Journal of Causal Inference Volume 1, Issue 1, pp 155--170, May 2013 (DOI: 10.1515/jci-2013-0003).

TECHNICAL REPORT
R-409
May 2013

Linear Models: A Useful "Microscope" for Causal Analysis

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 reviews basic techniques of linear path analysis and demonstrates, using simple examples, how causal phenomena of non-trivial character can be understood, exemplified and analyzed using diagrams and a few algebraic steps. The techniques allow for swift assessment of how various features of the model impact the phenomenon under investigation. This includes: Simpson's paradox, case-control bias, selection bias, missing data, collider bias, reverse regression, bias amplification, near instruments, and measurement errors.

1 Introduction

Many concepts and phenomena in causal analysis were first detected, quantified and exemplified in linear structural equation models (SEM) before they were understood in full generality and applied to nonparametric problems. Linear SEM's can serve as a "microscope" for causal analysis; they provide simple and visual representation of the causal assumptions in the model and often enable us to derive close-form expressions for quantities of interest which, in turns, can be used to assess how various aspects of the model affect the phenomenon under investigation. Likewise, linear models can be used to test general hypotheses and to generate counter-examples to over-ambitious conjectures.

Despite their ubiquity, however, techniques for using linear models in that capacity have all but disappeared from the main SEM literature, where they have been replaced by matrix algebra on the one hand and software packages on the other. Very few analysts today are familiar with traditional methods of path tracing (Wright, 1921; Duncan, 1975; Kenny, 1979; Heise, 1975) which, for small problems, can provide both intuitive insight and easy derivations using elementary algebra.

This note attempts to fill this void by introducing the basic techniques of path analysis to modern researchers, and demonstrating, using simple examples, how concepts and issues in modern causal analysis can be understood and analyzed in SEM. These will include: Simpson's paradox, case-control bias, selection bias, collider bias, reverse regression, bias amplification, near instruments, measurement errors, and more.

2 Preliminaries

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 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

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

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:

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

ρ^2_{yz}

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

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

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

$$\beta_{yx\cdot z} = \frac{\beta_{yx} - \beta_{yz}\beta_{zx}}{1-\beta^2_{zx}\sigma^2_x/\sigma^2_z} = \frac{\sigma^2_z\sigma_{yx} - \sigma_{yz}\sigma_{zx}}{\sigma^2_x\sigma^2_z - \sigma^2_{xz}} = \frac{\sigma_y}{\sigma_x}\frac{\rho_{yx} - \rho_{yz}\cdot\rho_{zx}}{1-\rho^2_{xz}} \tag{6}$$

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 an assignment 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,

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

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

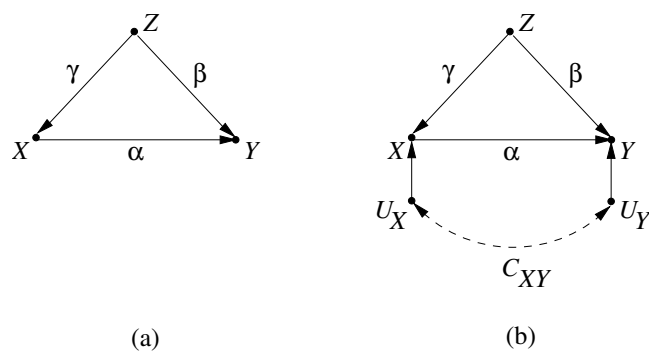

Figure 1: Path diagrams capturing the directionality of the assignment process of Eqs. (7)–(9) as well as possible correlations among omitted factors.

The directionality of this assignment process is captured by a *path-diagram*, in which the nodes represent variables, and the arrows represent the 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.[1]

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

[1]Readers familiar with *do*-calculus (Pearl, 1995) can interpret α as the experimental slope

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

while those familiar with counterfactual logic can write $\alpha = \frac{\partial}{\partial x} Y_{xz}(u)$. The latter implies the former, and the two coincide in linear models, where causal effects are homogenous (i.e., unit-independent.)

the diagram (Wright, 1921). Wright's method consists of equating the (standardized[2]) covariance $\sigma_{xy} = \rho_{xy}$ between any pair of variables to the 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 $X \to Y \leftarrow Z$).

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, Z), we get $\sigma_{xz} = \gamma$ since the path $X \to Y \leftarrow Z$ 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 need 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 $X \to Y$ and Z serves as the root for $X \leftarrow Z \to Y$. 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 Reading partial correlations from 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 read partial correlations directly from the diagram. For example, to read the partial regression coefficient $\beta_{xy\cdot z}$, 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

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

[2]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.

