Forthcoming, *Journal of Causal Inference* (Causal, Casual, and Curious Section).

TECHNICAL REPORT
R-459
March 2017

A Linear "Microscope" for Interventions and Counterfactuals

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
(310) 825-3243
judea@cs.ucla.edu

Abstract

This note illustrates, using simple examples, how causal questions of non-trivial character can be represented, analyzed and solved using linear analysis and path diagrams. By producing closed form solutions, linear analysis allows for swift assessment of how various features of the model impact the questions under investigation. We discuss conditions for identifying total and direct effects, representation and identification of counterfactual expressions, robustness to model misspecification, and generalization across populations.

Keywords: causal inference, structural equation models, counterfactuals, generalization, robustness.

1 Introduction

Two years ago, I wrote a paper entitled "Linear Models: A Useful 'Microscope' for Causal Analysis" (Pearl, 2013) in which linear structural equations models (SEM), were used as "microscopes" to illuminate causal phenomenon that are not easily managed in nonparametric models. In particular, linear SEMs enable us to derive close-form expressions for causal parameters of interest and to easily test or refute conjectures about the behavior of those parameters and what aspects of the model control this behavior. I now venture to leverage the simplicity of linear SEMs to illuminate interventions and counterfactuals, also called "potential outcomes," which often present a formidable challenge to non-parametric analysis.

After reviewing the basic notions of path analysis and counterfactual logic, we will demonstrate, using simple examples, how concepts and issues in modern counterfactual analysis can be understood and analyzed in SEM. These include: Causal effect identification, mediation, the mediation fallacy, unit-specific effects, the effect of treatment on the treated (ETT), generalization across populations, and more.

Section 2 reviews the fundamentals of path analysis as summarized in (Pearl, 2013). Section 3 introduces d-separation and the graphical definitions of interventions and counterfactuals, and provides the basic tools for the identification of interventional predictions and counterfactual expressions in linear models. Section 4 proceeds to demonstrate how these tools help to illuminate specific problems of causal and counterfactual nature, including mediation, sequential identification, robustness, and ignorability tests.

2 Preliminaries[1]

2.1 Covariance, regression, and correlation

We start with the standard definition of variance and covariance on a pair of variables X and Y. The *variance* of X is defined as

$$\sigma_x^2 = E[X - E(X)]^2$$

and measures the degree to which X deviates from its mean $E(X)$.

The *covariance* of X and Y is defined as

$$\sigma_{xy} = E[(X - E(X))(Y - E(Y))]$$

and measures the degree to which X and Y covary.

Associated with the covariance, we define two other measures of association: (1) the regression coefficient β_{yx} and (2) the correlation coefficient ρ_{yx}. The relationships between the three is given by the following equations:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{1}$$

$$\beta_{yx} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{\sigma_y}{\sigma_x}\rho_{xy}. \tag{2}$$

We note that $\rho_{xy} = \rho_{yx}$ is dimensionless and is confined to the unit interval; $0 \leq \rho_{xy} \leq 1$. The regression coefficient, β_{yx}, represents the slope of the least square error line in the prediction of Y given X

$$\beta_{yx} = \frac{\partial}{\partial x} E(Y|X = x).$$

2.2 Partial correlations and regressions

Many questions in causal analysis concern the change in a relationship between X and Y conditioned on a given set Z of variables. The easiest way to define this change is through the *partial regression coefficient* $\beta_{yx \cdot z}$ which is given by

$$\beta_{yx \cdot z} = \frac{\partial}{\partial x} E(Y|X = x, Z = z).$$

[1] This section is taken from (Pearl, 2013) and can be skipped by readers familiar with multiple regression, path diagrams and Wright's rules of path tracing.

In words, $\beta_{yx\cdot z}$ is the slope of the regression line of Y on X when we consider only cases for which $Z = z$.

The partial correlation coefficient $\rho_{xy\cdot z}$ can be defined by normalizing $\beta_{yx\cdot z}$:

$$\rho_{xy\cdot z} = \beta_{yx.z}\sigma_{x\cdot z}\sigma_{y\cdot z}.$$

A well known result in regression analysis (Crámer, 1946) permits us to express $\rho_{xy\cdot z}$ recursively in terms of pair-wise regression coefficients. When Z is singleton, this reduction reads:

$$\rho_{yx\cdot z} = \frac{\rho_{yx} - \rho_{yz}\rho_{xz}}{[(1-\rho_{yz}^2)(1-\rho_{xz}^2)]^{\frac{1}{2}}}. \tag{3}$$

Accordingly, we can also express $\beta_{yx\cdot z}$ and $\sigma_{yx\cdot z}$ in terms of pair-wise relationships, which gives:

$$\sigma_{yx\cdot z} = \sqrt{\sigma_{xx} - \sigma_{xz}^2/\sigma_z^2}\sqrt{\sigma_{yy} - \sigma_{yz}^2/\sigma_z^2} \quad \rho_{yx\cdot z} \tag{4}$$

$$\sigma_{yx\cdot z} = \sigma_x^2[\beta_{yx} - \beta_{yz}\beta_{zx}]\sigma_{yx} - \frac{\sigma_{yz}\sigma_{zx}}{\sigma_z^2} \tag{5}$$

$$\beta_{yx\cdot z} = \frac{\beta_{yx} - \beta_{yz}\beta_{zx}}{1 - \beta_{zx}^2\sigma_x^2/\sigma_z^2} = \frac{\sigma_z^2\sigma_{yx} - \sigma_{yz}\sigma_{zx}}{\sigma_x^2\sigma_z^2 - \sigma_{xz}^2} = \frac{\sigma_y}{\sigma_x}\frac{\rho_{yx} - \rho_{yz}\cdot\rho_{zx}}{1 - \rho_{xz}^2}. \tag{6}$$

Note that none of these conditional associations depends on the level z at which we condition variable Z; this is one of the features that makes linear analysis easy to manage and, at the same time, limited in the spectrum of relationships it can capture.

