The Limits of Causal Inference from Observational Data

Peter Spirtes
Carnegie Mellon University
ps7z@andrew.cmu.edu

1. Introduction

The following quotation from Rosenbaum (1995) expresses a commonly held view about the problem of potential confounders, and how they can be dealt with. (We will take a "confounder" of treatment and response to be a variable that is a cause of both treatment and response.)

> An observational study is an empirical investigation of treatments, policies, or exposures and the effect they cause, but it differs from an experiment in that the investigator cannot control the assignment of treatments to subjects. … Analytical adjustments are widely used in observational studies to remove overt biases, that is, differences between treated and control groups, present before treatment, that are visible in the data at hand. … If treated and control groups differed before treatment in ways not recorded, there would be a hidden bias. … sensitivity analyses … ask how the findings of a study might be altered by hidden biases of various magnitudes. It turns out that observational studies vary markedly in their sensitivity to hidden bias. The degree of sensitivity to hidden bias is an important consideration in judging whether the treatment caused its ostensible effects or alternatively whether these seeming effects could merely reflect a hidden bias. Rosenbaum (1995) pp. vii-viii.

There are two parts to Rosenbaum's recommendation for dealing with potential confounding. The first part is to condition on the measured potential confounders to adjust for the effect of confounding by measured variables. The second part is to do a sensitivity analysis to estimate how much hidden variables could affect the results of an analysis. It is widely supposed that both of these parts can be done using only information about the temporal order of the measured variables. However, I will show that a variety of methods that "adjust" for measured variables that are potential confounders (e.g. conditioning, matching, matching on propensity scores) by some form of conditioning can actually turn consistent estimates into inconsistent estimates, and increase bias. Whether conditioning on a measured potential covariate makes an estimate less biased or more biased depends upon causal relationships among the measured and unmeasured covariates, and not just the temporal order in which the variables occurred. Second, I will show that conditioning on measured potential confounders can increase the association between a pair of measured variables due to hidden variables. Again, whether conditioning on a measured potential covariate increases or decreases the association between a pair of measured variables due to hidden variables depends upon causal relationships among the measured and unmeasured covariates, and not just the temporal order in which the measured covariates occurred.. I will argue that this implies that conditioning on a measured potential covariate either can increase the sensitivity of the analysis to hidden biases, or the sensitivity to hidden bias was large to begin with.

In contrast to the causal inference methods mentioned above, there is a class of alternative methods which are not subject to the problems that I will point out here. Some of these alternative methods, such as instrumental variables, or worst-case bounds (Manski 1995), require that the user have some substantive background knowledge about the causal relationships among the measured and unmeasured covariates. Other techniques (e.g. the FCI algorithm described in Spirtes et al. 1993, 1995, 1999, and a variety of Bayesian search algorithms described in Heckerman et al. 1999) use the data to discover enough about the causal relationships among the measured and unmeasured covariates to make inferences about the strength of a causal connection between two variables possible.

2. Conditioning

Conditioning on measured potential confounders can increase, rather than decrease the bias of an estimate, or turn a consistent estimate into an inconsistent estimate. (See Spirtes et al. 1999, and Robins ??). I will illustrate this in the context of linear structural equation models (SEMs), but the general conclusions do not at all depend upon the linearity assumption, and can also be applied to non-linear directed acyclic graph models (Pearl, 1988).

SEMs are widely used in sociology, econometrics, biology, and other sciences. A SEM (without free parameters) has two parts: a probability distribution (in the Gaussian case specified by a set of linear structural equations and a covariance matrix among the “error” or “disturbance” terms), and an associated path diagram corresponding to the causal relations among variables specified by the structural equations and the correlations among the error terms (Bollen, 1989). The path diagram contains a directed edge from *B* to *A* if and only if there is a non-zero coefficient for *B* in the equation for *A*; and there is a double-headed arrow between *A* and *B* if and only if the error term for *A* and the error term for *B* have a non-zero correlation. The path diagram associated with a SEM may contain directed cycles (representing feedback), and double-headed arrows (representing correlated errors.) I will call a path diagram which contains no double-headed arrows a **directed graph**. (I place sets of variables and defined terms in boldface.) In a SEM *M*, I will denote the correlation matrix among the non-error variables by **Σ(M)**, and the corresponding path diagram by **G(M)**.

It is common knowledge among practicing social scientists that for the coefficient of X in the regression of Y upon X to be interpretable as the effect of X on Y there should be no "confounding" variable Z which is a cause of both X and Y:

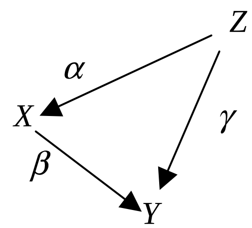


Figure 1

Simple calculations confirm this conclusion (using the notation in Figure 1):

$$\mathrm{Cov}(X,Y) = \beta\mathrm{V}(X) + \alpha\gamma\mathrm{V}(Z)$$

Hence

$$\frac{\mathrm{Cov}(X,Y)}{\mathrm{V}(X)} = \frac{\beta\mathrm{V}(X)+\alpha\gamma\mathrm{V}(Z)}{\mathrm{V}(X)} \neq \beta$$

.

Thus the coefficient from the regression of Y on X alone will be a consistent estimator only if either α or γ is equal to zero. Further, observe that the bias term $\alpha\gamma V(Z)/V(X)$ may be either positive or negative, and of arbitrary magnitude.

However, $\mathrm{Cov}(X,Z) = \alpha\mathrm{V}(Z)$ and $\mathrm{Cov}(Y,Z) = (\alpha\beta + \gamma)\mathrm{V}(Z)$, and hence

$$\mathrm{Cov}(X,Y \mid Z) = \mathrm{Cov}(X,Y) - \frac{\mathrm{Cov}(X,Z)\mathrm{Cov}(Y,Z)}{\mathrm{V}(Z)} =$$
$$\beta\mathrm{V}(X) + \alpha\gamma\mathrm{V}(Z) - \alpha\mathrm{V}(Z)(\alpha\beta+\gamma) = \beta(\mathrm{V}(X) - \alpha^2\mathrm{V}(Z))$$

and

$$\mathrm{V}(X \mid Z) = \mathrm{V}(X) - \frac{\mathrm{Cov}(X,Z)^2}{\mathrm{V}(Z)} = \mathrm{V}(X) - \alpha^2\mathrm{V}(Z),$$

so the coefficient of X in the regression of Y on X and Z is a consistent estimator of β since $\mathrm{Cov}(X,Y|Z)/V(X|Z) = \beta$.

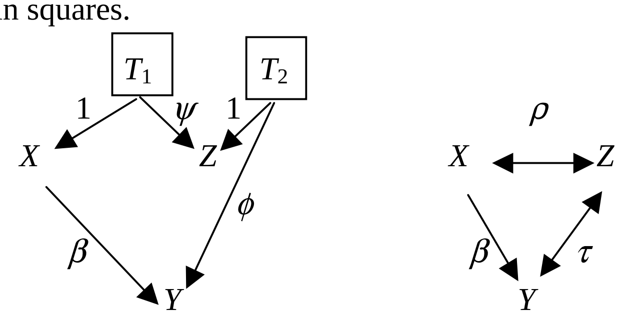
The danger presented by failing to include confounding variables is well understood by social scientists and statisticians. Indeed, it is often used as the justification for considering a long “laundry list” of “potential confounders” for inclusion in a given regression equation.