Armed with the ability to read partial regressions, we are now prepared to demonstrate some peculiarities of causal analysis.

3 The Microscope at Work: Examples and their Implications

3.1 Simpson's paradox

Simpson's paradox describes a phenomenon whereby an association between two variables reverses sign upon conditioning on a third variable, regardless of the value taken by the latter. The history of this paradox and the reasons it evokes surprise and disbelief are described in Chapter 6 of (Pearl, 2009a).

The conditions under which association reversal appears in linear models can be seen directly in Fig. 1(a). Comparing (12) and (14) we obtain

$$\beta_{yx} = \alpha + \beta\gamma \qquad \beta_{yx\cdot z} = \alpha$$

Thus, if α has a different sign from $\beta\gamma$, it is quite possible to have the regression of Y on X, β_{yx}, change sign upon conditioning on $Z = z$, for every z. The magnitude of the change depends on the product $\beta\gamma$ which measures the extent to which X and Y are confounded in the model.

3.2 Conditioning on intermediaries and their proxies

Conventional wisdom informs us that, in estimating the effect of one variable on another, one should not adjust for any covariate that lies on the pathway between the two (Cox, 1958). It took decades for epidemiologists to discover that similar prohibition applies to proxies of intermediaries (Weinberg, 1993). The amount of bias introduced by such adjustment can be assessed from Fig. 2.

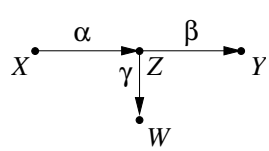

Figure 2: Path diagram depicting an intermediate variable (Z) and its proxy (W). Conditioning on W would distort the regression of Y on X.

Here, the effect of X on Y is simply $\alpha\beta$ as is reflected by the regression slope $\beta_{yx} = \alpha\beta$. If we condition on the intermediary Z, the regression slope vanishes, since

the equality $\sigma_{yx} = \alpha\beta = \sigma_{yz}\sigma_{zx}$ renders $\beta_{xy\cdot z}$ zero in Eq. (10). If we condition on a proxy W of Z, Eq. (10) yields

$$\beta_{yx\cdot w} = \frac{\beta_{yx} - \beta_{yw}\beta_{wx}}{1 - \beta_{wx}^2} = \frac{\alpha\beta - \beta\gamma^{\cancel{2}}\alpha}{1 - \alpha^2\gamma^2} = \frac{\alpha\beta(1-\gamma^2)}{1 - \alpha^2\gamma^2} \tag{15}$$

$\boxed{\gamma^2}$

which unveils a bias of size

$$\beta_{yx\cdot w} - \alpha\beta = \alpha\beta\gamma^2(1-\alpha^2)/(1-\alpha^2\gamma^2)$$

$\boxed{\beta_{yx\cdot w}}$

As expected, the bias disappears for $\gamma = 0$ and intensifies for $\gamma = 1$, where conditioning on W amounts to suppressing all variations in Z.

Speaking of suppressing variations, the model in Fig. 3 may carry some surprise.

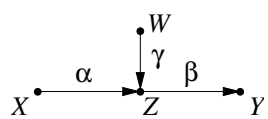

Figure 3: Conditioning on W does not distort the regression of Y on X.

Conditioning on W in this model also suppresses variations in Z, especially for high γ and, yet, it introduces no bias whatsoever; the partial regression slope is (Eq. 10):

$$\beta_{yx\cdot w} = \frac{\sigma_{yx} - \sigma_{yw}\sigma_{xw}}{1 - \sigma_{xw}} = \frac{\alpha\beta - 0}{1 - 0} = \alpha\beta \tag{16}$$

which is precisely the causal effect of X on Y. It seems as though no matter how tightly we "clamp" Z by controlling W, the causal effect of X on Y remains unaltered. Appendix I explains this counter-intuitive result.