2.3 Path diagrams and structural equation models

A linear structural equation model (SEM) is a system of linear equations among a set V of variables, such that each variable appears on the left hand side of at most one equation. For each equation, the variable on its left hand side is called the *dependent* variable, and those on the right hand side are called *independent* or *explanatory* variables. For example, the equation below

$$Y = \alpha X + \beta Z + U_Y \tag{7}$$

declares Y as the dependent variable, X and Z as explanatory variables, and U_Y as an "error" or "disturbance" term, representing all factors omitted from V that, together with X and Z determine the value of Y. A structural equation should be interpreted as a natural process, i.e., to determine the value of Y, nature consults the value of variables X, Z and U_Y and, based on their linear combination in (7), assigns a value to Y.

This interpretation renders the equality sign in Eq. (7) non-symmetrical, since the values of X and Z are not determined by inverting (7) but by other equations, for example,

$$X = \gamma Z + U_X \tag{8}$$

$$Z = U_Z. \tag{9}$$

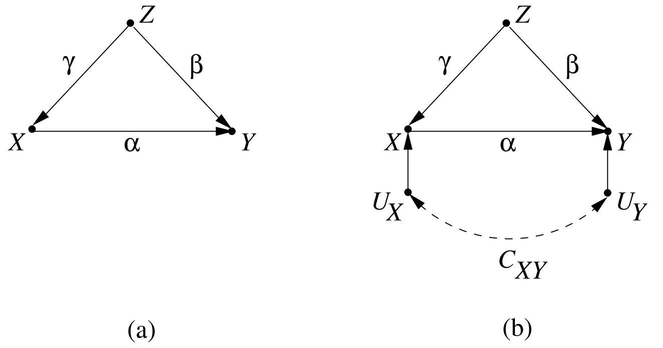

Figure 1: Path diagrams capturing the directionality of the assignment process of Eqs. (7)–(9) as well as possible correlations among omitted factors.

The directionality of this assignment process is captured by a *path-diagram*, in which the nodes represent variables, and the arrows represent the (potentially) non-zero coefficients in the equations. The diagram in Fig. 1(a) represents the SEM equations of (7)–(9) and the assumption of zero correlations between the U variables,

$$\sigma_{U_X,U_Y} = \sigma_{U_X,U_Z} = \sigma_{U_Z,U_Y} = 0.$$

The diagram in Fig. 1(b) on the other hand represents Eqs. (7)–(9) together with the assumption

$$\sigma_{U_X,U_Z} = \sigma_{U_Z,U_Y} = 0$$

while $\sigma_{U_X,U_Y} = C_{XY}$ remains undetermined.

The coefficients α, β, and γ are called *path coefficients*, or *structural parameters* and they carry causal information. For example, α stands for the change in Y induced by raising X one unit, while keeping all other variables constant.[2]

The assumption of linearity makes this change invariant to the levels at which we keep those other variables constant, including the error variables; a property called "effect homogeneity." Since errors (e.g., U_X, U_Y, Y_Z) capture variations among individual units (i.e., subjects, samples, or situations), effect homogeneity amounts to claiming that all units react equally to any treatment, which may exclude applications with profoundly heterogeneous subpopulations.

2.4 Wright's path-tracing rules

In 1921, the geneticist Sewall Wright developed an ingenious method by which the covariance σ_{xy} of any two variables can be determined swiftly, by mere inspection of

[2] In Section 3 we will give α a more formal definition using both interventional interpretation:

$$\alpha = \frac{\partial}{\partial x} E[(Y|do(x), do(z))]$$

and a counterfactual interpretation $\alpha = \frac{\partial}{\partial x} Y_{xz}(u)$.

the diagram (Wright, 1921). Wright's method consists of equating the (standardized[3]) covariance $\sigma_{xy} = \rho_{xy}$ between any pair of variables with a sum of products of path coefficients and error covariances along all d-connected paths between X and Y. A path is d-connected if it does not traverse any collider (i.e., head-to-head arrows, as in $X \rightarrow Y \leftarrow Z$).

For example, in Fig. 1(a), the standardized covariance σ_{xy} is obtained by summing α with the product $\beta\gamma$, thus yielding $\sigma_{xy} = \alpha + \beta\gamma$, while in Fig. 1(b) we get: $\sigma_{xy} = \alpha + \beta\gamma + C_{XY}$. Note that for the pair X and Z, we get $\sigma_{xz} = \gamma$ since the path $X \rightarrow Y \leftarrow Z$ is not d-connected.

The method above is valid for standardized variables, namely, variables normalized to have zero mean and unit variance. For non-standardized variables the method needs to be modified slightly, multiplying the product associated with a path p by the variance of the variable that acts as the "root" for path p. For example, for Fig. 1(a) we have $\sigma_{xy} = \sigma_x^2\alpha + \sigma_z^2\beta\gamma$, since X serves as the root for path $X \rightarrow Y$ and Z serves as the root for $X \leftarrow Z \rightarrow Y$. In Fig. 1(b), however, we get $\sigma_{xy} = \sigma_x^2\alpha + \sigma_z^2\beta\gamma + C_{XY}$ where the double arrow $U_X \leftrightarrow U_Y$ serves as its own root.

