

Non-parametric Mechanisms and Causal Modeling ^{*}

Adam Glynn[†]

Kevin Quinn[‡]

July 15, 2007

Abstract

Political scientists tend to think about causality in terms of mechanisms. In this paper we argue that non-parametric structural equation models are consistent with how many empirical political scientists think about causality and are consistent with the powerful and well-respected Neyman-Rubin Causal Model. Furthermore, using examples we demonstrate that two important practical questions are more easily addressed within the mechanistic framework: What (if any) set or sets of conditioning variables will allow the identification of average causal effects in a regression or matching model? When unmeasured confounding is present, what (if any) adjustment will non-parametrically identify the average causal effect?

^{*}The authors thank Thomas Richardson for introducing them to literature of graphical causal models. In addition, Quinn thanks the National Science Foundation (grants SES 03-50613 and BCS 05-27513) and the Center for Advanced Study in the Behavioral Sciences for its hospitality and support. The usual caveat applies.

[†]Department of Government and The Institute for Quantitative Social Sciences Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. aglynn@iq.harvard.edu

[‡]Department of Government and The Institute for Quantitative Social Sciences Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. kevin.quinn@harvard.edu

1 Introduction

Political scientists tend to think about causality in terms of mechanisms. Nearly all theoretically-informed work in political science has some discussion of the causal mechanisms hypothesized to have generated the data available to researchers. Notable examples include, among others, the “funnel of causality” of the Michigan model of voting (Campbell et al., 1960), Fearon’s (1995) work on rationalist explanations of war, and Cox (1997) on the mechanisms that produce the effective number of parties. Indeed, some have argued for an explicit focus on causal mechanisms as opposed to universal laws and grand theories (Elster, 1989a,b, 1998; Collier and Brady, 2004). The advantages of a focus on mechanisms include the transparency of assumptions and the possibility of highly contextual or contingent effects. Nevertheless, much recent work on causal inference by political methodologists has taken an essentially agnostic approach to the specification of underlying causal mechanisms.

In recent years, most political scientists who have been thinking seriously about causal inference have made extensive use of the Neyman-Rubin model (Neyman et al., 1935; Rubin, 1974, 1978). This framework provides a very powerful and general framework to consider issues of causality. It provides a coherent working definition of causal effects in terms of potential outcomes as well as a general statement of the assumptions sufficient to make causal inferences possible— even with observational data. Unfortunately, because of its generality the standard Neyman-Rubin model operates at a level of abstraction that is several steps away from the underlying mechanisms and processes that account for how observational data are generated. While such generality makes the Neyman-Rubin model very powerful, its agnosticism about the underlying causal mechanisms can make it difficult to apply in settings that are not close to a well-designed experiment.¹

The difficulties are two-fold. First, the standard Neyman-Rubin model provides little guidance

¹While many would argue that this is a good thing in that it *should* prevent researchers from drawing causal inferences from data that are far from the experimental ideal, it is easy to see that researchers continue to attempt causal inferences in difficult settings. Further, one could argue that such difficult inferences should continue to be made as long as they are clearly conditional on precisely stated assumptions and the assumptions are easily understandable by subject matter experts.

about what covariates should be adjusted for in order to eliminate confounding bias. The framework correctly states that conditional ignorability is sufficient for many causal inferences, but other than noting that variables caused by the treatment variable should not be conditioned on, very little concrete is said about which adjustments should and shouldn't be made. This would not be a serious problem if correct judgements about conditional ignorability were easy to make. Unfortunately, accurate judgements about conditional ignorability are often difficult to make. As we hope to demonstrate with some examples later in this paper, there are situations where, using standard rules of thumb and perhaps their intuition, many researchers would reason incorrectly about whether conditional ignorability holds.

A second problem arises in situations in which there is unmeasured confounding. Here conditional ignorability does not hold and the standard adjustment methods will not produce consistent estimates of causal effects. This does not mean it is impossible to make accurate causal inferences. Nonetheless, other than advocating sensitivity analysis—which can often be a good idea even in situations where unmeasured confounding is not likely—the standard Neyman-Rubin model is silent on how to make causal inferences when confronted with unmeasured confounding (but see Angrist et al. (1996)).

In this paper we advocate a hybrid approach that is consistent with the Neyman-Rubin framework but that specifies underlying mechanisms in a non-parametric fashion. More specifically, we show how the non-parametric structural equations models (NPSEMs) of Pearl (2000) add enough structure to the Neyman-Rubin model so that the issues above can be dealt with more easily. Unlike traditional linear structural equations models that make strong functional form assumptions, Pearl's NPSEMs only make assumptions about which variables determine the potential outcomes of other variables. The inclusion of mechanisms helps applied researchers think about the assumptions that are necessary to make causal inference and clarifies some aspects of procedures that have come out of the standard Neyman-Rubin framework.

This article proceeds as follows. In Section 2 we briefly review the key components of the

Neyman-Rubin Causal Model. In the next section we describe Pearl’s non-parametric structural equations model. Here we pay special attention to points of commonality and complementarity between the Neyman-Rubin model and the Pearl model. Section 4 examines the covariate selection problem with two examples that are designed to make clear some of the problems with standard approaches to choosing adjustment variables and how graphical methods can provide additional leverage on this difficult problem. The next section demonstrates how the formal rules governing the Pearl model can be used to non-parametrically identify causal effects in situations in which unmeasured confounding is present. Again, this is illustrated with two examples. The final section concludes.

2 The Neyman-Rubin Model

In the Neyman-Rubin model, causal effects are defined in terms of potential outcomes: $Y_x(u)$ (i.e. the potential outcome in unit u if X would have been set equal to x) (see Rubin (1978), Rosenbaum and Rubin (1983), Holland (1986)). Here u is thought of as unit-specific index, and therefore captures any individual-specific effects. It is tempting to think of units as individuals, schools, etc., but in actuality it is more accurate to think of the units as individuals, schools, etc. under a particular set of exogenous background conditions. Thus an individual at 9:00AM and the same individual at 10:00AM may very well be considered different units. If the value of X received by one unit does not affect the outcomes for other units, then given x and u , $Y_x(u)$ is completely determined. This assumption of non-interference is sometimes called SUTVA (see Angrist et al. (1996)).

2.1 Unit-Specific Causal Effects

In the Neyman-Rubin model, the potential outcomes are used to define unit-specific causal effects. For simplicity in presentation, we assume that X can only take on the values zero and one. Therefore, the unit-specific causal effect of $X = 1$ on Y relative to the effect of $X = 0$ in unit u is

calculated by comparing $Y_1(u)$ to $Y_0(u)$. A common means of comparison is the difference::

$$Y_1(u) - Y_0(u).$$

The key idea is that if it were possible to observe $Y_0(u)$ and $Y_1(u)$ for the two levels of the treatment variable (e.g. control and treatment), then we could observe the unit specific causal effect.

If we assume consistency (Robins, 1986) of the observed outcomes, then we may observe one of these two outcomes for each individual. This assumption requires that the observed outcome for each unit $Y(u)$ matches the potential outcome for unit u for the observed value of X . Formally, this can be written as the following:

$$\mathbf{X}(u) = x \implies \mathbf{Y}(u) = \mathbf{Y}_x(u).$$

Unfortunately, consistency does not allow the unit-specific causal effect to be directly observed since u only gets one of either $X = 0$ or $X = 1$ but never both. Holland (1986) calls this the fundamental problem of causal inference.

2.2 Population Causal Effects

Given the impossibility of observing individual causal effects, inference is usually confined to the characteristics of populations (sometimes the observed sample of individuals is taken as the entire population). For simplicity, we assume throughout this paper that the parameter of interest is the *average causal effect* from $X = 0$ to $X = 1$ is defined as

$$ACE \equiv \mathbb{E}[Y_1 - Y_0],$$

where the expectation merely represents an average over the population of interest. This parameter has a number of useful properties including the usual decomposition of the expectation of sums which allows us to separately consider the average potential outcomes under treatment and control.

$$ACE \equiv \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$$

Unfortunately, these averages are not observed in general. Instead we observe averages of potential outcomes over the subpopulations that actually received treatment and control. Hence we can identify the potentially similar parameter that Holland (1986) calls the *prima facie* average causal effect:

$$ACE^{pf} \equiv \mathbb{E}[Y_1|X = 1] - \mathbb{E}[Y_0|X = 0]$$

2.3 Ignorability and the Identification of Population Causal Effects

The Neyman-Rubin model makes clear that the following will not hold in general,

$$\mathbb{E}[Y_0] = \mathbb{E}[Y_0|X = 0] \tag{1}$$

$$\mathbb{E}[Y_1] = \mathbb{E}[Y_1|X = 1], \tag{2}$$

because averages over subpopulations need not match averages over the population. However, it is sufficient to assume the equalities in (1) and (2) in order to identify the ACE. This assumption sometimes known as mean ignorability² is usually hard to justify, because the subpopulation that receives treatment is often quite different from the subpopulation that receives control. Random treatment assignment for a large population is an example where the subpopulations will be similar.

It is often possible to “weaken” ignorability assumptions by conditioning on a set of background variables which we will denote as \mathbf{Z} . Hence, even if (1) and (2) do not hold, we may believe that,

$$\mathbb{E}[Y_0|\mathbf{Z}] = \mathbb{E}[Y_0|X = 0, \mathbf{Z}] \tag{3}$$

$$\mathbb{E}[Y_1|\mathbf{Z}] = \mathbb{E}[Y_1|X = 1, \mathbf{Z}], \tag{4}$$

hold for some set (or sets) of \mathbf{Z} . The equalities in (1) and (2) allow the identification of average causal effects within the strata defined by \mathbf{Z} , and these can then be combined through a weighted average to identify the overall ACE. When \mathbf{Z} lives in a high dimensional space, this averaging can present considerable practical difficulty, so in order to confine the discussion to the issues considered

²Although we focus on identification in this paper, there are other inferential goals, and hence it is often necessary to make stronger ignorability assumptions. Rosenbaum and Rubin (1983) describes sufficient ignorability assumptions for a variety of inferential tasks.

in this paper, we assume throughout that \mathbf{Z} is discrete and has low dimension or that the joint distribution of all variables has a simple parametric form.

2.4 Two Open Questions for Causal Modelers

Applied causal modelers are faced with at least two difficult problems that are not explicitly addressed within the Neyman-Rubin framework. First, they must find a set (or sets) of variables \mathbf{Z} such that (3) and (4) hold. Many rules of thumb have been proposed to solve the first problem, but these rules provide vague and sometimes misleading advice (an example is provided in Section 4). Second, they must decide how to proceed when some of the variables in \mathbf{Z} cannot be measured. Even when a sufficient set \mathbf{Z} can be specified, the causal modeler will often recognize that some important element of \mathbf{Z} is missing from the data set, and hence conditional ignorability is not possible. In these circumstances, the Neyman-Rubin model provides little advice on how to proceed. Some treatments propose sensitivity analyses (Rosenbaum, 1995) while others suggest instrumental variables (Angrist et al., 1996), but neither approach identifies the ACE (or even the average causal effect on the treated).