Typically, the only consideration raised against conditioning on many variables is that it may create sampling problems. According to Rosenbaum (1995):

> Increasing the number of covariates used in adjustment increases costs and complexities, and may make it more difficult to adjust for the most important covariates. As more covariates are collected and analyzed, it becomes increasingly difficult to ensure that all covariates meet high standards of accuracy and completeness, and increasingly difficult to ensure that each covariate receives the needed attention when used in modeling or matching. If there are many covariates, each with some missing data, there may be few subjects with complete data on all covariates, and this may make the analysis more difficult than it would otherwise be. (Rosenbaum, 1995, pp. 63-64.)

However, what is generally not pointed out is that conditioning on potential confounders can create problems that have nothing to do with sampling. The belief that conditioning on as many measured potential confounders as possible “adjusts for” confounding is a mistaken inference from a sound intuition. The sound intuition is that (barring deterministic relationships among the measured variables) conditioning on *all* covariates occurring prior to the treatment eliminates the problem of unmeasured confounders, in the sense that regressing Y on X conditional on all possible covariates will always produce a consistent estimate of the edge coefficient, and hence never turn a consistent estimate into an inconsistent estimate. The mistaken inference is that in the population, conditioning on all *measured* confounders also cannot turn a consistent estimate into an inconsistent estimate.

Suppose that Z temporally precedes both X and Y (and since T_1 and T_2 temporally precede Z, they also precede X and Y.) Let ε_X, ε_Y, and ε_Z be the error terms in the model of Figure 2(a), and ε'_X, ε'_Y, and ε'_Z be the error terms in the model of Figure 2(b). Latent variables are enclosed in squares.



(a) (b)

Figure 2

In the path diagram depicted in Figure 2(a) there are two unmeasured confounders T_1 and T_2, which are uncorrelated with one another. The set of potential confounders is $\{T_1, T_2, Z\}$, while the set of measured potential confounders is $\{Z\}$. The intuition that if we conditioned on the set of all potential confounders then the coefficient of X in the regression equation of Y on X is consistent is correct.

Suppose, however, that we condition only on the *measured* potential confounder Z. Any SEM with this path diagram may be converted into a SEM with the path diagram depicted in Figure 2(b), letting ρ = Cov(X,Z) = ψV(T_1), $\tau = \phi$V(T_2), , $V(\varepsilon'_Y) = V(\varepsilon_Y) + \psi^2 V(T_1) + V(T_2)$ and $V(\varepsilon'_Z) = V(\varepsilon_Z) + \phi^2 V(T_2)$. Note that the regression of Y on X yields a consistent estimate of β since Cov(X,Y) = βV(X). However,

$$\frac{\mathrm{Cov}(X,Y\,|\,Z)}{\mathrm{V}(X\,|\,Z)} = \frac{\mathrm{Cov}(X,Y)\mathrm{V}(Z) - \mathrm{Cov}(X,Z)\mathrm{Cov}(Y,Z)}{\mathrm{V}(X)\mathrm{V}(Z) - \mathrm{Cov}(X,Z)^2} =$$

$$\frac{\beta\mathrm{V}(X)\mathrm{V}(Z) - \rho(\rho\beta + \tau)}{\mathrm{V}(X)\mathrm{V}(Z) - \rho^2} = \beta - \frac{\rho\tau}{\mathrm{V}(X)\mathrm{V}(Z) - \rho^2}$$

Hence the coefficient of X in the regression of Y on X and Z is not a consistent estimate of β, (unless $\rho = 0$ or $\tau = 0$), and may even have a completely different sign. In the case where $\beta = 0$, the coefficient of X in the regression of Y on X will be zero in the population, but will become non-zero once Z is included.

Similarly, in non-linear models, estimation techniques which produce consistent estimates of the strength of the influence of X on Y when Z is not conditioned on will produce inconsistent estimates when Z is conditioned on, and can increase the bias of an estimate rather than decrease it.

The following (non-linear) example gives an extreme example of how conditioning increases the bias of an estimate, in a known causal structure. In Figure 2, let T_1 = *gender*, T_2 = *age in 2000*, X = *shampoo usage in 2000*, Y = *number of fillings in 2000*, and Z = *shaved face in 1990*. Assume that there is no edge between X and Y, i.e. there is no causal relation between X = *shampoo usage in 2000* and Y = *number of fillings in 2000*. (The possibility that conditioning increases the association does not depend upon the absence of this edge or other edges in Figure 2 – their absence simply makes the example more vivid.) Suppose the sample is taken from people between the ages of 20 and 40 (as of the year 2000.) Given the causal path diagram in Figure 2 and no edge between *shampoo usage in 2000* and *number of fillings in 2000*, the two variables are independent. The set of people who did not shave in 1990 consists almost entirely of women who in 1990 were between 10 and 30, and men who in 1990 were under 20 (and hence in this sample between 10 and 20). If low shampoo usage is associated with being male, then those with low shampoo usage and who did not shave in 1990 are almost all currently under 30. On the other hand if high shampoo usage is associated with being female, then high shampoo usage among those who did not shave in 1990 has the same age distribution as the sample as a whole. Thus *shampoo usage in 2000* is independent of *number of fillings in 2000*, but dependent conditional on *shaved face in 1990*. So if one were to estimate the treatment effect of *shampoo usage in 2000* on *number of fillings in 2000* with the association between *shampoo usage in 2000* on *number of fillings in 2000* without conditioning on *shaved face in 1990*, one would correctly conclude that there is no treatment effect. If one were to estimate the treatment effect of *shampoo usage in 2000* on *number of fillings in 2000* with the association between *shampoo usage in 2000* on *number of fillings in 2000* conditioned on *shaved face in 1990*, one would incorrectly conclude that there is a treatment effect. The conclusion to be drawn from these examples is that there is no sense in which one is "playing safe" by including rather than excluding "potential confounders" in the conditioning set; conditioning on these variables could change a consistent estimate into an inconsistent estimate.

The situation is also made somewhat worse by the use of misleading definitions of 'confounder': sometimes a confounder is said to be a variable that is strongly correlated with both X and Y, or even a variable whose inclusion changes the coefficient of X in the regression. Since, for sufficiently large τ and ρ, Z in Figure 2 would qualify as a confounder under either of these definitions, it follows that under either definition including confounding variables in a regression may make a hitherto consistent estimator inconsistent.

