X = 0, Z = 1) \cdot P(X = 0|Z = 1) \cdot P(Z = 1)\\
&= 0 \cdot \frac{3}{4} \cdot \frac{2}{5}\\
&+ 1 \cdot \frac{1}{4} \cdot \frac{2}{5}\\
&+ \frac{2}{3} \cdot \frac{5}{12} \cdot \frac{3}{5}\\
&+ \frac{2}{3} \cdot \frac{7}{12} \cdot \frac{3}{5}\\
&= \frac{1}{2}
\end{aligned}$$

We get that the $A\hat{T}E$ is:

$$A\hat{T}E = \frac{27}{40} - \frac{1}{2} = \frac{7}{40} = 0.175$$