Given the difficulty of these two problems, applied modelers may find it worthwhile to add structure to the causal model. The Non-parametric Structural Equation Model (NPSEM) of Pearl (2000) provides simple and concrete rules for the formulation of the sufficient sets of conditioning variables and the non-parametric identification of causal effects in the presence of unmeasured confounding.

3 The Non-parametric Structural Equation Model

3.1 A Deterministic Causal Model

Definition 1 (Causal Model (Pearl, 2000, p. 203)) *A causal model is a triple*

$$M = \langle \mathbf{U}, \mathbf{V}, \mathbf{F} \rangle$$

where:

1. \mathbf{U} is a set of background variables, (also called exogenous), that are determined by factors outside the model;

2. \mathbf{V} is a set $\{V_1, V_2, \dots, V_n\}$ of variables called endogenous, that are determined by variables in the model—that is, variables in $\mathbf{U} \cup \mathbf{V}$; and
3. \mathbf{F} is a set of functions $\{f_1, f_2, \dots, f_n\}$ such that each f_i is a mapping from (the respective domains of) $\mathbf{U} \cup (\mathbf{V} \setminus V_i)$ to V_i and such that the entire set \mathbf{F} forms a mapping from \mathbf{U} to \mathbf{V} . In other words, each f_i tells us the value of V_i given the values of all the other variables in $\mathbf{U} \cup \mathbf{V}$, and the entire set \mathbf{F} has a unique solution $\mathbf{V}(u)$. Symbolically, the set of equations \mathbf{F} can be represented by writing

$$v_i := f_i(\text{pa}(v_i), u_i), \quad i = 1, \dots, n,$$

where $\text{pa}(v_i)$ is any realization of the unique minimal set of variables $PA(V_i)$ in $\mathbf{V} \setminus V_i$ (connoting parents) sufficient for representing f_i . Likewise, $\mathbf{U}_i \subseteq \mathbf{U}$ stands for the unique minimal set of variables in \mathbf{U} sufficient for representing f_i .³

Given this definition, consider the following simple example:

$$z := f_Z(u_1)$$

$$x := f_X(z, u_2)$$

$$y := f_Y(x, z, u_3)$$

z is a *deterministic* function of u_1 , x is *deterministic* function of z and u_2 , and y is a *deterministic* function of x , z , and u_3 . The $:=$ notation above is to make clear that the equality in these equations is asymmetric, and we label the entire model M .

The model M is non-parametric in that no assumptions are made about f_Z, f_X, f_Y, U_1, U_2 and U_3 . Note that any dependencies between the exogenous u variables needs to be modeled explicitly with additional equations (or possibly through the redefinition of the u variables). For example, we could write the previous model to include possible dependence between U_1 and U_2 as the following model:

³A set of variables \mathbf{X} is *sufficient* for representing a function $y = f(x, z)$ if f is trivial in Z —that is, if for every x, z, z' we have $f(x, z) = f(x, z')$.

$$\begin{aligned}
u_1 &:= f_{U_1}(u_{12}) \\
u_2 &:= f_{U_2}(u_{12}) \\
z &:= f_Z(u_1) \\
x &:= f_X(z, u_2) \\
y &:= f_Y(x, z, u_3),
\end{aligned}$$

but in order to maintain the u variables as exogenous we will simply write the model:

$$\begin{aligned}
z &:= f_Z(u_{12}) \\
x &:= f_X(z, u_{12}) \\
y &:= f_Y(x, z, u_3),
\end{aligned}$$

in which the definitions of f_Z and f_X have changed, and the inclusion of the u_{12} in both f_Z and f_X shows a particular type of dependence between these variables. Absence of the extra equations and distinct u variables in the functions f_Z , f_X , and f_Y imply a type of conditional independence which we explore in Section 2.3.

Let U denote the collection of exogenous factors $\{U_1, U_2, U_3\}$ and let u denote a realization of U . For simplicity, we assume that the model is recursive (i.e. no causal cycles) in that all endogenous variables can be explicitly defined by a single function of the exogenous variables. For example, in our simple model, y can be written as a function of all other variables and functions:

$$\begin{aligned}
z &:= f_Z(u_1) \\
x &:= f_X(z, u_2) \\
y &:= f_Y(x, z, u_3) \\
y &:= f_Y(f_X(f_Z(u_1), u_2), f_Z(u_1), u_3).
\end{aligned}$$

Therefore, $Y_M(u)$ denotes the unique value of Y generated by model M given u .

Now consider intervening in the system to set variable X equal to a particular value x without *directly* disturbing any of the other variables in the system. This involves creating a submodel in which the equation for X is removed.

Definition 2 (Submodel (Pearl, 2000, p. 204)) *Let M be a causal model, \mathbf{X} a set of variables in \mathbf{V} and x a particular realization of \mathbf{X} . A submodel M_x of M is the causal model*

$$M_x = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}_x \rangle$$

where

$$\mathbf{F}_x = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} = x\}$$

We denote this intervention (or action) with $do(x)$. What this intervention amounts to is creating a new set of structural equations in which the equation for x becomes some value. For example, consider the simple example again with x set to zero.

$$z := f_Z(u_1)$$

$$x := 0$$

$$y := f_Y(x, z, u_3)$$

We call this system of equations submodel M_x .

Within a submodel, we can define potential outcomes that are analogous to the potential outcomes from the Neyman-Rubin model by solving the set of equations for Y under the submodel.

Definition 3 (Potential Outcome (Pearl, 2000, p. 204)) *Let \mathbf{X} and \mathbf{Y} be two subsets of variables in \mathbf{V} . The potential response of \mathbf{Y} to action $do(\mathbf{X} = x)$, denoted $\mathbf{Y}_x(u)$ is the solution for \mathbf{Y} of the set of equations F_x .*

Hence, we let $Y_x(u)$ denote the unique value of Y that results from a particular realized value u of U in model M_x .

If u were observed, the model M and submodel M_x would define unit-specific causal effects analogous to unit-specific causal effects in the Neyman-Rubin model. Consider the following simple non-parametric example:

Suppose that a medical study randomly assigns stroke victims to either treatment (blood thinner) or control (placebo), and after a period of time the patients are observed to be alive or dead. Unbeknownst to the study designers, the stroke victims suffer from two different types of strokes. Some of the victims suffer from a “clotting” type of stroke, and others suffer from a “leaking” type of stroke. Hence a blood thinner will have different effects on the two groups. If we let X be treatment assignment, Y be patient status, U_1 be the exogenous treatment randomization mechanism, and U_2 be the exogenous variable that describes each patient’s potential response to treatment, then this scenario can be conceptualized within the NPSEM framework with the following set of equations:

$$x := f_X(u_1)$$

$$y := f_Y(x, u_2)$$

With binary treatment and outcome, the domain of U_2 (and hence stroke victims) can be partitioned into four groups that describe potential response to treatment: those whose condition is so mild, that they would be alive regardless of whether they received treatment or control (Always), those whose condition is so severe that they would be dead regardless of whether they received treatment or control (Never), those with a treat-able “clotting” condition who would be alive with the blood thinning treatment and dead without (Helped), and those with a mild “leaking” condition who would be dead with the blood thinning treatment but alive with the control (Hurt).

In this example as in most cases, u is not observed, and hence we do not observe the unit-specific causal effects. The solution is to shift inferential focus to population causal effects as was done in the Neyman-Rubin framework.

3.2 A Population Causal Model

We can create a population causal model from the deterministic causal model of the previous subsection by assuming a distribution over U .

Definition 4 (Probabilistic Causal Model (Pearl 2000, p. 205)) *A probabilistic causal model is a pair*

$$\langle M, P(u) \rangle$$

where M is a causal model and $P(u)$ is a probability function defined over the domain of U .

The model M along with an assumption as to the distribution of U generates the “pre-intervention” distribution $P(U, V)$, and given the assumption of a recursive causal model, this joint distribution is uniquely defined. *One can estimate the marginal distribution of the observed variables in V directly from observational data without making untestable assumptions.* Continuing the example from the previous section, the discrete nature of the collection U allows the interpretation of $P(u)$ as the population proportions of individuals. Hence, $P(u)$ describes the proportions of (*Always, Never, Helped, Hurt*) individuals in the population, and the proportions of (*treatment, control*) individuals in the population. Furthermore, these distributions in combination with the causal model define the proportions of (*alive, dead*) individuals in the population.

If we assume that $P(u)$ remains unchanged for the submodel M_x , then we can ask what the probability distribution of $Y_x(u)$ is for a u randomly drawn from the population distribution of U . In other words, what is the distribution of Y in the population after the intervention $do(x)$. This quantity is denoted $P(y|do(x))$ and is called the post-intervention distribution of Y . *Post-intervention distributions are not directly directly estimable from observational data without untestable causal assumptions.* With probability distributions defined over post-intervention distributions, the expectations and average causal effects are easy to derive:

$$\mathbb{E}[Y|do(x)] \equiv \mathbb{E}[Y_x],$$

and obviously

$$ACE \equiv \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \mathbb{E}[Y|do(1)] - \mathbb{E}[Y|do(0)]$$

As in the Neyman-Rubin model, we would like to establish situations in which the observable pre-intervention distribution identifies averages over the unobserved post-intervention distribution. This task is simplified by representing NPSEMs as Directed Acyclic Graphs.

3.3 Representing NPSEMs as Directed Acyclic Graphs (DAGs)

We begin with some basic terminology. A *graph* $G = \{\mathbf{V}, \mathbf{E}\}$ is a collection of *vertices* $V \in \mathbf{V}$ and *edges* $E \in \mathbf{E}$. Each edge connects two vertices. In what follows, each edge will be *directed* in that it represents an asymmetric relationship between the vertices it links. A directed edge from vertex V_1 to vertex V_2 is denoted $V_1 \rightarrow V_2$. In what follows, we will think of the vertices as variables and edges as causal relationships between variables. A *path* is a sequence of edges E_1, E_2, \dots, E_k in which the end vertex of E_i is the start vertex of E_{i+1} for $i = 1, \dots, (k - 1)$. The direction of the edges does not matter. For instance, a path from V_1 to V_3 exists in each of the following.

$$V_1 \rightarrow V_2 \rightarrow V_3 \tag{5}$$

$$V_1 \leftarrow V_2 \rightarrow V_3 \tag{6}$$

$$V_1 \rightarrow V_2 \leftarrow V_3 \tag{7}$$

$$V_1 \leftarrow V_2 \leftarrow V_3 \tag{8}$$

While all of these relationships represent paths from V_1 to V_3 , it is useful to make some distinctions between these types of paths and vertices. A *directed path* from V_1 to V_k requires that all edges point toward V_k along the path. For instance, (5) depicts a directed path from V_1 to V_3 . A *collider* on a path is a vertex for which both edges on the path are in-pointing. For instance, in (7), V_2 is a collider. In (5), (6), and (8), V_2 is known as a non-collider.