2.5 Computing partial correlations using path diagrams

The reduction from partial to pair-wise correlations summarized in eqs. (4)–(6), when combined with Wright's path-tracing rules permits us to extend the latter so as to compute partial correlations using both algebraic and path tracing methods. For example, to compute the partial regression coefficient $\beta_{yx\cdot z}$, we start with a standardized model where all variances are unity (hence $\sigma_{xy} = \rho_{xy} = \beta_{xy}$), and apply Eq. (6) with $\sigma_x = \sigma_z = 1$ to get:

$$\beta_{yx\cdot z} = \frac{(\sigma_{yx} - \sigma_{yz}\sigma_{zx})}{(1 - \sigma_{xz}^2)} \tag{10}$$

At this point, each pair-wise covariance can be computed from the diagram through path-tracing and, substituted in (10), yields an expression for the partial regression coefficient $\beta_{yx\cdot z}$.

To witness, the pair-wise covariances for Fig. 1(a) are:

$$\sigma_{yx} = \alpha + \beta\gamma \tag{11}$$

$$\sigma_{xz} = \gamma \tag{12}$$

$$\sigma_{yz} = \beta + \alpha\gamma \tag{13}$$

Substituting in (10), we get

$$\begin{aligned}\beta_{yx\cdot z} &= [(\alpha + \beta\gamma) - (\beta + \gamma\alpha)\gamma]/(1 - \gamma^2) \\ &= \alpha(1 - \gamma^2)/(1 - \gamma^2) \\ &= \alpha \end{aligned} \tag{14}$$

[3] Standardized parameters refer to systems in which (without loss of generality) all variables are normalized to have zero mean and unit variance, which significantly simplifies the algebra.

Indeed, we know that, for a confounding-free model like Fig. 1(a) the direct effect α is identifiable and given by the partial regression coefficient $\beta_{xy \cdot z}$. Repeating the same calculation on the model of Fig. 1(b) yields:

$$\beta_{yx \cdot z} = \alpha + C_{XY}$$

leaving α non-identifiable.

2.6 Reading vanishing partials from path diagrams

When considering a set $Z = Z_1, Z_2, \ldots, Z_k$ of regressors the partial correlation $\rho_{yx \cdot z_1, z_2, \ldots, z_k}$ can be computed by applying Eq. (3) recursively. However, when the number of regressors is large, the partial correlation becomes unmanageable. Vanishing partial correlations, however, can be readily identified from the path diagram without resorting to algebraic operations. This reading, which is essential for the analysis of interventions, is facilitated through a graphical criterion called d-separation (Pearl, 1988). In other words, the criterion permits us to glance at the diagram and determine when a set of variables $Z = Z_1, Z_2, \ldots, Z_k$ renders the equality $\rho_{yx \cdot z} = 0$.

The idea of d-separation is to associate zero correlation with separation; namely, the equality $\rho_{yx \cdot z} = 0$ would be valid whenever the set Z "separates" X and Y in the diagram. The only twist is to define separation in a way that takes proper account of the directionality of the arrows in the diagram.

Definition 1 (*d*-Separation) *A path p is blocked by a set of nodes Z if and only if*

1. *p contains a chain of nodes $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$ such that the middle node B is in Z (i.e., B is conditioned on), or*

2. *p contains a collider $A \rightarrow B \leftarrow C$ such that the collision node B is not in Z, and no descendant of B is in Z.*

If Z blocks every path between two nodes X and Y, then X and Y are d-separated, conditional on Z, and then the partial correlation coefficient $\rho_{yx \cdot z}$ vanishes (Pearl, 2009).

Armed with the ability to read vanishing partials, we are now prepared to demonstrate some peculiarities of interventions and counterfactuals.

3 Interventions and Counterfactuals in Linear Systems

3.1 Interventions and their effects

Consider an experiment in which we intervene on variable X and set it to constant $X = x$. Let $E[Y|do(x)]$ denote the expected value of outcome Y under such an intervention. The relationship between $E[Y|do(x)]$ and the parameters of any given model can readily be obtained by explicating how an intervention modifies the data-generating process. In particular the intervention $do(x)$ overrides all preexisting causes of X and, hence, transforms the graph G into a modified graph $G_{\overline{X}}$ in which all arrows entering X are eliminated, as shown in Fig. 2(b).

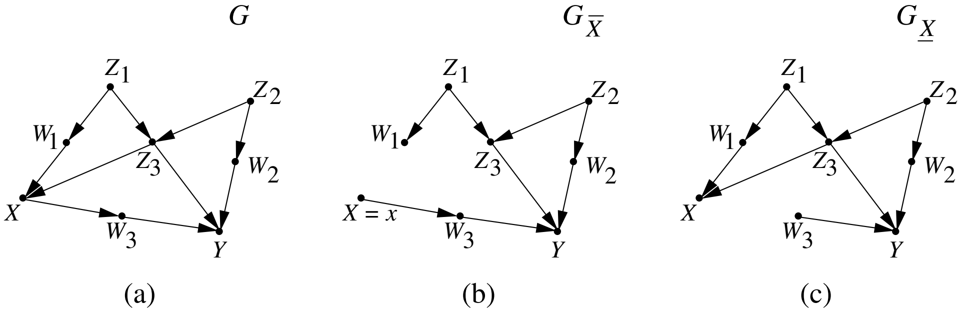

Figure 2: Illustrating the graphical reading of interventions. (a) The original graph. (b) The modified graph $G_{\overline{X}}$ representing the intervention $do(x)$. (c) The modified graph $G_{\underline{X}}$ in which separating X from Y represents non-confoundedness.

Thus, the interventional expectation $E[Y|do(x)]$ is given by the conditional expectation $E(Y|X = x)$ evaluated in the modified model $G_{\overline{X}}$. Applying Wright's Rule to this model, we obtain a well known result in path analysis: $E[Y|do(x)] = \tau x$, where τ stands for the sum of products of path coefficients along all paths directed from X to Y.

Likewise, the average causal effect (ACE) of X on Y, defined by the difference

$$ACE = E[Y|do(x + 1)] - E[Y|do(x)] \tag{15}$$

is a constant, independent of x, and is given by τ,

In the early days of path analysis, total effects were estimated by first estimating all path coefficients along the causal paths and then summing up products along those paths. The d-separation criterion of Definition 1 simplifies this computation significantly and leads to the following theorem.