3.3 Case-control bias

In the last section, we explained the bias introduced by conditioning on an intermediate variable (or its proxy) as a restriction on the flow of information between X and Y. This explanation is not entirely satisfactory, as can be seen from the model of Fig. 4. Here, Z is not on the pathway between X and Y, and one might surmise that

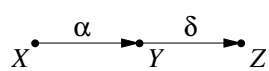

Figure 4: Conditioning on Z, a descendant of Y, biases the regression of Y on X.

no bias would be introduced by conditioning on Z, but analysis dictates otherwise. Path tracing combined with Eq. (10) gives:

$$\begin{aligned}\beta_{yx\cdot z} &= (\sigma_{yx} - \sigma_{yz}\sigma_{zx})/(1 - \sigma_{xz}^2) \\ &= (\alpha - \delta^2\alpha)/(1 - \alpha^2\delta^2) \\ &= \alpha(1-\delta^2)/(1-\alpha^2\delta^2)\end{aligned} \tag{17}$$

and yields the bias

$$\beta_{yx\cdot z} - \alpha = \alpha\delta^2(\alpha^2 - 1)/(1 - \alpha^2\delta^2) \tag{18}$$

This bias reflects what economists called "selection bias" (Heckman, 1979) and epidemiologists "case-control bias" (Robins, 2001), which occurs when only patients for whom the outcome Y is evidenced (e.g., a complication of a disease) are counted in the database. An intuitive explanation of this bias (invoking virtual colliders) is given in (Pearl, 2009a, p. 339). In contrast, conditioning on a proxy of the explanatory variable X, as in Fig. 5, introduces no bias, since

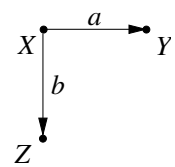

Figure 5: Conditioning on Z, a descendant of X, does not bias the regression of Y on X.

$$\beta_{yx\cdot z} = \frac{(\sigma_{yx} - \sigma_{yz}\sigma_{zx})}{(1 - \sigma_{xz}^2)} = \frac{a - (ab)b}{1 - b^2} = a \tag{19}$$

This can also be deduced from the conditional independence $Z \perp\!\!\!\perp Y | X$ which is implied by the diagram in Fig. 5, but not in Fig. 4. However, to assess the size of the induced bias, as we did in Eq. (18), requires an algebraic analysis of path tracing.

3.4 Sample Selection Bias

The two examples above are special cases of a more general phenomenon called "selection bias" which occurs when samples are preferentially selected to the data set, depending on the values of some variables in the model (Bareinboim and Pearl, 2012; Daniel et al., 2011; Geneletti et al., 2009; Pearl, 2012). In Fig. 6, for example, if $Z = 1$ represents inclusion in the data set, and $Z = 0$ exclusion, the selection decision

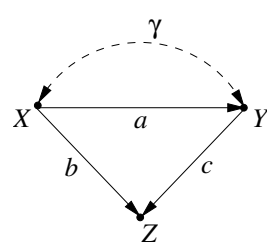

Figure 6: Conditioning on $Z = 1$ represents inclusion in the dataset and biases the regression of Y on X, unless $c = 0$.

is shown to be a function of both X and Y. Since inclusion ($Z = 1$) amounts to

conditioning on Z, we may ask what the regression of Y on X is in the observed data, $\beta_{yx\cdot z}$, compared with the regression in the entire population, $\beta_{yx} = a + \gamma$.

Applying our path-tracing analysis in (10) we get:

$$\beta_{yx\cdot z} = \frac{\sigma_{yx} - \sigma_{yz}\sigma_{zx}}{1 - \sigma_{zx}^2} = \frac{(a+\gamma) - [(a+\gamma)b + c][b + (a+\gamma)c]}{1 - [b + (a+\gamma)c]^2} = \frac{(a+\gamma)[1 - b^2 - c^2] - bc}{1 - [b + (a+\gamma)c]^2}. \tag{20}$$

We see that a substantial bias may result from conditioning on Z, persisting even when X and Y are not correlated, namely, when $\sigma_{xy} = a + \gamma = 0$. Note also that the bias disappears for $c = 0$, as in Fig. 5, but not for $b = 0$, which returns us to the case-controlled model of Fig. 4.

Selection bias is symptomatic of a general phenomenon associated with conditioning on collider nodes (Z in our example). The phenomenon involves spurious associations induced between two causes upon observing their common effect, since any information refuting one cause should make the other more probable. It has been known as Berkson Paradox (Berkson, 1946), "explaining away" (Kim and Pearl, 1983) or simply "collider bias."[3]

3.5 Missing data

In contrast to selection bias, where exclusion ($S = 0$) removes an entire unit from the dataset, in missing data problems a unit may have each of its variables masked independently of the others (Little and Rubin, 1987, p. 89). Therefore, the diagram representing the missingness process should assign each variable V_i a "switch" R_i, called "missingness mechanism" which determines whether V_i is observed ($R_i = 0$) or masked ($R_i = 1$). The arrows pointing to R_i tells us which variables determine whether R_i fires ($R_i = 1$) or not ($R_i = 0$). In Fig. 7(a), for example, the missingness of X, denoted R_x, depends only on the latent variable L, while the missingness of Y