The *parents* of V_i (denoted $pa(V_i)$) is the set of vertices from which an arrow goes directly to V_i . The *ancestors* of V_i (denoted $an(V_i)$) is the set of vertices from which a directed path exists to V_i .

Similarly, The *children* of V_i (denoted $ch(V_i)$) is the set of vertices to which an arrow goes directly from V_i , and the *descendants* of V_i (denoted $de(V_i)$) is the set of vertices to which a directed path exists from V_i . For simplicity of definitions, it is often helpful to include the vertex V_i in the sets $an(V_i)$ and $de(V_i)$, and we will do so unless otherwise stated.

A graph in which all the edges are directed (i.e. single headed arrows) and no edges of the form $V_i \rightarrow V_i$ exist is said to be a *directed graph*. A directed graph that does not have cycles (possibly involving long paths) of the form $V_i \rightarrow V_j, V_j \rightarrow V_i$ is said to be a *directed acyclic graph (DAG)*. In what follows, we will restrict our attention to DAGs.

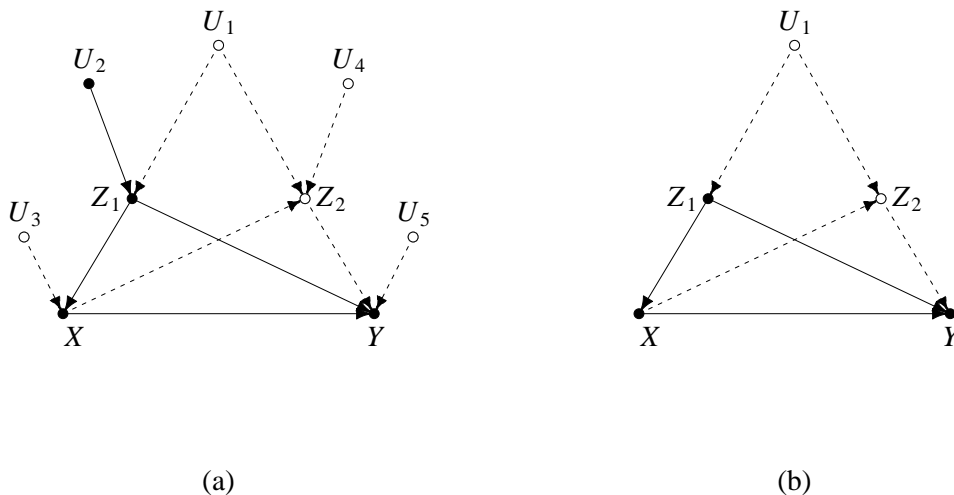


Figure 1: *Graphical Model Consistent with Structural Equations 9 - 12*. Panel (a) shows all exogenous variables and their associated edges. Panel (b) removes superfluous exogenous variables and their edges. Note that observability is neither necessary nor sufficient for a variable to be exogenous. Here U_1, \dots, U_5 are the exogenous variables. U_2 is observed but the other U variables are not. Further, one of the endogenous variables (Z_2) is unobserved while the other endogenous variables are observed.

We use the edges in a DAG to represent the inputs to the functions of a corresponding NPSEM. In this type of causal model, given $U = u$, we can uniquely determine the value of all other variables. Furthermore, if we specify a probability distribution over U , then the joint distribution over the observed variables is also uniquely determined. The rules for forming a DAG G_M from a NPSEM M are the following:

1. Represent each unobserved variable with an open vertex.
2. Represent each observed variable with a closed vertex.
3. For each equation in M draw an edge from each variable on the right-hand-side of the $:=$ operator to the variable on the left-hand-side using a solid line when both vertices are observed and a dashed line when one vertex is unobserved.

In the interest of graphical simplicity, many authors will delete nodes and edges that do not affect the results of graphical tests of interest. For example, exogenous variables that point into a single endogenous variable can often be removed. However, such practice may obscure assumptions inherent in the model. In particular, an exogenous variable that points into a single endogenous variable implies a type of independence between the exogenous variables.⁴ Furthermore, the removal of exogenous variables from the graph may obscure the fact that the NPSEM defines individual causal effects.

A simple example will make this process more clear. Consider again the following structural model M :

$$z_1 := f_{Z_1}(u_1, u_2) \tag{9}$$

$$x := f_X(z_1, u_3) \tag{10}$$

$$z_2 := f_{Z_2}(x, u_1, u_4) \tag{11}$$

$$y := f_Y(x, z_1, z_2, u_5) \tag{12}$$

In Figure 1 (a), G_M is constructed using rules 1-3. A pruned version of G_M is drawn in Figure 1 (b) in which vertices and edges that are unnecessary for identifying the effect of X on Y have been dropped. The exact interpretation of the graphs in Figure 1 will wait until we discuss d -Separation in the next section.

3.4 Factorization of the Joint Distribution and d -Separation

Given an NPSEM model M with k variables (observed and unobserved), we rename the variables V_1, \dots, V_k in order to allow a more general formulation. Using the procedure of Section 3.3, we form

⁴In some treatments of this material, exogenous variables always point into a single endogenous variable and dependencies among the variables are represented by dashed arcs (Pearl, 2000, Ch. 3).

the DAG G_M with vertices V_1, \dots, V_k . The causal Markov condition holds (Pearl, 2000, p.30), and we can factor the joint distribution of the variables according to the following rule:

$$P(v_1, \dots, v_k) = \prod_{i=1}^k P(v_i | pa(v_i)),$$

where $pa(v_i)$ are the realized values from the set of random variables $pa(V_i)$. Furthermore, given this factorization, we can read conditional independence relations from such a model with the concept of *d-Separation*.

If we write \mathbf{Z} to be a set of conditioning variables, then the two variables V_i and V_j are conditionally independent given the set of variables \mathbf{Z} if the vertices V_i and V_j are *d-separated* by \mathbf{Z} in the DAG G_M . We say that vertices V_i and V_j are *d-Separated* by \mathbf{Z} , when all paths from V_i to V_j are blocked by \mathbf{Z} . A path from v_i to v_j is blocked by \mathbf{Z} when at least one of the following two conditions holds:

1. A non-collider on the path from V_i to V_j is an element of \mathbf{Z} .
2. There exists a collider on the path from V_i to V_j and no descendant of the collider is an element of \mathbf{Z} .

A stylized example will help to show why the *d-Separation* condition relates to conditional independence in a NPSEM. Suppose that Joe's lawn is either wet ($Y = 1$) or not wet ($Y = 0$), and Joe's lawn only gets wet when either it rains ($X = 1$), or the sprinkler is turned on ($Z = 1$). Furthermore, assume that Joe is an eccentric fellow who decides whether to turn his sprinkler on everyday by flipping a coin. Because of the random assignment of the sprinkler, the variables X and Z are marginally independent of each other, however, conditioning on Y induces dependence (If we know the grass is wet, and we know the sprinkler wasn't on, then we know it must have rained). Therefore, X and Z are dependent conditional on Y . The NPSEM associated with this

scenario is the following:

$$\begin{aligned} z &:= f_Z(u_1) \\ x &:= f_X(u_2) \\ y &:= f_Y(x, z). \end{aligned}$$

where u_1 represents the result of the coin flip that Joe makes and u_2 represents the result of the coin flip that nature makes. In this example, y depends deterministically on z and x , hence no u_3 is needed. Furthermore, this model can be represented by the following DAG:

$$U_1 \rightarrow Z \rightarrow Y \leftarrow X \leftarrow U_2.$$

Following the rules of d -separation, we see that the path from Z to X is blocked by the collider Y as long as we don't condition on descendants of Y . Therefore Z and X are d -Separated conditional on the empty set. If we instead condition on Y , then we have conditioned on a descendant of Y , and the path from Z to X is no longer blocked. Hence the d -Separation rules recover the conditional independence relations that we intuited above.

3.5 The Rules of the *do*-Calculus

Using the NPSEM and d -Separation as described in the previous sections, Pearl (2000) provides graphical criteria for the general identification of population causal effects.

Definition 5 (Rules of *do*-Calculus (Pearl (2000) p. 85)) *Let G be the directed acyclic graph associated with a causal model as in Definition 1, and let $P(\cdot)$ stand for the probability distribution induced by that model. For any disjoint subsets of variables X, Y, Z , and W . We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X . Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X . To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$. Finally, the expression $P(y|do(x), z) \equiv P(y, z|do(x))/P(z|do(x))$ stands for the probability of $Y = y$ given that X is held constant at x and that (under this condition) $Z = z$ is observed. We have the following rules.*

1. (Insertion/deletion of observations)

$$P(y|do(x), z, w) = P(y|do(x), w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}}}$$

2. (Action/observation exchange)

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}\underline{Z}}}$$

3. (Insertion/deletion of actions)

$$P(y|do(x), do(z), w) = P(y|do(x), w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{XZ}(W)}}$$

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

Intuitively, the first rule describes situations where adjustment for an additional variable is not important for an intervention. The second rule describes situations analogous to conditional ignorability, where the standard adjustment methods will identify causal effects. Pearl (2000) refers to this rule as the back door criterion. The third rule describes situations where intervening on a variable will not affect the outcome of interest. By repeatedly applying the three rules of the *do*-Calculus, we can sometimes identify average causal effects by representing unobservable post-intervention distributions with observable pre-intervention distributions. The next two sections present examples to demonstrate the use of the three rules in order to answer the two open questions for causal modelers.

4 The Relevance of Seemingly Irrelevant Covariates

The choice of what variables to adjust for when attempting to make causal inferences from observational data is one of the most important decisions made by applied researchers. Unfortunately, it has also received less scholarly attention than the question of how to perform an adjustment *given* the appropriate set of covariates. Misunderstandings abound.