Theorem 1 *(Identification of total effects) The total effect of X on Y can be identified in graph G whenever a set Z of observed variables exists, non-descendants of Y, that d-separates X and Y in the graph $G_{\underline{X}}$ in which all arrows emanating from X are removed. Moreover, whenever such a set Z exists, the causal effect is given by the (partialed) regression slope*

$$ACE = \beta_{yx \cdot z}. \tag{16}$$

This identification condition is known as *backdoor*, and it is written as $(X \perp\!\!\!\perp Y|Z)_{G_{\underline{X}}}$. In Fig. 2(c), for example, the set $Z = \{Z_3, W_2\}$ satisfies the backdoor condition, while the set $Z = \{Z_3\}$ does not. Note that Eq. (16) is valid even if some of the parameters along the causal paths cannot be estimated. In Fig. 2(c), the path coefficient along the arrow $W_3 \to Y$ need not be estimable, the total effect will still be given by Eq. (16).

A modification of Theorem 1 is required whenever the target quantity is the *direct*, rather than the *total* effect of X on Y. In this case, the parameter α on the arrow connecting X and Y can be identified using the following Theorem (Pearl, 2009, Ch. 5.3.1).

Theorem 2 *(Single-door Criterion) Let G be any acyclic causal graph in which α is the coefficient associated with arrow $X \to Y$, and let G_α denote the diagram that results when $X \to Y$ is deleted from G. The coefficient α is identifiable if there exists a set of variables Z such that (i) Z contains no descendant of Y and (ii) Z d-separates X from Y in G_α. If Z satisfies these two conditions, then α is equal to the regression coefficient $\beta_{yx \cdot z}$. Conversely, if Z does not satisfy these conditions, then $\beta_{yx \cdot z}$ is not a consistent estimand of α (except in rare instances of measure zero).*

In Fig. 1(a), for example, the parameter α is identified by $\beta_{yx \cdot z}$ because Z d-separates X from Y in G_α. In Fig. 1(b), on the other hand, Z fails to d-separate X from Y in G_α and, hence, α is not identifiable by regression.

Usually, to identify a direct effect α the set Z needs to include descendants of X. For example, if α stands for the direct effect of Z_3 on Y in Fig. 2(a), then the set Z needs to include descendants of Z_3, to block the path $Z_3 \to X \to W_3 \to Y$. However, $Z = \{X, Z_2\}$ is admissible, as well as $Z = \{W_3, W_2\}$, but not $Z = \{X, W_3\}$.

A full account of identification conditions in linear systems is given in Chen and Pearl (2015).

There is one more interventional concept that deserves our attention before we switch to discuss counterfactuals: covariate-specific effect. Assume we are interested in predicting the interventional expectation of Y for a subset of individuals for whom $Z = z$, where Z is a pre-intervention set of characteristics. We write this expectation as $E[Y|do(x), z]$, and define it as the conditional expectation of Y, given z, in the modified post-intervention model, depicted by $G_{\overline{X}}$. Formally,

$$P(y|do(x), z) = P(y, z|do(x))/P(z|do(x)). \tag{17}$$

Since $Z = z$ is pre-intervention event, it will not be affected by the intervention, so $P(z|do(x)) = P(z)$. Therefore, $E[Y|do(x), z]$ reduces to $\tau x + cz$ where c is the regression slope of Y on Z in $G_{\overline{X}}$. For example, in the model of Fig. 1 we have

$$\begin{aligned} c &= \beta &&= \beta_{yz \cdot x} \\ \tau &= \alpha &&= \beta_{yx \cdot z} \end{aligned}$$

hence

$$E[Y|do(x), z] = \beta_{yx \cdot z} x + \beta_{yz \cdot x} z.$$

We see that, in general, the z-specific causal effect $E[Y|do(x), z]$ is identifiable if and only if the total effect τ is identifiable. This stands in sharp contrast to non-linear models where conditioning on Z may prevent the identification of the z-specific causal effect (Pearl, 2015a).

If however Z is affected by the intervention and our interest lies in the expected outcome of individuals currently at level $Z = z$ had they been exposed to intervention $X = x$, Eq. (17) no longer represents the desired quantity, and we must use counterfactual analysis instead (see Section 4.4).

3.2 The Graphical representation of counterfactuals

The *do*-operator facilitates the estimation of average causal effects, with the average ranging either over the entire population or over the z-specific sub-population. In contrast, counterfactual analysis deals with behavior of individuals for which we have certain observations, or evidence (e). A counterfactual query asks, "Given that we observe $E = e$ for an individual u, what would we expect the value of Y to be for that individual if X had been x?" For example, given that Joe's salary is $Y = y$, what would his salary be had he had x years of education ($X = x$)? This expectation is denoted $E[Y_x|Y = y]$. The conditioning event $Y = y$ represents the observed evidence (e) while the subscript x represents a hypothetical condition specified by the counterfactual antecedent. Structural equation models are able to answer counterfactual queries of this sort, using a model modification operation similar to the *do*-operator.

Let M_x stand for the modified version of M, with the equation of X replaced by $X = x$. The formal definition of the counterfactual $Y_x(u)$ reads

$$Y_x(u) = Y_{M_x}(u) \tag{18}$$

In words: The counterfactual $Y_x(u)$ in model M is defined as the solution for Y in the "surgically modified" submodel M_x. Equation (18) was called "The Fundamental Law of Counterfactuals" (Pearl et al., 2016) for it allows us to take our scientific conception of reality, M, and use it to answer all counterfactual questions of the type "What would Y be had X been x?"