Finally, it is worth reiterating the well-known fact that in certain circumstances there may be no regression which will estimate the parameter of interest, (although some other consistent estimator may exist):

$$\begin{tikzcd} & \boxed{T} \arrow[dl, "1"'] \arrow[dr, "\phi"] & \\ W \arrow[r, "\alpha"'] & X \arrow[r, "\beta"'] & Y \end{tikzcd}$$

Figure 3

In the SEM shown in Figure 3, Cov(X,Y) = $\beta V(X) + \phi V(T)$; hence the coefficient of X in the regression of Y on X is not a consistent estimator of β. Further

$$\frac{\mathrm{Cov}(X,Y\,|\,W)}{\mathrm{V}(X,W)} = \beta + \frac{\phi \mathrm{V}(T)}{\mathrm{V}(X) - \alpha^2 \mathrm{V}(W)} = \beta + \frac{\phi \mathrm{V}(T)}{\mathrm{V}(T) + \mathrm{V}(X)}$$

hence including W in the regression does not help matters. However, a consistent estimator exists, the so-called Instrumental Variable estimator:

$$\frac{\mathrm{Cov}(Y,W)}{\mathrm{Cov}(X,W)} = \frac{\alpha\beta \mathrm{V}(W)}{\alpha \mathrm{V}(W)} = \beta$$

The instrumental variable estimator requires background causal knowledge that there is no causal path from W to Y except through X, and no unmeasured confounder of W and Y. Both of these assumptions are untestable in the data.

These are not simply theoretical worries; conditioning in real data sets can and does increase associations between variables, and can make a large difference to an analysis. For example, by measuring the concentration of lead in a child's baby teeth, Herbert Needleman was the first epidemiologist to even approximate a reliable measure of cumulative lead exposure. His work helped convince the United States to eliminate lead from gasoline and most paint. In their 1985 article in *Science*, Needleman, Geiger and Frank gave results for a multivariate linear regression of children's IQ on lead exposure. Having started their analysis with almost 40 covariates, they were faced with a variable selection problem to which they applied backwards elimination regression, arriving at a final regression equation involving lead and five covariates. The covariates were measures of genetic contributions to the child's IQ (the parent's IQ), the amount of environmental stimulation in the child's early environment (the mother's education), physical factors that might compromise the child's cognitive endowment (the number of previous live births), and the parent's age at the birth of the child, which might be a proxy for many factors. The measured variables they used are as follows:

ciq	- child's verbal IQ score	*piq*	- parent's IQ scores
lead	- measured concentration in baby teeth	*mab*	- mother's age at child's birth
med	- mother's level of education in years	*fab*	- father's age at child's birth
nlb	- number of live births previous to the sampled child		

The standardized regression solution[1] is as follows, with t-ratios in parentheses. Except for *fab*, which is significant at 0.1, all coefficients are significant at 0.05, and $R^2 = .271$.

$$\hat{ciq} = -.143\, lead + .219\, med + .247\, piq + .237\, mab - .204\, fab - .159\, nlb \qquad [1]$$
$$(2.32) \qquad (3.08) \qquad (3.87) \qquad (1.97) \qquad (1.79) \qquad (2.30)$$

The standardized regression coefficient for *mab* in effect conditions on all of the other measured variables, and is significant. However, although *ciq* is conditionally dependent on *mab* given all of the other variables, it is unconditionally independent of *ciq*. The same is true of *fab* and *nlb*. Leaving *mab*, *fab*, and *nlb* substantially changes the conclusions drawn from the study (Scheines 1999).

These examples raise the following general questions.

(a) *If Y is regressed on a set of variables W, including X, in which SEMs will the partial regression coefficient of X be a consistent estimate of the structural coefficient β associated with the $X \rightarrow Y$ edge?*

(b) *If Y is regressed on the set **W**, including X, in which SEMs will the partial regression coefficient of X be zero if there is no edge between X and Y?*

(c) *Given a particular SEM, with path diagram G, in which there is an edge $X \rightarrow Y$, with coefficient β, is it possible to find a subset **W** of observed variables, (including X), such that when Y is regressed on the set **W**, the coefficient of X in the regression is a consistent estimate of β?*

In order to answer these questions, it is first necessary to define a graphical relationship called d-separation and related path diagram terminology. The concepts defined here are illustrated in Figure 4. A path diagram consists of two parts, a set of vertices **V** and a set of edges **E**. Each edge in **E** is between two distinct vertices in **V**. There are two kinds of edges in **E**, directed edges $A \rightarrow B$ or $A \leftarrow B$, and double-headed edges $A \leftrightarrow B$; in either case A and *B* are **endpoints** of the edge; further, A and *B* are said to be **adjacent**. There may be multiple edges between vertices. In Figure 4 the set of vertices is $\{A,B,C,D,E\}$ and the set of edges is $\{A \leftrightarrow B, B \rightarrow C, C \rightarrow D, D \rightarrow C, E \rightarrow D\}$. For a directed edge $A \rightarrow B$, *A* is the **tail** of the edge and *B* is the **head** of the edge, *A* is a **parent** of *B*, and *B* is a **child** of *A*.

[1] The covariance data for this reanalysis was originally obtained from Needleman by Steve Klepper, who generously forwarded it. In this, and all subsequent analyses, the correlation matrix is used.

An **undirected path** U between X_a and X_b is a sequence of edges $<E_1,...,E_m>$ such that one endpoint of E_1 is X_a, one endpoint of E_m is X_b, and for each pair of consecutive edges E_i, E_{i+1} in the sequence, $E_i \neq E_{i+1}$, and one endpoint of E_i equals one endpoint of E_{i+1}. In Figure 4, $A \leftrightarrow B \rightarrow C \leftarrow D$ is an example of an undirected path between A and D. A **directed path** P between X_a and X_b is a sequence of directed edges $<E_1,...,E_m>$ such that the tail of E_a is X_1, the head of E_m is X_b, and for each pair of edges E_i, E_{i+1} adjacent in the sequence, $E_i \neq E_{i+1}$, and the head of E_i is the tail of E_{i+1}. For example, $B \rightarrow C \rightarrow D$ is a directed path. A **vertex occurs on a path** if it is an endpoint of one of the edges in the path. The set of vertices on $A \leftrightarrow B \rightarrow C \leftarrow D$ is $\{A, B, C, D\}$. A path is **acyclic** if no vertex occurs more than once on the path. $C \rightarrow D \rightarrow C$ is a cyclic directed path. The following is a list of all the acyclic directed paths in Figure 4: $B \rightarrow C$, $C \rightarrow D$, $E \rightarrow D$, $D \rightarrow C$, $B \rightarrow C \rightarrow D$, $E \rightarrow D \rightarrow C$.

A vertex A is an **ancestor** of B (and B is a **descendant** of A) if and only if either there is a directed path from A to B or $A = B$. Thus the ancestor relation is the transitive, reflexive closure of the parent relation. The following table lists the child, parent, descendant and ancestor relations in Figure 4.