The standard econometric advice is that when a linear model is appropriate for causal inference and one is unsure whether a measured pre-treatment variable should be included or not it is best to err on the side of including the potentially irrelevant covariate. Kmenta (1986) sums this up nicely:⁵

The conclusion, then, is that if the specification error consists of including some irrelevant explanatory variables in the regression equation, the least squares estimators of the regression coefficients are unbiased but not efficient. The estimators of the variances are also unbiased, so that, in the absence of other complications, the usual tests of significance and confidence intervals for the regression coefficients are valid. (p. 449)

⁵For similar sentiments see also Fox (1997, p. 235 *fn* 50), Pindyck and Rubinfeld (1998, p. 187), and Greene (2000, p. 338)

This sort of argument leads many researchers to adjust for a very large number (often upwards of 15-20) measured pre-treatment variables without thinking carefully about how these might or might not be related to the outcome, treatment assignment, and each other causally.

Many epidemiologists have a slightly different view. In assessing whether a particular set of pre-treatment variables \mathbf{Z} acts as a confounder (i.e., failing to adjust for \mathbf{Z} would bias causal effect estimates) they tend to focus on whether various associations in the observed pre-intervention distribution are non-zero. More specifically, they judge the effect of X on Y to be confounded by \mathbf{Z} if

1. $X \not\perp\!\!\!\perp \mathbf{Z}$; and
2. $Y \not\perp\!\!\!\perp \mathbf{Z} | X$

For examples see Schlesselman (1982); Rothman (1986); Rothman and Greenland (1998) along with the discussion in Chapter 6 of Pearl (2000). Because the relevant associations are directly observable from observational data such a criterion has great appeal. The criterion can easily be checked, and if \mathbf{Z} is judged to be a confounder, then it is adjusted for in some way, e.g., via regression, stratification, matching, etc.

Much of the empirical work in leading political science journals tends to (at least implicitly) take one of these two positions when defending the choice of adjustment variables.⁶ Unfortunately, neither of these approaches is guaranteed to identify sets of adjustment variables that are sufficient to control confounding— *even when some subset of measured pre-treatment variables available to the researcher is sufficient to control confounding*. This is true regardless of the adjustment method.

The fundamental problem with both of these approaches is that they seek either implicitly (as with

⁶Clarke (2005) provides examples of political science articles that include a large number of adjustment variables seemingly for fear of omitted variable bias. He then goes on to correctly note that, within the context of linear regression, bias is not necessarily monotonically decreasing in the number of background variables adjusted for. Much of the discussion in that article is focused on whether a covariate is “relevant” which we assume means whether the covariate exerts a causal effect on the outcome variable. The results we describe below are more general in that they apply to all recursive causal models— not just causal linear regression models. Further, we will see that there are situations where adjusting for an “irrelevant” variable, i.e., a variable that exerts no causal effect on either the outcome or treatment assignment, will bias estimates of causal effects.

the econometric approach) or explicitly (as with the epidemiologists approach) to use characteristics of the pre-intervention distribution without additional causal assumptions to determine the appropriate set of adjustment variables.

The Neyman-Rubin model offers a major step forward in that it focuses a researcher’s attention on whether conditional ignorability (a causal assumption) holds for a given set of adjustment variables. The major advantage here is that if conditional ignorability does hold given \mathbf{Z} then adjustment for \mathbf{Z} is *guaranteed* to be sufficient to control confounding. The major drawback is that, by itself, the Neyman-Rubin framework provides little guidance as to what sets of background variables are likely to produce conditional ignorability. Conditional ignorability is a global assumption that is defined for potential outcomes and is thus not strictly testable. The reliance on potential outcomes and the associated lack of testability is not specific to the Neyman-Rubin framework— as noted above all genuine causal models (including the Pearl model) suffer from the same problem. However, the reliance on a single global assumption is not a property of all causal models. By replacing this single large assumption with a series of local assumptions the deterministic structural equations models of Pearl offer researchers additional means of assessing the adequacy of various adjustment strategies.

As noted above, the rules of Pearl’s *do*-Calculus give rise to a simple graphical criterion called the back-door criterion that can be checked to see if a given set \mathbf{Z} is sufficient to control confounding bias. This criterion can be stated as follows.

Definition 6 (Back-Door Criterion (Pearl (2000, p. 79))) *Given a causal model M and associated causal graph G_M , A set of covariates \mathbf{Z} satisfies the back-door criterion for a causal variable X and outcome Y if:*

1. *no element of \mathbf{Z} is a descendant of X ; and*
2. *X is d -separated from Y by \mathbf{Z} in the graph $G_{\underline{X}}$ formed by deleting all edges out of X from G_M .*

If \mathbf{Z} satisfies the back-door criterion then the potential outcome distribution can be calculated

using the standard stratification adjustment (Cochran, 1968; Rubin, 1977):

$$P(Y_x = y) = \sum_{\mathbf{z}} P(y|x, \mathbf{z})P(\mathbf{z})$$

where \mathbf{z} may be multivariate. Pearl refers to this as the *back-door* adjustment. Since if \mathbf{Z} satisfies the back-door criterion the standard stratification adjustment is appropriate, it follows that matching or stratifying on $P(x|\mathbf{z})$ (the propensity score given a realized value \mathbf{z} of \mathbf{Z}), along with related adjustments that make use of conditional ignorability, will also be appropriate (Rosenbaum and Rubin, 1983, 1984). As we will see below, this is true regardless of whether all (or even any) of the variables that affect treatment assignment are in \mathbf{Z} — all that is required is that conditional ignorability hold given \mathbf{Z} .⁷

Again, the major advantage of this graphical approach to the identification of causal effects is that it is framed in terms of a series of local assumptions about causal mechanisms. These local assumptions are often easier to consider, debate, and possibly reject as unbelievable than the single global assumption of conditional ignorability.

The following examples illustrate what can go wrong when researchers use either the traditional econometric or epidemiological approaches to select adjustment variables. In addition, the examples show how the back-door criterion can be easily employed to determine an appropriate adjustment strategy.

4.1 A Simple Non-Parametric Example

Consider the following stylized example.⁸ Researchers are interested in determining the effect of a novel two-week within-school program on inter-ethnic cooperative behavior. It is assumed that there are two mutually exclusive and exhaustive ethnic groups. The researchers adopt the following research design. The population of interest consists of all students in two large high

⁷We note in passing the obvious point that the results of Rosenbaum and Rubin (1983) show that if conditional ignorability holds given \mathbf{Z} then using $P(x|\mathbf{z})$ or any other balancing score as an adjustment covariate is appropriate. They do not show that $P(x|\mathbf{z})$ or any other balancing score for arbitrary \mathbf{Z} implies conditional ignorability.

⁸Note that while this is obviously not an example of a well-designed study it is not so different from many studies that are actually conducted in the social sciences.

schools. Students from these two schools are randomly selected to participate in the study. To keep matters simple, it is assumed that all selected students participate and that the number of students is large enough that sampling variability can be ignored. In what follows, the implicit treatment variable X is the exposure to the novel program (measured as unexposed = 0, exposed = 1).

The program is implemented in only one of the two schools and all students in that school are exposed. The school that implements the program has a longstanding reputation as the more ethnically inclusive of the two schools under study. The schools are alike in all other relevant respects. Each selected student is taken to a mobile experimental laboratory where s/he is told that s/he will play the role of the first-mover in a one-shot divide the dollar game. Via a computer terminal, the student makes a single take-it-or-leave-it offer $y \in \{\$0.01, \$10, \$19.99\}$ to the second (unseen) player. Before playing the game, the first-mover is told that if the second player accepts the offer the first-mover will receive $\$20 - y$ and if the second player rejects the offer the first-mover will receive $\$0$. Unknown to the first-mover, the second “player” is computer program that randomly chooses to either reject or accept the offer. Nonetheless, the first-mover is told that the second player is a member of the ethnic group different from the first-mover’s group. The goal of the study is to ascertain the effects of exposure to the novel program (X) on altruistic inter-ethnic behavior (Y).

Because the program was implemented in the school known to be more ethnically inclusive, there is concern among the researchers that the ethnocentrism of a student’s parents might exert an effect on the school the child attends (and hence exposure to the program) as well as the child’s ethnocentrism and ultimately the child’s behavior in the experimental game. In the hope of developing a proxy for the parent’s ethnocentrism, the researchers administer a short survey to each student before they are exposed (or could have been exposed) to the new program. Each student’s survey responses are used to construct a measure of his/her latent ethnocentrism. This variable is labeled Z (measured as low ethnocentrism = 0, high ethnocentrism = 1). After the study has

been completed, the researchers find (as they had expected) that there is a significant association between Z and X and a significant association between Z and Y within levels of X . Table 1 represents the joint probability function of X, Y , and Z . We emphasize the obvious point that even though we can ignore sampling variability and we thus know the joint population distribution of X, Y , and Z , it is not possible to determine whether conditional ignorability (or mean ignorability) holds here.

$Z = 0$ (low ethnocentrism)		
	$X = 0$ (not exposed)	$X = 1$ (exposed)
$Y = \$0.01$	0.1728	0.2219
$Y = \$10.00$	0.1224	0.1868
$Y = \$19.99$	0.0144	0.0220

$Z = 1$ (high ethnocentrism)		
	$X = 0$ (not exposed)	$X = 1$ (exposed)
$Y = \$0.01$	0.0772	0.0281
$Y = \$10.00$	0.1012	0.0368
$Y = \$19.99$	0.0120	0.0044

Table 1: *Joint Probability Function For Variables in Ethnic Cooperation Example.*

Nonetheless, since $Z \not\perp X$, $Z \not\perp Y|X$, and Z is temporally prior to X , the researchers judge Z to be a confounder and decide to adjust for it using exact stratification (the method of adjustment is not important for this example). This yields the potential outcome distributions in the two leftmost columns of Table 2. From this table it appears that the program exerts a modest, positive effect on cooperative behavior. The counterfactual probability of the most selfish offer ($Y = \$0.01$) decreases by 3 percentage points under treatment while the counterfactual probability of the most equitable offer ($Y = \$10.00$) increases by the same amount.

Unknown to the researchers, the world works slightly, but importantly, differently than they assume. Specifically, the actual causal mechanisms are consistent with with the causal graph in Figure 2. As the researchers expect, parental ethnocentrism (U_1) does affect the survey measure of student ethnocentrism (Z) and exposure to the program (X). What the researchers failed to

	Putative $P(Y_0 = y)$	Putative $P(Y_1 = y)$	True $P(Y_0 = y)$	True $P(Y_1 = y)$
$y = \$0.01$	0.52	0.49	0.50	0.50
$y = \$10.00$	0.43	0.46	0.45	0.45
$y = \$19.99$	0.05	0.05	0.05	0.05

Table 2: *Putative Potential Outcome Distribution Based on Data in Table 1 After Adjusting for Z along with True Potential Outcome Distribution.*

recognize is that the survey measure of ethnocentrism is also affected by a second latent variable—student IQ (U_2)—and this latent variable also has an effect on the outcome variable (Y). The mechanisms here are the following. Students with higher IQs are better able to grasp what the survey instruments are probing. If students realize that the instrument is attempting to measure a socially undesirable trait and they possess that trait they will misreport their true beliefs on the survey. Further, students with higher IQs are more likely to realize the monetary incentives embedded in the experimental game and offer \$0.01. Importantly, the missing edges from Z to X and from Z to Y imply the lack of the corresponding direct effects.

