Contents

Bayesian Networks

1 Introduction to Bayesian networks

Identification of causality in Bayesian networks draws on Pearl's *do*-calculus. *do*-calculus stems from the idea that causality can be inferred by intervention (say, to explore counterfactuals) combined with evidence rather than from evidence alone. Principal ingredients include Bayes sum and product rules, causal graphs, *d*-separation, back-door adjustment, and front-door adjustment. These ideas are discussed and illustrated below.

One of the many challenges associated with causal inference involves framing the causal connections. Pearl [2010] argues this is a strength of Bayesian networks. In his November 1996 public lecture for the UCLA faculty research leadership program reproduced in his book [2010, p. 425] Pearl offers the following encouragement, "There is no need to panic when someone tells us: 'you did not take this or that factor into account.' On the contrary, the graph welcomes such new ideas, because it is so easy to add factors and measurements into the model. Simple tests are now available that permit an investigator to merely glance at the graph and decide if one can compute the effect of one variable on another."

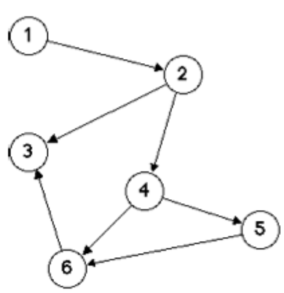
1.1 Causal graphs

Causal graphs are graphical representations or encodings of causal relations (the thought experiment in causal inference). A path $X \rightarrow Y$ implies X causes Y. A directed acyclic graph (DAG) is the simplest variety as causality is defined for each node and there are no cycles or feedback loops.[1] More generally, undi-

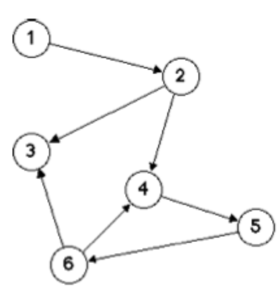
[1] A simple, informal algorithm for distinguishing an acyclic graph from a cyclic graph follows:

1. If the graph has no nodes, it is acyclic.
2. If the graph has no leafs, it is cyclic. A leaf is a node with no descendants (targets or arcs going out).
3. Choose a leaf, delete it and all arcs coming into the leaf to form a new graph.
4. Return to 1 and repeat.

If we eliminate all nodes, the graph is acyclic. Alternatively, if we eliminate all leafs and the graph is not empty, the graph is cyclic.



an acyclic graph



a cyclic graph

rected paths indicate the causal direction is unknown while dashed arcs or paths (possibly, bidirected) indicate connections between observables and unobserved variables.

Subgraphs are formed from DAGs when arcs are dropped, say by interventions, as depicted in figure 1.1.

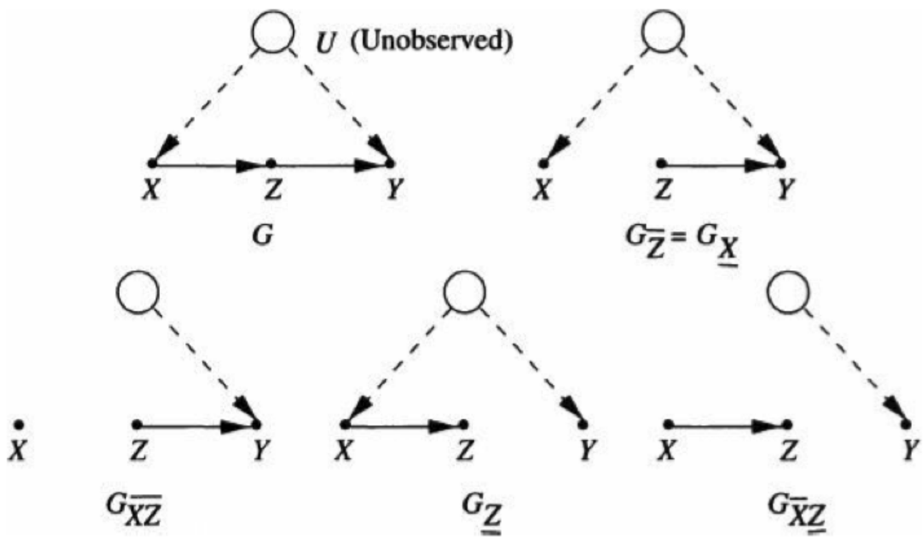


Figure 1.1: DAG G and its subgraphs

1.2 do-calculus and causal effects

Assignment $X = x$ (not the same thing as conditioning on $X = x$)[2] is denoted $do(x)$ and such interventions nonparametrically define the causal effect of X on Y, $\Pr(Y \mid do(x))$. Paths from an ancestor are effectively removed from the graph when a descendant variable is assigned a value by $do(\cdot)$ (intervention) because any effect of a parent on a child is negated.

1.3 *d*-separation

d-separation in a graph represents probabilistic independence or conditional independence.

[2] $do(x)$ is a thought experiment or action while conditioning on $X = x$ is evidentiary or observation.

Formally, a path p is d-separated (blocked) by a set of nodes Z (including the null set $\emptyset$) if and only if

1. p contains a chain $i \to m \to j$ or fork $i \leftarrow m \to j$ such that the middle node m is in Z, or

2. p contains an inverted fork (collider) $i \to m \leftarrow j$ such that the middle node m is not in Z and no descendant of m is in Z.

A set Z d-separates X and Y if and only if Z blocks every path from X to Y.

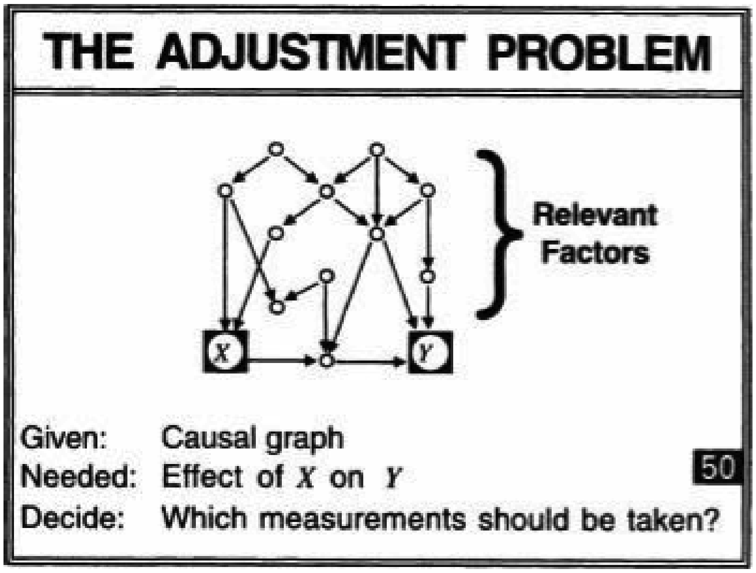
Theorem 1 (d-separation and conditional independence) *If sets X and Y are d-separated by Z in a DAG G, then X is independent of Y conditional on Z in every distribution consistent with G. Conversely, if X and Y are not d-separated by Z in a DAG G, then X and Y are dependent conditional on Z in at least one distribution consistent with G.*

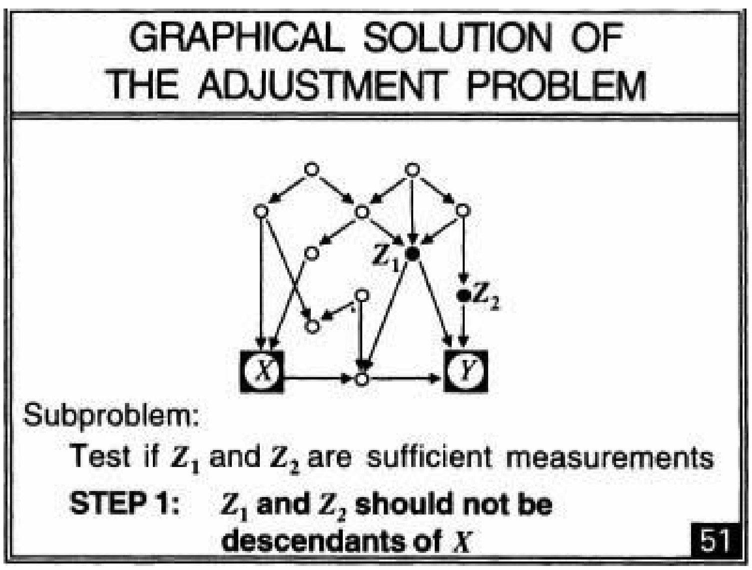
The converse part is actually much stronger. If X and Y are not blocked then they are dependent in almost all distributions consistent with G. Independence of unblocked paths requires precise parameter tuning that is unlikely. Hence, if we condition on a collider node (resulting in d-connection) we likely create dependence of unintended variety. The next section further explores this issue under the guise of covariate selection and Simpson's paradox.

2 Covariate selection

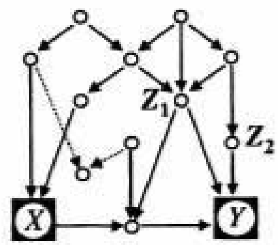
This section deals with semi-Markovian or DAGs (directed acyclic graph) models. Simpson's paradox (results can dramatically change, including sign reversal, when conditioning on additional covariates) indicates the importance and subtlety of covariate selection. A simple algorithm applied to a causal graph indicates when inclusion of covariates produces consistent estimates and otherwise likely produces inconsistent estimates of quantities (causal effects) of interest. Pearl [2010] refers to this as the adjustment problem and presents this via a series of adjusted graphs.

2.1 The adjustment problem





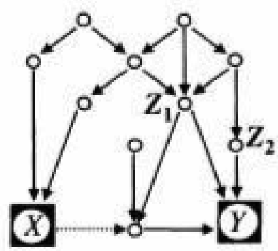
GRAPHICAL SOLUTION OF THE ADJUSTMENT PROBLEM (Cont.)



STEP 2: Delete all non-ancestors of $\{X, Y, Z\}$

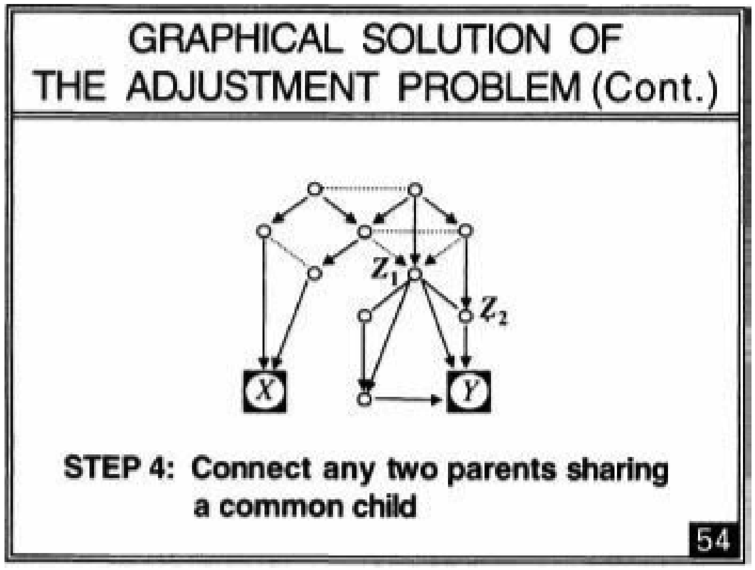
52

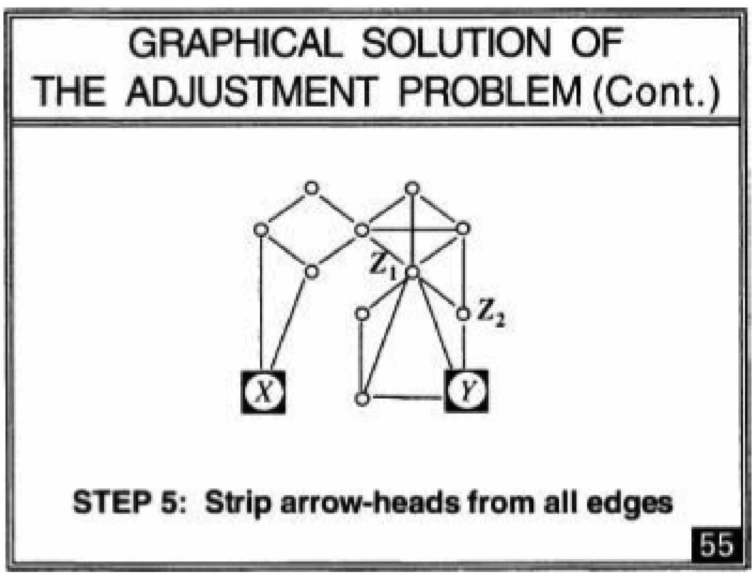
GRAPHICAL SOLUTION OF THE ADJUSTMENT PROBLEM (Cont.)



STEP 3: Delete all arcs emanating from X

53





While Z_1 and Z_2 are not parents of X,[3] adjusting by the parents of X ($pa\,(x)$) is always sufficient for identifying the causal effect of X on Y. The

[3] Z_1 and Z_2 block all back-door paths into X connecting to Y (as do the parents of X). This makes them a sufficient set to eliminate confounding of the causal effect of X on Y.

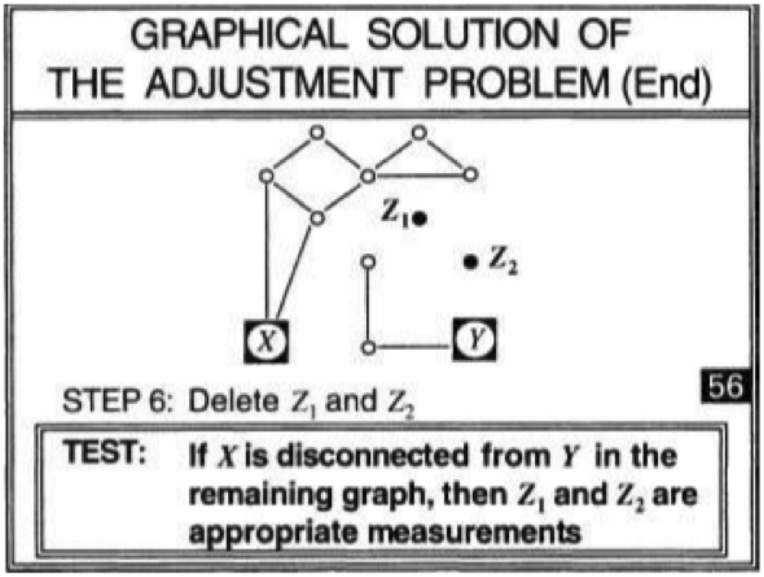


Figure 2.1: Adjustment Problem

adjustment formula is

$$\Pr\left(Y=y\mid do\left(X=x\right)\right)=\sum_{z}\Pr\left(pa\left(x\right)=z\right)\Pr\left(Y=y\mid X=x,pa\left(x\right)=z\right)$$

However, some parents of X may be latent or unobservable making the adjustment formula insufficient for identification. Fortunately, this can sometimes be remedied with instrumental variables, back-door adjustment, and/or front-door adjustment. These ideas are discussed next.

2.2 Instrumental variables

The following *DAGs* depict instrumental variables associated with causal effects but confounded by (unobservable) omitted, correlated variables. Figure (a) depicts a typical instrumental variables setting. The causal effect of X on Y is confounded by correlated omitted variables (the hidden unobservables depicted by the dashed arc connecting through the latent variable L). The instrument(s) Z are related to the causal variable X but independent of outcome Y in the modified graph deleting the $X \to Y$ path (X is a collider with respect to the hidden unobservable and Z blocking the path between Z and Y). Linear IV first regresses both X and Y on Z producing $r_{xz}=a$ and $r_{yz}=ab$. Then the causal effect of X on Y is recovered from the ratio.

$$\frac{r_{yz}}{r_{xz}}=\frac{ab}{a}=b$$

Figure (b) depicts a conditional instrumental variable strategy. Instrument(s) Z is related to the causal variable X but conditionally independent (d-separated) of outcome Y (given covariate(s) W) in the modified graph deleting the $X \to Y$ path. Here, linear IV regresses both X and Y on Z and W producing $r_{xz \cdot w} = r_{xz} = a$ (Z blocks the path between X and W where Y is a collider rendering conditioning on W mute) and $r_{yz \cdot w} = ab$ (conditioning on W is essential to the IV strategy for $r_{yz \cdot w}$). Then the causal effect of X on Y is recovered from the conditional ratio.

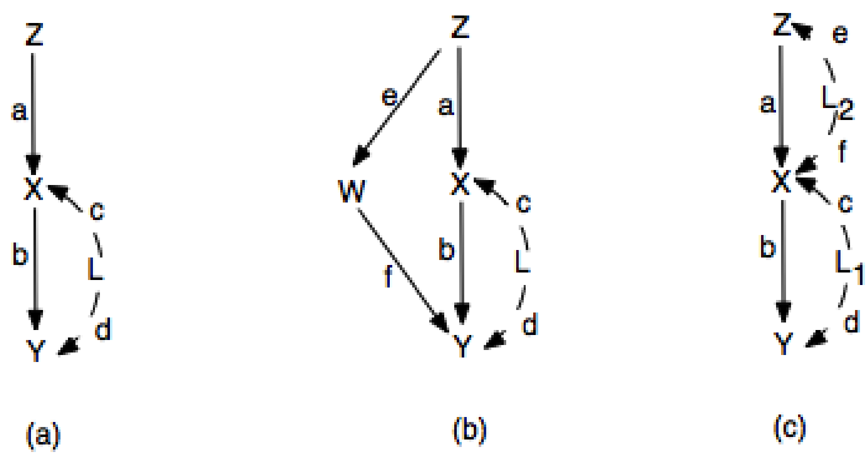
$$\frac{r_{yz \cdot w}}{r_{xz \cdot w}} = \frac{ab}{a} = b$$

Figure (c) is identified analogously to figure (a) in spite of the $Z \to X$ effect being confounded by L_2. As above, the causal effect of X on Y is recovered from the ratio where U_z is unobserved forces (implicit in the graph and other than L_2) causing Z.

$$\frac{r_{yz}}{r_{xz}} = \frac{br_{xz}}{r_{xz}} = b$$

where

$$br_{xz} = \frac{ae^2 Var\left[L_2\right] + efVar\left[L_2\right] + aVar\left[U_z\right]}{e^2 Var\left[L_2\right] + Var\left[U_z\right]}$$



Instrumental variables

Next, we discuss the back-door adjustment.

2.3 Back-door

Formally, a set of variables Z is a back-door to the ordered pair (X, Y) if

(i) no node in Z is a descendant of X, and

(ii) Z blocks every path between X and Y that contains an arrow into X. Consider the causal graph in figure 2.2.

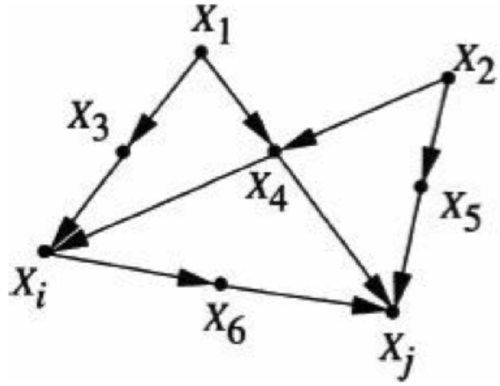


Figure 2.2: DAG for $X_i \to X_j$

The set $Z = \{X_3, X_4\}$, $Z = \{X_4, X_5\}$, or $Z = \{X_3, X_4, X_5\}$ is a back-door to the ordered pair (X_i, X_j) while $Z = \{X_4\}$ or $Z = \{X_6\}$ is not a back-door to (X_i, X_j). Since a back-door d-separates other nodes in the graph from (X_i, X_j), other nodes can be ignored in identifying the causal effect of X_i on X_j and the back-door adjustment for identifying the causal effect is

$$\Pr(y \mid do(x)) = \sum_z \Pr(y \mid x, z) \Pr(z) \qquad \text{(back-door adj)}$$

For figure 2.2 employing either $Z = \{X_4\}$ or $Z = \{X_6\}$ would bias the estimate of the causal effect as the back-door is left unblocked (X_i and X_j are not disconnected when the adjustment problem algorithm is applied). On the other hand, $Z = \{X_3, X_4\}$, $Z = \{X_4, X_5\}$, or $Z = \{X_3, X_4, X_5\}$ blocks everything between (X_i, X_j) except the path $X_i \to X_6 \to X_j$ (X_i and X_j are disconnected when the adjustment problem algorithm is applied) while $Z = \{X_3, X_4, X_6\}$ or $Z = \{X_4, X_5, X_6\}$ d-separates X_i from X_j obliterating the very effect in which we are interested (X_6 is a descendant of X_i — a violation of the adjustment problem algorithm). However, in the next section we find X_6 can be utilized as a front-door to $X_i \to X_j$.