Vertex	Children	Parents	Descendants	Ancestors
A	$\emptyset$	$\emptyset$	$\{A\}$	$\{A\}$
B	$\{C\}$	$\emptyset$	$\{B,C,D\}$	$\{B\}$
C	$\{D\}$	$\{B,D\}$	$\{C,D\}$	$\{B,C,D,E\}$
D	$\{C\}$	$\{C,E\}$	$\{C,D\}$	$\{B,C,D,E\}$
E	$\{D\}$	$\emptyset$	$\{C,D,E\}$	$\{E\}$

A vertex X is a **collider** on undirected path U if and only if U contains a subpath $Y \leftrightarrow X \leftrightarrow Z$, or $Y \rightarrow X \leftrightarrow Z$, or $Y \rightarrow X \leftarrow Z$, or $Y \leftrightarrow X \leftarrow Z$; otherwise if X is on U it is a **non-collider** on U. For example, C is a collider on $B \rightarrow C \leftarrow D$ but a non-collider on $B \rightarrow C \rightarrow D$. X is an **ancestor of a set** of vertices **Z** if X is an ancestor of some member of **Z**.

For disjoint sets of vertices, **X**, ***Y***, and **Z**, **X** is **d-connected** to ***Y*** given **Z** if and only if there is an acyclic undirected path U between some member X of **X**, and some member Y of **Y**, such that every collider on U is an ancestor of **Z**, and every non-collider on U is not in **Z**. For disjoint sets of vertices, **X**, ***Y***, and **Z**, **X** is **d-separated** from ***Y*** given **Z** if and only if **X** is not d-connected to ***Y*** given **Z**.

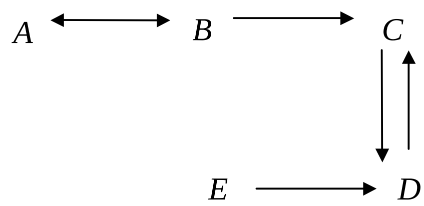


Figure 4

For example, the path $E \rightarrow D \rightarrow C$ d-connects E and C given $\varnothing$; it also d-connects E and C given $\{A\}$, $\{B\}$, or $\{A,B\}$. $E \rightarrow D \leftarrow C$ d-connects E and C given$\{D\}$, given $\{D,B\}$, $\{D,A\}$, or $\{D,A,B\}$. The following is a list of all the pairwise d-separation relations in Figure 4 (where each pair is followed by a list of all of the sets that d-separate them):

$\{A\}$ and $\{C\}$ are d-separated given: $\{B\}$, $\{B,D\}$, $\{B,E\}$, $\{B,D,E\}$

$\{A\}$ and $\{D\}$ are d-separated given: $\{B\}$, $\{B,C\}$, $\{B,E\}$, $\{B,C,E\}$

$\{A\}$ and $\{E\}$ are d-separated given: $\varnothing$, $\{B\}$, $\{B,C\}$, $\{B,D\}$, $\{B,C,D\}$, $\{C,D\}$

$\{B\}$ and $\{E\}$ are d-separated given: $\varnothing$, $\{C,D\}$

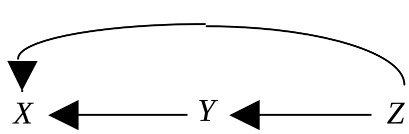
The first theorem states that d-separation in a path diagram G is a sufficient condition for G to entail that $\rho(X,Y.\mathbf{Z}) = 0$ (i.e. in every SEM with path diagram G, the partial correlation of X and Y given $\mathbf{Z}$ equals 0.) The following theorems are from Spirtes (1995), and are an extension of results in Pearl (1988) to graphs which contain directed cycles and double-headed arrows. Theorem 1 also was proved in Koster (1996). $\rho(X,Y.\mathbf{Z})$ is the partial correlation of X and Y given $\mathbf{Z}$.

Theorem 1: If M is a SEM, and $\{X\}$ and $\{Y\}$ are d-separated given $\mathbf{Z}$ in $G(M)$, then $\rho(X,Y.\mathbf{Z}) = 0$ in Σ(M).

The second theorem states that d-separation is a necessary condition for a path diagram to entail a zero partial correlation.

Theorem 2: If $\{X\}$ and $\{Y\}$ are not d-separated given $\mathbf{Z}$ in path diagram G, then there is a SEM M such that $G(M) = G$, and $\rho(X,Y.\mathbf{Z}) \neq 0$ in Σ(M).

Theorem 2 does *not* say that there might not be an individual SEM M with "extra" zero partial correlations among variables that are not d-separated in $G(M)$, as the following example shows.



$$X = .3\,Y + .6\,Z + \varepsilon_X$$
$$Y = -2\,Z + \varepsilon_Y$$
$$Z = \varepsilon_Z$$

Figure 5

(The errors are uncorrelated because there are no double-headed arrows in the path diagram.) In this case X and Y are independent, i.e. $\rho(X,Y) = 0$, even though $\{X\}$ and $\{Y\}$ are not d-separated given $\emptyset$. However, this zero correlation holds because of the particular linear coefficients. Thus, according to Theorem 2 there is some other SEM M such with the same path diagram in which $\rho(X,Y) \neq 0$. It has been shown (Spirtes et. al 1993) that the set of parameters which produce conditional independence relations among variables which are not d-separated in G has zero Lebesgue measure over the parameter space.

The answers to the questions that I asked earlier can now be stated. They are simple applications of d-separation and appear in (Spirtes et al 1998).

(a) *If Y is regressed on a set of variables* **W**, *including X, in which SEMs will the partial regression coefficient of X be a consistent estimate of the structural coefficient* β *associated with the* $X \rightarrow Y$ *edge?*

The coefficient of X is a consistent estimator of β if **W** does not contain any descendant of Y in G, and X is d-separated from Y given **W** in $G\backslash\{X \rightarrow Y\}$.[2] If this condition does not hold, then for almost all instantiations of the parameters in the SEM, the coefficient of X will fail to be a consistent estimator of β.

It follows directly from this that (almost surely) β cannot be estimated consistently via any regression equation if either there is an edge $X \leftrightarrow Y$ (i.e. ε_X and ε_Y are correlated) or if X is a descendant of Y (so that the path diagram is cyclic). And as the example of Figure 2 shows, there are cases where β cannot be estimated consistently via the regression equation even though only X and potential confounders of X and Y are regressed on, and there is no actual hidden confounding (in the sense that there are no hidden variables causing both X and Y.)

Moreover, the answer to (a) validates the intuition that conditioning on all potentials confounders produces an unbiased estimate. The set of potential confounders does not contain any descendant of Y, because a potential confounder must occur prior to Y; hence

[2]Note this criterion is similar to Pearl's back door criterion (Pearl, 1995), except that the back-door criterion was proposed as a means of estimating the *total* effect of X on Y.

no descendant of Y is conditioned on. Suppose, contrary to the hypothesis that there is a path U d-connecting X and Y given the set of potential confounders. U either contains an edge into Y, or an edge out of Y. If U contains an edge out of Y, then because Y is not an ancestor of X, U contains a collider. Because U d-connects X and Y given the potential confounders, there is a potential confounder which is a descendant of each collider on U. But then there is a potential confounder C that is a descendant of the collider on U closest to Y on U. It follows that C is a descendant of Y, which is a contradiction. On the other hand, if U contains an edge into Y, then the parent of Y on U is a potential confounder, and conditioned on. Hence U does not d-connect X and Y given the potential confounders, which is a contradiction.