Confronted with this information, many researchers would still adjust for Z when estimating the effect of X on Y . Several types of arguments might be advanced to support such a decision—“ Z is a proxy for multiple unmeasured confounders (and hence there is even more reason to condition on it)” and “conditioning on a pre-treatment variable cannot increase bias” are two such arguments—but all such arguments are incorrect. Using the back-door criterion to check whether adjustment for Z is sufficient to eliminate confounding we see that it is not— X and Y are d -connected given Z via the back-door path $X \leftarrow U_1 \rightarrow Z \leftarrow U_3 \rightarrow Y$. Nonetheless, things are not hopeless in this situation, for it turns out that the null set \emptyset satisfies the back-door criterion. In other words, no adjustment is necessary. If assumptions embodied by the causal graph in Figure 2 are correct, then the observed $P(y|x)$ is a consistent estimate of $P(y|do(x)) \equiv P(Y_x = y)$. The true counterfactual outcome distributions are shown in the two rightmost columns of Table 2. Here we see that there is no causal effect of the program on outcomes.

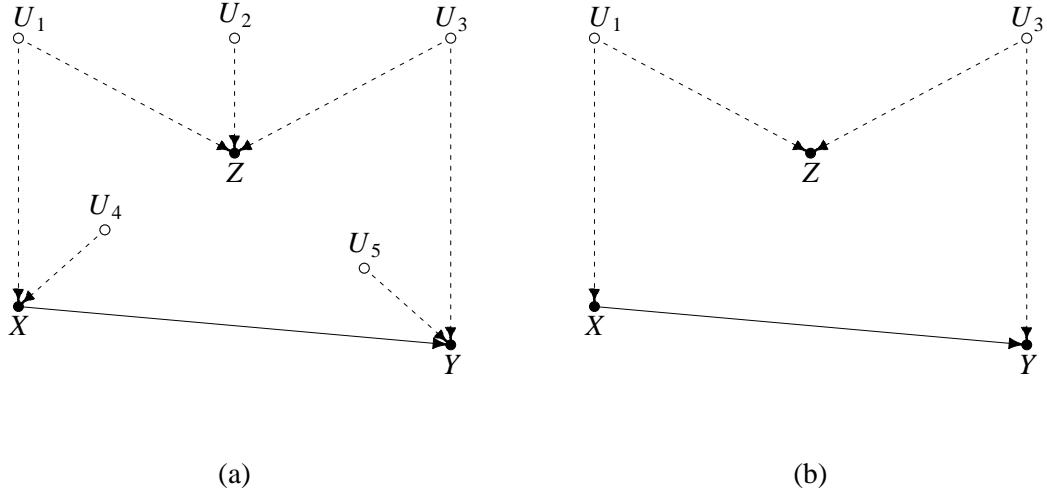


Figure 2: *Causal Graphs Consistent with the True Data Generating Process Behind Ethnic Cooperation Example (Data in Table 1).* U_1 is the ethnocentrism of parents, U_3 is student IQ, Z is the survey measure of student ethnocentrism, X is exposure to the program, and Y is the offer made in the experimental game. In the example, the edge from X to Y is technically missing because there is no causal effect of the program on behavior. Nonetheless, we include the edge because the goal of the analysis is to estimate this relationship rather than assume it. Panel (a) contains all exogenous variables while panel (b) excludes nodes and edges that are unnecessary for determining if and how the causal effect of X on Y is identified.

It is important to note that the consistency of $P(y|x)$ for $P(Y_x = y)$ and the inconsistency⁹ of estimators that adjust for Z does not depend on the type of adjustment method, particular parametric assumptions, or unlikely cancellations of effects. It is a general result for all causal models that can be written in the form of

$$z := f_Z(u_1, u_2, u_3)$$

$$x := f_X(u_1, u_4)$$

$$y := f_Y(x, u_3, u_5)$$

with U_1, \dots, U_5 all mutually independent; or even more generally for all causal models in which the back-door criterion holds for \emptyset but not for \mathbf{Z} .

⁹Inconsistency statements of this type require the additional assumption of the *faithfulness* (Spirtes et al., 1993) of the graph (i.e. the graph encodes all conditional independence relations in the population distribution). But in this context, such an assumption is actually conservative and will hold for all but the most contrived distributions.

4.2 A More Complicated Parametric Example

As noted above, the back-door criterion is relevant for a wide range of commonly-used adjustment methods within a wide range of observational and / or experimental studies. This is true even when the underlying causal relationships are much more complicated than the stylized example in the previous subsection. In this subsection, we demonstrate these points by using the back-door criterion to identify sets of covariates sufficient to control confounding in a more complicated example. We then show how two widely used methods of adjustment—linear regression and subclassification on the estimated propensity score—can be used to estimate average causal effects as well as how these methods fall prey to the same problems discussed in the previous subsection.

Consider the graphical causal model in Figure 3. Here we see four measured background variables ($Z_1, Z_2, Z_3,$ and Z_4), a single treatment variable (X), an intermediate outcome variable (Z_5), and a final outcome variable (Y). In addition, there are numerous unmeasured exogenous variables (U_1, U_2, \dots, U_{11}).

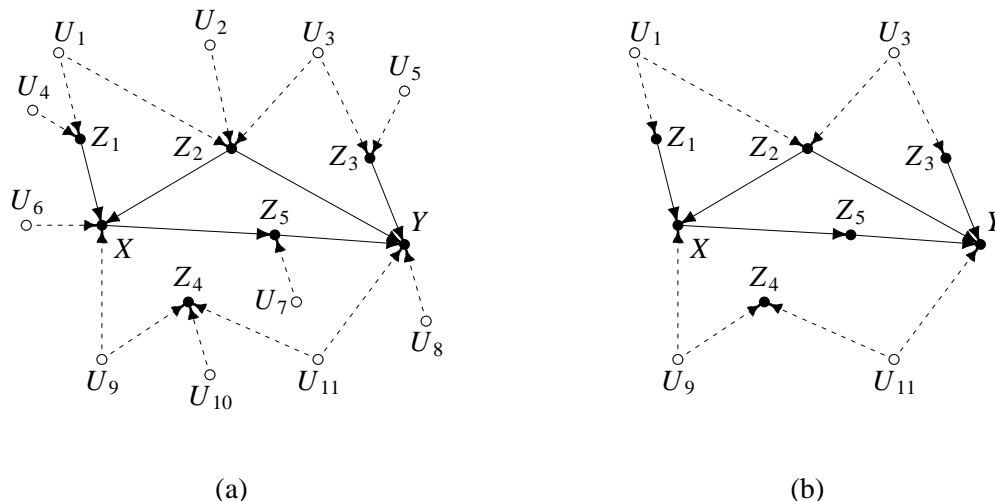


Figure 3: *Causal Graphs Consistent with Linear Structural Equations in Figure 4.* Panel (a) contains all exogenous variables while panel (b) excludes nodes and edges that are unnecessary for determining if and how the causal effect of X on Y is identified.

The research question of interest is to determine the total causal effect of X on Y . To identify

the sets of observed covariates that are sufficient for the control of confounding we check the back-door criterion for all subsets of the measured covariates. Here we see that $\{Z_1, Z_2\}$, $\{Z_2, Z_3\}$, and $\{Z_1, Z_2, Z_3\}$ all satisfy the back-door criterion. Thus adjusting for $\{Z_1, Z_2\}$, $\{Z_2, Z_3\}$, or $\{Z_1, Z_2, Z_3\}$ will provide a consistent estimate of the causal effect of X on Y .¹⁰ Note that Z_4 is not a member of any of the sets of adjustment variables above. Thus it should never be conditioned on— *even though it is associated with both X and Y given X and could be a pre-treatment variable.*

To make this discussion more concrete, we assume particular linear forms for the structural equations that are consistent with the causal graph in Figure 3. These linear structural equations are presented in Figure 4. In what follows, we generate 1,000,000 independent replicates from this system of equations and then use linear regression and subclassification on the estimated propensity score to estimate the causal effect of X on Y . More specifically, we show how in each case adjustment for $\{Z_1, Z_2\}$, $\{Z_2, Z_3\}$, or $\{Z_1, Z_2, Z_3\}$ provides an accurate estimate of the causal effect of X on Y , and how in each case conditioning on an incorrect set of adjustment variables produces a very inaccurate estimate of the effect of interest. Before proceeding, we note that the linearity of the equations in Figure 4 allows us to easily derive the true average causal effect $\mathbb{E}[Y|do(X = 1)] - \mathbb{E}[Y|do(X = 0)]$ from the parameter values in these equations. Here we see that the true average causal effect is 0.5.

Consider the estimation of $\mathbb{E}[Y|do(X = 1)] - \mathbb{E}[Y|do(X = 0)]$ using linear regression. While linear regression will typically not produce a consistent estimate of the potential outcome distribution, it will produce a consistent estimate of the average causal effect if the system is linear and the back-door criterion is satisfied for some subset of measured covariates (Pearl, 2000, Chapter 5).

¹⁰The structure of the causal graph in Figure 3 also suggests a fourth way to identify the causal effect of X on Y by using the intermediate outcome Z_5 . This is the so-called *front-door adjustment*. This is discussed in the next section.

$$\begin{array}{ll}
u_1 \stackrel{ind}{\sim} \mathcal{N}(0, 0.75^2) & u_7 \stackrel{ind}{\sim} \mathcal{N}(0.6, 1^2) \\
u_2 \stackrel{ind}{\sim} \mathcal{N}(0, 3^2) & u_8 \stackrel{ind}{\sim} \mathcal{N}(1, 1^2) \\
u_3 \stackrel{ind}{\sim} \mathcal{N}(5, 0.4^2) & u_9 \stackrel{ind}{\sim} \mathcal{N}(0, 5^2) \\
u_4 \stackrel{ind}{\sim} \mathcal{N}(0, 1^2) & u_{10} \stackrel{ind}{\sim} \mathcal{N}(0, 0.05^2) \\
u_5 \stackrel{ind}{\sim} \mathcal{N}(1, 0.8^2) & u_{11} \stackrel{ind}{\sim} \text{Binom}(4, 0.55) \\
u_6 \stackrel{ind}{\sim} \mathcal{N}(0, 5^2) & \\
\\
z_1 := u_1 + u_4 & z_3 := u_3 + u_5 \\
z_2 := 0.5u_1 + u_2 + u_3 & z_4 := -0.2u_9 + u_{10} - 0.25u_{11} \\
\\
x := \begin{cases} 1 & \text{if } z_1 + 2.5z_2 + u_6 - u_9 > 12.5 \\ 0 & \text{otherwise} \end{cases} \\
\\
z_5 := x + u_7 \\
y := 0.5z_5 + z_2 - 1.5z_3 + u_8 - 2u_{11}
\end{array}$$

Figure 4: *Specific Linear Structural Equations Consistent with Figure 3.* The results discussed in this section are based on 1,000,000 observations generated from this model.