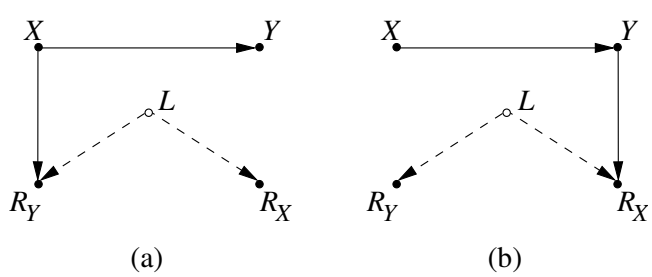

Figure 7: Missingness diagrams in which conditioning on $R_x = 0$ or $R_y = 0$ represents unmasking the values of X and Y, respectively. The parameter σ_{xy}, σ_x^2 and σ_y^2 can all be estimated bias-free from data generated by either model, through each model requires a different estimation procedure.

[3]It has come to my attention recently, and I feel responsibility to make it public, that seasoned reviewers for highly reputable journals reject papers because they are not convinced that such bias can be created; it defies, so they claim, everything they have learned from statistics and economics. A typical resistance to accepting Berkson's Paradox is articulated in (Pearl, 2009b; Rubin, 2009).

is shown to depend on both L and X.

Assume we wish to estimate the covariance σ_{xy} from partially observed data generated by the model of Fig. 7(a); can we obtain an unbiased estimate of σ_{xy}? The question boils down to expressing σ_{xy} in terms of the information available to us, namely, the values of X and Y that are revealed to us whenever $R_x = 0$ or $R_y = 0$ (or both). If we simply estimate σ_{xy} from samples in which both X and Y are observed, that would amount to conditioning on both $R_x = 0$ and $R_y = 0$ which would introduce a bias since the pair $\{X, Y\}$ is not independent of the pair $\{R_x, R_y\}$ (owed to the unblocked path from Y to R_y).

The graph reveals, however, that σ_{xy} can nevertheless be estimated bias-free from the information available, using two steps. First, we note that X is independent of its missingness mechanism R_x, since the path from X to R_x is blocked (by the collider at R_y). Therefore, $\sigma_x^2 = (\sigma_x^2 | R_x = 0)$.[4] This means that we can estimate σ_x from the samples in which X is observed, regardless of whether Y is missing. Next we note that the regression slope β_{yx} can be estimated (e.g., using OLS) from samples in which both X and Y are observed. This is because conditioning on $R_x = 0$ and $R_y = 0$ is similar to conditioning on Z in Fig. 5, where Z is a proxy of the explanatory variable X.

Putting the two together (using Eq. (2)) we can write:

$$\sigma_{xy} = \sigma_x^2 \beta_{yx} = (\sigma_x^2 | R_x = 0) \cdot (\beta_{yx} | R_x = 0, R_y = 0)$$

which guarantees that the product of the two estimates on the right hand side would result in an unbiased estimate of σ_{xy}. Note that a similar analysis of Fig. 7(b) would yield

$$\sigma_{xy} = (\sigma_y^2 | R_y = 0) \cdot (\beta_{xy} | R_x = 0, R_y = 0)$$

which instructs us to estimate σ_y^2 using samples in which Y is observed and estimate the regression of X on Y from samples in which both X and Y are observed. Remarkably, the two models are statistically indistinguishable and yet each dictates a different estimation procedure, thus demonstrating that no model-blind estimator can guarantee to deliver an unbiased estimate, even when such exists. If the path diagram permits no decomposition of σ_{xy} into terms conditioned on $R_x = 0$ and $R_y = 0$ (as would be the case, for example, if an arrow existed from X to R_x in Fig. 7(a)) we would conclude then that σ_{xy} is not estimable by any method whatsoever. A general analysis of missing data problems using causal graphs is given in (Mohan et al., 2013).

3.6 The M-bias

The M-bias is another instant of Berkson's paradox where the conditioning variable, Z, is a pre-treatment covariate, as depicted in Fig. 8. The parameters γ_1 and γ_2

[4] $(\sigma_x^2 | R_x = 0)$ stands for the conditional variance of X given $R_x = 0$. We take the liberty of treating R_y as any other variable in the linear system, even though it is binary, hence the relationship $X \to R_y$ must be nonlinear. The linear context simplifies the intuition and the results hold in nonparametric systems as well.