In this case, the back-door adjustment identifies the causal effect of X_i on X_j as

$$\Pr(x_j \mid do(x_i)) = \sum_{x_3} \sum_{x_4} \Pr(x_j \mid x_i, x_3, x_4) \Pr(x_3, x_4)$$

or

$$\Pr(x_j \mid do(x_i)) = \sum_{x_4} \sum_{x_5} \Pr(x_j \mid x_i, x_4, x_5) \Pr(x_4, x_5)$$

The back-door identification follows from Bayes chain rule for the joint distribution exploiting conditional independence in the graph (conditional on its parents everything else is independent of a node)

$$\begin{aligned}\Pr(x_j, x_1, x_2, x_3, x_4, x_5, x_6, x_i) &= \Pr(x_j \mid x_4, x_5, x_6) \Pr(x_6 \mid x_i) \Pr(x_i \mid x_3, x_4) \\ &\quad \Pr(x_3, x_4, x_5 \mid x_1, x_2) \Pr(x_1) \Pr(x_2)\end{aligned}$$

Utilize conditional independence in the graph to write the causal effect of X_i on X_j for the back-door $Z = \{X_3, X_4\}$. The back-door blocks paths from X_i to X_j involving X_1, X_2, X_5 reducing the above joint distribution (in other words, integrating out X_1, X_2, X_5).

$$\begin{aligned}\Pr(x_j, x_3, x_4, x_6, x_i) &= \Pr(x_j \mid x_3, x_4, x_6, x_i) \Pr(x_6 \mid x_i) \Pr(x_i \mid x_3, x_4) \\ &\quad \times \Pr(x_3, x_4) \\ &= \Pr(x_j, x_6 \mid x_3, x_4, x_i) \Pr(x_i \mid x_3, x_4) \Pr(x_3, x_4)\end{aligned}$$

$do(x_i)$ removes the path from X_3, X_4 to X_i and $\Pr(x_i \mid x_3, x_4)$ from the joint distribution (actually sets $\Pr(x_i \mid x_3, x_4) = 1$) leading to

$$\Pr(x_j, x_6 \mid do(x_i)) = \sum_{x_3} \sum_{x_4} \Pr(x_j, x_6 \mid x_3, x_4, x_i) \Pr(x_3, x_4)$$

Summing over X_6 produces the causal effect for X_i on X_j.

$$\begin{aligned}\sum_{x_6} \Pr(x_j, x_6 \mid do(x_i)) &= \sum_{x_6} \sum_{x_3} \sum_{x_4} \Pr(x_j, x_6 \mid x_3, x_4, x_i) \Pr(x_3, x_4) \\ \Pr(x_j \mid do(x_i)) &= \sum_{x_3} \sum_{x_4} \Pr(x_j \mid x_3, x_4, x_i) \Pr(x_3, x_4)\end{aligned}$$

The derivation of the causal effect X_i on X_j for $Z = \{X_4, X_5\}$ is analogous.

$$\begin{aligned}\Pr(x_j, x_3, x_4, x_5, x_6, x_i) &= \Pr(x_j \mid x_3, x_4, x_5, x_6, x_i) \Pr(x_6 \mid x_i) \Pr(x_i \mid x_3, x_4) \\ &\quad \times \Pr(x_3, x_4, x_5) \\ &= \Pr(x_j, x_6 \mid x_3, x_4, x_5, x_i) \Pr(x_i \mid x_3, x_4) \Pr(x_3, x_4, x_5)\end{aligned}$$

$$\begin{aligned}\Pr(x_j, x_6 \mid do(x_i)) &= \sum_{x_4} \sum_{x_5} \Pr(x_j, x_6 \mid x_3, x_4, x_5, x_i) \Pr(x_3, x_4, x_5) \\ &= \sum_{x_4} \sum_{x_5} \Pr(x_j, x_3, x_6 \mid x_4, x_5, x_i) \Pr(x_4, x_5)\end{aligned}$$

Summing over X_3 and X_6 produces the causal effect for X_i on X_j.

$$\begin{aligned}\sum_{x_3} \sum_{x_6} \Pr(x_j, x_3, x_6 \mid do(x_i)) &= \sum_{x_3} \sum_{x_6} \sum_{x_4} \sum_{x_5} \Pr(x_j, x_3, x_6 \mid x_4, x_5, x_i) \\ &\quad \times \Pr(x_4, x_5) \\ \Pr(x_j \mid do(x_i)) &= \sum_{x_4} \sum_{x_5} \Pr(x_j \mid x_4, x_5, x_i) \Pr(x_4, x_5)\end{aligned}$$

We return to this discussion following exposition of *do*-calculus rules.

Action, $\Pr(Y \mid do(x))$, and observation, $\Pr(Y \mid X = x)$, are typically different. Action and observation are only equivalent when X d-separates its parents from Y. The example demonstrates the typical case where action and observation differ.

Example 1 (back-door adjustment — observation $\neq$ action) *Suppose X_i, X_j, and X_1 through X_6 are binary with conditional distributions (consistent with the back-door adjustment DAG in figure 2.2)*

$\Pr(X_j = 1 \mid X_4 = 0, X_5 = 0, X_6 = 0)$	0.001
$\Pr(X_j = 1 \mid X_4 = 0, X_5 = 0, X_6 = 1)$	0.001
$\Pr(X_j = 1 \mid X_4 = 0, X_5 = 1, X_6 = 0)$	0.001
$\Pr(X_j = 1 \mid X_4 = 0, X_5 = 1, X_6 = 1)$	0.999
$\Pr(X_j = 1 \mid X_4 = 1, X_5 = 0, X_6 = 0)$	0.001
$\Pr(X_j = 1 \mid X_4 = 1, X_5 = 0, X_6 = 1)$	0.999
$\Pr(X_j = 1 \mid X_4 = 1, X_5 = 1, X_6 = 0)$	0.999
$\Pr(X_j = 1 \mid X_4 = 1, X_5 = 1, X_6 = 1)$	0.999

$\Pr(X_6 = 1 \mid X_i = 0)$	0.001
$\Pr(X_6 = 1 \mid X_i = 1)$	0.999
$\Pr(X_i = 1 \mid X_3 = 0, X_4 = 0)$	0.001
$\Pr(X_i = 1 \mid X_3 = 0, X_4 = 1)$	0.999
$\Pr(X_i = 1 \mid X_3 = 1, X_4 = 0)$	0.999
$\Pr(X_i = 1 \mid X_3 = 1, X_4 = 1)$	0.999

$\Pr(X_3 = 1 \mid X_1 = 0)$	0.001
$\Pr(X_3 = 1 \mid X_1 = 1)$	0.999
$\Pr(X_4 = 1 \mid X_1 = 0, X_2 = 0)$	0.001
$\Pr(X_4 = 1 \mid X_1 = 0, X_2 = 1)$	0.001
$\Pr(X_4 = 1 \mid X_1 = 1, X_2 = 0)$	0.999
$\Pr(X_4 = 1 \mid X_1 = 1, X_2 = 1)$	0.999
$\Pr(X_5 = 1 \mid X_2 = 0)$	0.001
$\Pr(X_5 = 1 \mid X_2 = 1)$	0.999

$\Pr(X_1 = 1)$	0.2
$\Pr(X_2 = 1)$	0.6

$$
\begin{aligned}
E[X_j] &= 0.202590419 \\
E[X_i] &= 0.202195802 \\
E[X_1] &= 0.2 \\
E[X_2] &= 0.6 \\
E[X_3] &= 0.2006 \\
E[X_4] &= 0.2006 \\
E[X_5] &= 0.5998 \\
E[X_6] &= 0.20279141
\end{aligned}
$$

Then, the back-door adjustment identifies the causal effect

$$
\begin{aligned}
& \Pr(X_j = 1 \mid do(X_i = 0)) \\
= & \sum_{x_3, x_4} \Pr(X_j = 1 \mid x_3, x_4, X_i = 0) \Pr(x_3, x_4) \\
= & \sum_{x_4, x_5} \Pr(X_j = 1 \mid x_4, x_5, X_i = 0) \Pr(x_4, x_5) \\
= & \sum_{x_3, x_4, x_5} \Pr(X_j = 1 \mid x_3, x_4, x_5, X_i = 0) \Pr(x_3, x_4, x_5) \\
= & \; 0.121637881 \\
\neq & \Pr(X_j = 1 \mid X_i = 0) = 0.001749062
\end{aligned}
$$

$$
\begin{aligned}
& \Pr(X_j = 1 \mid do(X_i = 1)) \\
= & \sum_{x_3, x_4} \Pr(X_j = 1 \mid x_3, x_4, X_i = 1) \Pr(x_3, x_4) \\
= & \sum_{x_4, x_5} \Pr(X_j = 1 \mid x_4, x_5, X_i = 1) \Pr(x_4, x_5) \\
= & \sum_{x_3, x_4, x_5} \Pr(X_j = 1 \mid x_3, x_4, x_5, X_i = 1) \Pr(x_3, x_4, x_5) \\
= & \; 0.679161319 \\
\neq & \Pr(X_j = 1 \mid X_i = 1) = 0.995050378
\end{aligned}
$$

$$
\begin{aligned}
& E[X_j \mid do(X = 1)] - E[X_j \mid do(X_i = 0)] \\
= & \; 0.557523438
\end{aligned}
$$

As expected, action ($\Pr(Y = 1 \mid do(X = 0))$*) differs from observation (*$\Pr(Y = 1 \mid X = x)$*).*

To further explore the robustness of the back-door adjustment, next we consider a more varied data generating process (DGP) but still consistent with the DAG in figure 2.2.

Example 2 (Back-door adjustment — more varied DGP) *Suppose X_i, X_j, and X_1 through X_6 are binary with more varied conditional distributions (but consistent with the back-door adjustment DAG in figure 2.2)*

$\Pr(X_j = 1 \mid X_4 = 0, X_5 = 0, X_6 = 0)$	0.2
$\Pr(X_j = 1 \mid X_4 = 0, X_5 = 0, X_6 = 1)$	0.3
$\Pr(X_j = 1 \mid X_4 = 0, X_5 = 1, X_6 = 0)$	0.3
$\Pr(X_j = 1 \mid X_4 = 0, X_5 = 1, X_6 = 1)$	0.2
$\Pr(X_j = 1 \mid X_4 = 1, X_5 = 0, X_6 = 0)$	0.6
$\Pr(X_j = 1 \mid X_4 = 1, X_5 = 0, X_6 = 1)$	0.5
$\Pr(X_j = 1 \mid X_4 = 1, X_5 = 1, X_6 = 0)$	0.5
$\Pr(X_j = 1 \mid X_4 = 1, X_5 = 1, X_6 = 1)$	0.6

$\Pr(X_6 = 1 \mid X_i = 0)$	0.02
$\Pr(X_6 = 1 \mid X_i = 1)$	0.99
$\Pr(X_i = 1 \mid X_3 = 0, X_4 = 0)$	0.1
$\Pr(X_i = 1 \mid X_3 = 0, X_4 = 1)$	0.2
$\Pr(X_i = 1 \mid X_3 = 1, X_4 = 0)$	0.4
$\Pr(X_i = 1 \mid X_3 = 1, X_4 = 1)$	0.5

$\Pr(X_3 = 1 \mid X_1 = 0)$	0.1
$\Pr(X_3 = 1 \mid X_1 = 1)$	0.8
$\Pr(X_4 = 1 \mid X_1 = 0, X_2 = 0)$	0.5
$\Pr(X_4 = 1 \mid X_1 = 0, X_2 = 1)$	0.4
$\Pr(X_4 = 1 \mid X_1 = 1, X_2 = 0)$	0.2
$\Pr(X_4 = 1 \mid X_1 = 1, X_2 = 1)$	0.1
$\Pr(X_5 = 1 \mid X_2 = 0)$	0.03
$\Pr(X_5 = 1 \mid X_2 = 1)$	0.98

$\Pr(X_1 = 1)$	0.2
$\Pr(X_2 = 1)$	0.6

$$
\begin{aligned}
E[X_j] &= 0.368262963 \\
E[X_i] &= 21 \\
E[X_1] &= 0.2 \\
E[X_2] &= 0.6 \\
E[X_3] &= 0.24 \\
E[X_4] &= 0.38 \\
E[X_5] &= 0.6 \\
E[X_6] &= 0.2237
\end{aligned}
$$

Then, the back-door adjustment identifies the causal effect

$$\begin{aligned}
& \Pr\left(X_j = 1 \mid do\left(X_i = 0\right)\right) \\
= & \sum_{x_3, x_4} \Pr\left(X_j = 1 \mid x_3, x_4, X_i = 0\right) \Pr\left(x_3, x_4\right) \\
= & \sum_{x_4, x_5} \Pr\left(X_j = 1 \mid x_4, x_5, X_i = 0\right) \Pr\left(x_4, x_5\right) \\
= & \sum_{x_3, x_4, x_5} \Pr\left(X_j = 1 \mid x_3, x_4, x_5, X_i = 0\right) \Pr\left(x_3, x_4, x_5\right) \\
= & \; 0.3706816 \\
\neq & \Pr\left(X_j = 1 \mid X_i = 0\right) = 0.365823733
\end{aligned}$$

$$\begin{aligned}
& \Pr\left(X_j = 1 \mid do\left(X_i = 1\right)\right) \\
= & \sum_{x_3, x_4} \Pr\left(X_j = 1 \mid x_3, x_4, X_i = 1\right) \Pr\left(x_3, x_4\right) \\
= & \sum_{x_4, x_5} \Pr\left(X_j = 1 \mid x_4, x_5, X_i = 1\right) \Pr\left(x_4, x_5\right) \\
= & \sum_{x_3, x_4, x_5} \Pr\left(X_j = 1 \mid x_3, x_4, x_5, X_i = 1\right) \Pr\left(x_3, x_4, x_5\right) \\
= & \; 0.3571792 \\
\neq & \Pr\left(X_j = 1 \mid X_i = 1\right) = 0.377439116
\end{aligned}$$

$$\begin{aligned}
& E\left[X_j \mid do\left(X_i = 1\right)\right] - E\left[X_j \mid do\left(X_i = 0\right)\right] \\
= & -0.0135024 \\
\neq & E\left[X_j \mid X_i = 1\right] - E\left[X_j \mid X_i = 0\right] \\
= & \; 0.01161538
\end{aligned}$$

Again, action ($\Pr\left(Y = 1 \mid do\left(X = x\right)\right)$) differs from observation ($\Pr\left(Y = 1 \mid X = x\right)$).

2.4 Front-door

Consider the causal graph in figure 2.3.

The above back-door approach is not directly accessible as U is unobservable (whereas $Z = \{X_3, X_4\}$ or $Z = \{X_3, X_4, X_5\}$ is observable in the example 1), so we employ a different (front-door) strategy. The joint distribution for this graph is

$$\Pr\left(u, x, y, z\right) = \Pr\left(y \mid u, x, z\right) \Pr\left(z \mid u, x\right) \Pr\left(x \mid u\right) \Pr\left(u\right)$$

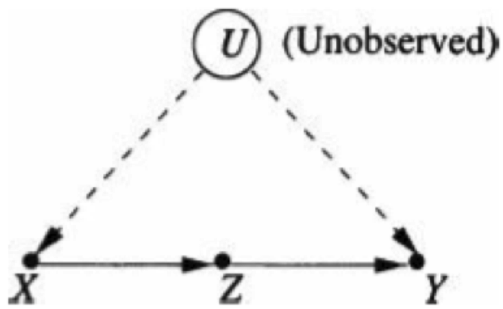


Figure 2.3: Front-Door DAG

However, the graph indicates Z only depends on U through X and Z mediates between X and Y. Hence, $\Pr(z \mid u, x) = \Pr(z \mid x)$, $\Pr(y \mid u, x, z) = \Pr(y \mid u, z)$, and

$$\Pr(u, x, y, z) = \Pr(y \mid u, z) \Pr(z \mid x) \Pr(x \mid u) \Pr(u)$$

$do(x)$ removes the path $U \to X$ so $\Pr(x \mid u)$ drops out.[4]

$$\Pr(u, y, z \mid do(x)) = \Pr(y \mid u, z) \Pr(z \mid x) \Pr(u)$$

Summing over U and Z gives

$$\Pr(y \mid do(x)) = \sum_z \Pr(z \mid x) \sum_u \Pr(y \mid u, z) \Pr(u)$$

Now, we utilize the conditional independence encoded in the graph

$$\begin{aligned} \Pr(u \mid x, z) &= \Pr(u \mid x) \\ \Pr(y \mid u, x, z) &= \Pr(y \mid u, z) \end{aligned}$$

by writing

$$\begin{aligned} \sum_u \Pr(y \mid u, z) \Pr(u) &= \sum_x \sum_u \Pr(y \mid u, z) \Pr(u \mid x) \Pr(x) \\ &= \sum_x \sum_u \Pr(y \mid u, x, z) \Pr(u \mid x, z) \Pr(x) \\ &= \sum_x \sum_u \Pr(y, u \mid x, z) \Pr(x) \\ &= \sum_x \Pr(y \mid x, z) \Pr(x) \end{aligned}$$

[4]Conditioning on Z does not d-separate X and Y as the path through U is unblocked (or d-connected via its fork).

Then, the causal effect of X on Y

$$\Pr(y \mid do(x)) = \sum_z \Pr(z \mid x) \sum_u \Pr(y \mid u, z) \Pr(u)$$

can be expressed, via substitution, purely in terms of observables.

$$\Pr(y \mid do(x)) = \sum_z \Pr(z \mid x) \sum_x \Pr(y \mid x, z) \Pr(x) \qquad \text{(front-door adj)}$$

This expression is the front-door adjustment. Whenever we have a mediating variable Z that meets the conditions $\Pr(x, z) > 0$, $\Pr(u \mid x, z) = \Pr(u \mid x)$, and $\Pr(y \mid u, x, z) = \Pr(y \mid u, z)$ we have a ready nonparametric estimator for the causal effect of X on Y from observable quantities.

Formally, a set of variables Z is defined a front-door for the ordered pair (X, Y) if

(i) Z intercepts all directed paths from X to Y,

(ii) there is no unblocked back-door path from X to Z, and

(iii) all back-door paths from Z to Y are blocked by X.

The above front-door adjustment can be considered a two-step application of the back-door adjustment.[5] First, the causal effect of X on Z is

$$\Pr(z \mid do(x)) = \Pr(z \mid x)$$

since there exists no back-door path to Z. Next, we consider the causal effect of Z on Y. The back-door path to Z is blocked (d-separated) by X. Hence, the back-door adjustment is

$$\Pr(y \mid do(z)) = \sum_x \Pr(y \mid x, z) \Pr(x)$$

Combining the two yields the above front-door adjustment for the causal effect of X on Y.

$$\begin{aligned} \Pr(y \mid do(x)) &= \sum_z \Pr(z \mid do(x)) \Pr(y \mid do(z)) \\ &= \sum_z \Pr(z \mid x) \sum_x \Pr(y \mid x, z) \Pr(x) \end{aligned}$$

Example 3 (front-door adjustment) *Suppose U, X, Y, and Z are binary with conditional distributions (consistent with the above front-door adjustment DAG*

[5] That the front-door adjustment can be described as a two-step application of the back-door adjustment allows the front-door adjustment to require no exception to step one in the adjustment problem described in figure 2.1 (Z should not be a descendant of X).