Thus we would expect that the estimates of β_1 in the following regression equations

$$y = \beta_0 + \beta_1 x + \beta_2 z_1 + \beta_3 z_2 + \epsilon$$

$$y = \beta_0 + \beta_1 x + \beta_2 z_2 + \beta_3 z_3 + \epsilon$$

$$y = \beta_0 + \beta_1 x + \beta_2 z_1 + \beta_3 z_2 + \beta_4 z_3 + \epsilon$$

would all be close to the true value of 0.5. Looking at Table 3 we see that this is in fact the case.

We also see that, as we would expect, omitting Z_2 from $\{Z_1, Z_2, Z_3\}$ and only conditioning on Z_1 and Z_3 results in an estimated causal effect that is far from the truth.

Finally, note that when all of the observed covariates are included as right-hand-side variables in a regression model that the sign of the estimated causal effect is the opposite of that of the truth. While the sign and magnitude of this bias depend on the specific forms of the structural equations in Figure 4, the fact that conditioning on all the observed covariates ($\{Z_1, Z_2, Z_3, Z_4\}$) results in

an inconsistent estimate of the average causal effect only depends on the general relationships embodied in the graphical structure in Figure 3.

Adjustment Model	$\hat{\mathbb{E}}[Y do(X = 1)] - \hat{\mathbb{E}}[Y do(X = 0)]$
$y = \beta_0 + \beta_1x + \beta_2z_1 + \beta_3z_2 + \epsilon$	0.504
$y = \beta_0 + \beta_1x + \beta_2z_2 + \beta_3z_3 + \epsilon$	0.505
$y = \beta_0 + \beta_1x + \beta_2z_1 + \beta_3z_2 + \beta_4z_3 + \epsilon$	0.505
$y = \beta_0 + \beta_1x + \beta_2z_1 + \beta_3z_3 + \epsilon$	4.063
$y = \beta_0 + \beta_1x + \beta_2z_1 + \beta_3z_2 + \beta_4z_3 + \beta_5z_4 + \epsilon$	-0.179
$P(X = 1 z) = \Phi(\alpha_0 + \alpha_1z_1 + \alpha_2z_2)$	0.521
$P(X = 1 z) = \Phi(\alpha_0 + \alpha_1z_2 + \alpha_2z_3)$	0.531
$P(X = 1 z) = \Phi(\alpha_0 + \alpha_1z_1 + \alpha_2z_2 + \alpha_3z_3)$	0.522
$P(X = 1 z) = \Phi(\alpha_0 + \alpha_1z_1 + \alpha_2z_3)$	4.063
$P(X = 1 z) = \Phi(\alpha_0 + \alpha_1z_1 + \alpha_2z_2 + \alpha_3z_3 + \alpha_4z_4)$	-0.218
Truth	0.500

Table 3: *Adjustment Methods and Associated Estimated Average Causal Effects for Example Consistent with Figures 4 and 3.* The first five rows correspond to various regression adjustments. The second five rows correspond to various probit regressions for the estimated propensity scores. In these rows, the estimated causal effect is calculated by subclassifying on the estimated propensity score. The true average causal effect is in the last row.

Some intuition about what is happening in this linear example is the following. First note that the disturbance in the equation for Y always includes U_{11} . Marginally, X and U_{11} are independent. However, *given* Z_4 , X and U_{11} are generically *dependent*. More specifically, given the equations for X and Z_4 in Figure 4 we see that Z_4 will likely take a relatively large positive value when either U_9 , U_{11} , or both take a large negative value. Further, the probability of $X = 1$ is decreasing in realized values of U_9 . Thus, conditioning on a large positive value of Z_4 and observing that $X = 0$ provides evidence that U_{11} has taken a large negative value. Thus, when Z_4 is included as a right-hand-side variable in a regression model of Y on X the disturbance term (which is always defined conditionally on all covariates in the model) is no longer independent of X . As a result, the OLS estimator of the coefficient on X is no longer consistent for the total causal effect.

The same general patterns of bias emerge when one estimates the average causal effect of X on Y by subclassifying on the estimated propensity score. Since the entire system of equations is linear (and the true data generating process for X is consistent with a probit mode), we use

generic probit regression with main effects for the conditioning variables to estimate the propensity score. The various specifications for the propensity score model are given in the second five rows of Table 3. In general, the patterns of bias are identical to those seen in the equivalent linear model adjustments.¹¹ This should be of no surprise, since when the data are generated according to a linear model both methods are doing very similar things. Nevertheless, there are some interesting points to note here. First, as most researchers realize (and is made clear in the seminal work of Rosenbaum and Rubin (1983)) the propensity score model need not be consistent with the true data generating process for the treatment variable or even include that true data generating process as a special case of the estimated propensity score model. It is easy to prove that all that is required is that conditional ignorability holds given \mathbf{z} and that the statistical model used to estimate $P(x|\mathbf{z})$ is sufficiently flexible to accurately estimate this distribution. The fact that subclassifying on the estimated propensity scores from the propensity score model:

$$P(X = 1|z) = \Phi(\alpha_0 + \alpha_1 z_2 + \alpha_2 z_3)$$

provide essentially the same estimate as subclassifying on the propensity scores from the probit model that is consistent with the data generating process confirms this point.

What will be more surprising to many readers is the fact that subclassifying on the propensity scores from the model:

$$P(X = 1|z) = \Phi(\alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \alpha_4 z_4)$$

produces a badly biased estimate with the wrong sign. This propensity score model includes the true data generating model for X as a special case (with α_3 and α_4 both equal to 0) as well as all the other propensity score models that would allow one to construct consistent estimates of the causal effect of interest. Further, all of the covariates can be thought of as pre-treatment variables with no loss of generality. Thus many researchers might conclude there is no real harm to conditioning

¹¹The slight upward bias of most of the propensity score estimates is likely do to the fact that these estimates were based on a less than optimal subclassification plan.

on all the observed covariates. This is not correct. Conditioning on all the observed covariates produces a misleadingly biased estimate of the causal effect in this example.

In this section we have shown how a simple graphical criterion (the back door criterion) can allow one to examine a set of local causal assumptions to determine whether a particular causal effect is identified. Along the way, we have seen how a number of widely held beliefs are not entirely accurate. For example, *none* of the following statements are strictly true.

- Bias cannot increase as an additional pre-treatment variable is adjusted for.
- All pre-treatment variables that are associated with treatment assignment as well as with the outcome given treatment assignment are confounders and need to be adjusted for.
- Balance on all measured pre-treatment variables is necessary for consistent estimation of causal effects.
- Balance on all measured pre-treatment variables is sufficient for consistent estimation of causal effects.

A major strength of the Pearl model is that it not only provides a principled method of determining sufficient sets of adjustment variables but that it also helps clarify why the statements above are false. It does this by putting statements in a causal language that are at once consistent with the Neyman-Rubin framework and the mechanistic accounts of causation that many researchers carry around in their heads.

5 Is the Front Door Open?

In the previous section, we demonstrated a method for determining sufficient sets of conditioning variables in order to identify causal effects through the standard adjustments. In this section we demonstrate a method for identifying causal effects when none of the standard approaches work. In particular when some elements of the sufficient conditioning set(s) are not measurable, then conditional ignorability and the back door criterion do not hold, and we are forced to consider other options. Sensitivity analyses (Rosenbaum, 1995) provide one reasonable (and honest) approach to bounding causal effects, although there is some disagreement over which form of sensitivity analysis is most useful (see Rosenbaum (2002) and discussion). Traditional instrumental variable techniques

provide an approach to identification but are known to be sensitive to violations of the assumptions (Bound et al., 1995), while more recent instrumental variables techniques identify subpopulation causal effects (Angrist et al., 1996), however, there is disagreement over the importance of these subpopulations (see discussion of Angrist et al. (1996)). In contrast, the front door criterion (Pearl, 2000, p. 81) provides an alternative approach for *non-parametrically* identifying the ACE.

Definition 7 (Front-Door Criterion (Pearl (2000, p. 81))) *A set of variables \mathbf{Z} is said to satisfy the front-door criterion relative to an ordered pair of variables (X, Y) if :*

1. \mathbf{Z} intercepts all directed paths from X to Y ;
2. there is no back-door path from X to Z ; and
3. all back-door paths from Z to Y are blocked by X .

Graphically, the conditions of the front door criterion are depicted in Figure 5(b). In the following two examples, we demonstrate the front-door criterion and a more a general technique for the identification of causal effects that goes beyond the standard approaches based on conditional ignorability.

5.1 A Simple Non-parametric Example

As a stylized example we consider estimating the ACE of SAT coaching on admission to an arbitrary university. Suppose we have data on whether applicants partook in an SAT coaching class. Each applicant is partially characterized by two observed variables. Let Y be admission to the program, and let X be the use of SAT coaching.

$$Y \in \{0 \text{ (Not Admitted)}, 1 \text{ (Admitted)}\}$$

$$X \in \{0 \text{ (Not Coached)}, 1 \text{ (Coached)}\}$$

Suppose that we want to estimate the average causal effect of coaching on admission. This can be achieved by deriving the two distributions:

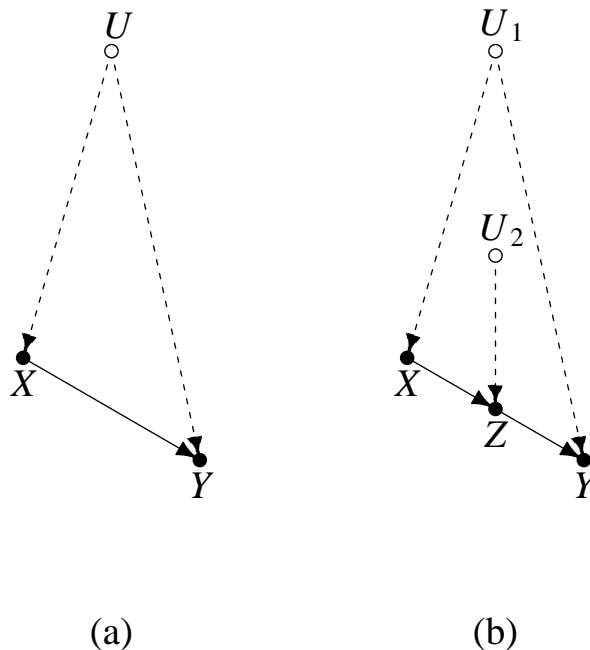


Figure 5: *DAGs consistent with nonignorable treatment assignment.* (a) DAG consistent for a model with only treatment and outcome variables measured. (b) DAG consistent with the front-door criterion.

$$P(y|do(X = 0)) \equiv P(Y_0 = y)$$

$$P(y|do(X = 1)) \equiv P(Y_1 = y),$$

and using these probability distributions for the potential outcomes to define the ACE or other population causal effects. However, in this example it will improve understanding if we further define the individual causal effects. Because both variables are binary, we can define the individual causal effects through a single latent variable (U) whose domain can be partitioned into four groups: those that achieve admittance to college regardless of whether they are coached ($U = \textit{Always}$), those that never achieve admittance to college regardless of whether they are coached ($U = \textit{Never}$), those that achieve admittance with coaching but don't without ($U = \textit{Helped}$), and those that don't achieve admittance with coaching but do without ($U = \textit{Hurt}$). Of course, we may wonder whether

there are really any of the “Hurt” individuals in the population, but a key feature of the front door procedure, is that we do not need to make such an assumption.