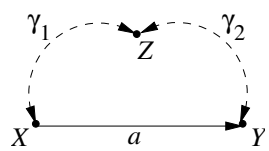

Figure 8: Adjusting for Z, which may be either pre-treatment or post-treatment covariate, introduces bias where none exists. The better the predictor the higher the bias.

represent error covariances C_{XZ} and C_{ZY}, respectively, which can be generated, for example, by latent variables effecting each of these pairs.

To analyze the size of this bias, we apply Eq. (10) and get:

$$\beta_{yx\cdot z} = \frac{a - (\gamma_2 + a\gamma_1)\gamma_1}{1 - \gamma_1^2} = a - \frac{\gamma_1\gamma_2}{1 - \gamma_1^2} \tag{21}$$

Thus, the bias induced increases substantially when γ_1 approaches one, that is, when Z becomes a good predictor of X. Ironically, this is precisely when investigators have all the textbook reasons to adjust for Z. Being pre-treatment, the collider Z cannot be distinguished from a confounder (as in Fig. 1(a)) by any statistical means, and has alluded some statisticians to conclude that "there is no reason to avoid adjustment for a variable describing subjects before treatment" (Rosenbaum, 2002, p. 76).

3.7 Reverse Regression

Is it possible that men would earn a higher salary than equally qualified women, and *simultaneously*, men are more qualified than women doing equally paying job? This counter-intuitive condition can indeed exist, and has given rise to a controversy called "Reverse Regression;" some sociologists argued that, in salary discrimination cases, we should not compare salaries of equally qualified men and women, but, rather, compare qualifications of equally paid men and women (Goldberger, 1984).

The phenomenon can be demonstrated in Fig. 9. Let X stand for gender (or

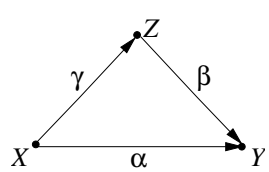

Figure 9: Path diagram in which Z acts as a mediator between X and Y, demonstrating negative reverse regression $\beta_{zx\cdot y}$ for positive α, β and γ.

age, or socioeconomic background), Y for job earnings and Z for qualification. The partial regression $\beta_{yx\cdot z}$ encodes the differential earning of males ($X = 1$) over females ($X = 0$) having the same qualifications ($Z = z$), while $\beta_{zx\cdot y}$ encodes the differential qualification of males ($X = 1$) over females ($X = 0$) earning the same salary (y).

For the model in Fig. 9, we have

$$\beta_{yx\cdot z} = \alpha$$
$$\beta_{zx\cdot y} = (\sigma_{zx} - \sigma_{zy}\sigma_{yx})/(1-\sigma_{zy}^2) \quad = [(\gamma - (\beta+\gamma\alpha)(\alpha+\beta\gamma)]/[1-(\beta+\gamma\alpha)^2]$$

Surely, for any $\alpha > 0$ and $\beta > 0$ we can choose γ so as to make $\beta_{zx\cdot y}$ negative. For example, the combination $\alpha = \beta = 0.8$ and $\gamma = 0.1$ yields

$$\beta_{zx\cdot y} = [(0.1 - (0.8 + 0.1 \times 0.8)(0.8 + 0.8 \times 0.1)]/[1 - (0.8 + 0.1 \times 0.9)^2] = -5.8545$$

Thus, there is no contradiction in finding men earning a higher salary than equally qualified women, and simultaneously, men being more qualified than women doing equally paying job. A negative $\beta_{zx\cdot y}$ may be a natural consequence of male-favoring hiring policy ($\alpha > 0$), male-favoring training policy ($\gamma > 0$) and qualification-dependent earnings ($\beta > 0$).

The question of whether standard or reverse regression is more appropriate for proving discrimination is also clear. The equality $\beta_{yx\cdot z} = \alpha$ leaves no room for hesitation, because α coincides with the counterfactual definition of "direct effect of gender on hiring had qualification been the same," which is the court's definition of discrimination.

The reason the reverse regression appeals to intuition is because it reflects a model in which the employer decides on the qualification needed for a job on the basis of both its salary level and the applicant sex. If this were a plausible model, it would indeed be appropriate to persecute an employer who demands higher qualifications from men as opposed to women. But such a model should place Z as a post-salary variable e.g., $X \to Z \leftarrow Y$.

3.8 Bias Amplification

In the model of Fig. 10, Z acts as an instrumental variable, since $\sigma_{zu} = 0$. If U is unobserved, however, Z cannot be distinguished from a confounder, as in Fig. 1(a), in the sense that for every set of parameters (α, β, γ) in Fig. 1(a) one can find a set (a, b, c, d) for the model in Fig. 10 such that the observed covariance matrices of the two models are the same. This indistinguishability, together with the fact that Z may be a strong predictor of X may lure investigators to condition on Z to obtain an unbiased estimate of d (Hirano and Imbens, 2001). Recent work has shown however that such adjustment would amplify the bias created by U (Bhattacharya and Vogt, 2007; Pearl, 2010a; Wooldridge, 2009).