2.3)

$\Pr(Y=1 \mid Z=0, U=0)$	0.45
$\Pr(Y=1 \mid Z=0, U=1)$	0.65
$\Pr(Y=1 \mid Z=1, U=0)$	0.4
$\Pr(Y=1 \mid Z=1, U=1)$	0.6
$\Pr(X=1 \mid U=0)$	0.6
$\Pr(X=1 \mid U=1)$	0.4
$\Pr(Z=1 \mid X=0)$	0.2
$\Pr(Z=1 \mid X=1)$	0.7
$\Pr(U=1)$	0.6

$$\begin{aligned} E[Y] &= 0.492 \\ E[Z] &= 0.44 \\ E[X] &= 0.48 \end{aligned}$$

Then, the front-door adjustment identifies the causal effect

$$\begin{aligned} \Pr(Z=1 \mid do(X=0)) &= \Pr(Z=1 \mid X=0) \\ &= 0.2 \\ \Pr(Z=1 \mid do(X=1)) &= \Pr(Z=1 \mid X=1) \\ &= 0.7 \end{aligned}$$

$$\begin{aligned} \Pr(Y=1 \mid do(Z=0)) &= \sum_x \Pr(Y=1 \mid x, Z=0)\Pr(x) \\ &= 0.58 \\ \Pr(Y=1 \mid do(Z=1)) &= \sum_x \Pr(Y=1 \mid x, Z=1)\Pr(x) \\ &= 0.38 \end{aligned}$$

$$\begin{aligned} \Pr(Y=1 \mid do(X=0)) &= E[Y \mid do(X=0)] \\ &= (0.8)(0.58) + (0.2)(0.38) \\ &= 0.54 \\ &\neq E[Y \mid X=0] \\ &= 0.544615385 \end{aligned}$$

$$\begin{aligned} \Pr(Y=1 \mid do(X=1)) &= E[Y \mid do(X=1)] \\ &= (0.3)(0.58) + (0.7)(0.38) \\ &= 0.44 \\ &\neq E[Y \mid X=1] \\ &= 0.435 \end{aligned}$$

$$E[Y \mid do(X=1)] - E[Y \mid do(X=0)] = -0.10$$

If U is observable then U provides a back-door from which the same causal effects of X on Y are determined.

$$\begin{aligned}\Pr(Y=1 \mid do(X=0)) &= \sum_u \Pr(Y=1 \mid u, X=0)\Pr(u) \\ &= 0.54\end{aligned}$$

$$\begin{aligned}\Pr(Y=1 \mid do(X=1)) &= \sum_u \Pr(Y=1 \mid u, X=1)\Pr(u) \\ &= 0.44\end{aligned}$$

Figure 2.3 is similar to figure 2.2 if the upper portion is considered unobservable (or ignored). This suggests utilizing X_6 as a front-door also identifies the causal effect of X_i on X_j. We demonstrate the result with an example.[6]

Example 4 (front-door adjustment for example 1) *Return to example 1. Utilize X_6 as a front-door to the causal effect of X_i on X_j.*

$$\begin{array}{c}\sum_{x_i}\Pr(X_j=1 \mid X_6=0, x_i)\Pr(x_i) \\ \sum_{x_i}\Pr(X_j=1 \mid X_6=1, x_i)\Pr(x_i) \\ \Pr(X_6=1 \mid X_i=0) \\ \Pr(X_6=1 \mid X_i=1)\end{array}$$

Then, the front-door adjustment identifies the causal effect

$$\begin{aligned}\Pr(X_6=1 \mid do(X_i=0)) &= \Pr(X_6=1 \mid X_i=0) \\ &= 0.001 \\ \Pr(X_6=1 \mid do(X_i=1)) &= \Pr(X_6=1 \mid X_i=1) \\ &= 0.999\end{aligned}$$

$$\begin{aligned}\Pr(X_j=1 \mid do(X_6=0)) &= \sum_{x_i}\Pr(X_j=1 \mid x_i, X_6=0)\Pr(x_i) \\ &= 0.12107924 \\ \Pr(X_j=1 \mid do(X_6=1)) &= \sum_{x_i}\Pr(X_j=1 \mid x_i, X_6=1)\Pr(x_i) \\ &= 0.67971996\end{aligned}$$

$$\begin{aligned}\Pr(X_j=1 \mid do(X_i=0)) &= E[X_j \mid do(X_i=0)] \\ &= (0.999)(0.12107924)+(0.001)(0.67971996) \\ &= 0.121637881 \\ &\neq \Pr(Y=1 \mid X=0) \\ &= 0.001749062\end{aligned}$$

[6] Of course, this is true for every probability distriubution consistent with DAG 2.2 including example 2.

$$\begin{aligned}
\Pr(X_j = 1 \mid do(X_i = 1)) &= E[X_j \mid do(X_i = 1)] \\
&= (0.001)(0.12107924) + (0.999)(0.67971996) \\
&= 0.679161319 \\
&\neq \Pr(X_j = 1 \mid X_i = 1) \\
&= 0.995050378
\end{aligned}$$

$$E[X_j \mid do(X_i = 1)] - E[X_j \mid do(X_i = 0)] = 0.557523438$$

Of course, the causal effect identified utilizing $Z = X_6$ as a front-door to $X_i \to X_j$ is the same effect identified in example 1 by the back-door $Z = \{X_3, X_4\}, Z = \{X_4, X_5\}$, or $Z = \{X_3, X_4, X_5\}$.

2.5 do-calculus rules

The front-door criteria above are actually too stringent. Conditions (ii) and (iii) can be violated provided there is a covariate to block back-door paths. For example, in the graph in figure 2.4, Z_2 serves as a front-door-like criterion relative to (X, Z_3) provided we condition on Z_1. Z_1 blocks (d-separates) any back-door path from X to Z_2 as well as from Z_2 to Z_3. To better accommodate such variations, Pearl [1995] provides a theorem of do-calculus inference rules.

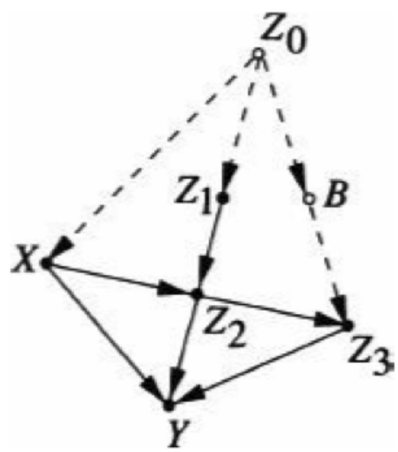


Figure 2.4: do-calculus DAG

Theorem 1 (do-calculus rules) *Let G be the DAG associated with a causal model and let* $\Pr(\cdot)$ *be the probability distribution induced by the model. For any disjoint set of variables X, Y, Z, and W the following rules apply.*

Rule 1 (insertion/deletion of observations):

$$\Pr\left(y \mid do\left(x\right), z, w\right) = \Pr\left(y \mid do\left(x\right), w\right) \quad if \ \ \left(Y \perp Z \mid X, W\right)_{G_{\overline{X}}}$$

where $\perp$ refers to stochastic independence or d-separation in the graph.

Rule 2 (action/observation exchange):

$$\Pr\left(y \mid do\left(x\right), do\left(z\right), w\right) = \Pr\left(y \mid do\left(x\right), z, w\right) \quad if \ \ \left(Y \perp Z \mid X, W\right)_{G_{\overline{X}\underline{Z}}}$$

Rule 3 (insertion/deletion of actions):

$$\Pr\left(y \mid do\left(x\right), do\left(z\right), w\right) = \Pr\left(y \mid do\left(x\right), w\right) \quad if \ \ \left(Y \perp Z \mid X, W\right)_{G_{\overline{X}, \overline{Z(W)}}}$$

where $Z\left(W\right)$ is the set of Z-nodes that are not ancestors of any W-nodes in $G_{\overline{X}}$.

Rule 1 affirms d-separation of Z and Y leaves Y conditionally independent of Z following intervention $X = x$ which corresponds to the subgraph $G_{\overline{X}}$. Figure 2.5 provides a simple illustration of do-calculus rule 1 where X d-separates its parent(s) Z from Y.

The subgraph $G_{\overline{X}\underline{Z}}$ only differs from the subgraph $G_{\overline{X}}$ by eliminating the direct path $Z \rightarrow Y$ but leaves the same back-door paths from Z to Y. Rule 2 effectively says that intervention by $Z = z$ has no different effect on Y than passive conditioning on the evidence $Z = z$ when $\{X, W\}$ blocks all back-door paths from Z to Y (in the subgraph $G_{\overline{X}}$). Figure 2.6 provides a simple illustration of do-calculus rule 2.

Consistency of rules 1 and 2 can be tested if we eliminate the bow in figure 2.6 so that X and Z are independent (due to collider Y) but both are causal to Y (see figure 2.7). Rule 1 affirms the independence of X and Z while rule 2 indicates $\Pr\left(Y \mid do\left(X = x\right)\right) = \Pr\left(Y \mid X = x\right)$. To demonstrate internal consistency, suppose we apply the back-door adjustment (even though there is no back-door into X) to identify action (the causal effect)

$$\Pr\left(Y \mid do\left(X = x\right)\right) = \sum_{z} \Pr\left(z\right) \Pr\left(Y \mid x, z\right)$$

while outcome conditional on observation x is

$$\Pr\left(Y \mid X = x\right) = \sum_{z} \Pr\left(z \mid x\right) \Pr\left(Y \mid x, z\right)$$

However, rule 1 indicates $\Pr\left(z \mid x\right) = \Pr\left(z\right)$ which affirms rule 2's implication in figure 2.7.

$$\Pr\left(Y \mid do\left(X = x\right)\right) = \Pr\left(Y \mid X = x\right)$$

The subgraph $G_{\overline{X}, \overline{Z(W)}}$ corresponds to deleting all equations relating to the variables Z. Therefore, when this condition is satisfied intervention by $Z = z$

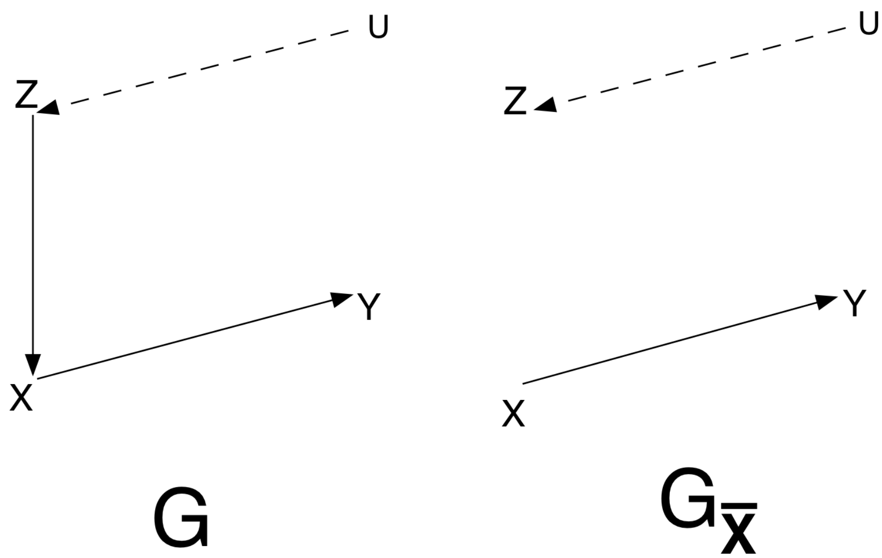


Figure 2.5: Rule 1 DAGs

can be eliminated (inserted) without altering Y as stated in rule 3. Figure 2.8 provides a simple illustration of do-calculus rule 3.

Let's review the relation between our simple DAGs and the do-calculus rules.[7] Since the DAG in figure 2.5 satisfies both rule 1 and rule 2, logically it also satisfies rule 3 (as is the case). Hence, the probability of Y given $\{X, W\}$ is not affected by insertion/deletion of Z observation, exchange of Z observation/action, or insertion/deletion of Z action.

On the other hand, the DAG in figure 2.6 satisfies rule 2 but not rule 1 or rule 3. There is a direct path from Z to Y in both $G_{\overline{X}}$ and $G_{\overline{X}\,\overline{Z}}$. Consequently, the probability of Y given $\{X, W\}$ is not affected by exchange of Z observation/action, but it is not necessarily immune to insertion/deletion of Z observation or Z action.

Similarly, the simple DAG in figure 2.8 satisfies rule 3 but not rule 1 or rule 2. There is an unblocked back-door path from Z to Y ($Z \leftarrow U \rightarrow Y$) in both $G_{\overline{X}}$ and $G_{\overline{X}\underline{Z}}$. This implies the probability of Y given $\{X, W\}$ is not affected by insertion/deletion of Z action, but it is not necessarily immune to

[7] For simplicity, $W = \emptyset$ in these DAGs.

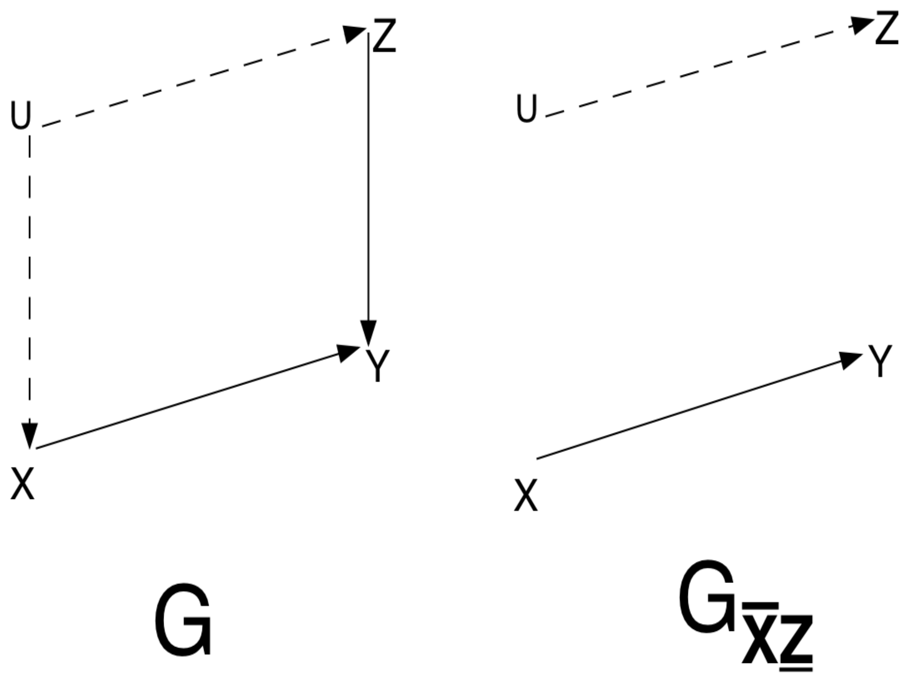


Figure 2.6: Rule 2 DAGs

insertion/deletion of Z observation or exchange of Z observation/action.[8]

2.6 Illustration of do-calculus rules

Causal effects are defined in terms of do-calculus but identified only if the intervention probability can be mapped into probabilities over observables only.

2.6.1 do-calculus for figure 2.3

Now, we illustrate the do-calculus rules applied to the identification of various causal effects in the above (U, X, Z, Y) front-door adjustment DAG in figure 2.3.

Front-door adjustment Task one: $\Pr(z \mid do(x))$

Rule 2 applies as $X \perp Z$ in $G_{\underline{X}}$ where the path $X \longleftarrow U \rightarrow Y \longleftarrow Z$ is blocked at the collider Y. Hence, we can directly conclude

$$\Pr(z \mid do(x)) = \Pr(z \mid x)$$

[8]Further discussion of rules 2 and 3 by reference to augmented DAGs is reported in the appendix.

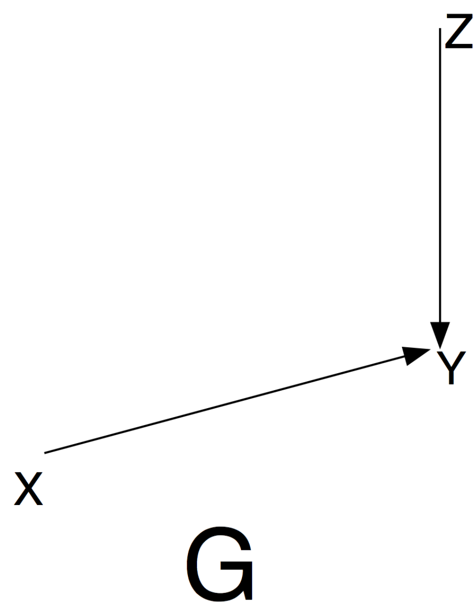


Figure 2.7: DAG illustrating rules 1 and 2

Task two: $\Pr(y \mid do(z))$

We cannot exchange $do(z)$ with z as rule 2 does not directly apply since there is a back-door path from Z to Y in $G_{\underline{Z}}$: $Z \longleftarrow X \longleftarrow U \longrightarrow Y$. We can block this path by conditioning and summing over all values of X.

$$\Pr(y \mid do(z)) = \sum_x \Pr(y \mid x, do(z)) \Pr(x \mid do(z))$$

For the latter term on the right-hand side, we employ rule 3 for action deletion

$$\Pr(x \mid do(z)) = \Pr(x)$$

since X and Z are d-separated in $G_{\overline{Z}}$ where again the path $X \longleftarrow U \rightarrow Y \longleftarrow Z$ is blocked at the collider Y. For the former term on the right-hand side, we utilize rule 2

$$\Pr(y \mid x, do(z)) = \Pr(y \mid x, z)$$

since X d-separates Z from Y in $G_{\underline{Z}}$. Putting this together gives

$$\begin{aligned} \Pr(y \mid do(z)) &= \sum_x \Pr(y \mid x, z) \Pr(x) \\ &= E_X[\Pr(y \mid x, z)] \end{aligned}$$

Task three: $\Pr(y \mid do(x))$

$$\Pr(y \mid do(x)) = \sum_z \Pr(y \mid z, do(x)) \Pr(z \mid do(x))$$

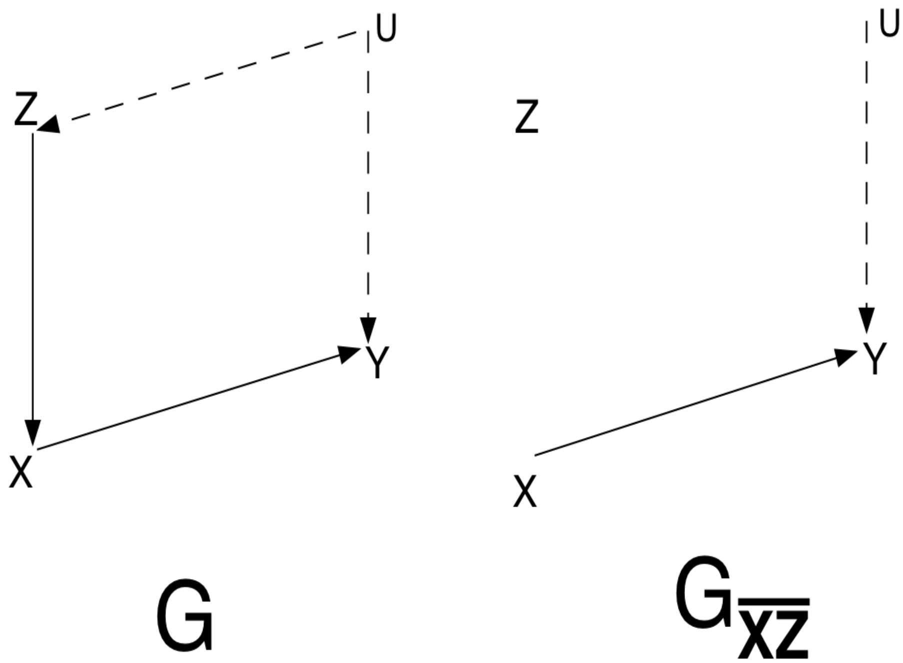


Figure 2.8: Rule 3 DAGs

Task one reduced the latter term on the right-hand side

$$\Pr(z \mid do(x)) = \Pr(z \mid x)$$

but we have no rule to eliminate $do(x)$ in the former term.[9] However, rule 2 allows us to add $do(z)$ since the condition $(Y \perp Z \mid X)\, G_{\overline{X}\underline{Z}}$ is satisfied. Then, we can delete the action $do(x)$ using rule 3 since $(Y \perp X \mid Z)\, G_{\overline{XZ}}$ applies. This leads to

$$\begin{aligned} \Pr(y \mid z, do(x)) &= \Pr(y \mid do(z), do(x)) \\ &= \Pr(y \mid do(z)) \end{aligned}$$

Task two indicates

$$\Pr(y \mid do(z)) = \sum_x \Pr(y \mid x, z) \Pr(x)$$

[9] Suppose rule 3 did not include the provision $\overline{Z(W)}$ but rather was simply $\overline{Z}$. Then, $\Pr(y \mid z, do(x)) = \Pr(y \mid z)$ and $\Pr(y \mid do(x))$ would reduce to $\sum_z \Pr(z \mid x) \Pr(y \mid z)$. An inconsistency that fails to account for the unblockable back-door unless the bow is eliminated from the DAG.

Putting everything together we have the front-door adjustment

$$\Pr(y \mid do(x)) = \sum_z \Pr(z \mid x) \sum_{x'} \Pr\left(y \mid x', z\right) \Pr\left(x'\right)$$

Task four: $\Pr(y, z \mid do(x))$

$$\Pr(y, z \mid do(x)) = \Pr(y \mid z, do(x)) \Pr(z \mid do(x))$$

Both right-hand side terms were utilized in task three from which we obtain

$$\Pr(y, z \mid do(x)) = \Pr(z \mid x) \sum_{x'} \Pr\left(y \mid x', z\right) \Pr\left(x'\right)$$

Task five: $\Pr(y, x \mid do(z))$

$$\begin{aligned} \Pr(y, x \mid do(z)) &= \Pr(y \mid x, do(z)) \Pr(x \mid do(z)) \\ &= \Pr(y \mid x, z) \Pr(x \mid z) \end{aligned}$$

The first term on the right-hand side derives from rule 2 in subgraph $G_{\underline{Z}}$ and the second term from rule 3 as applied in task two.

Back-door adjustment If U is observable then we can employ a back-door adjustment where U supplies the back-door.[10]

$$\Pr(y \mid do(x)) = \sum_u \Pr(y \mid u, do(x)) \Pr(u \mid do(x))$$

Task one: $\Pr(u \mid do(x))$

By rule 3 $(U \perp X \mid \emptyset) G_{\overline{X}}$ as Y is a collider and d-separates U and X in $G_{\overline{X}}$ or $X \to Z \to Y \leftarrow U$. Hence, $\Pr(u \mid do(x)) = \Pr(u)$.