Given this partition, the following holds:

$$\mathbb{E}[Y_0] = 1 \cdot P(\textit{Always}) + 0 \cdot P(\textit{Never}) + 0 \cdot P(\textit{Helped}) + 1 \cdot P(\textit{Hurt})$$

$$\mathbb{E}[Y_1] = 1 \cdot P(\textit{Always}) + 0 \cdot P(\textit{Never}) + 1 \cdot P(\textit{Helped}) + 0 \cdot P(\textit{Hurt}).$$

Hence it is clear that the ACE is the difference in the helped and hurt proportions. Furthermore, because individuals self select into Not Coached/Coached, the probability of coaching is likely to be dependent on these latent types. This dependence means that the treatment assignment is not ignorable, and the prima facie causal effect will not identify the true ACE. Figure 5(a) shows a DAG that is consistent with this example, and one can easily check that the back-door criterion is not satisfied.

However, suppose we also observe the SAT scores (Z) for individuals (for illustrative purposes, we discretize scores into “good” ($Z = 0$) and “bad” ($Z = 1$) SAT scores). This is certainly a post-treatment variable, and hence we most likely should not adjust for it in the normal ways. However, the front door criterion provides another method for using this variable in estimation of the ACE of coaching on admission. Instead of characterizing individuals into the four previously defined latent types, we will more finely classify students into sixteen types as described by two four level latent variables (U_1, U_2). The latent variable U_1 describes the potential outcomes for admission based on an SAT score “treatment”. Therefore, students can be divided into four categories: those that achieve admittance to college regardless of SAT score ($U_1 = \textit{Always}$), those the never achieve admittance to college regardless of SAT score ($U_1 = \textit{Never}$), those that achieve admittance with good SAT scores but don’t with bad SAT scores ($U_1 = \textit{Helped}$), and those that achieve admittance with bad SAT scores but don’t with good SAT scores ($U_1 = \textit{Hurt}$). Again, it may stretch the imagination to assume that there are “Hurt” individuals in the population, but generally, we need not assume that $P(U_1 = \textit{Hurt}) = 0$. Similar to U_1 , the latent variable U_2 describes the potential

U_1 (Score \rightarrow Admission)	U_2 (Coaching \rightarrow Score)	U (Coaching \rightarrow Admission)
Always	Any	Always
Never	Any	Never
Helped	Always	Always
	Never	Never
	Helped	Helped
	Hurt	Hurt
Hurt	Always	Never
	Never	Always
	Helped	Hurt
	Hurt	Helped

Table 4: Mapping from the domains of U_1 and U_2 from Figure 5(b) into the domain of U from Figure 5(a)

outcomes for SAT score based on a coaching treatment. Again, students can be divided into four categories: those that receive good SAT scores regardless of whether they are coached ($U_2 = \textit{Always}$), those that receive bad SAT scores regardless of whether they are coached ($U_2 = \textit{Never}$), those that receive good SAT scores with coaching but bad scores without ($U_2 = \textit{Helped}$), those that receive bad SAT scores with coaching but good scores without ($U_2 = \textit{Hurt}$). If we assume that SAT coaching has an effect on college admission only through its effect on SAT score, then the original latent variable U is a function of U_1 and U_2 (Table 4 describes this relationship). This assumption may be questionable in this example because SAT coaching might have effects on other aspects of the application package (e.g. the admissions essay), however, the assumption is analogous to the usual exclusion restriction from instrumental variables models. Graphically, this assumption is reflected in the missing arrow from X to Y in Figure 5(b).

Suppose that we also make the following assumptions. First, applicants self select into Not Coached/Coached based on U_1 and not U_2 . This assumption will be reasonable if applicants do not know how much coaching will help/hurt them, but they do know their GPA and hence have some idea what kind of SAT score will be necessary for them to be accepted at the college. Graphically, this assumption is represented by the lack of an arrow from U_2 to X in Figure 5(b). Second, applicants' SAT score potential outcomes do not affect their admission to college (i.e.

missing arrow from U_2 to Y in Figure 5(b)). Third, applicants' admission potential outcomes do not affect their SAT score (i.e. missing arrow from U_1 to Z in Figure 5(b)). Finally, we define a joint distribution over U_1 and U_2 in terms of the population proportions of Always, Never, Helped, and Hurt for these variables, and we assume that these two variables are independent with respect to this distribution. Graphically, this is represented by the fact that we have split the original U variable into these two distinct variables.¹² Intuitively, these assumptions (i.e. missing arrows) allow us to identify the intermediate causal effects, and then to obtain the ACE of coaching on admission by multiplying the intermediate causal effects. For this simple example, we can derive this from the potential outcomes. Recall that the ACE of coaching on admission is $P(U = Helped) - P(U = Hurt)$. Our assumptions allow us to identify from observable data the ACE of coaching on score ($P(U_2 = Helped) - P(U_2 = Hurt)$) and the ACE of score on admission ($P(U_1 = Helped) - P(U_1 = Hurt)$). Using Table 4 and the assumptions above we get the following result.

$$P(U = Helped) = P(U_1 = Helped)P(U_2 = Helped) + P(U_1 = Hurt)P(U_2 = Hurt) \quad (13)$$

$$P(U = Hurt) = P(U_1 = Helped)P(U_2 = Hurt) + P(U_1 = Hurt)P(U_2 = Helped) \quad (14)$$

$$\begin{aligned} P(U = Helped) - P(U = Hurt) &= \{P(U_1 = Helped) - P(U_1 = Hurt)\} \\ &\cdot \{P(U_2 = Helped) - P(U_2 = Hurt)\} \end{aligned} \quad (15)$$

Equations (13), (14), and (15) show that we do not need treatment assignment (coaching) to be ignorable for the admissions potential outcomes in order to identify the ACE. In particular, if students select into coaching based on their U_1 classification, then the traditional ignorability assumption does not hold, and we cannot identify the ACE of coaching on admissions with any sort of regression or matching method based on the observed variables. However, we can identify the ACE of coaching on admissions by using (15) and the observed variables.

¹²In (Pearl, 2000, Ch. 3) this is represented by the lack of a dashed arc between the variables.

Of course, these assumptions cannot be justified for this illustrative example. It is unlikely that the potential outcomes described by U_1 and U_2 will be independent, and effects of coaching on SAT score or the effects of SAT score on admission will not be identified by the three observed variables in this example. However, the simple front door criterion is a special case of a far more general result. Intuitively, we may be able to identify the ACE by using adjustment methods to identify intermediate effects.

While it was easy to couch the assumptions and adjustments for this simple example in the language of the Neyman-Rubin model, for more complicated identification schemes, the rules of the *do*-Calculus are more convenient. The next subsection presents a more complicated non-parametric example that appeals to the rules of the *do*-Calculus.

5.2 A More Complicated Parametric Example

As stated above, front-door techniques generalize to more complicated models than the stylized model from the previous subsection. In this subsection, we demonstrate these points by using the rules of the *do*-Calculus to identify causal effects in a model where there is no sufficient set for conditional ignorability and the assumptions of the previous section do not hold. Of course, a new set of assumptions must hold in order for identification to be possible, but as with the switch from ignorability to conditional ignorability in Section 2, we may find the new assumptions to be more palatable than the assumptions from the previous subsection.

Consider the graphical model in Figure 6. In addition to the measured X , Z , and Y variables from Figure 5, we additionally measure the variables Q and W . In this graph, Q functions as an intermediate outcome in the same way that Z did in the previous subsection, while W functions as a confounder for the effects of X on Z and X on Q . In addition to the observed variables, we add the exogenous variables U_3 and U_4 that point into the new measured variables. Notice that this graph implicitly assumes independence between the U variables, but that by pointing U_1 into both X and Y , we assume an unmeasured confounder for the affect of X on Y .

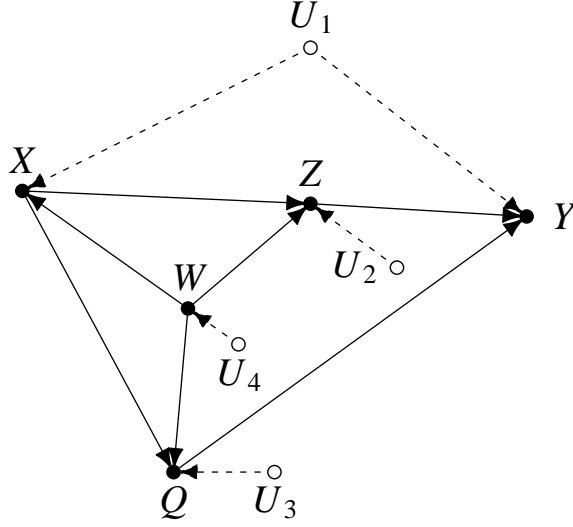


Figure 6: DAG consistent with the linear structural model in Figure 7.

Recall that the U variables must uniquely define all the endogenous variables within this model. Therefore, the definitions of the exogenous variables U_1 and U_2 have changed now that we have included other variables in the model. In order for the exogenous variables to uniquely define the individual causal effects, they must now specify potential variable values across a much larger domain. For example, U_1 must now describe the potential outcomes for Y given all possible cross classifications of Z and Q and the potential outcomes for X given all possible values of W .