Eq. (18) also tells us how we can find the potential outcome variable Y_x in the graph. If we modify model M to obtain the submodel M_x, then the outcome variable Y in the modified model is the counterfactual Y_x in the original model.

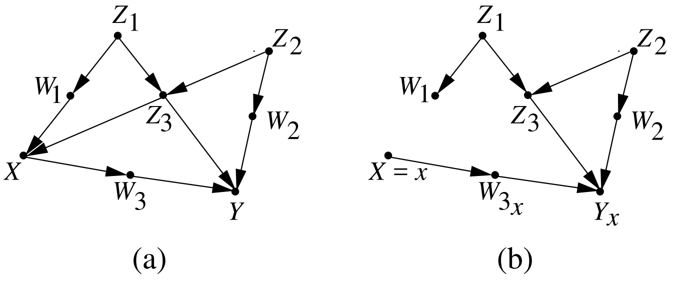

Figure 3: Illustrating the graphical reading of counterfactuals. (a) The original model. (b) The modified model M_x in which the node labeled Y_x represents the potential outcome predicated on $X = x$.

Since modification calls for removing all arrows entering the variable X, as illustrated in Fig. 3(b), we see that the node associated with Y serves as a surrogate for Y_x, with the understanding that the substitution is valid only under the modification.

This temporary visualization of counterfactuals is sufficient to describe the statistical properties of Y_x and how those properties depend on other variables in the model. In particular, the statistical variations of Y_x are governed by all exogenous variables capable of influencing Y when X is held constant, as in Fig. 2(b). Under such conditions, the set of variables capable of transmitting variations to Y are the parents of Y, (observed and unobserved) as well as parents of nodes on the pathways between X and Y. In Fig. 2(b), for example, these parents are $\{Z_3, W_2, U_3, U_Y\}$, where U_Y and U_3, the error terms of Y and W_3, are not shown in the diagram. Any set of variables that blocks a path to these parents also blocks that path to Y_x, and will result therefore in a conditional independence for Y_x. In particular, if we have a set Z of covariate that satisfies the backdoor criterion in M (see Definition 1), that set also blocks all paths between X and those parents, and consequently, it renders X and Y_x independent in every stratum $Z = z$.

These considerations are summarized formally in Theorem 3.

Theorem 3 (Counterfactual interpretation of backdoor) *If a set Z of variables satisfies the backdoor condition relative to (X, Y) then, for all x, the counterfactual Y_x is conditionally independent of X given Z*

$$P(Y_x|X, Z) = P(Y_x|Z). \tag{19}$$

The condition of Theorem 3, sometimes called "conditional ignorability" implies that $P(Y_x = y) = P(Y = y|do(X = x))$ is identifiable by adjustment over Z. In other words, in linear systems, the average causal effect is given by the partial regression coefficient $\beta_{yx\cdot z}$ (as in Eq. (16)), whenever Z is backdoor admissible.

3.3 Counterfactuals in linear models

In linear Gaussian models any counterfactual quantity is identifiable whenever the model parameters are identified. This is because the parameters fully define the model's functions, with the help of which we can define M and M_x in Eq. (17). The question remains whether counterfactuals can be identified in observational studies, when some of the model parameters are not identified. It turns out that any counterfactual of the form $E[Y_{X=x}|E = e]$, with e an arbitrary set of events is identified whenever $E[Y|do(X = x)]$ is identified (Pearl, 2009, p. 389). The relation between the two is summarized in Theorem 4, which provides a shortcut for computing counterfactuals.

Theorem 4 *Let τ be the slope of the total effect of X on Y,*

$$\tau = E[Y|do(x + 1)] - E[Y|do(x)]$$

then, for any evidence $E = e$, we have:

$$E[Y_{X=x}|E = e] = E[Y|E = e] + \tau(x - E[X|E = e]). \tag{20}$$

This provides an intuitive interpretation of counterfactuals in linear models: $E[Y_{X=x}|E = e]$ can be computed by first calculating the best estimate of Y conditioned on the evidence e, $E[Y|e]$, and then adding to it whatever change is expected in Y when X is shifted from its current best estimate, $E[X|E = e]$, to its hypothetical value, x.

Methodologically, the importance of Theorem 4 lies in enabling researchers to answer hypothetical questions about individuals (or sets of individuals) from population data. The ramifications of this feature in legal and social contexts will be explored in the following sections. In the situation illustrated by Fig. 4, we will demonstrate how Theorem 4 can be used in computing the *effect of treatment on the treated* (Shpitser and Pearl, 2009)

$$ETT = E[Y_1 - Y_0|X = 1]. \tag{21}$$

Substituting the evidence $e = \{X = 1\}$ in Eq. (20) we get:

$$\begin{aligned} ETT &= E[Y_1|X = 1] - E[Y_0|X = 1] \\ &= E[Y|X = 1] - E[Y|X = 1] + \tau(1 - E[X|X = 1]) - \tau(0 - E[X|X = 1]) \\ &= \tau \\ &= c + ab \end{aligned}$$

In other words, the effect of treatment on the treated is equal to the effect of treatment on the entire population. This is a general result in linear systems that can be seen directly from Eq. (20); $E[Y_{x+1} - Y_x|e] = \tau$, independent on the evidence e. Things are different when a multiplicative (i.e., non-linear) interaction term is added to the output equation (Pearl et al., 2016), but this takes us beyond the linear sphere.