Note that this proof does not work if one conditions on only the *measured* potential confounders. In that case the parent of Y on U may be unmeasured, and hence not conditioned on. Referring again to the causal path diagram of Figure 2, X and Y are d-separated by the empty set, and by $\{T_1,T_2,Z\}$, but not by $\{Z\}$. Thus conditioning on no variables ($\emptyset$) produces a consistent estimate, and conditioning on the set of all potential confounders ($\{T_1,T_2,Z\}$) produces a consistent estimate, but conditioning on the set of all measured potential confounders ($\{Z\}$) does not produce a consistent estimate.

Note that conditioning on a variable can turn a consistent estimate of β into an inconsistent estimate only if it is a descendant of Y or a collider on a path between X and Y. If background knowledge indicates that the variable is exogenous (e.g. age or gender) then it cannot be a collider on a path or a descendant of another variable, and (apart from sampling problems) there is no reason not to condition on it.

(b) *If Y is regressed on the set* **W**, *including X, in which SEMs will the partial regression coefficient of X be zero if there is no edge between X and Y?*

The coefficient of X will be zero if X and Y are d-separated given $\mathbf{W}\backslash\{X\}$. (See Scheines (1994) and Glymour (1994)). This follows directly from the fact that the coefficient of X in the regression equation is proportional to $\rho(X,Y.\mathbf{W}\backslash\{X\})$, which in turn will be zero if $\{X\}$ is d-separated from $\{Y\}$ given $\mathbf{W}\backslash\{X\}$. As before, if $\{X\}$ and $\{Y\}$ are not d-separated given $\mathbf{W}\backslash\{X\}$, then, even if there is no edge between X and Y, for almost all assignments of values to the model parameters the coefficient of X will be non-zero.

(c) *Given a particular SEM, with path diagram G, in which there is an edge* $X \rightarrow Y$, *with coefficient* β, *is it possible to find a subset* **W** *of observed variables, (including X), such that when Y is regressed on the set* **W**, *the coefficient of X in the regression is a consistent estimate of* β?

From (a), if there is a subset **W** of the observed variables which contains no descendant of Y, but which d-separates X from Y in $G\backslash\{X \rightarrow Y\}$, then the regression coefficient of X in the regression of Y on **W** will be a consistent estimate of β.

The general conclusion to be drawn from this is that not only is it not possible to get a consistent estimate of the edge coefficient of $X \rightarrow Y$ without knowing how much unmeasured confounding there is, it is not even possible to know whether conditioning on a potential confounder reduces the amount of bias in an estimate without knowing details of the correct causal path diagram. Moreover, this conclusion applies to other methods typically used to "control for" confounding. Matching on measured covariates, case-control studies, matching on propensity scores, and blocking may all make estimates more biased rather than less, depending upon what the true causal path diagram is.

3. Sensitivity Analyses

3.1. Single Sided Sensitivity Analyses

Sensitivity analyses attempt to deal with the problem of potential hidden confounders by estimating an upper bound to how much association between measured variables could be produced in the worst case by hidden variables. It is possible to do sensitivity analyses for the results of statistical tests or confidence intervals. I will concentrate on the case of statistical tests. In all examples discussed below, Z represents treatment, R represents response, and **X** represents the set of measured covariates of treatment (i.e. variables occurring prior to treatment).

The intutition behind sensitivity analyses is quite simple. A dependency bewteen Z and R could either be due to Z causing R, or due to some measured variables **X** and unmeasured variables **H**. First consider the case where there are no measured covariates. For example in G_2 in Figure 6 there are no measured covariates, and unmeasured covariates U and X are confounders of Z and R that produce a dependence between Z and R, so $\mathbf{H} = \{U,X\}$. If we could condition on **H** then Z and R would be (conditionally) independent; but because **H** is unmeasured, it cannot be conditioned on, and hence Z and R are (unconditionally) dependent. If we can estimate an upper bound on the amount of dependence between Z and R due to **H**, then any observed dependence between Z and R that is much larger than the upper bound due to **H** is produced either by sampling variation or a direct effect of Z on R. If the sampling variation explanation is unlikely enough, the conclusion is that at least some of the dependence between Z and R is due to Z causing R, and the null hypothesis can be rejected.

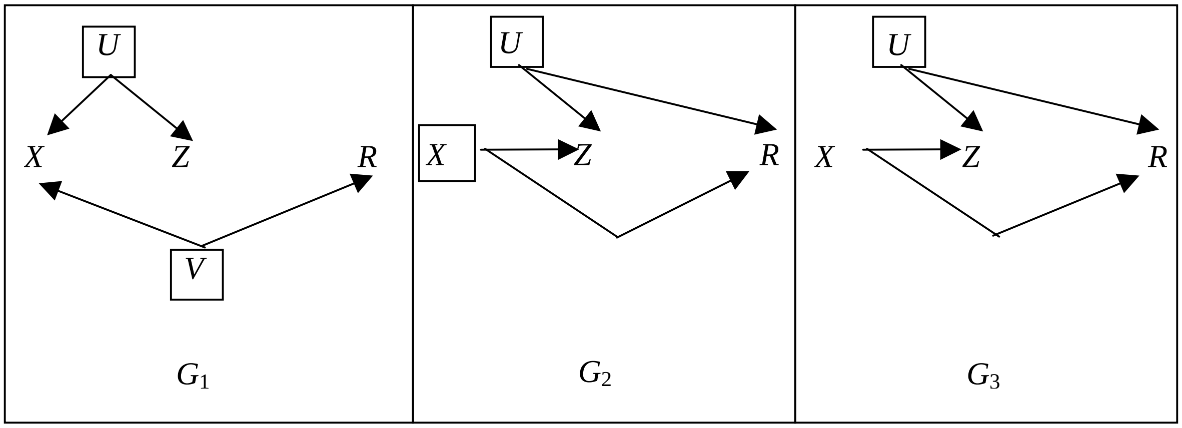


Figure 6

An upper bound to the dependence of Z and R due to **H** can be measured by defining a measure Γ of the combined effect of **H** on treatment, and a measure Δ of the combined effect of **H** on R. An upper bound on the amount of dependence between Z and R due to **H** is then a function of Γ and Δ. In a single sided sensitivity analysis, as a worst case it is assumed that R is perfectly dependent on **H**.

Now consider the case where there are measured covariates, as in G_3 of Figure 6, where X is measured but U is not. Under the null hypothesis a dependence between Z and R is due to measured covariates **X** (equal to $\{X\}$ in this example), and unmeasured variables **H** (equal to $\{U\}$ in this example). However, after conditioning on measured covariates **X**, any residual dependence between Z and R is due to **H**. We can now take $\Gamma_{\mathbf{X}}$ to be a measure of the effect of the unmeasured variables on Z conditonal on **X**; and similarly $\Delta_{\mathbf{X}}$ is a measured of the combined effect of the unmeasured variables on R conditional on **X**. An upper bound on the amount of dependence due to unmeasured variables is then a function of $\Gamma_{\mathbf{X}}$ and $\Delta_{\mathbf{X}}$. In a single sided sensitivity analysis, as a worst case it is assumed that the unmeasured variables are perfectly associated with the response conditional on **X**.