Given this set up, we repeatedly apply the three rules of the *do*-Calculus along with the rules of *d*-Separation and the law of total probability to attempt to identify the ACE of X on Y . Specifically, we must write the unobserved post-intervention distribution in terms of observable pre-intervention distributions. Using the law of total probability as defined over submodels, $P(y|do(x)) = \sum_{z,q,w} P(y|z, q, w, do(x)) \cdot P(z, q|w, do(x)) \cdot P(w|do(x))$. Hence we divide the task into three parts: identifying $P(y|z, q, w, do(x))$, identifying $P(z, q|w, do(x))$, and identifying $P(w|do(x))$.

Identifying $P(y|z, q, w, do(x))$:

1. $P(y|z, q, w, do(x)) = P(y|z, q, do(x))$ by Rule 1
2. $P(y|z, q, do(x)) = P(y|do(z), do(q), do(x))$ by Rule 2
3. $P(y|do(z), do(q), do(x)) = P(y|do(z), do(q))$ by Rule 3
4. $P(y|do(z), do(q)) = P(y|do(z), do(q), w, x)$ by Rule 1
5. $P(y|do(z), do(q)) = \sum_{x,w} P(y|do(z), do(q), x, w) \cdot P(x, w, |do(z), do(q))$ by law of total probability
6. $\sum_{x,w} P(y|do(z), do(q), x, w) \cdot P(x, w)$ by Rule 3
7. $\sum_{x,w} P(y|z, q, x, w) \cdot P(x, w)$ by Rule 2

Identifying $P(z, q|w, do(x))$:

1. $P(z, q|w, do(x)) = P(z, q|w, x)$ by Rule 2
2. $P(z, q|w, x) = P(z|w, x) \cdot P(q|w, x)$ by d -Separation

Identifying $P(w|do(x))$:

1. $P(w|do(x)) = P(w)$ by Rule 3

Therefore, the overall causal effect of X on Y can be written as

$$\begin{aligned}
P(y|do(x)) &= \sum_{z,q,w} P(y|z, q, w, do(x)) \cdot P(z, q|w, do(x)) \cdot P(w|do(x)) \\
&= \sum_{z,q,w} \left[\sum_{x',w'} \{P(y|z, q, x', w') \cdot P(x', w')\} P(z|w, x) \cdot P(q|w, x) \cdot P(w) \right] \quad (16)
\end{aligned}$$

The formula in (16) contains only observable distributions and in principle can be estimated non-parametrically. In practice, this task can be quite difficult due to the difficulty of non-parametrically estimating joint and conditional distributions with lots of variables. To make this discussion more

$$\begin{array}{ll}
u_1 \stackrel{ind}{\sim} \mathcal{N}(0, 1) & u_3 \stackrel{ind}{\sim} \mathcal{N}(0, 1) \\
u_2 \stackrel{ind}{\sim} \mathcal{N}(0, 1) & u_4 \stackrel{ind}{\sim} \mathcal{N}(0, 1) \\
\\
w := u_4 & x := \begin{cases} 1 & \text{if } w + u_1 > 0 \\ 0 & \text{otherwise} \end{cases} \\
z := w + x + u_2 & q := w + x + u_3 \\
\\
y := z + q + 2u_1 &
\end{array}$$

Figure 7: *Specific Linear Structural Equations Consistent with Figure 6.* The results discussed in this section are based on 1,000,000 observations generated from this model.

concrete, we assume a particular parametric formulation for equations that are consistent with the causal graph in Figure 6. These forms are presented in Figure 7.

Using the distributions and structural equations from Figure 7, we generated 1,000,000 observations in order to validate a number of estimation methods. Table 5 contains estimates of the ACE with the first nine rows corresponding to all possible additive regressions that include X , and the next four rows corresponding to the front door adjustment. While the true ACE is -2 , all nine regression adjustments result in positive estimates. However, for this linear additive example, the front door adjustment can be derived from the coefficients from three regressions: the regression of Z on X and W , the regression of Q on X and W , and the regression of Y on Z , Q , X , and W . Intuitively, the front door formula presented in the second to last line of the table reflects the X effect on Y through Z , added to the X effect on Y through Q . However, the reader should note that this simplification is special to linear structural equation models, and in general the formula in (16) must be used for a causal model with a structure defined by Figure 6.

This section has demonstrated that conditional ignorability is not necessary for the non-parametric identification of causal effects. The front door criterion and the general identification rules of the *do*-Calculus provide for a wealth of identification possibilities outside of the standard adjustments. Furthermore, the while identification through (15) for the simple non-parametric example does not

Estimation Method	$\hat{\mathbb{E}}[Y do(X = 1)] - \hat{\mathbb{E}}[Y do(X = 0)]$
$y = \beta_0 + \beta_1x + \epsilon$	2.507
$y = \beta_0 + \beta_1x + \beta_2z + \epsilon$	2.377
$y = \beta_0 + \beta_1x + \beta_2q + \epsilon$	2.377
$y = \beta_0 + \beta_1x + \beta_2w + \epsilon$	1.304
$y = \beta_0 + \beta_1x + \beta_2z + \beta_3q + \epsilon$	2.322
$y = \beta_0 + \beta_1x + \beta_2z + \beta_3w + \epsilon$	2.306
$y = \beta_0 + \beta_1x + \beta_2q + \beta_3w + \epsilon$	2.310
$y = \beta_0 + \beta_1x + \beta_2z + \beta_3q + \beta_4w + \epsilon$	3.311
$z = \gamma_0 + \gamma_1x + \epsilon$	
$q = \delta_0 + \delta_1x + \epsilon$	
$y = \beta_0 + \beta_1x + \beta_2z + \beta_3q + \beta_4w + \epsilon$	
$\gamma_1 \cdot \beta_2 + \delta_1 \cdot \beta_3$	-2.007
True ACE	-2

Table 5: *Estimation Methods and Associated Estimated Average Causal Effects for an Example Consistent with Figures 7 and 6.* The first nine rows correspond to the putative causal effects from additive linear regressions on all possible combinations of the observed variables (including X). All of these regression formulations produce positive estimates of the average causal effect. The second four rows define the front door adjustment in this simplified additive linear framework. The front door estimate in the fourth line comes from the regression estimates in the previous three lines, and this estimate corresponds to the true ACE in the last row.

require the NPSEM or graphical models¹³, and other such identifications may be possible without the approach presented in Pearl (2000), the *do*-Calculus provides a general strategy for the identification of causal effects from observational data. We are not aware of any analogous technique.

6 Conclusion

In this paper we have attempted to make the case that the NPSEMs of Pearl (a) are consistent with how many empirical political scientists think about causality within their subject-matter area, (b) are consistent with the powerful and well-respected Neyman-Rubin Causal Model, and (c) offer the potential to improve the practice of causal inference in political science. Political scientists tend to think in terms of causal mechanisms and the Pearl model deals explicitly in mechanisms. The Pearl model is a model of deterministic potential outcomes and many of the key ideas and results from the Neyman-Rubin model have direct analogies in the Pearl framework. The Pearl model

¹³We found this simple formulation by starting with the graph and defining the subpopulations. We would not have known to look without the graphical criteria.

offers several advantages to applied researchers. We discuss each of these below.

The Pearl framework provides explicit rules for covariate selection based on the causal graph. Given a graph (or a set of plausible graphs), the back-door criterion supplies the collection of sufficient conditioning sets. In more general cases with possible unmeasured confounding, the rules of the *do*-Calculus provide a powerful strategy for the identification of causal effects. Finally, the transparent local assumptions about mechanisms in the Pearl model are easier to consider, debate, and potentially reject than a global assumption of conditional ignorability. Causal inference by its very nature relies on untestable causal assumptions. As such it is imperative that any method for causal inference allow researchers to state these causal assumptions as plainly and comprehensibly as possible so that subject-matter experts can easily weigh in on the plausibility of these assumptions.

While the Pearl model provides the advantages described above, we would certainly not advocate abandoning the Neyman-Rubin framework that has been so useful to political scientists. Furthermore, while some assumptions are easier to assess in Pearl model, others will be easier to assess in the Neyman-Rubin model (see the discussion of Pearl (1995)). A hybrid approach that utilizes the strengths of both models seems appropriate.

References

- Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91:444–455.
- Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association* 90:443–450.
- Campbell, Angus, Philip Converse, Warren E. Miller, and Donald Stokes. 1960. *The American Voter*. New York: John Wiley.
- Clarke, Kevin A. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22:341–352.
- Cochran, William G. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrics* 24(2):295–313.
- Collier, David, and Henry E. Brady. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield.
- Cox, Gary W. 1997. *Making Votes Count: Strategic Coordination in the World's Electoral Systems*. Cambridge: Cambridge University Press.
- Elster, Jon. 1989a. *The Cement of Society*. Cambridge: Cambridge University Press.
- Elster, Jon. 1989b. *Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.
- Elster, Jon. 1998. "A Plea for Mechanisms." In *Social Mechanisms: An Analytical Approach to Social Theory* (Peter Hedström, and Richard Swedberg, editors), Cambridge: Cambridge University Press.
- Fearon, James D. 1995. "Rationalist Explanations for War." *International Organization* 49(3):379–414.
- Fox, John. 1997. *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage.
- Greene, William H. 2000. *Econometric Analysis*. New York: Macmillan, fourth edition.
- Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.
- Kmenta, Jan. 1986. *Elements of Econometrics*. New York: Macmillan, second edition.
- Neyman, Jerzy, (with K. Iwazskiewicz, and S. Kolodziejczyk). 1935. "Statistical Problems in Agricultural Experimentation." *Supplement of Journal of the Royal Statistical Society* 2:107–180.
- Pearl, Judea. 1995. "Causal Diagrams for Empirical Research." *Biometrika* 82:669–710.
- Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.
- Pindyck, Robert S., and Daniel L. Rubinfeld. 1998. *Economic Models and Economic Forecasts*. Boston: Irwin McGraw Hill, fourth edition.

- Robins, J.M. 1986. "A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect." *Mathematical Modeling* 7:1393–1512.
- Rosenbaum, Paul R. 1995. *Observational Studies*. New York: Springer.
- Rosenbaum, Paul R. 2002. "Covariance Adjustment in Randomized Experiments and Observational Studies." *Journal of the American Statistical Association* 90:443–450.
- Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.
- Rosenbaum, Paul R., and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79:516–524.
- Rothman, Kenneth J. 1986. *Modern Epidemiology*. Boston: Little Brown.
- Rothman, Kenneth J., and Sander Greenland. 1998. *Modern Epidemiology*. Philadelphia: Lippincott-Raven.
- Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5):688–701.
- Rubin, Donald B. 1977. "Assignment to Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2(1):1–26.
- Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6(1):34–58.
- Schlesselman, James J. 1982. *Case-Control Studies: Design Conduct Analysis*. New York: Oxford University Press.
- Spirtes, P., C. Glymour, and R. Scheines. 1993. *Causation, Prediction, and Search*. New York: Springer.