Task two: $\Pr(y \mid u, do(x))$

By rule 2 $(Y \perp X \mid U) G_{\underline{X}}$ as conditioning on U d-separates Y and X in $G_{\underline{X}}$ or $Z \to Y \leftarrow U \to X$. Therefore, $\Pr(y \mid u, do(x)) = \Pr(y \mid u, x)$

Task three: $\Pr(y \mid do(x))$

Putting tasks one and two together gives the back-door adjustment

$$\begin{aligned} \Pr(y \mid do(x)) &= \sum_u \Pr(y \mid u, do(x)) \Pr(u \mid do(x)) \\ &= \sum_u \Pr(y \mid u, x) \Pr(u) \end{aligned}$$

[10] We only present this hypothetical case to illustrate consistency of the do-calculus rules. U is unobservable ruling out this identification strategy in practice.

2.6.2 do-calculus for figure 2.2

Back-door adjustment We present the do-calculus rules for the back-door adjustment involving the two minimal sets $Z = \{X_3, X_4\}$ and $Z = \{X_4, X_5\}$ or their union $Z = \{X_3, X_4, X_5\}$ for identifying the causal effect of X_i on X_j.

$$\Pr(x_j \mid do(x_i)) = \sum_z \Pr(x_j \mid z, do(x_i)) \Pr(z \mid do(x_i))$$

Task one: $\Pr(z \mid do(x_i))$

By rule 3 $(Z \perp X \mid \emptyset)\, G_{\overline{X}}$ as Y is a collider and d-separates $Z = \{X_3, X_4\}, Z = \{X_4, X_5\}$, or $Z = \{X_3, X_4, X_5\}$ and X in $G_{\overline{X}}$ (refer to the subgraphs in figure 1.1). Hence, $\Pr(x_3, x_4 \mid do(x_i)) = \Pr(x_3, x_4)$.

Task two: $\Pr(x_j \mid z, do(x_i))$

By rule 2 $(X_j \perp X_i \mid Z)\, G_{\underline{X}}$ as conditioning on $Z = \{X_3, X_4\}, Z = \{X_4, X_5\}$, or $Z = \{X_3, X_4, X_5\}$ d-separates Y and X in $G_{\underline{X}}$. Therefore, $\Pr(x_j \mid z, do(x_i)) = \Pr(x_j \mid u, x_i)$

Task three: $\Pr(x_j \mid do(x_i))$

Putting tasks one and two together gives the back-door adjustment

$$\begin{aligned}\Pr(x_j \mid do(x_i)) &= \sum_z \Pr(x_j \mid z, do(x_i)) \Pr(z \mid do(x_i)) \\ &= \sum_{x_3} \sum_{x_4} \Pr(x_j \mid x_3, x_4, x_i) \Pr(x_3, x_4)\end{aligned}$$

or

$$\begin{aligned}\Pr(x_j \mid do(x_i)) &= \sum_z \Pr(x_j \mid z, do(x_i)) \Pr(z \mid do(x_i)) \\ &= \sum_{x_4} \sum_{x_5} \Pr(x_j \mid x_4, x_5, x_i) \Pr(x_4, x_5)\end{aligned}$$

or

$$\begin{aligned}\Pr(x_j \mid do(x_i)) &= \sum_z \Pr(x_j \mid z, do(x_i)) \Pr(z \mid do(x_i)) \\ &= \sum_{x_3} \sum_{x_4} \sum_{x_5} \Pr(x_j \mid x_3, x_4, x_5, x_i) \Pr(x_3, x_4, x_5)\end{aligned}$$

2.6.3 Front-door adjustment

The front-door do-calculus for figure 2.2 is the same as tasks one through three for figure 2.3 where X_j replaces Y, X_i replaces X, X_6 replaces Z, and X_1 through X_5 are ignored like U.

2.7 Identifiable and nonidentifiable

Identification of the causal effect of X on Y is essentially determined by back-door and front-door adjustments. The graphs in figure 2.9 represent identifiable

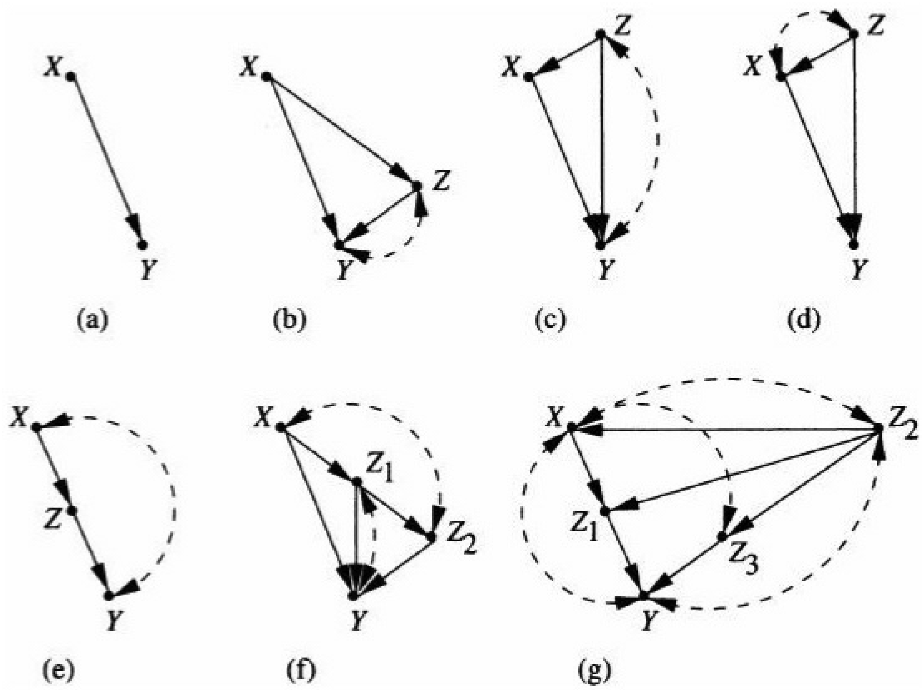


Figure 2.9: DAGs involving identifiable causal effects of X on Y

causal effects of X on Y. Dashed arcs (bows) represent confounding by unobserved variables while Z and W represent observed covariates.

For the graphs in figure 2.10 with bows, the causal effect of X on Y is nonidentifiable.

Note, the difference between the nonidentifiable DAG in figure 2.10 (a) and the identifiable DAG in figure 2.9 (e) is the front-door adjustment (or two rounds of back-door adjustments) utilizes covariate Z for identification whereas Z is missing when the causal effect is nonidentifiable. This typifies the nonparametric identification problem. Additional details follow.

2.7.1 Identifiable DAGs

The causal effect of X on Y in figure 2.9(a) is identified by do-calculus rule 2 where X is d-separated from Y in $G_{\underline{X}}$ leading to

$$\Pr(y \mid do(x)) = \Pr(y \mid x)$$

Also, X d-separates its parents, $pa(X) = \emptyset$, from Y (an application of rule 2) implying observation equals action.

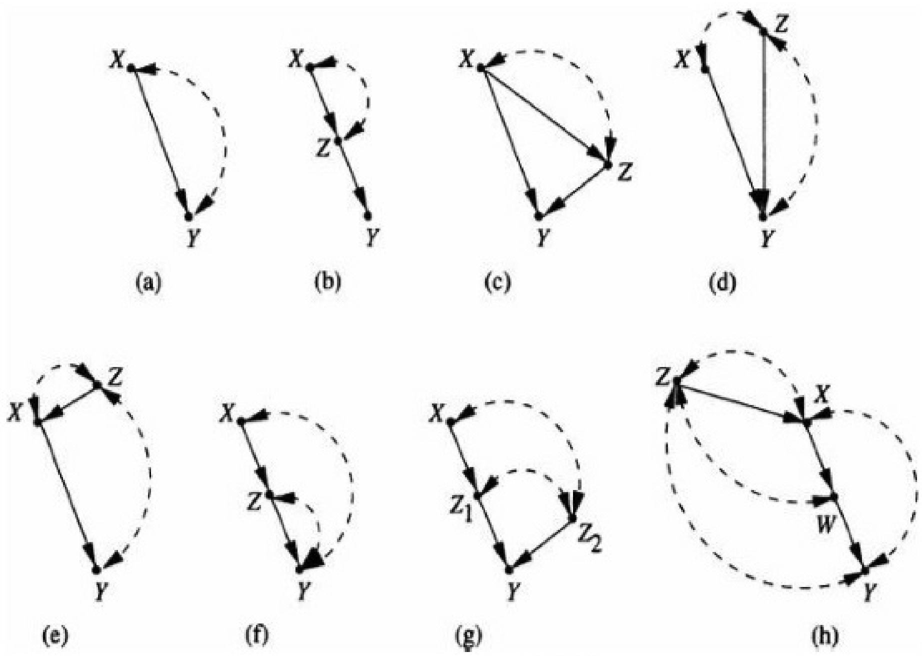


Figure 2.10: DAGs involving nonidentifiable causal effects of X on Y

The causal effect of X on Y in figure 2.9(b) is also identified by do-calculus rule 2 where X is d-separated from Y in $G_{\underline{X}}$ leading to

$$\Pr(y \mid do(x)) = \Pr(y \mid x)$$

Again, X is a root and d-separates its parents, $pa(X) = \emptyset$, from Y implying observation equals action.

The causal effect of X on Y in figures 2.9(c) and (d) involve a classic back-door adjustment utilizing Z.

$$\begin{aligned}\Pr(y \mid do(x)) &= \sum_z \Pr(y \mid do(x), z) \Pr(z \mid do(x)) \\ &= \sum_z \Pr(y \mid x, z) \Pr(z)\end{aligned}$$

where $\Pr(z \mid do(x)) = \Pr(z)$ by rule 3 since Y is a collider making $(Z \perp X \mid \emptyset)_{G_{\underline{X}}}$ and $\Pr(z \mid do(x)) = \Pr(z)$. Also, by rule 2 $(Y \perp X \mid Z)_{G_{\overline{X}}}$ as Z d-separates X and Y. implying $\Pr(y \mid do(x), z) = \Pr(y \mid x, z)$. Substitution gives the back-door adjustment.

The causal effect of X on Y in figure 2.9(e) is resolved, as discussed in detail earlier, by front-door adjustment utilizing Z.

$$\Pr(y \mid do(x)) = \sum_z \Pr(z \mid x) \sum_{x'} \Pr\left(y \mid x', z\right) \Pr\left(x'\right)$$

The causal effect of X on Y in figure 2.9(f) involves a back-door adjustment utilizing covariates $\{Z_1, Z_2\}$.

$$\begin{aligned}\Pr(y \mid do(x)) &= \sum_{z_1}\sum_{z_2} \Pr(y \mid do(x), z_1, z_2) \Pr(z_2 \mid do(x), z_1) \\ &\quad \times \Pr(z_1 \mid do(x))\end{aligned}$$

Step 1: Use rule 2 to exchange $do(X)$ with X where $(Y \perp X \mid Z_1, Z_2)_{G_{\underline{X}}}$ as Z_1 and Z_2 d-separate Y and X. This makes

$$\Pr(y \mid do(x), z_1, z_2) = \Pr(y \mid x, z_1, z_2)$$

Step 2a: We can't directly eliminate $do(x)$ in $\Pr(z_2 \mid do(x), z_1)$[11] but we can insert $do(z_1)$ by rule 2. $(Z_2 \perp Z_1 \mid X)_{G_{\overline{X}\underline{Z_1}}}$ where Y (a collider) d-separates Z_1 and Z_2 in $G_{\overline{X}\underline{Z_1}}$ implies

$$\Pr(z_2 \mid do(x), z_1) = \Pr(z_2 \mid do(x), do(z_1))$$

Step 2b: Now, rule 3 allows elimination of $do(x)$ as $(Z_2 \perp X \mid Z_1)_{G_{\overline{Z_1 X}}}$ where again Y is a collider producing d-separation of X and Z_2 in $G_{\overline{Z_1 X}}$. This produces

$$\Pr(z_2 \mid do(x), do(z_1)) = \Pr(z_2 \mid do(z_1))$$

Step 2c: We can't directly replace action with observation as there is an unblocked back-door path between Z_1 and Z_2 in $G_{\underline{Z_1}}$. However, we can block this path by conditioning and summing over X.

$$\Pr(z_2 \mid do(z_1)) = \sum_x \Pr(z_2 \mid x, do(z_1)) \Pr(x \mid do(z_1))$$

Rule 2 now applies as $(Z_1 \perp Z_2 \mid X)_{G_{\underline{Z_1}}}$ so that the first term is

$$\Pr(z_2 \mid x, do(z_1)) = \Pr(z_2 \mid x, z_1)$$

The second term is resolved by rule 3 as $(X \perp Z_1)_{G_{\overline{Z_1}}}$. Thus,

$$\Pr(x \mid do(z_1)) = \Pr(x)$$

and

$$\Pr(z_2 \mid do(z_1)) = \sum_x \Pr(z_2 \mid x, z_1) \Pr(x)$$

Step 3:Rule 2 replaces action with observation as $(Z_1 \perp X)_{G_{\underline{X}}}$ where Z_2 and Y serve as colliders with respect to X and Z_1. Hence,

$$\Pr(z_1 \mid do(x)) = \Pr(z_1 \mid x)$$

[11] Since X is an ancestor to Z_1 and Z_2 rule 3 involves $G_{\overline{X(Z)}}$ equivalent to G.

Step 4: Combining and summing produces the back-door adjustment

$$\begin{aligned}\Pr(y \mid do(x)) &= \sum_{z_1}\sum_{z_2} \Pr(y \mid do(x), z_1, z_2) \Pr(z_1 \mid do(x)) \\ &\quad \times \Pr(z_2 \mid do(x), z_1) \\ &= \sum_{z_1}\sum_{z_2} \Pr(y \mid x, z_1, z_2) \Pr(z_1 \mid x) \sum_{x'} \Pr\left(z_2 \mid x', z_1\right) \Pr\left(x'\right)\end{aligned}$$

The causal effect of X on Y in figure 2.9(g) involves a front-door adjustment utilizing $Z = \{Z_1\}$ and back-door blocking covariates $W = \{Z_2, Z_3\}$.

Step 1: Apply rule 2 to $G_{\underline{X}}$ for the $X \to Z$ component. Since Y (a collider) and W combine to d-separate X from Z, $(Z \perp X \mid W)_{G_X}$, then

$$\Pr(z \mid do(x), w) = \Pr(z \mid x, w)$$

Step 2: The $Z \to Y$ component is

$$\Pr(y \mid do(z), w) = \sum_x \Pr(y \mid x, do(z), w) \Pr(x \mid do(z), w)$$

The second term is resolved by rule 3 where $(Z \perp X \mid W)_{G_{\overline{Z}}}$, thus

$$\Pr(x \mid do(z), w) = \Pr(x \mid w)$$

The first term is

$$\Pr(y \mid x, do(z), w) = \Pr(y \mid x, z, w)$$

by rule 2 where $(Y \perp Z \mid W)_{G_{\underline{Z}}}$ as X and W block all back-door paths between Z and Y. Hence,

$$\Pr(y \mid do(z), w) = \sum_x \Pr(y \mid x, z, w) \Pr(x \mid w)$$

The causal effect of interest is

$$\begin{aligned}\Pr(y \mid do(x)) &= \sum_w \sum_z \Pr(y \mid do(x), z, w) \Pr(z \mid do(x), w) \\ &\quad \times \Pr(w \mid do(x))\end{aligned}$$

Step 3: The last term is

$$\Pr(w \mid do(x)) = \Pr(w)$$

by rule 3 where $(W \perp X \mid \emptyset)_{G_{\overline{X}}}$ as Z_1 is a collider with respect to X and Z_2.

Step 4: The leading component summed over Z is

$$\Pr(y \mid do(x), w) = \sum_z \Pr(y \mid do(x), z, w) \Pr(z \mid do(x), w)$$

The latter term we earlier resolved by rule 2

$$\Pr(z \mid do(x), w) = \Pr(z \mid x, w)$$

Rule 2 provides

$$\Pr(y \mid do(x), z, w) = \Pr(y \mid do(x), do(z), w)$$

and rule 3 indicates X is d-separated from Y or $(Y \perp X \mid Z, W)_{G_{\overline{Z}\overline{X}}}$, thus

$$\Pr(y \mid do(x), do(z), w) = \Pr(y \mid do(z), w)$$

Earlier we resolved

$$\Pr(y \mid do(z), w) = \sum_{x} \Pr(y \mid x, z, w) \Pr(x \mid w)$$

Step 5: Putting everything together we have identication of the causal effect of X on Y via the front-door adjustment.

$$\begin{aligned} \Pr(y \mid do(x)) &= \sum_{w} \sum_{z} \Pr(y \mid do(x), z, w) \Pr(z \mid do(x), w) \Pr(w \mid do(x)) \\ &= \sum_{w} \sum_{z} \Pr(z \mid x, w) \sum_{x'} \Pr(y \mid x', z, w) \Pr(x' \mid w) \Pr(w) \end{aligned}$$

2.7.2 Nonidentifiable DAGs

The causal effect of X on Y in figure 2.10(a) is not identifiable as there is no front-door and because U (the bidirectional bow) is an unobservable back-door that confounds identifying the direct effect of X on Y. Further, there are no other observables that might allow identification of conditional causal effects.

The causal effect of X on Y in figure 2.10(b) is not identifiable as the portion of the front-door adjustment referring to $\Pr(Z \mid do(x))$ is confounded by the unobservable bow similar to figure 2.10(a). However, combined or conditional action $do(X = x)$ and $do(Z = z)$ on Y is identified. By rule 2 $(Y \perp X \mid Z)_{G_{\overline{Z}\underline{X}}}$ and $(Y \perp Z \mid X)_{G_{\underline{Z}}}$ so that

$$\begin{aligned} \Pr(y \mid do(x), do(z)) &= \Pr(y \mid do(x), z) \\ &= \Pr(y \mid x, do(z)) \\ &= \Pr(y \mid x, z) \end{aligned}$$

We refer to this as a conditional causal effect $(X \to Y \mid Z)$.

The causal effect of X on Y in figure 2.10(c) is not identifiable because of the unblockable back-door bow into X. While similar to figure 2.9(b) the difference in unobservables (or bows) confounds identification in figure 2.10(c). It may appear that Z can be employed to block the back-door. However, action $\Pr(z \mid do(x))$ cannot be translated into observation (by either rule 2 or rule 3) so point identification of $\Pr(y \mid do(x))$ via the back-door adjustment fails.

Nonetheless, similar to figure 2.10(b), conditional action $do\,(X = x)$ and $do\,(Z = z)$ on Y, or conditional causal effect $(X \to Y \mid Z)$, is identified for figure 2.10(c). By rule 2 $(Y \perp X \mid Z)_{G_{\overline{Z}\underline{X}}}$ and $(Y \perp Z \mid X)_{G_{\underline{Z}}}$ so that

$$\begin{aligned}\Pr(y \mid do\,(x)\,, do\,(z)) &= \Pr(y \mid do\,(x)\,, z) \\ &= \Pr(y \mid x, do\,(z)) \\ &= \Pr(y \mid x, z)\end{aligned}$$

The causal effect of X on Y in figure 2.10(d) is not identifiable since the two bows combine to create a back-door into X that cannot be blocked by Z, a collider with respect to the two bows. The confounding is completed by the bow between X and Z along with $Z \to Y$ creates a back-door path which can only be blocked by conditioning on Z but conditioning on Z d-connects the back-door path created by the two bows. The causal effect of X on Y is confounded.

We cannot even identify a conditional causal effect for figure 2.10(d). Conditional on $do\,(Z = z)$, we can exchange action $do\,(X = x)$ with observation as $(Y \perp X \mid Z)_{G_{\overline{Z}\underline{X}}}$. However, we cannot exchange action $do\,(Z = z)$ with observation even conditional on $do\,(X = x)$ since there is no subgraph in which Y and Z are conditionally independent.

The causal effect of X on Y in figure 2.10(e) is confounded similarly to that in figure 2.10(d). $Z \to X$ along with the bow between Z and Y creates a back-door path blocked by conditioning on Z but conditioning on Z (a collider with respect to the two bows) creates a back-door path through the two bows. Hence, the causal effect of X on Y is confounded.

Again, we cannot identify a conditional causal effect for figure 2.10(e). Conditional on $do\,(Z = z)$, we can exchange action $do\,(X = x)$ with observation as $(Y \perp X \mid Z)_{G_{\overline{Z}\underline{X}}}$. However, we cannot exchange action $do\,(Z = z)$ with observation even conditional on $do\,(X = x)$. We could delete $do\,(Z = z)$ conditional on $do\,(X = x)$ but we need to delete $do\,(Z = z)$ conditional on observation $X = x$ but cannot since Z is ancestor to X leaving subgraph $G_{\overline{Z(X)}}$ equivalent to G.

The causal effect of X on Y in figure 2.10(f) is confounded similarly to figure 2.10(b). In this case it is the portion of the front-door adjustment to $\Pr(Y \mid do\,(z))$ in which a back-door bow exists to prevent identification.

We cannot identify a conditional causal effect for figure 2.10(f). No subgraphs permit replacement of action $do\,(X = x)$ or $do\,(Z = z)$ with observation. While we can delete action $do\,(X = x)$ conditional on $do\,(Z = z)$, we cannot replace action $do\,(Z = z)$.