How are the measures of the effects of **H** on Z and R defined? In Rosenbaum's analysis of binary treatment and response, for a given stratum $\mathbf{X} = \mathbf{x}$, $\Gamma_{\mathbf{x}}$ is the ratio of the maximum odds ratio of treatment for someone within stratum **x**, to the minimum odds ratio of treatment for someone within stratum **x**. $\Gamma_{\mathbf{X}}$ is the maximum over all strata **x** of $\Gamma_{\mathbf{x}}$.[3] $\Delta_{\mathbf{X}}$ can be defined analogously. So

[3] Rosenbaum (1995) does not explicitly introduce **X** as a subscript of Γ. I include the subscript in order to remind the reader that Γ depends upon **X**.

$$\Gamma_{\mathbf{x}} = \frac{\max_{\mathbf{h}} P(Z=1 \mid \mathbf{H}=\mathbf{h}, \mathbf{X}=\mathbf{x}) / P(Z=0 \mid \mathbf{H}=\mathbf{h}, \mathbf{X}=\mathbf{x})}{\min_{\mathbf{h}} P(Z=1 \mid \mathbf{H}=\mathbf{h}, \mathbf{X}=\mathbf{x}) / P(Z=0 \mid \mathbf{H}=\mathbf{h}, \mathbf{X}=\mathbf{x})} \qquad \Gamma_{\mathbf{X}} = \max_{\mathbf{x}} \Gamma_{\mathbf{x}}$$

Table 1 is an example of the results of a sensitivity analysis for a study of lead in children's blood. The original study is in Morton et al. (1982). The "treatment" is whether or not the parents of the children work in a battery factory where lead is present, and the response is the level of lead in the child's blood. Researchers investigated the hypothesis that workers in the battery factory causes elevated lead levels in their children by bringing traces of lead home from work. The study matched pairs of children on neighborhood and age. The results of the sensitivity analysis are upper and lower bounds for the p-values of a statistical test as a function of Γ. Note that the statistical test does not fail to reject at the 0.05 significance level until Γ = 5. This is fairly insensitive to hidden bias. According to Rosenbaum (1995), if a null hypothesis can be rejected at Γ = 6 this is "a high degree of insensitivity to hidden bias – in many other studies, biases smaller than Γ = 6 could explain the association between treatment and response."

Γ	Minimum	Maximum
1	<0.0001	<0.0001
2	<0.0001	0.0019
3	<0.0001	0.0143
4	<0.0001	0.0401
4.25	<0.0001	0.0485
5	<0.0001	0.0764

Table 1: Sensitivity Analysis for Lead in the Blood of Children (Rosenbaum, p. 97)

According to Rosenbaum (1995), sensitivity analyses "play an important role in judging whether the treatment caused its ostensible effects or alternatively whether these seeming effects could merely reflect a hidden bias" (where in a single-sided sensitivity analysis, Γ is the measure of "bias".) However, if sensitivity analyses always lead to the conclusion that the conclusion is sensitive to an amount of bias that cannot be ruled out, it just indicates that there is not much point to doing observational studies (at least if one is interested in judging the existence of size of causal effects). For example, in order to reliably reject the null hypothesis in the lead study of Table 1 the amount of bias (Γ) has to be less than 5, and experts need to be able to judge that an upper bound to the bias is less than 5. I will now argue that in the typical case Γ is orders of magnitude larger than 5 by examining which variables need to be included in **H**.

What are the members of the set of unmeasured covariates **H**, whose values are being maximized over? So far I have not said what kind of unmeasured variables produce a dependence between *Z* and *R*. As long as there are no measured covariates, the only kind of unmeasured variables that produce a dependence between *Z* and *R* are unmeasured variables that cause both *Z* and *R*, i.e. confounders of *Z* and *R*. So as long as there are no measured covariates, the only unmeasured variables that need be used to define $\Gamma_\varnothing$ and $\Delta_\varnothing$ are confounders of *Z* and *R*.

The problem with not measuring any covariates is that typically an estimate of an upper limit to $\Gamma_\varnothing$ will be large, because there are many potential confounders in **H**. In G_2, *H* would contain both *U* and *X*. In Morton's lead study, if neighborhood and age were not conditioned on, then in the worst case, $\Gamma_\varnothing$ might, for all we know, be quite large due to those two factors alone.

One way to try solve the problem of a large upper limit to $\Gamma_\varnothing$ is to condition on measured covariates **X**. What kind of unmeasured variables can produce a dependence between *Z* and *R conditional on* **X**? As long as the causal path diagram is of the kind shown in G_3 of Figure 6, once again the only unmeasured variables that produce a dependence between *Z* and *R* conditional on **X** are the unmeasured confounders of *Z* and *R* (e.g. in G_3 the variable *U*). *Z* and *R* are independent conditional on the set of unmeasured confounders of *Z* and *R*. So in causal path diagrams like G_3, the only unmeasured variables that need to be used to define $\Gamma_{\mathbf{X}}$ and $\Delta_{\mathbf{X}}$ are the unmeasured confounders of *Z* and *R*. In all cases, as long as the causal path diagram is of the kind shown in G_3 of Figure 6, if $\mathbf{X} \subset \mathbf{X'}$, then the dependence between *Z* and *R* conditional on **X** is greater than or equal to the dependence between *Z* and *R* conditional on **X'**, $\Gamma_{\mathbf{X'}} \leq \Gamma_{\mathbf{X}}$, and $\Delta_{\mathbf{X'}} \leq \Delta_{\mathbf{X}}$. Hence, for causal path diagrams like those of G_3 of Figure 6, apart from sampling problems, it is always advantageous to condition on as many covariates of treatment as possible.

The problems with sensitivity analyses arise because of causal path diagrams like G_1 of Figure 6 (which is a subgraph of Figure 2 (a).) Suppose that the joint distribution is the following, where ε is a positive number near 0:

$R = V$	$P(V = 1) = ½$	$Z = U$	$P(U = 1) = ½$
$P(X = 1 \mid U = 0, V = 0) = 1$		$P(X = 1 \mid U = 0, V = 1) = 1 - \varepsilon$	
$P(X = 0 \mid U = 1, V = 0) = 1$		$P(X = 1 \mid U = 1, V = 1) = 1$	

There are no confounders of Z and R at all in G_1. If the covariate X is not conditioned on, then Z and R are independent, and $\Gamma_{\varnothing} = \Delta_{\varnothing} = 1$. However, conditional on X, Z and R are dependent, because there is a path that d-connects Z and R given X. In this case, Z and R are not independent conditional on the measured covariates and the (empty) set of confounders of Z and R. Although U and V are not confounders of Z and R, and do not produce a dependence between Z and R, U and V do produce a dependence between Z and R conditional on X (in the sense that they lie along a path that d-connects Z and R given X.) Z and R are independent conditional on X and either U or V. So in this case either U or V is in **H**. If $\mathbf{H} = \{U\}$ or $\mathbf{H} = \{V\}$ or $\mathbf{H} = \{U,V\}$ then $\Gamma_{\mathbf{X}} = \Delta_{\mathbf{X}} = \infty$. This example shows that although $\varnothing \subset \{X\}$, the dependence between Z and R conditional on X is greater than the dependence between Z and R conditional on $\varnothing$, $\Gamma_{\varnothing} \leq \Gamma_X$, and $\Delta_{\varnothing} \leq \Delta_X$.