4 The Microscope at Work

4.1 The mediation fallacy

In Fig. 4, the effect of X on Y consists of two parts, the direct effect, c, and the indirect effect mediated by Z, and quantified by the product ab. Attempts to disentangle the two by regression methods has led to a persistent fallacy among pre-causal analysts. Define the direct effect of X on Y as "the increase we would see in Y given a unit increase in X while holding Z constant," analysts interpreted the latter as the partial regression coefficient of Y on X, controlling for Z, or

$$c = \beta_{yx,z}.$$

But this can't be true because, using Wright's rule, we get (using Eq. (6)):

$$\beta_{yx\cdot z} = c - da/(1 - a^2)$$

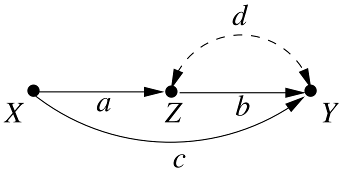

Figure 4: Demonstrating the mediation fallacy; "controlling for" the mediator Z does not give the direct effect c.

which coincides with c only when $d = 0$.

The discrepancy also reveals itself through the fact that Z does not satisfy the single-door condition of Theorem (2). Conditioning on Z opens the path $X \to Z \leftrightarrow Y$.

The fallacy comes about from the habit of translating "holding Z constant" to "conditioning on Z". The correct translation is "set Z to a constant by intervention," namely using the *do*-operator $do(Z = z)$. Unfortunately statistics proper does not provide us with an operator of "holding a variable constant." Lacking such operator, statisticians have resorted to the only operator in their disposal, conditioning, and ended up with a fallacy that has lingered on for almost a century (Burks, 1926a,b; Cole and Hernán, 2002; Pearl, 2012).

Thus, the correct definition of the direct effect of X on Y is (Pearl, 1998, Definition 8)

$$c = \frac{\partial}{\partial x} E(y|do(x), do(z)) \tag{22}$$

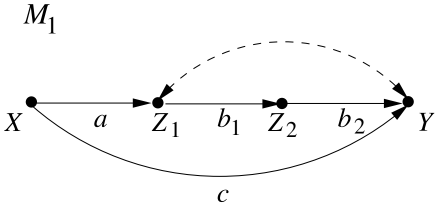

Figure 5: A model in which c cannot be estimated by one-shot OLS; it requires sequential backdoor adjustments.

Readers versed in causal mediation will recognize this expression as the "controlled direct effect" (Robins and Greenland, 1992; Pearl, 2001) which, for linear systems, coincides with the natural direct effect.

4.2 Sequential identification

It often happens that both the backdoor or single door conditions cannot be applied in one shot, but sequential application of them leads us to the right result.

Consider the problem depicted in Fig. 5, in which we require to estimate the direct effect, c, in a model containing two mediators, Z_1 and Z_2.

Clearly, we cannot identify c by OLS, because there is no set of variables that satisfies the single door criterion relative to G_c. Conditioning on any set of mediators would open the path $X \to Z_1 \leftrightarrow Y$. However, since the total effect is identifiable, we can write

$$\tau = c + ab_1b_2 = \beta_{yx}.$$

We further notice that each of a, b_1, b_2 can be identified by the single door condition, using the conditioning sets:

$$\{0\} \text{ for } a, \ \{0\} \text{ for } b_1, \text{ and } \{Z_1\} \text{ for } b_2.$$

Thus we can write

$$a = \beta_{z_1x} \quad , \quad b_1 = \beta_{z_2z_1} \quad , \quad b_2 = \beta_{yz_2 \cdot z_1}$$

and c becomes

$$\begin{aligned} c &= \tau - ab_1b_2 \\ &= \beta_{yx} - \beta_{z_1x}\beta_{z_2z_1}\beta_{yz_2 \cdot z_1}. \end{aligned}$$

This problem is the linear version of the sequential decision problem treated in (Pearl and Robins, 1995) and given a nonparametric solution using a sequential application of the backdoor condition. (See also *Causality*, 2009, p. 352.) An attempt to solve this problem without the *do*-operator was made in Wermuth and Cox (2008; 2014) where it was called "indirect confounding" (Pearl, 2015b).

4.3 Robustness to model misspecification

In his seminal book "Introduction to Structural Equation Models" (1975), Otis Duncan devotes a chapter (8) to Specification Error. He asks: Suppose the model I used is wrong and the correct model is given by another path diagram. Can we "salvage" some of the effects estimated on the basis of the wrong model, so as to give us unbiased estimates for the true model?

Duncan was fascinated by the possibility of salvaging some unbiased estimates despite the wrongness of the working model. He goes through six different pairs of models and asks: "Show that the OLS estimator of b_{ij} [the causal parameter] in Model 1 estimates b_{ij} in Model 2 without bias."

Duncan's analysis was based on Wright's rules which is not very efficient. It requires that we derive the estimates in the two models, and then compare them to decide if they are algebraically identical, in light of other model assumptions.

Using the single door criterion (Theorem 2), we can solve Duncan's puzzle by inspection. We simply enumerate the sets of admissible covariates in each of the two models and check if there is a match.

To illustrate, consider the four models in Fig. 6.

The admissible sets for c in each of the four models are:

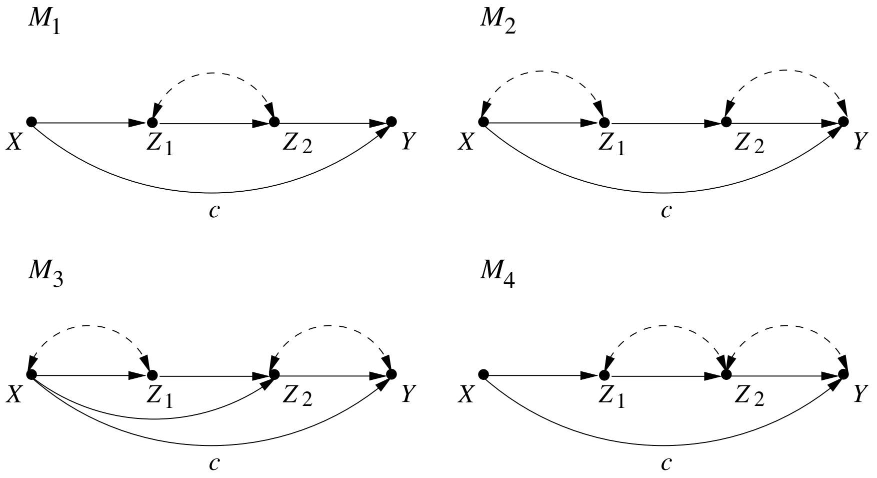

Figure 6: The estimate $c = \beta_{yx \cdot z_1 z_2}$, obtained for M_1, is also valid for M_2 and M_3, but not for M_4.