The causal effect of X on Y in figure 2.10(g) is not identifiable similarly to figure 2.10(b). The front-door Z_1 requires a back-door covariate Z_2 but Z_2 is a collider with respect to the two bows and conditioning on Z_2 creates a back-door through the two bows with respect to $\Pr(Z_1 \mid do\,(x))$. Hence, the causal effect of X on Y is inescapably confounded.

Similar to figures 2.10(b) and (c), conditional action $do\,(X = x)$ and $do\,(Z = z)$ on Y, or conditional causal effect $(X \to Y \mid Z)$, is identified for figure 2.10(g) where $Z = \{Z_1, Z_2\}$. By rule 2 $(Y \perp X \mid Z)_{G_{\overline{Z}\underline{X}}}$ and $(Y \perp Z \mid X)_{G_{\underline{Z}}}$ so that

$$\begin{aligned}\Pr(y \mid do(x), do(z)) &= \Pr(y \mid do(x), z) \\ &= \Pr(y \mid x, do(z)) \\ &= \Pr(y \mid x, z)\end{aligned}$$

The causal effect of X on Y in figure 2.10(h) is confounded similarly to figure 2.10(g). The front-door W requires a back-door covariate Z but Z is a collider with respect to the two bows and conditioning on Z creates a back-door through the two bows with respect to $\Pr(W \mid do(x))$. Hence, the causal effect of X on Y is inescapably confounded. Further, no combination of action X and action or covariate W and/or Z identifies a conditional causal effect for figure 2.10(h).

To this point, we've confined attention to DAGs and nonparametric identification of causal effects. Next, we explore causal effects for a parametric, linear model involving a cyclic graph.

3 Structural models and counterfactuals

Structural models are fundamentally linked to causal graphs. Invariant functional relations are represented by arcs in the graph. Pearl [2010] describes three hierarchical elements to causal effects: action, prediction, and counterfactual. We illustrate these ideas via a simple model of economic equilibrium.

Consider the following system of Gaussian-linear equations depicting supply and demand for a product (Goldberger [1992]).

$$\begin{aligned}q &= b_1 p + d_1 i + u_1 \\ p &= b_2 q + d_2 w + u_2\end{aligned}$$

where q is household demand for the product, p is the unit price for the product, i is household income, w is the wage rate for producing the product, and u_1 and u_2 are error terms — unobserved or omitted terms associated with supply and demand for the product. The cyclic graph represented by this system is in figure 3.1.

As I and W are exogenous while Q and P are endogenous, we can rewrite the system of equations[12]

$$Y = AY + \varepsilon$$

or

$$(I_2 - A) Y = \varepsilon$$

[12] Since household income is denoted I, the identify matrix (also denoted I) is subscripted by its dimension.

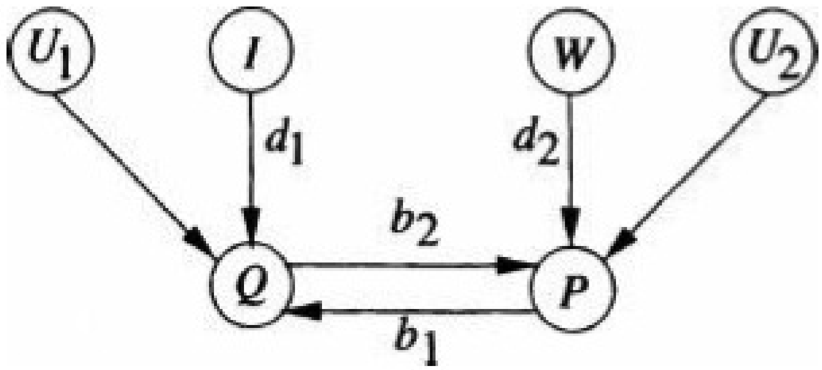


Figure 3.1: Equilibrium DAG

where

$$Y = \begin{bmatrix} Q \\ P \end{bmatrix},$$

$$A = \begin{bmatrix} 0 & b_1 \\ b_2 & 0 \end{bmatrix},$$

and

$$\varepsilon = D_I (X + U) = \begin{bmatrix} d_1 I + U_1 \\ d_2 W + U_2 \end{bmatrix}$$

for

$$\begin{aligned} D_I &= \begin{bmatrix} D & I_2 \end{bmatrix} \\ &= \begin{bmatrix} d_1 & 0 & 1 & 0 \\ 0 & d_2 & 0 & 1 \end{bmatrix}, \\ X &= \begin{bmatrix} I \\ W \end{bmatrix}, \\ U &= \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} \end{aligned}$$

Then,

$$Y = (I_2 - A)^{-1} \varepsilon$$

where

$$E[Y] = (I_2 - A)^{-1} E[\varepsilon]$$

and

$$Var[Y] = \Sigma_{YY} = (I_2 - A)^{-1} D_I \Sigma_{\varepsilon\varepsilon} D_I^T \left((I_2 - A)^{-1} \right)^T$$

for

$$\begin{aligned}
\Sigma_{\varepsilon\varepsilon} &= \begin{bmatrix} \Sigma_{XX} & \Sigma_{XU} \\ \Sigma_{UX} & \Sigma_{UU} \end{bmatrix} \\
&= \begin{bmatrix} \sigma_I^2 & 0 & 0 & 0 \\ 0 & \sigma_W^2 & 0 & 0 \\ 0 & 0 & \sigma_{U_1}^2 & 0 \\ 0 & 0 & 0 & \sigma_{U_2}^2 \end{bmatrix}
\end{aligned}$$

3.1 Joint distribution for (Y, X, U)

The joint distribution of $\begin{bmatrix} Y \\ X \\ U \end{bmatrix}$ is multivariate normal with mean vector $\begin{bmatrix} E[Q] \\ E[P] \\ E[I] \\ E[W] \\ E[U_1] \\ E[U_2] \end{bmatrix}$

and variance $\begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} & \Sigma_{YU} \\ \Sigma_{XY} & \Sigma_{XX} & \Sigma_{XU} \\ \Sigma_{UY} & \Sigma_{UX} & \Sigma_{UU} \end{bmatrix}$ where

$$\begin{aligned}
& \Sigma_{YY} \\
= \; & (I_2 - A)^{-1} D_I \Sigma_{\varepsilon\varepsilon} D_I^T \left((I_2 - A)^{-1} \right)^T \\
= \; & \frac{1}{(1 - b_1 b_2)^2} \times \\
& \begin{bmatrix} \sigma_{U_1}^2 + b_1^2 \sigma_{U_2}^2 + d_1^2 \sigma_I^2 + b_1^2 d_2 \sigma_W^2 & b_2 \sigma_{U_1}^2 + b_1 \sigma_{U_2}^2 + b_2 d_1^2 \sigma_I^2 + b_1 d_2^2 \sigma_W^2 \\ b_2 \sigma_{U_1}^2 + b_1 \sigma_{U_2}^2 + b_2 d_1^2 \sigma_I^2 + b_1 d_2^2 \sigma_W^2 & b_2^2 \sigma_{U_1}^2 + \sigma_{U_2}^2 + b_2^2 d_1^2 \sigma_I^2 + d_2^2 \sigma_W^2 \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
\Sigma_{YX} &= (I_2 - A)^{-1} D \Sigma_{XX} \\
&= \frac{1}{1 - b_1 b_2} \begin{bmatrix} d_1 \sigma_I^2 & b_1 d_2 \sigma_W^2 \\ b_2 d_1 \sigma_I^2 & d_2 \sigma_W^2 \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
\Sigma_{YU} &= (I_2 - A)^{-1} \Sigma_{UU} \\
&= \frac{1}{1 - b_1 b_2} \begin{bmatrix} \sigma_{U_1}^2 & b_1 \sigma_{U_2}^2 \\ b_2 \sigma_{U_1}^2 & \sigma_{U_2}^2 \end{bmatrix}
\end{aligned}$$

Then, for example, conditioning on observation of $I = i$ and $W = w$ leads to

$$E[Y \mid I = i, W = w] = E[Y] + \Sigma_{YX} \Sigma_{XX}^{-1} \left(\begin{bmatrix} i \\ w \end{bmatrix} - E \begin{bmatrix} I \\ W \end{bmatrix} \right)$$

3.2 Action, prediction, counterfactual queries

We pose three questions.

1. What is expected demand Q if price is controlled at $P = p_0$? [action]
2. What is expected demand Q if price is reported at $P = p_0$? [prediction]
3. Given a current price $P = p_0$, what is expected demand Q if we were to control the price at $P = p_1$? [counterfactual]

The first query (intervention prior to observation) involves $do\,(P = p_0)$ and changes the equations (eliminates paths from ancestors to P in the graph; that is, $Q \overset{b_2}{\to} P$, $W \overset{d_2}{\to} P$, and $U_2 \to P$ are effectively removed from the graph) to

$$\begin{aligned} q &= b_1 p_0 + d_1 i + u_1 \\ p &= p_0 \end{aligned}$$

Expected demand given $I = i$ and $do\,(P = p_0)$ is

$$E\left[Q \mid do\,(P = p_o)\,, I = i\right] = b_1 p_0 + d_1 i + E\left[U_1 \mid I = i\right]$$

Since U_1 and I are independent (depicted in the graph),

$$\begin{aligned} E\left[U_1 \mid I = i\right] &= E\left[U_1\right] \\ E\left[U_1\right] &= E\left[Q\right] - b_1 E\left[P\right] - d_1 E\left[I\right] \end{aligned}$$

and

$$E\left[Q \mid do\,(P = p_o)\,, I = i\right] = E\left[Q\right] + b_1\left(p_0 - E\left[P\right]\right) + d_1\left(i - E\left[I\right]\right) \qquad \text{(action)}$$

The second query (observation without intervention) is more standard in the econometrics literature and fundamentally different from the first.

$$E\left[Q \mid P = p_o, I = i, W = w\right] = b_1 p_0 + d_1 i + E\left[U_1 \mid P = p_o, I = i, W = w\right] \qquad \text{(prediction)}$$

where for the joint distribution assigned

$$\begin{aligned} & E\left[U_1 \mid P = p_o, I = i, W = w\right] \\ = \; & \frac{b_2 \sigma_{U_1}^2 \left\{b_2 E\left[Q\right] + p_0 - E\left[P\right] - d_2\left(w - E\left[W\right]\right) - b_2\left(b_1 p_0 + d_1 i\right)\right\}}{b_2^2 \sigma_{U_1}^2 + \sigma_{U_2}^2} \\ & + \frac{\sigma_{U_2}^2 \left\{E\left[Q\right] - b_1 E\left[P\right] - d_1 E\left[I\right]\right\}}{b_2^2 \sigma_{U_1}^2 + \sigma_{U_2}^2} \end{aligned}$$

Unlike query one, p_0 affects Q (through $E\left[U_1 \mid P = p_o, I = i, W = w\right]$) even when $b_1 = 0$.

Clearly, upon observing and conditioning on $P = p_0$, U is no longer independent of X. This is confirmed by d-connectedness/d-separation from the graph. For $Z = \emptyset$, U_1 is d-separated from I (as well as W and U_2) and U_2 is d-separated from W (as well as I and U_1). Hence, these variables are unconditionally uncorrelated as expressed in $\Sigma_{\varepsilon\varepsilon}$. However, P is a collider and conditioning on P

d-connects all variables. Consequently, all of these variables have nonzero correlation (unless the appropriate coefficient, b_1, d_1, or d_2, equals zero) as indicated by their variance conditional on P.

$$\Sigma_{IWU_1U_2|P} =$$

$$\frac{1}{b_2^2\sigma_{U_1}^2 + \sigma_{U_2}^2 + b_2^2 d_1^2 \sigma_I^2 + d_2^2 \sigma_W^2} \times \begin{bmatrix} \alpha_1 & -b_2 d_1 d_2 \sigma_I^2 \sigma_W^2 & -b_2^2 d_1 \sigma_{U_1}^2 \sigma_I^2 & -b_2 d_1 \sigma_{U_2}^2 \sigma_I^2 \\ -b_2 d_1 d_2 \sigma_I^2 \sigma_W^2 & \alpha_2 & -b_2 d_2 \sigma_{U_1}^2 \sigma_W^2 & -d_2 \sigma_{U_2}^2 \sigma_W^2 \\ -b_2^2 d_1 \sigma_{U_1}^2 \sigma_I^2 & -b_2 d_2 \sigma_{U_1}^2 \sigma_W^2 & \alpha_3 & -b_2 \sigma_{U_1}^2 \sigma_{U_2}^2 \\ -b_2 d_1 \sigma_{U_2}^2 \sigma_I^2 & -d_2 \sigma_{U_2}^2 \sigma_W^2 & -b_2 \sigma_{U_1}^2 \sigma_{U_2}^2 & \alpha_4 \end{bmatrix}$$

where

$$\begin{aligned} \alpha_1 &= \sigma_I^2 (b_2^2 \sigma_{U_1}^2 + \sigma_{U_1}^2 + d_2^2 \sigma_w^2) \\ \alpha_2 &= \sigma_W^2 (\sigma_{U_2}^2 + b_2^2 (\sigma_{U_1}^2 + d_1^2 \sigma_I^2)) \\ \alpha_3 &= \sigma_{U_1}^2 (\sigma_{U_2}^2 + b_2^2 d_1^2 \sigma_I^2 + d_2^2 \sigma_W^2) \end{aligned}$$

and

$$\alpha_4 = \sigma_{U_2}^2 (b_2^2 (\sigma_{U_1}^2 + d_1^2 \sigma_I^2) + d_2^2 \sigma_W^2)$$

The third query (observation followed by intervention) asks the counterfactual expectation of $Q_{P=p_1}$ conditional on observing $(P = p_o, I = i, W = w)$.

$$E[Q_{P=p_1} \mid P = p_o, I = i, W = w] = b_1 p_1 + d_1 i + E[U_1 \mid P = p_o, I = i, W = w] \quad \text{(counterfactual)}$$

where everything is the same as for query two except $b_1 p_0$ is replaced by $b_1 p_1$.

4 Partial compliance and bounding

Suppose treatment is randomly assigned and indicated by Z (independent of other factors U, X is a collider, — Z serves as an instrumental variable), X indicates treatment received, and Y is observed outcome as depicted by the causal graph in figure 4.1.

Imperfect compliance can lead to partial (rather than point) identification (upper and lower bounds) of the causal effect.

Consider the case where X, Y, and Z are binary, z_1 (z_0) indicates treatment is assigned (not assigned), x_1 (x_0) indicates treatment is administered (not administered), y_1 (y_0) indicates a positive (negative) response. U, on the other hand, may combine discrete and continuous (unspecified) random variables.

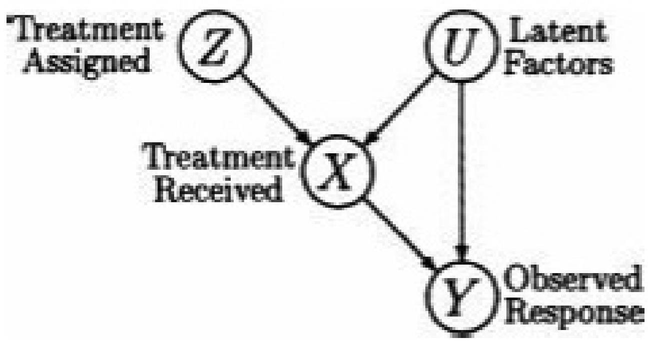


Figure 4.1: Partial compliance DAG

The joint distribution decomposes as

$$\Pr(y, x, z, u) = \Pr(y \mid x, z, u)\Pr(x \mid z, u)\Pr(z \mid u)\Pr(u)$$

Since Z is related to Y only through X, and Z and U are marginally independent (X is a collider) we have

$$\Pr(y \mid x, z, u) = \Pr(y \mid x, u)$$

and

$$\Pr(z \mid u) = \Pr(z)$$

Therefore, the decomposition becomes

$$\Pr(y, x, z, u) = \Pr(y \mid x, u)\Pr(x \mid z, u)\Pr(z)\Pr(u)$$

which is still unobservable as U is unobserved. However, the conditional distributions are observable

$$\Pr(y, x \mid z_j) = \sum_u \Pr(y \mid x, u)\Pr(x \mid z_j, u)\Pr(u), \quad j = 0, 1$$

With imperfect compliance, our challenge is to bound the quantities of interest based on the observable distributions $\Pr(y, x \mid z_0)$ and $\Pr(y, x \mid z_1)$. We try to utilize these distributions along with do-calculus rule 2 to address the unobservable causal effects

$$\begin{aligned}
\Pr(y_1 \mid do(x_j)) &= \sum_u \Pr(y_1 \mid do(x_j), u)\Pr(u \mid do(x_j)) \\
&= \sum_u \Pr(y_1 \mid do(x_j), u)\Pr(u) \\
&= \sum_u \Pr(y_1 \mid x_j, u)\Pr(u)
\end{aligned}$$

The second line follows from rule 3 as $(X \perp U \mid \emptyset)_{G_{\overline{X}}}$ where Y is a collider. The last line employs rule 2 where $(Y \perp X \mid U)_{G_{\underline{X}}}$. Then, we attempt to identify the average change in Y due to treatment (the average causal effect) from these distributions.

$$\begin{aligned} ACE(X \to Y) &= \Pr(y_1 \mid do(x_1)) - \Pr(y_1 \mid do(x_0)) \\ &= \sum_u \{\Pr(y_1 \mid x_1, u) - \Pr(y_1 \mid x_0, u)\} \Pr(u) \end{aligned}$$

Unfortunately, causal effects cannot be directly addressed as the (back-door adjustment) equation involves unobservables, U. As a result we're left to work with the observable conditional distributions. This connection involves multiple steps, bear with us.

The structural equation associated with two binary variables is

$$y = f(x, u)$$

which simplifies to four possible cases.

$$\begin{aligned} f_0 &: \ y = 0 \\ f_1 &: \ y = x \\ f_2 &: \ y \neq x \\ f_3 &: \ y = 1 \end{aligned}$$

where regardless of its rich composition, U simply combines with binary X to produce binary Y.

4.1 Canonical representation for U

Let the variable R_x represent compliance behavior where $r_x = 0, 1, 2, 3$ represents a never-taker, complier, defier, and always-taker, respectively. Then,

$$x = f_X(z, x_r) = \begin{array}{lll} x_0 & \text{if} & r_x = 0, \\ x_0 & \text{if} & r_x = 1 \text{ and } Z = z_0, \\ x_1 & \text{if} & r_x = 1 \text{ and } Z = z_1, \\ x_1 & \text{if} & r_x = 2 \text{ and } Z = z_0, \\ x_0 & \text{if} & r_x = 2 \text{ and } Z = z_1, \\ x_1 & \text{if} & r_x = 3 \end{array}$$

Likewise, the variable R_y represents response behavior to treatment where $r_y = 0, 1, 2, 3$ represents never-recover, helped, harmed, always-recover.

$$y = f_Y(x, y_r) = \begin{array}{lll} y_0 & \text{if} & r_y = 0, \\ y_0 & \text{if} & r_y = 1 \text{ and } X = x_0, \\ y_1 & \text{if} & r_y = 1 \text{ and } X = x_1, \\ y_1 & \text{if} & r_y = 2 \text{ and } X = x_0, \\ y_0 & \text{if} & r_y = 2 \text{ and } X = x_1, \\ y_1 & \text{if} & r_y = 3 \end{array}$$

These mapping variables replace U in the graph depicted in figure 4.2 where the double arrow (bow) between R_x and R_y indicates they may not be independent.