The magnitude of this bias and its relation to the pre-conditioning bias, ab, can be computed from the diagram of Fig. 10, as follows:

$$\beta_{yx\cdot z} = \frac{\sigma_{xy} - \sigma_{yz}\cdot\sigma_{xz}}{1-\sigma_{xz}^2} = \frac{(ab+\gamma_0) - c\gamma_0 c}{1-c^2} = \gamma_0 + \frac{ab}{1-c^2} \tag{22}$$

We see the the bias created, $\frac{ab}{(1-c^2)}$, is proportional to the pre-existing bias ab and increases with c; the better Z predicts X, the higher the bias. An intuitive explanation of this phenomenon is given in Pearl (2010a)

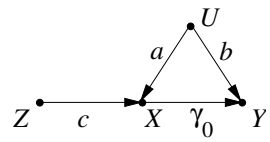

Figure 10: Bias amplification, $\beta_{yx\cdot z} - \gamma_0 > ab$, produced by conditioning on an instrumental variable (Z).

3.9 Near Instruments - amplifiers or attenuators?

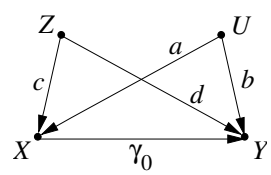

Figure 11: A diagram where Z acts both as an instrument and as a confounder.

The model in Fig. 11 is indistinguishable from that of Fig. 10 when U is unobserved. However, here Z acts both as an instrument and as a confounder. Conditioning on Z is beneficial in blocking the confounding path $X \leftarrow Z \rightarrow Y$ and harmful in amplifying the baseline bias $cd + ab$. The trade off between these two tendencies can be quantified by computing $\beta_{yx\cdot z}$, yielding

$$\begin{aligned}
\beta_{yx\cdot z} &= \frac{\sigma_{xy} - \sigma_{yz}\sigma_{zx}}{1 - \sigma_{xz}^2} \\
&= \frac{\gamma_0 + cd + ab - (d + c\gamma_0)c}{1 - c^2} \\
&= \frac{\gamma_0(1 - c^2) + ab}{1 - c^2} \\
&= \gamma_0 + \frac{ab}{1 - c^2}
\end{aligned} \tag{23}$$

We see that the baseline bias $ab + cd$ is first reduced to ab and then magnified by the factor $(1 - c^2)^{-1}$. For Z to be a bias-reducer, its effect on Y (i.e., d) must exceed its effect on X (i.e., c) by a factor $ab/(1 - c^2)$. This trade-off was assessed by simulations in (Myers et al., 2011) and analytically in (Pearl, 2011), including an analysis of multi-confounders, and nonlinear models.

3.10 The Butterfly

Another model in which conditioning on Z may have both harmful and beneficial effects is seen in Fig. 12. Here, Z is both a collider and a confounder. Conditioning on Z blocks the confounding path through α_1 and α_2 and, at the same time induces a virtual confounding path through the latent variables that create the covariances $C_{XZ} = \gamma_1$ and $C_{ZY} = \gamma_2$.

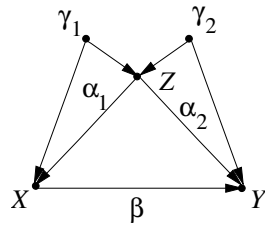

Figure 12: Adjusting for Z may be harmful or beneficial depending on the model's parameters.

This trade-off can be evaluated from our path-tracing formula Eq. (10) which yields

$$\beta_{yx\cdot z} = \frac{\beta_{yx} - \beta_{yz}\beta_{zx}}{1-\beta_{zx}^2} = \frac{[\beta + (\alpha_1+\gamma_1)\alpha_2 + \alpha_1\gamma_2] - [\alpha_2 + \gamma_2 + \beta(\gamma_1+\alpha_1)][\gamma_1+\alpha_1]}{1-(\alpha_1+\gamma_1)^2}$$
$$= \frac{\beta - \gamma_2\gamma_1 - \beta(\gamma_1+\alpha_1)^2}{1-(\alpha_1+\gamma_1)^2} \tag{24}$$

We first note that the pre-conditioning bias

$$\beta_{yx} - \beta = \alpha_2(\alpha_1+\gamma_1) + \alpha_1\gamma_2 \tag{25}$$

may have positive or negative values even when both $\sigma_{xz} = 0$ and $\sigma_{zy} = 0$. This refutes folklore wisdom, according to which a variable Z can be exonerated from confounding considerations if it is uncorrelated with both treatment (X) and outcome (Y).