M1: $\{Z_2\}, \{Z_1, Z_2\}$

M2: $\{Z_1\}, \{Z_1, Z_2\}$

M3: $\{Z_1, Z_2\}$

M4: none

Thus, if M_1 is our working model, we can salvage our estimate of $c = \beta_{yx \cdot z_1 z_2}$ if the true model is either M_2 or M_3. But if the true model is M_4, there is no match to M_1, and both of our options, $c = \beta_{yx \cdot z_1 z_2}$ or $c = \beta_{yx \cdot z_2}$ will be biased. M_4 still permits the identification of c (using generalized instrumental variables (Brito and Pearl, 2002)) but not by using OLS.

4.4 Mediator-specific effects

Consider the linear model depicted in Fig. 7, in which X stands for education, Z for skill level and Y for salary. Suppose our aim is to estimate $E[Y_x|Z = z]$ which stands for the expected salary of individuals with current skill level $Z = z$, had they received $X = x$ years of education.

Inspecting the graph, we see that salary depends only on skill level. In other words, education has no effect on salary once we know the employee's skill level. One might surmise, therefore, that the answer is $E[Y_x|Z = z] = bz$, independent of x. But this is the wrong answer because $E[Y_x|Z = z]$ asks not for the salary of individuals with skill $Z = z$ but for the salary of those who currently have skill $Z = z$ but would have attained a different skill had they obtained x years of education. The first question is captured by the expression $E(Y|do(x), z)$ while the second is captured by the counterfactual $E[Y_x|Z = z]$. The first evaluates indeed to bz, while the second should depends on x, since an increase in education would cause the skill level to increase beyond the current level of $Z = z$ and, consequently, the salary would increase as well.

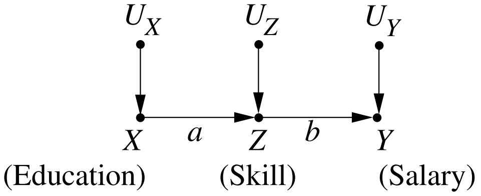

Figure 7: A model demonstrating how skill-specific salary depends on education.

We now compute $E[Y_x|Z = z]$. Using the counterfactual formula of Theorem 4

$$E[Y_x|e] = E[Y|e] + \tau(x - E[X|e]),$$

we insert $e = \{Z = z\}$, and obtain

$$E[Y_x|Z = z] = E[Y|z] + \tau(x - E[X|z]).$$

Assuming U_X and U_Z are standardized, we have

$$\begin{aligned} E[X|z] = \beta_{xz}z = \beta_{zx}\frac{\sigma_X^2}{\sigma_Z^2}z &= \frac{a}{cov(aX + U_Z)} \\ &= z\frac{a}{(1 + a^2)}. \end{aligned}$$

which gives

$$\begin{aligned} E[Y_x|Z = z] &= bz + ab(x - \frac{za}{(1 + a^2)}) \\ &= abx + \frac{bz}{1 + a^2}. \end{aligned}$$

We see that the skill-specific salary depends on education x.

4.5 Mediator-specific effects on the treated

Consider again the model of Fig. 7 and assume that we wish to assess the effect of education on salary for those individuals who have received $X = x'$ years of education and now possess skill $Z = z$. Inspecting the diagram, one might surmise again that, the salary depends on skill only, and not on the hypothetical education.

In the language of potential outcome this would amount to saying that treatment assignment is ignorable conditional on Z or $Y_x \perp\!\!\!\perp X|Z$. But is it? To answer this question we set out to compute $E[Y_x|X = x', Z = z]$ and examine whether it depends on x' and z.

Inserting $e = \{Z = z, X = x'\}$ in Eq. (19) we obtain

$$\begin{aligned} E[Y_x|Z = z, X = x'] &= E[Y|z, x'] + \tau(x - E[X|z, x']) \\ &= \beta z + \alpha\beta(x - x'). \end{aligned}$$

We see that $E[Y_x|Z = z, X = x']$ depends on x', hence $Y_x \not\perp\!\!\!\perp X|Z$.

This dependence can also be seen from the graph. Recalling that Y_x is none other but the exogenous variables (U_Z and U_Y) that affect Y when X is held constant, we note that, conditioned on Z, U_Z is indeed dependent on X. Hence Y_x depends on X conditioned on Z; $Y_x \not\perp\!\!\!\perp X|Z$.

4.6 Testing S-Ignorability

In generalizing experimental findings from one population (or environment) to another, a common method of estimation invokes *re-calibration* or *re-weighting* (Hotz et al., 2005; Cole and Stuart, 2010; Pearl and Bareinboim, 2014). The reasoning goes as follows: Suppose the disparity between the two populations can be attributed to a factor S such that the potential outcomes in the two population are characterized by $E(Y_x|S = 1)$ and $E(Y_x|S = 2)$, respectively. If we find a set of covariates Z such that

$$Y_x \perp\!\!\!\perp S|Z \tag{23}$$

then we can transfer the finding from population 1 to population 2 by writing

$$E(Y_x|S = 2) = \sum_z E(Y_x|S = 1, z)P(z|S = 2).$$

Thus, if we can measure the z-specific causal effect in population 1, the average causal effect in population 2 can be obtained by conditioning over the strata of Z and taking the average, re-weighted by $P(z|S = 2)$, the distribution of Z in the target population, where $S = 2$.