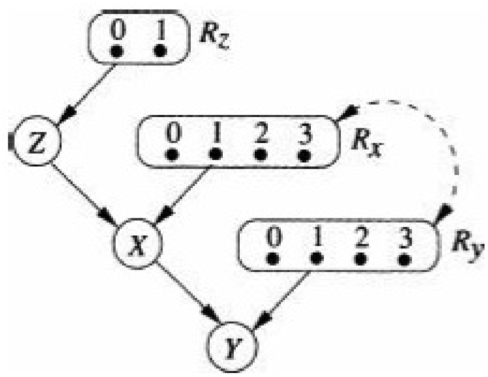


Figure 4.2: Canonical representation for U

Since

$$\Pr(y_1 \mid do(x_1)) = \Pr(r_y = 1) + \Pr(r_y = 3)$$

and

$$\Pr(y_1 \mid do(x_0)) = \Pr(r_y = 2) + \Pr(r_y = 3)$$

the average causal effect of treatment is

$$\begin{aligned} ACE(X \to Y) &= \Pr(y_1 \mid do(x_1)) - \Pr(y_1 \mid do(x_0)) \\ &= \Pr(r_y = 1) - \Pr(r_y = 2) \end{aligned}$$

4.2 Natural bounds

Pearl [1995] identifies natural bounds for the average causal effect, $ACE(X \to Y)$.[13]

$$\begin{aligned} & \Pr(y_1 \mid z_1) - \Pr(y_1 \mid z_0) - \Pr(y_1, x_0 \mid z_1) - \Pr(y_0, x_1 \mid z_0) \\ \le \ & ACE(X \to Y) \le \\ & \Pr(y_1 \mid z_1) - \Pr(y_1 \mid z_0) + \Pr(y_0, x_0 \mid z_1) + \Pr(y_1, x_1 \mid z_0) \end{aligned}$$

The natural lower bound is success following compliant treatment

$$\begin{aligned} \min \Pr(y_1 \mid do(x_1)) &= \Pr(y_1 \mid z_1) - \Pr(y_1, x_0 \mid z_1) \\ &= \Pr(y_1, x_1 \mid z_1) \end{aligned}$$

[13] See the appendix for a sketch of Pearl's proof for the natural bounds on $ACE(X \to Y)$.

less the complement of failure following compliant nontreatment

$$\begin{aligned}\max \Pr(y_1 \mid do(x_0)) &= \Pr(y_1 \mid z_0) + \Pr(y_0, x_1 \mid z_0) \\ &= 1 - \Pr(y_0, x_0 \mid z_0)\end{aligned}$$

The natural upper bound is analogous. It leads with the complement to failure following compliant treatment

$$\begin{aligned}\max \Pr(y_1 \mid do(x_1)) &= \Pr(y_1 \mid z_1) + \Pr(y_0, x_0 \mid z_1) \\ &= 1 - \Pr(y_0, x_1 \mid z_1)\end{aligned}$$

then subtracts success following compliant nontreatment

$$\begin{aligned}\min \Pr(y_1 \mid do(x_0)) &= \Pr(y_1 \mid z_0) - \Pr(y_1, x_1 \mid z_0) \\ &= \Pr(y_1, x_0 \mid z_0)\end{aligned}$$

The width of the natural bounds is determined by the rate of noncompliance $\Pr(x_1 \mid z_0) + \Pr(x_0 \mid z_1)$.

$$\begin{aligned}& \textit{upper bound} - \textit{lower bound} \\ = \ & \{\Pr(y_1 \mid z_1) - \Pr(y_1 \mid z_0) + \Pr(y_0, x_0 \mid z_1) + \Pr(y_1, x_1 \mid z_0)\} \\ & - \{\Pr(y_1 \mid z_1) - \Pr(y_1 \mid z_0) - \Pr(y_1, x_0 \mid z_1) - \Pr(y_0, x_1 \mid z_0)\} \\ = \ & \{\Pr(y_0, x_0 \mid z_1) + \Pr(y_1, x_1 \mid z_0)\} - \{-\Pr(y_1, x_0 \mid z_1) - \Pr(y_0, x_1 \mid z_0)\}\end{aligned}$$

Summing over Y produces the width on the bounds.

$$\begin{aligned}& \{\Pr(y_0, x_0 \mid z_1) + \Pr(y_1, x_0 \mid z_1)\} + \{\Pr(y_1, x_1 \mid z_0) + \Pr(y_0, x_1 \mid z_0)\} \\ = \ & \Pr(x_0 \mid z_1) + \Pr(x_1 \mid z_0)\end{aligned}$$

Or

$$\begin{aligned}& \textit{upper bound} - \textit{lower bound} \\ = \ & \{1 - \Pr(y_0, x_1 \mid z_1) - \Pr(y_1, x_0 \mid z_0)\} \\ & - \{\Pr(y_1, x_1 \mid z_1) - [1 - \Pr(y_0, x_0 \mid z_0)]\} \\ = \ & \{1 - \Pr(y_0, x_1 \mid z_1) - \Pr(y_1, x_1 \mid z_1)\} \\ & - \{\Pr(y_1, x_0 \mid z_0) - [1 - \Pr(y_0, x_0 \mid z_0)]\} \\ = \ & \{1 - \Pr(x_1 \mid z_1)\} - \{\Pr(x_0 \mid z_0) - 1\} \\ = \ & \{\Pr(x_0 \mid z_1)\} + \{\Pr(x_1 \mid z_0)\}\end{aligned}$$

4.3 Linear programming bounds

We can tighten these bounds if we formulate upper and lower bounds on the average causal effect of treatment via a linear objective function subject to linear constraints, thus, a linear programming formulation for the bounds (Balke and Pearl [1997]). The conditional distributions $\Pr(y, x \mid z_j), j = 0, 1$ can be written

$$\begin{array}{ll} p_{00.0} = \Pr(y_0, x_0 \mid z_0), & p_{00.1} = \Pr(y_0, x_0 \mid z_1), \\ p_{01.0} = \Pr(y_0, x_1 \mid z_0), & p_{01.1} = \Pr(y_0, x_1 \mid z_1), \\ p_{10.0} = \Pr(y_1, x_0 \mid z_0), & p_{10.1} = \Pr(y_1, x_0 \mid z_1), \\ p_{11.0} = \Pr(y_1, x_1 \mid z_0), & p_{11.1} = \Pr(y_1, x_1 \mid z_1) \end{array}$$

Now, let the joint probability $\Pr(r_x, r_y)$ be written

$$q_{jk} = \Pr(r_x = j, r_y = k), \quad j, k = 0, 1, 2, 3$$

This leads to

$$\begin{aligned} \Pr(y_1 \mid do(x_1)) &= \Pr(r_y = 1) + \Pr(r_y = 3) \\ &= q_{01} + q_{11} + q_{21} + q_{31} + q_{03} + q_{13} + q_{23} + q_{33} \end{aligned}$$

and

$$\begin{aligned} \Pr(y_1 \mid do(x_0)) &= \Pr(r_y = 2) + \Pr(r_y = 3) \\ &= q_{02} + q_{12} + q_{22} + q_{32} + q_{03} + q_{13} + q_{23} + q_{33} \end{aligned}$$

Equality constraints can be written $Rq = p$

$$\begin{array}{ll} p_{00.0} = q_{00} + q_{01} + q_{10} + q_{11}, & p_{00.1} = q_{00} + q_{01} + q_{20} + q_{21}, \\ p_{01.0} = q_{20} + q_{22} + q_{30} + q_{32}, & p_{01.1} = q_{10} + q_{12} + q_{30} + q_{32}, \\ p_{10.0} = q_{02} + q_{03} + q_{12} + q_{13}, & p_{10.1} = q_{02} + q_{03} + q_{22} + q_{23}, \\ p_{11.0} = q_{21} + q_{23} + q_{31} + q_{33}, & p_{11.1} = q_{11} + q_{13} + q_{31} + q_{33} \end{array}$$

The linear program for the lower bound on $\Pr(y_1 \mid do(x_1))$ is[14]

$$\begin{array}{ll} \min\limits_{q \geq 0} & q_{01} + q_{11} + q_{21} + q_{31} + q_{03} + q_{13} + q_{23} + q_{33} \\ s.t. & Rq = p \\ & \sum\limits_{j=0}^{3} \sum\limits_{k=0}^{3} q_{jk} = 1 \end{array}$$

The solution for the lower bound on $\Pr(y_1 \mid do(x_1))$ can be expressed

$$\Pr(y_1 \mid do(x_1)) \geq \max \left\{ \begin{array}{c} p_{11.1}, \\ p_{11.0}, \\ -p_{00.0} - p_{01.0} + p_{00.1} + p_{11.1}, \\ -p_{01.0} - p_{10.0} + p_{10.1} + p_{11.1} \end{array} \right\}$$

The upper bound on $\Pr(y_1 \mid do(x_1))$ is

$$\Pr(y_1 \mid do(x_1)) \leq \min \left\{ \begin{array}{c} 1 - p_{01.1}, \\ 1 - p_{01.0}, \\ p_{00.0} + p_{11.0} + p_{10.1} + p_{11.1}, \\ p_{10.0} + p_{11.0} + p_{00.1} + p_{11.1}, \end{array} \right\}$$

[14] There are seven linearly independent equations in the constraints for both the primal and dual programs. See the appendix for details on the derivation of the linear programming bounds.

The linear program for the lower bound on $\Pr(y_1 \mid do(x_0))$ is

$$\begin{array}{ll} \min\limits_{q \geq 0} & q_{02} + q_{12} + q_{22} + q_{32} + q_{03} + q_{13} + q_{23} + q_{33} \\ s.t. & Rq = p \\ & \sum_{j=0}^{3} \sum_{k=0}^{3} q_{jk} = 1 \end{array}$$

The solution for the lower bound on $\Pr(y_1 \mid do(x_0))$ can be expressed

$$\Pr(y_1 \mid do(x_0)) \geq \max \left\{ \begin{array}{c} p_{10.1}, \\ p_{10.0}, \\ p_{10.0} + p_{11.0} - p_{00.1} - p_{11.1}, \\ p_{01.0} + p_{10.0} - p_{00.1} - p_{01.1} \end{array} \right\}$$

The upper bound on $\Pr(y_1 \mid do(x_0))$ is

$$\Pr(y_1 \mid do(x_0)) \leq \min \left\{ \begin{array}{c} 1 - p_{00.1}, \\ 1 - p_{00.0}, \\ p_{01.0} + p_{10.0} + p_{10.1} + p_{11.1}, \\ p_{10.0} + p_{11.0} + p_{01.1} + p_{10.1}, \end{array} \right\}$$

Since the average causal effect is the difference between these two quantities, bounds for $ACE(X \to Y)$ can be found based on the above. The lower bound for $ACE(X \to Y)$ is

$$\begin{aligned} ACE(X \to Y)_{lower} &= \Pr(y_1 \mid do(x_1))_{lower} - \Pr(y_1 \mid do(x_0))_{upper} \\ &= \max \left\{ \begin{array}{c} p_{11.1}, \\ p_{11.0}, \\ -p_{00.0} - p_{01.0} + p_{00.1} + p_{11.1}, \\ -p_{01.0} - p_{10.0} + p_{10.1} + p_{11.1} \end{array} \right\} \\ &\quad - \min \left\{ \begin{array}{c} 1 - p_{00.1}, \\ 1 - p_{00.0}, \\ p_{01.0} + p_{10.0} + p_{10.1} + p_{11.1}, \\ p_{10.0} + p_{11.0} + p_{01.1} + p_{10.1}, \end{array} \right\} \end{aligned}$$

while the upper bound for $ACE(X \to Y)$ is

$$\begin{aligned} ACE(X \to Y)_{upper} &= \Pr(y_1 \mid do(x_1))_{upper} - \Pr(y_1 \mid do(x_0))_{lower} \\ &= \min \left\{ \begin{array}{c} 1 - p_{01.1}, \\ 1 - p_{01.0}, \\ p_{00.0} + p_{11.0} + p_{10.1} + p_{11.1}, \\ p_{10.0} + p_{11.0} + p_{00.1} + p_{11.1}, \end{array} \right\} \\ &\quad - \max \left\{ \begin{array}{c} p_{10.1}, \\ p_{10.0}, \\ p_{10.0} + p_{11.0} - p_{00.1} - p_{11.1}, \\ p_{01.0} + p_{10.0} - p_{00.1} - p_{01.1} \end{array} \right\} \end{aligned}$$

Alternatively, the average causal effect can be written

$$\begin{aligned}ACE\left(X \rightarrow Y\right) &= \Pr\left(r_y = 1\right) - \Pr\left(r_y = 2\right) \\ &= q_{01} + q_{11} + q_{21} + q_{31} - \left(q_{02} + q_{12} + q_{22} + q_{32}\right)\end{aligned}$$

Then, the linear program for the lower bound is

$$\begin{array}{ll}\min\limits_{q \geq 0} & q_{01} + q_{11} + q_{21} + q_{31} - \left(q_{02} + q_{12} + q_{22} + q_{32}\right) \\ s.t. & Rq = p \\ & \sum\limits_{j=0}^{3} \sum\limits_{k=0}^{3} q_{jk} = 1\end{array}$$

The solution for the lower bound of $ACE\left(X \rightarrow Y\right)$ can be expressed

$$ACE\left(X \rightarrow Y\right) \geq \max \left\{ \begin{array}{c} p_{11.1} + p_{00.0} - 1, \\ p_{11.0} + p_{00.1} - 1, \\ p_{11.0} - p_{11.1} - p_{10.1} - p_{01.0} - p_{10.0}, \\ p_{11.1} - p_{11.0} - p_{10.0} - p_{01.1} - p_{10.1}, \\ -p_{01.1} - p_{10.1}, \\ -p_{01.0} - p_{10.0}, \\ p_{00.1} - p_{01.1} - p_{10.1} - p_{01.0} - p_{00.0}, \\ p_{00.0} - p_{01.0} - p_{10.0} - p_{01.1} - p_{00.1} \end{array} \right\}$$

The upper bound program for $ACE\left(X \rightarrow Y\right)$ is analogous with solution

$$ACE\left(X \rightarrow Y\right) \leq \min \left\{ \begin{array}{c} 1 - p_{01.1} - p_{10.0}, \\ 1 - p_{01.0} + p_{10.1}, \\ -p_{01.0} + p_{01.1} + p_{00.1} + p_{11.0} + p_{00.0}, \\ -p_{01.1} + p_{11.1} + p_{00.1} + p_{01.0} + p_{00.0}, \\ p_{11.1} + p_{00.1}, \\ p_{11.0} + p_{00.0}, \\ -p_{10.1} + p_{11.1} + p_{00.1} + p_{11.0} + p_{10.0}, \\ -p_{10.0} + p_{11.0} + p_{00.0} + p_{11.1} + p_{10.1} \end{array} \right\}$$

Next, we append to this discussion a brief reference to the average treatment effect on the treated.

4.4 Treatment on the treated

When an analyst is interested in the effect of an existing program under its current incentive system and current participants then the quantity of interest is treatment on treated rather than the average treatment effect (which gauges the effect of introducing a new program randomly over the population). Treatment on the treated is quantified as

$$\begin{aligned}ETT\left(X \rightarrow Y\right) &= \Pr\left(Y_{\widehat{x}_1} = y_1 \mid x_1\right) - \Pr\left(Y_{\widehat{x}_0} = y_1 \mid x_1\right) \\ &= \sum_u \left[\Pr\left(y_1 \mid x_1, u\right) - \Pr\left(y_1 \mid x_0, u\right)\right] \Pr\left(u \mid x_1\right)\end{aligned}$$

where $Y_{\widehat{x}_j}$ refers to outcome when intervening with action x_j and conditioning on x_1 refers to the subpopulation observed in the treatment regime.

Pearl [1995] derives bounds on $ETT\,(X \to Y)$[15]

$$\begin{aligned} & \frac{\Pr(y_1 \mid z_1) - \Pr(y_1 \mid z_0)}{\Pr(x_1)/\Pr(z_1)} - \frac{\Pr(y_0, x_1 \mid z_0)}{\Pr(x_1)} \\ \leq \quad & ETT\,(X \to Y) \leq \\ & \frac{\Pr(y_1 \mid z_1) - \Pr(y_1 \mid z_0)}{\Pr(x_1)/\Pr(z_1)} + \frac{\Pr(y_1, x_1 \mid z_0)}{\Pr(x_1)} \end{aligned}$$

or

$$\begin{aligned} & \frac{(p_{11.1} + p_{10.1}) - (p_{11.0} + p_{10.0})}{[(p_{11.1} + p_{01.1})\Pr(z_1) + (p_{11.0} + p_{01.0})\Pr(z_0)]/\Pr(z_1)} \\ & - \frac{p_{01.0}}{(p_{11.1} + p_{01.1})\Pr(z_1) + (p_{11.0} + p_{01.0})\Pr(z_0)} \\ \leq \quad & ETT\,(X \to Y) \leq \\ & \frac{(p_{11.1} + p_{10.1}) - (p_{11.0} + p_{10.0})}{[(p_{11.1} + p_{01.1})\Pr(z_1) + (p_{11.0} + p_{01.0})\Pr(z_0)]/\Pr(z_1)} \\ & + \frac{p_{11.0}}{(p_{11.1} + p_{01.1})\Pr(z_1) + (p_{11.0} + p_{01.0})\Pr(z_0)} \end{aligned}$$

Alternatively, the bounds can be expressed[16]

$$\begin{aligned} & \frac{P(y_1) - P(x_1 \mid z_0) - P(y_1, x_0 \mid z_0)}{P(x_1)} \\ \leq \quad & ETT\,(X \to Y) \leq \\ & \frac{P(y_1) - P(y_1, x_0 \mid z_0)}{P(x_1)} \end{aligned}$$

[15] See the appendix for a sketch of Pearl's proof for the bounds on $ETT\,(X \to Y)$.

[16] The lower bound can be written

$$\begin{aligned} & \frac{\Pr(y_1 \mid z_1)\Pr(z_1) - \Pr(y_1 \mid z_0)[1 - \Pr(z_0)] - \Pr(y_0, x_1 \mid z_0)}{\Pr(x_1)} \\ = \quad & \frac{\Pr(y_1, z_1) + \Pr(y_1, z_0) - \Pr(y_1, x_1 \mid z_0) - \Pr(y_1, x_0 \mid z_0) - \Pr(y_0, x_1 \mid z_0)}{\Pr(x_1)} \\ = \quad & \frac{P(y_1) - P(x_1 \mid z_0) - P(y_1, x_0 \mid z_0)}{P(x_1)} \end{aligned}$$

By analogous derivation, the upper bound can be expressed

$$\begin{aligned} & \frac{\Pr(y_1 \mid z_1)\Pr(z_1) - \Pr(y_1 \mid z_0)[1 - \Pr(z_0)] + \Pr(y_1, x_1 \mid z_0)}{\Pr(x_1)} \\ = \quad & \frac{\Pr(y_1, z_1) + \Pr(y_1, z_0) - \Pr(y_1, x_1 \mid z_0) - \Pr(y_1, x_0 \mid z_0) + \Pr(y_1, x_1 \mid z_0)}{\Pr(x_1)} \\ = \quad & \frac{P(y_1) - P(y_1, x_0 \mid z_0)}{P(x_1)} \end{aligned}$$

Clearly, if $P(x_1 \mid z_0) = 0$ then $ETT(X \to Y)$ is point-identified.

Next, we consider an example to illustrate partial identification for both the average causal effect and treatment on the treated.

4.5 Partial identification example

Suppose the data generating process is as follows.

$$\begin{array}{cc} p_{00.0} = 0.919 & p_{00.1} = 0.315 \\ p_{01.0} = 0.000 & p_{01.1} = 0.139 \\ p_{10.0} = 0.081 & p_{10.1} = 0.073 \\ p_{11.0} = 0.000 & p_{11.1} = 0.473 \\ \Pr(z_1) = 0.500 & \Pr(z_0) = 0.500 \end{array}$$

The compliance rate

$$\begin{aligned} \Pr(x_1 \mid z_1) &= p_{11.1} + p_{01.1} \\ &= 0.473 + 0.139 \\ &= 0.612 \end{aligned}$$

the encouragement effect or intent to treat[17]

$$\begin{aligned} & \Pr(y_1 \mid z_1) - \Pr(y_1 \mid z_0) \\ =\ & (p_{11.1} + p_{10.1}) - (p_{11.0} + p_{10.0}) \\ =\ & (0.473 + 0.073) - (0.000 + 0.081) \\ =\ & 0.465 \end{aligned}$$

and the mean difference

$$\begin{aligned} & \Pr(y_1 \mid x_1) - \Pr(y_1 \mid x_0) \\ =\ & \frac{p_{11.1}\Pr(z_1) + p_{11.0}\Pr(z_0)}{(p_{11.1} + p_{01.1})\Pr(z_1) + (p_{11.0} + p_{01.0})\Pr(z_0)} \\ & - \frac{p_{10.1}\Pr(z_1) + p_{10.0}\Pr(z_0)}{(p_{10.1} + p_{00.1})\Pr(z_1) + (p_{10.0} + p_{00.0})\Pr(z_0)} \\ =\ & \frac{0.473 \cdot 0.5 + 0.0 \cdot 0.5}{(0.473 + 0.139)\,0.5 + (0.0 + 0.0)\,0.5} \\ & - \frac{0.073 \cdot 0.5 + 0.081 \cdot 0.5}{(0.073 + 0.315)\,0.5 + (0.081 + 0.919)\,0.5} \\ =\ & 0.661925 \end{aligned}$$

describe the data.