Second, we notice that conditioning on Z may either increase or decrease bias, depenending on the structural parameters. This can be seen by comparing (25) with the post-conditioning bias:

$$\beta_{yx\cdot z} - \beta = -\gamma_1\gamma_2/[1-(\alpha_1+\gamma_1)^2] \tag{26}$$

In particular, since Eq. (26) is independent on α_2, it is easy to choose values of α_2 that make (25) either higher of lower than (26).

3.11 Measurement error

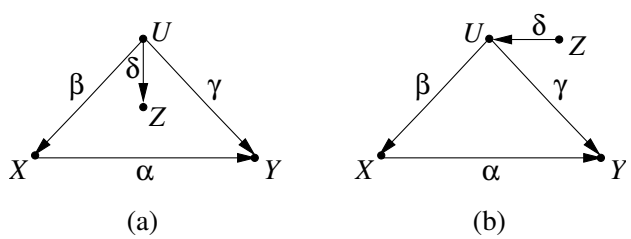

Figure 13: Conditioning on Z, a proxy for the unobserved confounder U, does not remove the bias ($\beta\gamma$),

Assume the confounder U in Fig. 13(a) is unobserved but we can measure a proxy Z of U. Can we assess the amount of bias introduced by adjusting for Z instead of U? The answer, again, can be extracted from our path-tracing formula, which yields

$$\begin{aligned}\beta_{yx\cdot z} &= \frac{\sigma_{yx} - \sigma_{yz}\sigma_{zx}}{1-\sigma_{zx}^2} = \frac{(\alpha+\beta\gamma) - (\gamma\delta + \alpha\beta\delta)\beta\delta}{1-\beta^2\delta^2} \\ &= \frac{\alpha + \beta\delta - \beta\delta^2(\gamma+\alpha\beta)}{1-\beta^2\delta^2} = \frac{\alpha(1-\beta^2\delta^2) + \gamma\beta(1-\delta^2)}{1-\beta^2\delta^2} \\ &= \alpha + \frac{\gamma\beta(1-\delta^2)}{1-\beta^2\delta^2} \end{aligned} \tag{27}$$

As expected, the bias vanishes when δ approaches unity, indicating a faithful proxy. Moreover, if δ can be estimated from an external pilot study, the causal effect α can be identified. (See Pearl, 2010b; Kuroki et al., 2013.) Remarkably, identical behavior emerges in the model of Fig. 13(b) in which Z is a driver of U, rather than a proxy.

The same treatment can be applied to errors in measurements of X or of Y and, in each case, the formula of $\sigma_{xy\cdot z}$ reveals what model parameters are the ones affecting the resulting bias.

4 Conclusions

We have demonstrated how path-analytic techniques can illuminate the emergence of several phenomena in causal analysis and how these phenomena depend on the structural features of the model. Although the techniques are limited to linear analysis, hence restricted to homogeneous populations with no interactions, they can be superior to simulation studies whenever conceptual understanding is of essence, and problem size is manageable.

Acknowledgment

This research was supported in parts by grants from NSF #IIS-1249822 and ONR #N00014-13-1-0153.

Appendix I

In linear systems, the explanation for the equality $\sigma_{yx} = \sigma_{yx\cdot w}$ in Fig. 3 is simple. Conditioning on W does not physically constrain Z, it merely limits the variance of Z in the subpopulation satisfying $W = w$ which was chosen for observations. Given that effect-homogeneity prevails of linear models, we know that the effect of X on Z remains invariant to the level w chosen for observation and, therefore, this w-specific effect reflects the effect of X on the entire population. This dictates (in a confounding-free model) $\beta_{xy\cdot w} = \beta_{xy}$.

But how can we explain the persistence of this phenomenon in nonparametric models, where we know (e.g., using *do*-calculus (Pearl, 2009a)) that adjustment for W does not have any effect on the resulting estimand? In other words, the equality

$$E[Y|X = x] = E_w E[Y|X = x, W = w]$$

will hold in the model of Fig. 3 even when the structural equations are nonlinear. Indeed, the independence of W and X, implies

$$\begin{aligned} E[Y|X = x] &= \sum_w E[Y|X = x, W = w]P(W = w|X = x) \\ &= \sum_w E[Y|X = x, W = w]P(W = w) \\ &= E_w E[Y|X = x, W = w] \end{aligned}$$

The answer is that adjustment for W involves averaging over W; conditioning on W does not. In other words, whereas the effect of X on Z may vary across strata of W, the average of this effect is none other but the effect over the entire population, i.e., $E[Y|do(X = x)]$, which equals $E[Y|X = x]$ in the non-confounding case.