What is the general rules for determining which variables are in **H**? In general **H** should include enough variables so that any dependence between Z and R conditional on **X** is due to the unmeasured variables in **H**, so that if both **X** and **H** (instead of just **X**) were conditioned on, then Z and R would be independent. This is guaranteed if Z and R are d-separated conditional on **X** and **H** in the true causal path diagram.[4] If the true causal path diagram were known, it would be possible to calculate from the causal path diagram a subset of the parents of Z that must be included in **H** in order to make Z and R d-separated given **X** and **H**. But in the absence of knowledge of the true causal path diagram, in order to ensure that Z and R are d-separated given **X** and **H**, regardless of what causal relationships exist among **X,** Z, and R, **H** must include all of the unmeasured parents of Z whether they are also causes of R or not. And if one does not know which covariates of treatment are actually parents of Z, then one must include in **H** *all* of the unmeasured covariates of treatment in order to guarantee that treatment Z and R are d-separated given **X** and **H**. (This implies that $\mathbf{X} \cup \mathbf{H}$ is a partition of the covariates of treatment.)

[4] Rosenbaum (1995) does not explicitly say which unmeasured covariates are to be conditioned on. He does say that he assumes that when the unmeasured covariates **H**, together with the measured covariates, are conditioned on, then treatment is "strongly ignorable". This guarantees that $P(R|Z,\mathbf{X},\mathbf{H})$ in the observed distribution is equal to $P(R|Z,\mathbf{X},\mathbf{H})$ in the distribution where Z is experimentally manipulated, thus connecting the causal null hypothesis (Z does not cause R) with a statistical null hypothesis (Z and R are independent given **X** and **H**.) In graphical terms, under the null hypothesis, this is equivalent to the assumption that there are no paths d-connecting Z and R given **H** and **X**.

But if the values of all of the unmeasured covariates of treatment (or even just the parents of treatment) are maximized over when calculating $\Gamma_{\mathbf{X}}$, in *any* deterministic setting $\Gamma_{\mathbf{X}} = \infty$, as long as the measured covariates do not perfectly predict treatment. For example, in the example of G_1 in Figure 6, if $\mathbf{X} = \{X\}$ then $\Gamma_{\mathbf{X}} = \infty$ because U is a parent of Z, and when $U = 1$, $P(Z = 1) = 1$, and when $U = 0$, $P(Z = 0) = 1$. So even if the measured covariates correctly predict treatment 99.999999% of the time, and none of the parents of treatment in the true causal path diagram are causal ancestors of the response (i.e. there is no directed path from a parent of treatment to the response) nor correlated with the response, if treatment is determined by the parents of treatment, $\Gamma_{\mathbf{X}} = \infty$. This is the case for every SEM, where a variable is a deterministic function of its parents in the causal path diagram and an error term (assuming the measured covariates do not perfectly predict treatment.) And even if $\Gamma_{\mathbf{X}}$ is not actually large, the estimate of an upper bound of $\Gamma_{\mathbf{X}}$ that is very large or infinite does not require having an explicit list of causes of treatment that one has not measured that determine treatment – it simply requires believing that there might be such a list of unmeasured causes.

Even if one does not believe that the parents of treatment in the causal path diagram completely determine treatment, unless there are large, sudden, highly random changes in the values of macroscopic random variables, $\Gamma_{\mathbf{X}}$ will be orders of magnitude larger than 5. Consider the example of the lead study of Morton et al. (1982). In the lead study the matching was done on neighborhood of residence and age. If we are allowed to include as unmeasured covariates such features as where the parents applied for jobs prior to the study, where the parents worked prior to the study, the parent's education, the socio-economic status of the family the parents grew up in, the complete state of the parents down to every cell in their body including an exact record of every neuron in his or her brain, we would expect $\Gamma_{\mathbf{X}}$ to be enormous. Granted some of these variables are unlikely to be causes of the response (except via working at the battery factory during the time of the study), and hence not confounders (in the sense of being unmeasured causes of both the treatment and the response), but in the kind of worst-case sensitivity analysis that Rosenbaum suggests, they should be considered in estimating $\Gamma_{\mathbf{X}}$; or if they are not to be considered then the grounds on which they are to be excluded in estimating $\Gamma_{\mathbf{X}}$ should be explained and justified.

It might be argued that causal path diagrams of the kind shown in G_1 in Figure 6 occur rarely in nature. However, the points about the problems presented by causal path diagrams of the kind shown in Figure 6 depend only upon G_1 in Figure 6 being a sub-path diagram of the true causal path diagram, i.e. that there is an unmeasured variable causing both treatment and covariates of treatment, and an unmeasured variable causing both response and covariates of treatment. The motivation of sensitivity analysis is the worry that even a large dependence between measured variables can be produced by unmeasured causes of both treatment and response. In the absence of background knowledge to the contrary, the same worry exists for the dependence between treatment and covariates of treatment, and the dependence between response and covariates of treatment.

3.2. Dual Sensitivity Analyses

A dual sensitivity analysis attempts to deal with the problem of large $\Gamma_{\mathbf{X}}$ by introducing a second parameter $\Delta_{\mathbf{X}}$, e.g. see Gastwirth, Krieger, and Rosenbaum (1998). In the case of binary variables, $\Delta_{\mathbf{x}}$ is the ratio of the largest probability of a response to the smallest probability of a response in a given stratum **x** of the covariates **X**, $\Delta_{\mathbf{X}}$ is the maximum of the $\Delta_{\mathbf{x}}$. When $\Delta_{\mathbf{X}} = 1$, the upper bound on the p-value is independent of $\Gamma_{\mathbf{X}}$, so even if $\Gamma_{\mathbf{X}}$ is quite large, if $\Delta_{\mathbf{X}}$ is around 1, it is sometimes possible to reject the null hypothesis. Similarly when $\Gamma_{\mathbf{X}} = 1$, the upper bound on the p-value is independent of $\Delta_{\mathbf{X}}$. However, when $\Gamma_{\mathbf{X}} = \Delta_{\mathbf{X}} = \infty$, then the upper bound on the p-value is always 1. A single-sided sensitivity analysis is the case where the upper limit placed on $\Delta_{\mathbf{X}}$ is ∞.

Gastwirth et al. are not explicit about the values of which sets of variables are being maximized over when calculating $\Delta_{\mathbf{X}}$, other than that they must occur prior to treatment. However, the same problems now arise for $\Delta_{\mathbf{X}}$ as arose for $\Gamma_{\mathbf{X}}$.