However, partial compliance implies the average causal effect cannot be point identified but substantial insight can be gained from its bounds.

$$0.392 \leq ACE(X \to Y) \leq 0.780$$

[17] Apparently, this is the measure employed by the Food and Drug Administration (FDA) in drug trials.

where

$$
\begin{aligned}
ACE\left(X \rightarrow Y\right)_{lower} &= \Pr\left(y_1 \mid do\left(x_1\right)\right)_{lower} - \Pr\left(y_1 \mid do\left(x_0\right)\right)_{upper} \\
&= \max\left\{\begin{array}{c} p_{11.1}, \\ p_{11.0}, \\ -p_{00.0} - p_{01.0} + p_{00.1} + p_{11.1}, \\ -p_{01.0} - p_{10.0} + p_{10.1} + p_{11.1} \end{array}\right\} \\
&\quad - \min\left\{\begin{array}{c} 1 - p_{00.1}, \\ 1 - p_{00.0}, \\ p_{01.0} + p_{10.0} + p_{10.1} + p_{11.1}, \\ p_{10.0} + p_{11.0} + p_{01.1} + p_{10.1}, \end{array}\right\} \\
&= \max\left\{\begin{array}{c} 0.473, \\ 0., \\ -0.131, \\ 0.465 \end{array}\right\} - \min\left\{\begin{array}{c} 0.685, \\ 0.081, \\ 0.627, \\ 0.293 \end{array}\right\} \\
&= 0.473 - 0.081 = 0.392
\end{aligned}
$$

and

$$
\begin{aligned}
ACE\left(X \rightarrow Y\right)_{upper} &= \Pr\left(y_1 \mid do\left(x_1\right)\right)_{upper} - \Pr\left(y_1 \mid do\left(x_0\right)\right)_{lower} \\
&= \min\left\{\begin{array}{c} 1 - p_{01.1}, \\ 1 - p_{01.0}, \\ p_{00.0} + p_{11.0} + p_{10.1} + p_{11.1}, \\ p_{10.0} + p_{11.0} + p_{00.1} + p_{11.1}, \end{array}\right\} \\
&\quad - \max\left\{\begin{array}{c} p_{10.1}, \\ p_{10.0}, \\ p_{10.0} + p_{11.0} - p_{00.1} - p_{11.1}, \\ p_{01.0} + p_{10.0} - p_{00.1} - p_{01.1} \end{array}\right\} \\
&= \min\left\{\begin{array}{c} 0.861, \\ 1., \\ 1.465, \\ 0.869 \end{array}\right\} - \max\left\{\begin{array}{c} 0.073, \\ 0.081, \\ -0.707, \\ -0.373 \end{array}\right\} \\
&= 0.861 - 0.081 = 0.780
\end{aligned}
$$

Remarkably, even though 38.8% failed to comply with treatment we can conclude that at least 39.2% of the population would benefit from treatment.

The natural bounds for $ACE(X \rightarrow Y)$ are the same as above.

$$\begin{aligned}
& \Pr(y_1 \mid z_1) - \Pr(y_1 \mid z_0) - \Pr(y_1, x_0 \mid z_1) - \Pr(y_0, x_1 \mid z_0) \\
= & \; p_{11.1} + p_{10.1} - (p_{11.0} + p_{10.0}) - p_{10.1} - p_{01.0} \\
= & \; 0.473 + 0.073 - (0.000 + 0.081) - 0.073 - 0.000 \\
= & \; 0.392 \\
\leq & \; ACE(X \rightarrow Y) \leq \\
& \Pr(y_1 \mid z_1) - \Pr(y_1 \mid z_0) + \Pr(y_0, x_0 \mid z_1) + \Pr(y_1, x_1 \mid z_0) \\
= & \; p_{11.1} + p_{10.1} - (p_{11.0} + p_{10.0}) + p_{00.1} + p_{11.0} \\
= & \; 0.473 + 0.073 - (0.000 + 0.081) + 0.315 + 0.000 \\
= & \; 0.780
\end{aligned}$$

Or

$$\begin{aligned}
& \Pr(y_1, x_1 \mid z_1) - [1 - \Pr(y_0, x_0 \mid z_0)] \\
= & \; p_{11.1} - 1 + p_{00.0} \\
= & \; 0.473 - 1 + 0.919 \\
= & \; 0.392 \\
\leq & \; ACE(X \rightarrow Y) \leq \\
& 1 - \Pr(y_0, x_1 \mid z_1) - \Pr(y_1, x_0 \mid z_0) \\
= & \; 1 - p_{01.1} - p_{10.0} \\
= & \; 1 - 0.139 - 0.081 \\
= & \; 0.780
\end{aligned}$$

The width of the interval equals the likelihood of noncompliance.

$$\begin{aligned}
\Pr(x_1 \mid z_0) + \Pr(x_0 \mid z_1) &= p_{11.0} + p_{01.0} + p_{10.1} + p_{00.1} \\
&= 0.000 + 0.000 + 0.073 + 0.315 \\
&= 0.780 - 0.392 = 0.388
\end{aligned}$$

Since $\Pr(x_1 \mid z_0) = 0$, $ETT(X \rightarrow Y)$, treatment effect on the treated, is point identified. The first term in the upper and lower bounds is the same, the only difference arises in the second term of each bound but when no one adopts treatment that was not assigned treatment this term vanishes in each bound

$(p_{01.0} = p_{11.0} = 0)$.

$$
\begin{aligned}
& \frac{(p_{11.1} + p_{10.1}) - (p_{11.0} + p_{10.0})}{[(p_{11.1} + p_{01.1}) \Pr(z_1) + (p_{11.0} + p_{01.0}) \Pr(z_0)] / \Pr(z_1)} \\
& - \frac{p_{01.0}}{(p_{11.1} + p_{01.1}) \Pr(z_1) + (p_{11.0} + p_{01.0}) \Pr(z_0)} \\
= \; & 0.759804 \\
\leq \; & ETT(X \to Y) \leq \\
& \frac{(p_{11.1} + p_{10.1}) - (p_{11.0} + p_{10.0})}{[(p_{11.1} + p_{01.1}) \Pr(z_1) + (p_{11.0} + p_{01.0}) \Pr(z_0)] / \Pr(z_1)} \\
& - \frac{p_{11.0}}{(p_{11.1} + p_{01.1}) \Pr(z_1) + (p_{11.0} + p_{01.0}) \Pr(z_0)} \\
= \; & 0.759804
\end{aligned}
$$

This indicates that 76% of current participants benefit from treatment.

4.6 Test of instruments

Conventional wisdom indicates unobservability makes empirical tests of instruments or model exogeneity unassailable. However, Balke and Pearl's bounds provide a test of severe model or instrument failure, that is, the test does not always identify poor instruments (or general model mis-specification) but in extreme cases it can be useful. The test requires that each of the conditions in the upper bound of $ACE(X \to Y)$ lie at or above each condition in the lower bound of $ACE(X \to Y)$. Balke and Pearl's conditions simplify as

$$
\begin{aligned}
p_{00.0} + p_{10.1} &\leq 1 \\
p_{10.0} + p_{00.1} &\leq 1 \\
p_{01.0} + p_{11.1} &\leq 1 \\
p_{11.0} + p_{01.1} &\leq 1
\end{aligned}
$$

when any of these conditions is violated the model is mis-specified.[18] Therefore, when one of these conditions fails the thought experiment reflected in our causal graph is inconsistent with the data.[19]

In particular, the first two upper and lower bounds on $\Pr(y_1 \mid do(x_0))$ yield two of the conditions

lower	*upper*
$p_{10.1}$	$1 - p_{00.0}$
$p_{10.0}$	$1 - p_{00.1}$

[18] When the observed probabilities (frequencies) are inconsistent with these conditions there is no feasible solution to the linear program (described above) from which the bounds on $ACE(X \to Y)$ are derived.

[19] Not only may the power of the test be somewhat lacking, but also sampling error can indicate model deficiency when there is none. In other words, as usual both kinds of errors (false model rejection and failure to detect model inadequacy) are possible.

and the other two conditions come from the first two upper and lower bounds on $\Pr(y_1 \mid do(x_1))$

lower	*upper*
$p_{11.1}$	$1 - p_{01.0}$
$p_{11.0}$	$1 - p_{01.1}$

where conditioning on z_0 is compared with conditioning on z_1 as otherwise the test is not diagnostic.

Instrument inequality defined by violation of any of the above conditions occurs when manipulation of the instrument has substantial impact on outcome but treatment remains unchanged. This is (weakly) consistent with traditional econometric intuition that forbids the instruments from impacting outcome conditional on treatment (and perhaps covariates).

These conditions generalize for multivalued X, Y, or Z as

$$\max_x \sum_y \left[\max_z \Pr(y, x \mid z)\right] \leq 1$$

and for continuous Y or Z (but X remains discrete[20]) as

$$\int_y \left[\max_z f(y \mid x, z) \Pr(x \mid z)\right] dy \leq 1 \quad \forall x$$

The instrument inequality can be tightened considerably if there are no defiers in the population, that is,

$$\Pr(x_1 \mid z_1, u) \geq \Pr(x_1 \mid z_0, u) \quad \forall u$$

Then, the instrument inequality is

$$\begin{array}{l} \Pr(y, x_1 \mid z_1) \geq \Pr(y, x_1 \mid z_0) \\ \Pr(y, x_0 \mid z_0) \geq \Pr(y, x_0 \mid z_1) \end{array} \forall y \in \{y_1, y_0\}$$

Violations of these conditions indicate selection bias, direct effect of Z on Y, or defiers in the population.

4.7 Gibbs sampler

Thus far we've discussed identification implications (or possibilities with unlimited sample size) of imperfect compliance. Now, we discuss a latent variable, finite-sample (of size n) estimation approach. Since identification revolves around 16 compliance-response (CR) pairs that are unobserved while only $p_{yx.z}$ is observed, the challenge is to recover $\Pr(v_{CR} \mid data = \{p_{yx.z}\})$ where v_{CR} refers to the frequency or probability of CR.

Following Chickering and Pearl [1997], we employ a two-step Gibbs sampler to address the latency of CR. A Gibbs sampler is a Markov chain Monte Carlo

[20] The density is unconstrained for continuous X negating the power of the test.

method that draws from the full set of conditional posterior distributions to eventually yield draws from the marginal posterior of interest.

The first conditional posterior is

$$\Pr\left(cr^i \mid v_{cr^i}, data = \{p_{yx.z}\}\right) \propto f\left(x^i, y^i \mid z^i, cr^i\right) v_{cr^i}$$

where the superscript refers to individual i in the sample and $f\left(x^i, y^i \mid z^i, cr^i\right)$ is an indicator function equal to one when x, y, z agrees with cr and zero otherwise. This is recognized as a multinomial distribution and generates values for the latent variable CR (the reason a Gibbs sampler is called upon in this setting; if CR values were observed we could simply employ posterior simulation).

Since v_{cr^i} is unknown we begin with some initial value and replace it in subsequent rounds with draws from the second conditional posterior distribution

$$\Pr\left(v_{CR} \mid cr^1, \ldots, cr^n\right) \propto \prod_{i=0}^{3}\prod_{j=0}^{3}\left(v_{cr_{ij}}\right)^{N_{ij}+N'_{ij}-1}$$

where N_{ij} refers to the number of draws corresponding to cr_{ij} from the first conditional posterior and N'_{ij} refers to the prior concentration parameter for a Dirichlet distribution (or can be thought of as the result of a previous experiment; our experiments employ $N'_{ij} = 1$ for all $v_{cr_{ij}}$ or uniform priors). This conditional posterior follows a Dirichlet distribution and generates v_{CR} draws. Since

$$\begin{aligned}
ACE\left(X \to Y\right) &= \sum_{i=0}^{3} v_{c=i,r=1} - \sum_{i=0}^{3} v_{c=i,r=2}, \\
\Pr\left(y_1 \mid do\left(x_1\right)\right) &= \sum_{i=0}^{3} v_{c=i,r=1} + \sum_{i=0}^{3} v_{c=i,r=3}, \\
\Pr\left(y_1 \mid do\left(x_0\right)\right) &= \sum_{i=0}^{3} v_{c=i,r=2} + \sum_{i=0}^{3} v_{c=i,r=3}
\end{aligned}$$

are deterministic functions of v_{CR} they can be simulated directly from v_{CR}.

Sampling is repeated a large number of times ($10,000$ in our case) and the first set of draws ($5,000$) are discarded as burn-in since some of them are unlikely to be representative of the distribution of interest.

We consider three examples. One in which the average causal effect is point-identified, a second in which $ACE\left(X \to Y\right)$ is partially-identified, and a third involving a counterfactual query regarding an individual in a partially-identified setting. We report histograms of $ACE\left(X \to Y\right), \Pr\left(y_1 \mid do\left(x_1\right)\right),$ and $\Pr\left(y_1 \mid do\left(x_0\right)\right)$ for samples of size $n = 100$ and $1,000$.

Example 5 (point-identified Gibbs sampler) *Suppose the data are generated from*

$$\begin{array}{cc}
p_{00.0} = 0.55 & p_{00.1} = 0.45 \\
p_{01.0} = 0.45 & p_{01.1} = 0. \\
p_{10.0} = 0. & p_{10.1} = 0. \\
p_{11.0} = 0. & p_{11.1} = 0.55 \\
\Pr\left(z_1\right) = 0.50 & \Pr\left(z_0\right) = 0.50
\end{array}$$

Then, the causal effects of interest are point-identified.

$$\begin{aligned} ACE\,(X \to Y) &= 0.55, \\ \Pr\,(y_1 \mid do\,(x_1)) &= 0.55, \\ \Pr\,(y_1 \mid do\,(x_0)) &= 0.0 \end{aligned}$$

The sample data generated as well as sample-implied bounds and estimated average causal effects from the Gibbs sampler simulations are as follows.

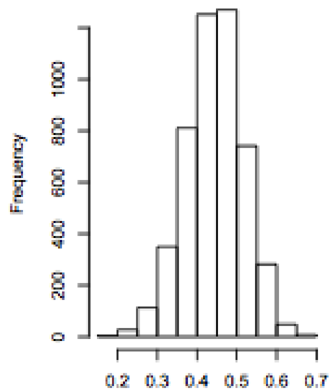
$N\,(y,x,z)$	$n = 100$	$n = 1,000$
$(0,0,0)$	30	295
$(0,1,0)$	23	224
$(1,0,0)$	0	0
$(1,1,0)$	0	0
$(0,0,1)$	20	221
$(0,1,1)$	0	0
$(1,0,1)$	0	0
$(1,1,1)$	27	260

	$n = 100$ *sample-implied bounds*[21]	$n = 100$ *mean estimate*	$n = 1,000$ *sample-implied bounds*	$n = 1,000$ *mean estimate*
$ACE\,(X \to Y)$	$(0.574, 0.558)$	0.444	$(0.541, 0.568)$	0.542
$\Pr\,(y_1 \mid do\,(x_1))$	$(0.574, 0.566)$	0.553	$(0.541, 0.568)$	0.554
$\Pr\,(y_1 \mid do\,(x_0))$	$(0.008, 0)$	0.109	0.	0.012

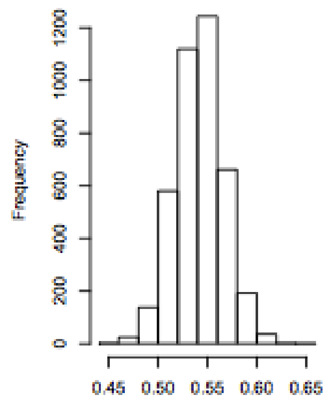
Since the samples are not balanced, the average causal effects with these large sample frequencies rather than the population frequencies (probabilities) would have narrow bounds (as indicated by the implied intervals) rather than be point-

[21] For the n=100 sample, frequency-implied bounds are reversed (with the lower bound slightly exceeding the upper bound) as a result of sampling variation (sampling error).

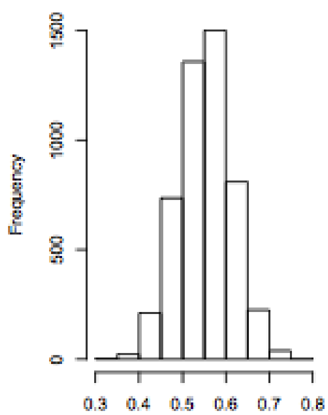
identified. Histograms for the average causal effects are below.



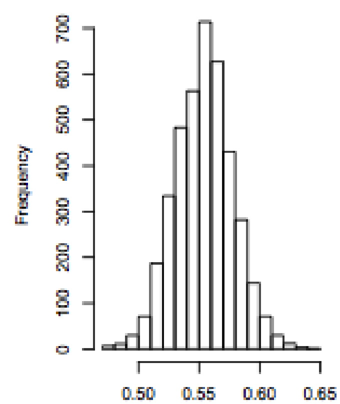
$ACE\,(X \rightarrow Y)\,, n = 100$



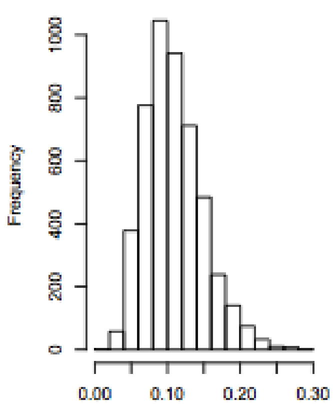
$ACE\,(X \rightarrow Y)\,, n = 1,000$



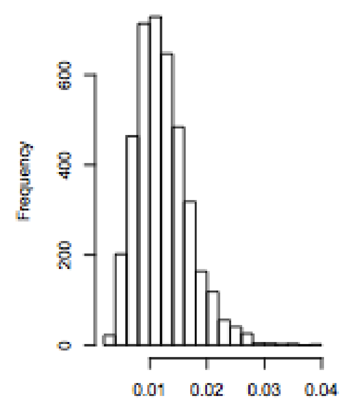
$\Pr\left(y_1 \mid do\left(x_1\right)\right), n = 100$



$\Pr\left(y_1 \mid do\left(x_1\right)\right), n = 1,000$



$\Pr\left(y_1 \mid do\left(x_0\right)\right), n = 100$



$\Pr\left(y_1 \mid do\left(x_0\right)\right), n = 1,000$

As expected, the intervals are much tighter for the larger samples but all are consistent with large-sample point-identification.

Example 6 (partially-identified Gibbs sampler) *Suppose the data are generated as in 4.4*

$$\begin{array}{cc} p_{00.0} = 0.919 & p_{00.1} = 0.315 \\ p_{01.0} = 0.000 & p_{01.1} = 0.139 \\ p_{10.0} = 0.081 & p_{10.1} = 0.073 \\ p_{11.0} = 0.000 & p_{11.1} = 0.473 \\ \Pr(z_1) = 0.500 & \Pr(z_0) = 0.500 \end{array}$$

Then, the causal effects of interest are partially-identified with bounds as reported earlier.

$$\begin{aligned} ACE(X \to Y) &= (0.392, 0.780), \\ \Pr(y_1 \mid do(x_1)) &= (0.473, 0.861), \\ \Pr(y_1 \mid do(x_0)) &= 0.081 \end{aligned}$$

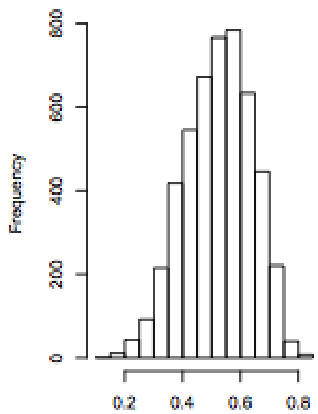
The sample data generated as well as sample-implied bounds and estimated average causal effects from the Gibbs sampler simulations are as follows.

$N(y,x,z)$	$n = 100$	$n = 1,000$
$(0,0,0)$	49	458
$(0,1,0)$	0	0
$(1,0,0)$	3	30
$(1,1,0)$	0	0
$(0,0,1)$	16	167
$(0,1,1)$	5	73
$(1,0,1)$	3	34
$(1,1,1)$	24	238

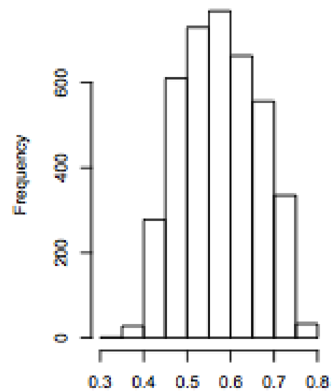
	$n = 100$ *sample-implied bounds*	$n = 100$ *mean estimate*	$n = 1,000$ *sample-implied bounds*	$n = 1,000$ *mean estimate*
$ACE(X \to Y)$	$(0.447, 0.829)$	0.523	$(0.408, 0.786)$	0.575
$\Pr(y_1 \mid do(x_1))$	$(0.505, 0.891)$	0.681	$(0.470, 0.852)$	0.657
$\Pr(y_1 \mid do(x_0))$ [22]	$(0.063, 0.058)$	0.157	$(0.066, 0.061)$	0.082

[22] These bounds are based on linear programming maxima and minima identified earlier. However, upper and lower bounds are reversed and with these frequencies there are no feasible solutions to the linear programs. This is suggestive of some model mis-specification or an unrepresentative sample. In this case, the inconsistency in the bounds, albeit a narrow discrepancy, is apparently due to sampling variation (or sampling error).

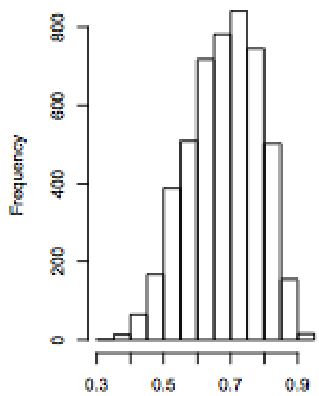
Histograms for the average causal effects are below.