Symbolically, we have

$$\begin{aligned} E[Y|do(X = x)] &= \sum_w E[Y|do(X = x), W = w]P[W = w|do(X = x)] \\ &= \sum_w E[Y|do(X = x), W = w]P(W = w) \\ &= \sum_w E[Y|X = x, W = w]P(W = w) \\ &= E(Y|X = x) \end{aligned}$$

The first reduction is licensed by the fact that X has no effect on W and the second by the back-door condition.

References

Bareinboim, E. and Pearl, J. (2012). Controlling selection bias in causal inference. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*. La Palma, Canary Islands, 100–108.

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin* **2** 47–53.

Bhattacharya, J. and Vogt, W. (2007). Do instrumental variables belong in propensity scores? Tech. Rep. NBER Technical Working Paper 343, National Bureau of Economic Research, MA.

Cox, D. (1958). *The Planning of Experiments*. John Wiley and Sons, NY.

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

Daniel, R. M., Kenward, M. G., Cousens, S. N. and Stavola, B. L. D. (2011). Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research* **21** 243–256.

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

Geneletti, S., Richardson, S. and Best, N. (2009). Adjusting for selection bias in retrospective, case-control studies. *Biostatistics* **10(1)**.

Goldberger, A. (1984). Reverse regression and salary discrimination. *The Journal of Human Resources* 293–318.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47** 153–161.

Heise, D. (1975). *Causal Analysis*. John Wiley and Sons, New York.

Hirano, K. and Imbens, G. (2001). Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services and Outcomes Research Methodology* **2** 259–278.

Kenny, D. (1979). *Correlation and Causality*. Wiley, New York.

Kim, J. and Pearl, J. (1983). A computational model for combined causal and diagnostic reasoning in inference systems. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI-83)*. Karlsruhe, Germany.

Kuroki, M., Tian, J. and Pearl, J. (2013). Measurement bias and effect restoration in causal inference. Tech. Rep. R-366, <http://ftp.cs.ucla.edu/pub/stat_ser/r366.pdf>, University of California Los Angeles, Computer Science Department, CA.

Little, R. J. and Rubin, D. B. (1987). *Statistical analysis with missing data*, vol. 4. Wiley, New York.

Mohan, K., Pearl, J. and Tian, J. (2013). Missing data as a causal inference problem. Tech. Rep. R-410, <http://ftp.cs.ucla.edu/pub/stat_ser/r410.pdf>, University of California Los Angeles, Computer Science Department, CA.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M. and Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology* 1213–1222.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710.

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

Pearl, J. (2009b). Myth, confusion, and science in causal analysis. Tech. Rep. R-348, University of California, Los Angeles, CA. <http://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf>.

Pearl, J. (2010a). On a class of bias-amplifying variables that endanger effect estimates. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. AUAI, Corvallis, OR, 417–424. <http://ftp.cs.ucla.edu/pub/stat_ser/r356.pdf>.

Pearl, J. (2010b). On measurement bias in causal inferences. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. AUAI, Corvallis, OR, 425–432. <http://ftp.cs.ucla.edu/pub/stat_ser/r357.pdf>.

Pearl, J. (2011). Invited commentary: Understanding bias amplification. *American Journal of Epidemiology* Published online, DOI: 10.1093/aje/kwr352.

Pearl, J. (2012). A solution to a class of selection-bias problems. Tech. Rep. R-405, <http://ftp.cs.ucla.edu/pub/stat_ser/r405.pdf>, Department of Computer Science, University of California, Los Angeles, CA.

Robins, J. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology* **12** 313–320.

Rosenbaum, P. (2002). *Observational Studies*. 2nd ed. Springer-Verlag, New York.

Rubin, D. (2009). Author's reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment group? *Statistics in Medicine* **28** 1420–1423.

Weinberg, C. (1993). Toward a clearer definition of confounding. *American Journal of Epidemiology* **137** 1–8.

Wooldridge, J. (2009). Should instrumental variables be used as matching variables? Tech. Rep. <https://www.msu.edu/∼ec/faculty/wooldridge/current%20research/treat1r6.pdf>, Michigan State University, MI.

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