The Achilles heal in this method is, of course, the task of finding a set Z that satisfies condition (23), sometimes called "S-ignorability." By and large, practitioners of re-calibration methods assume S-ignorability by default and rarely justify its plausibility. Remarkably, even students of graphical models may find this condition challenging.

Consider the model in Fig. 8(a). The structure of the model, again, shows the salary depending on skills alone, so one might surmize that Eq. (23) holds. However, leveraging our graphical representation of Y_x, we can easily verify that this is not the case. Since Y_x is a function of $\{S, U_Z, U_Y\}$, Z is a collider between S and U_Z. Therefore, when conditioning on Z, S becomes dependent on U_Z hence also on Y_x. This dependence ceases to exist in Fig. 8(b) because Z is no longer a collider. Another way to check ignorability conditions is to use Twin Networks, as in (Pearl, 2009).

S-ignorability can also be verified algebraically using Eq. (19). Substituting $e = \{Z = z, S = s\}$ we obtain

$$E[Y_x|z, s] = c_1, x + c_2 z + c_3 s$$

with $c_3 \neq 0$. Thus affirming the dependence of Y_x on S, given Z.

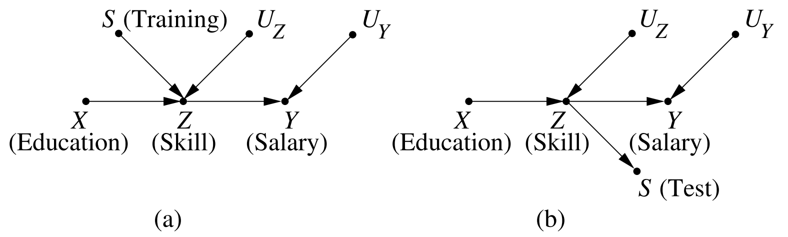

Figure 8: (a) The skill-specific potential outcome Y_x depends on S. (b) The skill-specific potential outcome Y_x is independent of S.

5 Conclusions

Linear models often allow us to derive counterfactuals in close mathematical form. This facility can be harnessed to test conjectures about interventions and counterfactuals that are not easily verifiable in nonparametric models. We have demonstrated the benefit of this facility in several applications, including testing for robustness of estimands and testing the soundness of re-weighting.

Acknowledgment

Portions of Section 4 build on Chapter 4 of (Pearl, Glymour, Jewell, *Causal Inference in Statistics: A Primer*, 2016). Ilya Shpitser contributed to the proof of Theorem 4.

This research was supported in parts by grants from NSF #IIS-1302448 and #1527490, ONR #N00014-17-S-B001, and DARPA #W911NF-16-1-0579.

References

Brito, C. and Pearl, J. (2002). Generalized instrumental variables. In *Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference* (A. Darwiche and N. Friedman, eds.). Morgan Kaufmann, San Francisco, 85–93.

Burks, B. (1926a). On the inadequacy of the partial and multiple correlation technique (part I). *Journal of Experimental Psychology* **17** 532–540.

Burks, B. (1926b). On the inadequacy of the partial and multiple correlation technique (part II). *Journal of Experimental Psychology* **17** 625–630.

Chen, B. and Pearl, J. (2015). Graphical tools for linear structural equation modeling. Tech. Rep. R-432, <http://ftp.cs.ucla.edu/pub/stat_ser/r432.pdf>, Department of Computer Science, University of California, Los Angeles, CA.

Cole, S. and Hernán, M. (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology* **31** 163–165.

Cole, S. and Stuart, E. (2010). Generalizing evidence from randomized clinical trials to target populations. *American Journal of Epidemiology* **172** 107–115.

Cox, D. and Wermuth, N. (2008). Distortion of effects caused by indirect confounding. *Biometrika* **95** 17–33.

Crámer, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press, Princeton, NJ.

Duncan, O. (1975). *Introduction to Structural Equation Models.* Academic Press, New York.

Hotz, V. J., Imbens, G. W. and Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics* **125** 241–270.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, CA.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods and Research* **27** 226–284.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence.* Morgan Kaufmann, San Francisco, CA, 411–420.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference.* 2nd ed. Cambridge University Press, New York.

Pearl, J. (2012). The causal mediation formula – a guide to the assessment of pathways and mechanisms. *Prevention Science* **13** 426–436, DOI: 10.1007/s11121–011–0270–1.

Pearl, J. (2013). Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference* **1** 155–170.

Pearl, J. (2015a). Detecting latent heterogeneity. *Sociological Methods and Research* DOI: 10.1177/0049124115600597, online 1–20.

Pearl, J. (2015b). Indirect confounding and causal calculus (on three papers by Cox and Wermuth). Tech. Rep. R-457, <http://ftp.cs.ucla.edu/pub/stat_ser/r457.pdf>, Department of Computer Science, University of California, Los Angeles, CA.

Pearl, J. and Bareinboim, E. (2014). External validity: From *do*-calculus to transportability across populations. *Statistical Science* **29** 579–595.

Pearl, J., Glymour, M. and Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer.* Wiley, New York.

Pearl, J. and Robins, J. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Uncertainty in Artificial Intelligence 11* (P. Besnard and S. Hanks, eds.). Morgan Kaufmann, San Francisco, 444–453.

Robins, J. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

Shpitser, I. and Pearl, J. (2009). Effects of treatment on the treated: Identification and generalization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Montreal, Quebec, 514–521.

Wermuth, N. and Cox, D. (2014). Graphical Markov models: Overview. In *International Encyclopedia of the Social and Behavioral Sciences* (Wright, ed.), vol. 10. Elsevier, Oxnard, 341–350.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20** 557–585.