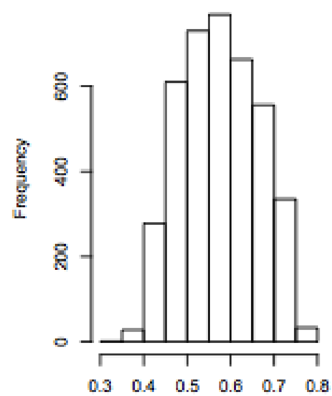
$ACE\left(X \rightarrow Y\right), n = 100$



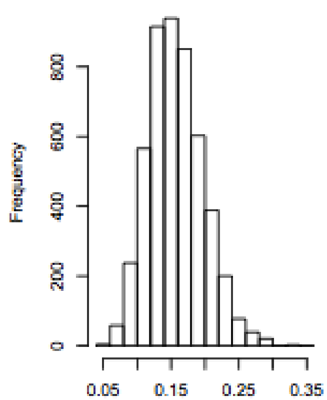
$ACE\left(X \rightarrow Y\right), n = 1,000$



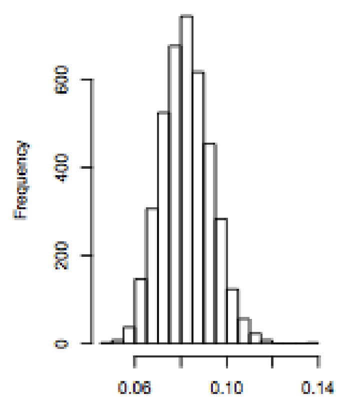
$\Pr\left(y_1 \mid do\left(x_1\right)\right), n = 100$



$\Pr\left(y_1 \mid do\left(x_1\right)\right), n = 1,000$



$\Pr\left(y_1 \mid do\left(x_0\right)\right), n = 100$



$\Pr\left(y_1 \mid do\left(x_0\right)\right), n = 1,000$

Again, the intervals are narrower for the larger samples but substantially wider than the point-identified example 5 which is consistent with large-sample identification predictions.

Example 7 (counterfactual query) *Suppose we're interested in evaluating counterfactual treatment for a subject in the experiment. Further, suppose the subject was assigned to the control group, complied with control assignment, and did not respond* (z_0, x_0, y_0)*. We ask what would be the subject's response to treatment (a counterfactual query)? First, given the subject's behavior in the experiment we know the individual is either a never-taker or a complier. Second, we know the individual's response is either never-recover or helped by treatment. Hence, the individual can be categorized in one of four compliance-response pairs:* $(cr_{00}, cr_{01}, cr_{10}, cr_{11})$*. Our question then translates into*

$$\Pr\left(y_1^{Z=z_1} \mid z_0, x_0, y_0\right)$$

which can be written as

$$g\left(v_{CR}\right) = \frac{v_{cr_{01}} + v_{cr_{11}}}{v_{cr_{00}} + v_{cr_{01}} + v_{cr_{10}} + v_{cr_{11}}}$$

This function $g\left(v_{CR}\right)$ *replaces* $ACE\left(X \to Y\right)$ *in a Gibbs sampler with other elements unchanged from example 6 except the draws are from the sample as opposed to multinomial draws from the population.*

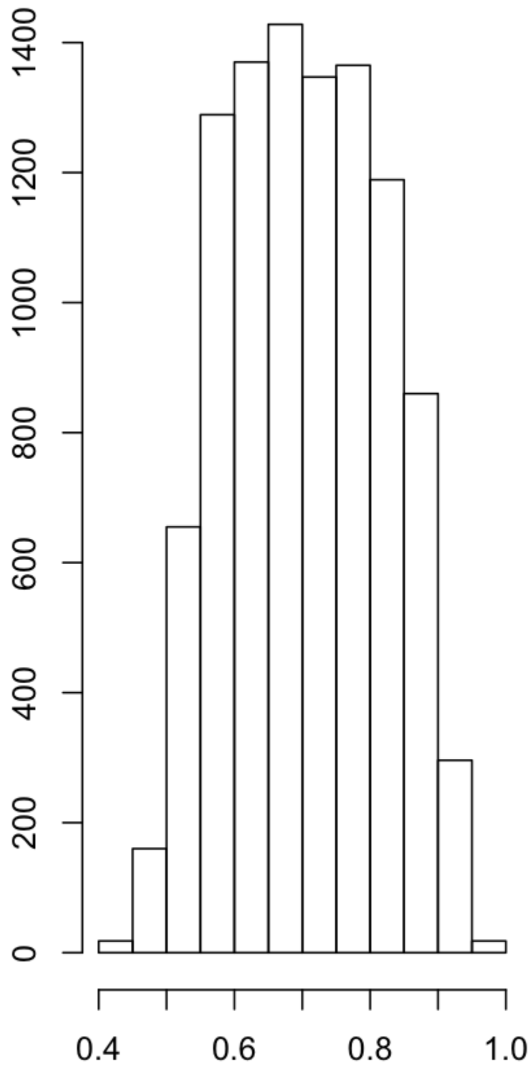
$N\left(y, x, z\right)$	$n = 337$
$(0, 0, 0)$	158
$(0, 1, 0)$	14
$(1, 0, 0)$	0
$(1, 1, 0)$	0
$(0, 0, 1)$	52
$(0, 1, 1)$	23
$(1, 0, 1)$	12
$(1, 1, 1)$	78

Partial-identification (large sample) bounds for $g\left(v_{CR}\right)$ *from a (nonlinear) program are*

$$0.506 \leq g\left(v_{CR}\right) \leq 0.857$$

The Gibbs sampler produces a mean for $g\left(v_{CR}\right) = 0.705$ *and a histogram con-*

sistent with the partial-identification bounds.



First, the histogram is largely consistent with the large-sample identification interval. Moreover, this is a rather remarkable expected benefit to treatment especially when we consider a 39% *noncompliance rate amongst those assigned to treatment and a sample of only* 337 *individuals.*

5 Appendix

5.1 Augmented DAGs and do-calculus

In this section, we attempt to crystalize understanding of the do-calculus theorem by augmenting the rule 2 DAG in figure 2.6 and the rule 3 DAG in figure 2.8. These augmented DAGs provide relatively simple illustrations of how rules 2 and 3 help to identify the causal effect of X on Y as well as when one rule aids identification and others fail to apply.

Do-calculus rule 2 is illustrated with an augmented DAG in figure 5.1. The

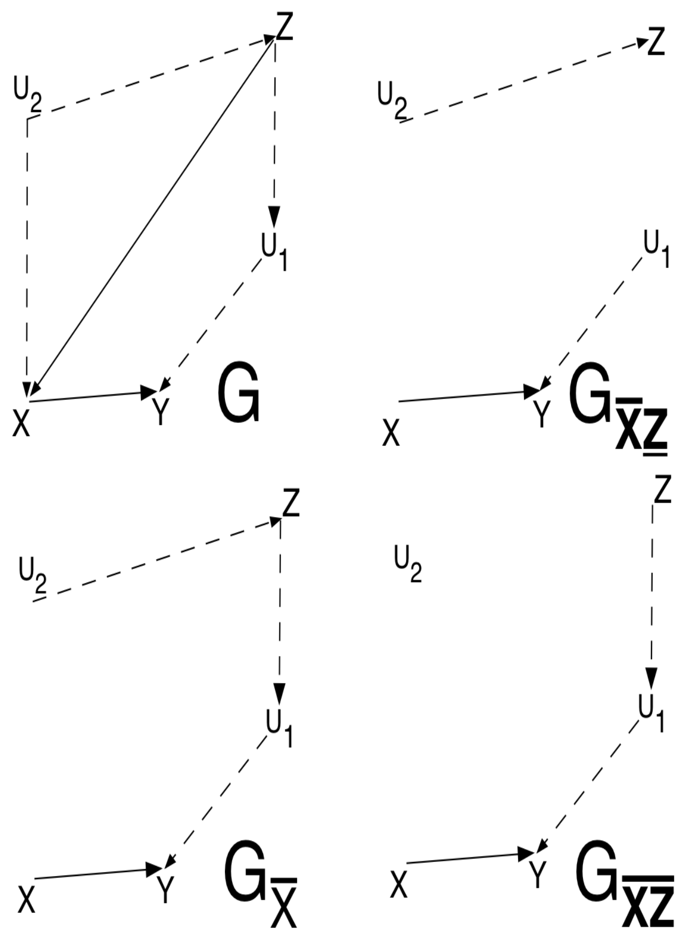


Figure 5.1: Augmented rule 2 DAG

back-door path from X to Y ($X \leftarrow Z \rightarrow U_1 \rightarrow Y$) is blocked (or d-separated) by

Z.[23] As indicated by rule 2, it makes no difference in assessing the probability of Y given X if Z is observed or set via action. However, any attempt to eliminate observed Z (rule 1) or action Z (rule 3) opens the back-door and confounds the causal effect of X on Y. Z is independent of Y given X only in DAG $G_{\overline{X}\underline{Z}}$ reinforcing rule 2.

Do-calculus rule 3 is illustrated with an augmented DAG in figure 5.2. The

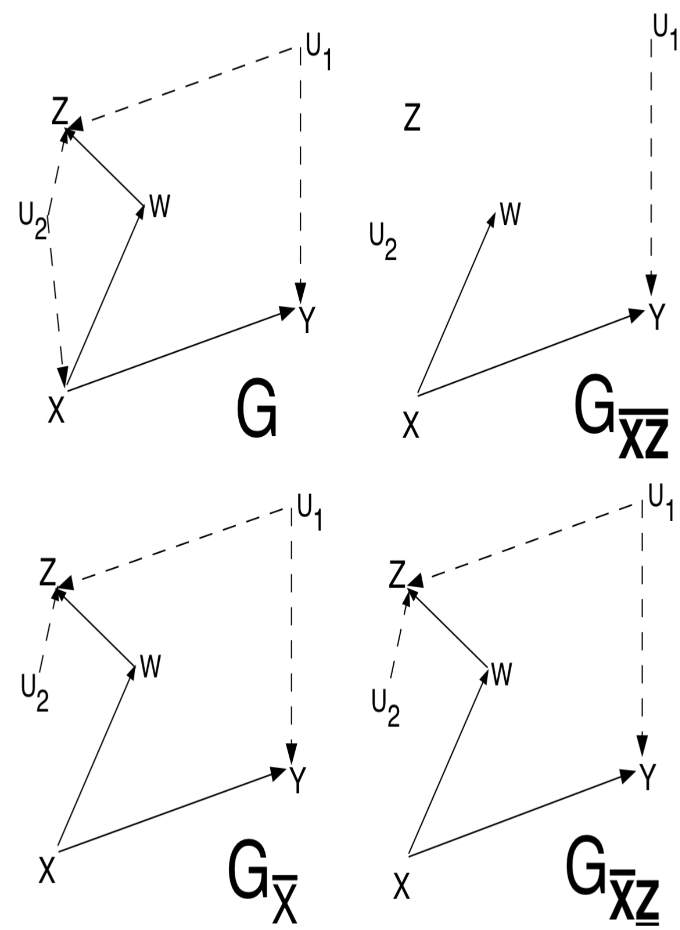


Figure 5.2: Augmented rule 3 DAG

[23]Identification of the causal effect of X on Y calls for action to be translated into observation. The intent of the current discussion is to illustrate the do-calculus rules rather than complete causal effect identification. When feasible, causal effect identification (associated with a DAG) is typically satisfied by back-door and/or front-door adjustments.

path $X \to W \to Z \leftarrow U_1 \to Y$ is blocked by the collider Z in $G_{\overline{X}}$ when action $do\,(x)$ is in play. Since a collider d-separates when we don't condition on it (else the causal effect of X on Y is confounded), we would drop conditioning on Z. Equivalently, the action do(Z) effectively eliminates the paths into Z, $W \to Z$ and $Z \leftarrow U_1$, thus d-separating the back-door from X to Y. Therefore, action Z can be inserted or deleted without impacting the probability of Y given $\{X, W\}$, as indicated by rule 3. On the other hand, rules 1 and 2 don't apply in this setting as conditioning on observed Z, a collider, is different than deletion (rule 1) or action (rule 2).

5.2 Linear programming bounds for ACE$(X \to Y)$

Bounds on causal effects, $\Pr(y_1 \mid do\,(x))$, or average causal effects

$$ACE\,(X \to Y) = \Pr(y_1 \mid do\,(x_1)) - \Pr(y_1 \mid do\,(x_0))$$

are derived from the dual program. For example, if the lower bound primal program is[24]

$$\begin{array}{cc} \min\limits_{q \geq 0} & \pi^T q \\ s.t. & Rq = p \end{array}$$

where π is the appropriate coefficient vector (comprised of zeroes and ones)[25] to address the causal effect of interest, then the dual program is

$$\begin{array}{cc} \max\limits_{\lambda} & p^T \lambda \\ s.t. & R^T \lambda \leq \pi \end{array}$$

Since λ is an 8-tuple that can be written in terms of 0 and ± 1, there are potentially $3^8 = 6,561$ λ-vectors. From these potential solutions, we determine those that are feasible, that is, those that satisfy $R^T \lambda \leq \pi$. Based on this reduced set of potential basic solutions, we find those that maximize $p^T \lambda$ and remain feasible (most potential solutions are dominated). This leads to the bounds described in the text.

Upper bounds are determined analogously where the dual program is[26]

$$\begin{array}{cc} \min\limits_{\lambda} & p^T \lambda \\ s.t. & R^T \lambda \geq \pi \end{array}$$

5.3 Natural bounds for ACE$(X \to Y)$

Define four functions:

$$\begin{array}{ll} f_0\,(u) = \Pr(y_1 \mid x_0, u) & g_0\,(u) = \Pr(x_1 \mid z_0, u) \\ f_1\,(u) = \Pr(y_1 \mid x_1, u) & g_1\,(u) = \Pr(x_1 \mid z_1, u) \end{array}$$

[24] Since $\sum_{j=0}^{3} \sum_{k=0}^{3} q_{jk} = 1$ is a redundant constratint we omit it in this discussion.

[25] The coefficient vector π is comprised of 0 and ± 1 for average causal effects.

[26] For the upper bound on $\Pr(y_1 \mid do\,(x))$, we can work with 8-tuple λ's comprised of 0 and 1 or $2^8 = 256$ potential vectors.

Then, we can depict components of $\Pr(y, x \mid z)$ as expectations of these functions:

$$\begin{aligned}
a &\equiv \Pr(y_1, x_0 \mid z_0) = E[f_0(1-g_0)] \\
b &\equiv \Pr(y_1, x_0 \mid z_1) = E[f_0(1-g_1)] \\
c &\equiv \Pr(x_1 \mid z_0) = E[g_0] \\
d &\equiv \Pr(x_1 \mid z_1) = E[g_1] \\
e &\equiv \Pr(y_1, x_1 \mid z_0) = E[f_1 g_0] \\
h &\equiv \Pr(y_1, x_1 \mid z_1) = E[f_1 g_1]
\end{aligned}$$

For any two random variables R and S such that $0 \leq R, S \leq 1$

$$1 + E[RS] - E[S] \geq E[R] \geq E[RS]$$

as $E[(1-R)(1-S)] \geq 0$. Further, the inequality holds for any pair of f, g functions since they are bounded between zero and one leading to

$$\begin{aligned}
1 + E[f_1 g_0] - E[g_0] &\geq E[f_1] \geq E[f_1 g_0] \\
1 + E[f_1 g_1] - E[g_1] &\geq E[f_1] \geq E[f_1 g_1] \\
1 + E[f_0(1-g_0)] - E[1-g_0] &\geq E[f_0] \geq E[f_0(1-g_0)] \\
1 + E[f_0(1-g_1)] - E[1-g_1] &\geq E[f_0] \geq E[f_0(1-g_1)]
\end{aligned}$$

or, can also be expressed

$$\begin{aligned}
\max\{h, e\} &\leq E[f_1] \leq \min\{(1+e-c), (1+h-d)\} \\
h &\leq E[f_1] \leq 1 + h - d \\
\max\{a, b\} &\leq E[f_0] \leq \min\{(a+c), (b+d)\} \\
a &\leq E[f_0] \leq a + c
\end{aligned}$$

where natural minima, maxima are indicated in each second line.

ACE$(X \to Y) = E[f_1] - E[f_0]$ and the lower bound is

$$\begin{aligned}
& \min E[f_1] - \max E[f_0] \\
= \;& h - (a+c) \\
= \;& \Pr(y_1, x_1 \mid z_1) - \Pr(y_1, x_0 \mid z_0) - \Pr(x_1 \mid z_0) \\
= \;& \Pr(y_1, x_1 \mid z_1) - [1 - \Pr(y_0, x_0 \mid z_0)]
\end{aligned}$$

while the upper bound is

$$\begin{aligned}
& \max E[f_1] - \min E[f_0] \\
= \;& 1 + h - d - a \\
= \;& 1 + \Pr(y_1, x_1 \mid z_1) - \Pr(x_1 \mid z_1) - \Pr(y_1, x_0 \mid z_0) \\
= \;& 1 - \Pr(y_0, x_1 \mid z_1) - \Pr(y_1, x_0 \mid z_0)
\end{aligned}$$

as indicated in the text.

5.4 Natural bounds for ETT$(X \to Y)$

Continue with definitions assigned in the appendix section 5.3 on natural bounds for $ACE\,(X \to Y)$ where

$$\Pr(y_1 \mid x_1, u) - \Pr(y_1 \mid x_0, u) = f_1(u) - f_0(u)$$

Then,

$$\begin{aligned}
ETT\,(X \to Y) &= E\left[f_1(u) - f_0(u) \mid X = x_1\right] \\
&= \sum_u \left[f_1(u) - f_0(u)\right] \Pr(u \mid x_1) \\
&= \frac{1}{\Pr(x_1)} \sum_u \left[f_1(u) - f_0(u)\right] \Pr(x_1 \mid u) \Pr(u) \\
&= \frac{1}{\Pr(x_1)} \sum_u \sum_z \left[f_1(u) - f_0(u)\right] \Pr(x_1 \mid z, u) \Pr(z) \Pr(u) \\
&= \frac{1}{\Pr(x_1)} \sum_u \left[f_1(u) - f_0(u)\right] \left[\Pr(z_1)\, g_1(u) + \Pr(z_0)\, g_0(u)\right] \Pr(u) \\
&= \frac{1}{\Pr(x_1)} E\left\{\left[f_1(u) - f_0(u)\right]\left[\Pr(z_1)\, g_1(u) + (1 - \Pr(z_1))\, g_0(u)\right]\right\} \\
&= \frac{1}{\Pr(x_1)} E\begin{bmatrix} \Pr(z_1)\, f_1 g_1 + (1 - \Pr(z_1))\, f_1 g_0 \\ -\Pr(z_1)\, f_0 g_1 - (1 - \Pr(z_1))\, f_0 g_0 \end{bmatrix} \\
&= \frac{1}{\Pr(x_1)} \left\{\begin{array}{c} \Pr(z_1)\, h + (1 - \Pr(z_1))\, e - \Pr(z_1)\left(E[f_0] - b\right) \\ -(1 - \Pr(z_1))\left(E[f_0] - a\right) \end{array}\right\} \\
&= \frac{1}{\Pr(x_1)} \left\{\Pr(z_1)(h + b) + (1 - \Pr(z_1))(e + a) - E[f_0]\right\} \\
&= \frac{1}{\Pr(x_1)} \left\{\begin{array}{c} \Pr(z_1)\left(E[f_1 g_1] + E[f_0(1 - g_1)]\right) \\ +(1 - \Pr(z_1))\left(E[f_1 g_0] + E[f_0(1 - g_0)]\right) - E[f_0] \end{array}\right\} \\
&= \frac{1}{\Pr(x_1)} \left\{\begin{array}{c} \Pr(z_1)\left(\Pr(y_1, x_1 \mid z_1) + \Pr(y_1, x_0 \mid z_1)\right) \\ +\Pr(z_0)\left(\Pr(y_1, x_1 \mid z_0) + \Pr(y_1, x_0 \mid z_0)\right) - E[f_0] \end{array}\right\} \\
&= \frac{1}{\Pr(x_1)} \left\{\Pr(z_1)\Pr(y_1 \mid z_1) + \Pr(z_0)\Pr(y_1 \mid z_0) - E[f_0]\right\} \\
&= \frac{1}{\Pr(x_1)} \left\{\Pr(y_1) - E[f_0]\right\}
\end{aligned}$$

Focusing on are noncompliance in experimental trials, $\Pr(x_1 \mid z_0)$, and we know from section 5.3

$$a \leq E[f_0] \leq a + c$$

or

$$\Pr(y_1, x_0 \mid z_0) \leq E[f_0] \leq \Pr(y_1, x_0 \mid z_0) + \Pr(x_1 \mid z_0)$$

Therefore,

$$\frac{\Pr(y_1) - \Pr(y_1, x_0 \mid z_0) - \Pr(x_1 \mid z_0)}{\Pr(x_1)} \leq ETT\,(X \to Y) \leq \frac{\Pr(y_1) - \Pr(y_1, x_0 \mid z_0)}{\Pr(x_1)}$$

as indicated in the text. The result affirms the treatment effect on the treated is point-identified if there is full compliance when no treatment is assigned, $\Pr(x_1 \mid z_0) = 0$.