If the values of all of the unmeasured covariates of treatment are maximized over in calculating $\Delta_{\mathbf{X}}$, then if **X’** and **X** are sets of measured covariates, and $\mathbf{X'} \subset \mathbf{X}$, then $\Delta_{\mathbf{X'}} \geq \Delta_{\mathbf{X}}$; that is the sensitivity cannot increase, and might decrease if more measured covariates of treatment are conditioned on.[5] In particular, in calculating $\Delta_{\mathbf{X}}$, in order to ensure that if $\mathbf{X'} \subset \mathbf{X}$, then $\Delta_{\mathbf{X'}} \geq \Delta_{\mathbf{X}}$, it is not enough to maximize over the values of unmeasured covariates of treatment that cause both treatment and response. It is also necessary to maximize over the values of unmeasured covariates of treatment that are not causes of treatment.

However, if in calculating $\Delta_{\mathbf{X}}$, the values of all of the unmeasured covariates of treatment are maximized over, then $\Delta_{\mathbf{X}}$ will be infinite if the past determines the present. In the example of G_1 of Figure 6, if $\mathbf{X} = \{X\}$ and the values of all of the unmeasured covariates of treatment are maximized over (U and V), then for a given value of X, some

---

[5] Note, unlike the case of Γ, for this to be true it is not enough to maximize over only the values of the parents of treatment; one must maximize over the values of all the covariates of treatment.

units have $V = 0$ and hence $R = 0$, while other units have $V = 1$ and hence $R = 1$. Hence Δ_X is infinite. Even in the absence of complete determinism Δ_X could be quite large, especially if the time interval between treatment and response is not large. In the lead study of Morton et al. (1982), one unmeasured covariate of treatment is whether the child has eaten lead paint. If there is a pair of children in which one has eaten lead paint prior to treatment, and one has not, even given the null hypothesis there could be large differences in the probability of a response. Because in calculating Δ_X the values of all of the unmeasured covariates of working in the battery factory are maximized over, it follows that Δ_X can be large even if the eating of lead paint did not cause the parent to work in the battery factory. Nor is it is necessary to know that a difference in paint-eating actually occurred in order for a conservative estimate to put a high upper bound on Δ_X; it is enough that it (or some other unmeasured cause of lead in the blood) might have occurred.

3.3. Other Problems with Sensitivity Analyses

I have argued that typically reasonable upper bound estimates of Γ, Δ, and the association between a pair of variables conditional on a set of potential covariates will be quite large. There is a separate question of whether people are capable of making reasonable estimates of these quantities at all. Each of them is a complicated functions of the true causal path diagram, the magnitudes of the parameters, the signs of the parameters, which measured variables are conditioned on, and which unmeasured variables have values that are maximized over. To estimate an upper bound on any of these quantities requires estimating the maximum values these can attain over many different causal path diagrams and parameters. Since many statisticians and social scientists are unaware that conditioning on measured variables that are colliders in a path diagram can induce dependence between variables, they are not capable of properly estimating bounds even in very simple cases where the causal structure is completely known.[6] This implies they are unlikely to be able to accurately estimate upper bounds with any accuracy in the much more complicated cases that occur in actual applications, and which require considering enormous number of alternative models.

---

[6] The well-known Monty Hall or Three Prisoners paradox seems paradoxical precisely because in the causal path diagram the door Monty shows you after your guess is a collider on a path between your guess and the door where the prize is. This makes your guess and where the prize is independent, but dependent conditional on the door Monty shows you.

A second problem with the Gastwirth et al. (1998) dual sensitivity analysis of matched pairs concerns the specific form of the distributions of Z and R that they assume. The form of the parameterization used in the Gastwirth et al. (1998) dual sensitivity analysis of matched pairs places restrictions on the joint distribution of Z and R which are not plausible in many domains. In their notation, let the index i indicate which matched pair the unit is in, and the index j indicates which member of the matched pair the unit is.

$$P(Z_{ij} = z|u_{ij}) = \exp\{\alpha_i(u_{ij}) + \beta_i(z) + \gamma z u_{ij}\}$$
$$P(R_{ij} = z|u_{ij}) = \exp\{\zeta_i(u_{ij}) + \kappa_i(r) + \delta r u_{ij}\}$$

$\alpha_i(u_{ij})$ and $\zeta_i(u_{ij})$ are normalizing constants. The form of the first equation places no restriction on the distribution of Z; it simply reparameterizes the distribution in terms of an unmeasured variable u. The variable u is defined so that $u = 0$ is assigned to the unit with the smallest probability of $Z = 1$, and $u = 1$ is the assigned to the unit with the highest probability of $Z = 1$. In general the value are assigned to u in such a way that if γ is positive then $u_{ij} > u_{kl}$, then $P(Z_{ij} = 1|u_{ij}) > P(Z_{kl} = 1|u_{kl})$. Similarly, the form of the second equation by itself places no constraints on the distribution of R. However, together they place strong constraints upon the joint distribution of Z and R that are implausible in many domains. For instance, under the null hypothesis, if γ and $\delta > 0$, then if $u_{ij} > u_{kl}$, then $P(Z_{ij} = 1|u_{ij}) > P(Z_{kl} = 1|u_{kl})$ and $P(R_{ij} = 1|u_{ij}) > P(R_{kl} = 1|u_{kl})$, i.e. the units with the highest probabilities of $Z = 1$ also have the highest probabilities of $R = 1$. Other combinations of signs for γ and δ place similar constraints.

4. Conclusion

Potential confounders raise problems for inferences about the strength of causal connections between measured variables. Known methods that assume only the temporal order of the measured variables and attempt to deal with potential measured confounders by conditioning risk turning consistent estimates into inconsistent estimates or greatly increasing bias. This suggests that reliable estimation of the strength of causal connections requires employing techniques that either make stronger causal assumptions (e.g. Manski 1995), or that use the data to make inferences about the causal relations among the measured variables (e.g. Spirtes et al. 1999 or Heckerman et al. 1999.)

5. Bibliography

Bollen, K., 1989, Structural Equations with Latent Variables. (Wiley, New York).

Gastwirth, J., Krieger, A., and Rosenbaum, P. (1998) Dual and simultaneous sensitivity analyses for matched pairs, *Biometrika*, 85, pp. 907-920.

Glymour, C., Scheines, R., Spirtes, P., and Kelly, K. (1987) Discovering Causal Structure: Artificial Intelligence, Philosophy and Statistical Modeling, Academic Press, San Diego, CA.

Glymour, C., Spirtes, P., and Scheines, R. (1994), In place of regression (in Patrick Suppes: Scientific Philosopher, Paul Humphreys (editor), Vol. 1, Kluwer Academic Publishers, Dordrecht, Holland.

Heckerman, D., Meek, C., and Cooper, G. (1999) A Bayesian approach to causal discovery. *Computation, Causation and Discovery*. C. Glymour and G. Cooper. Cambridge, MA, MIT Press.

Koster, J. (1996). Markov properties of non-recursive causal models. *Annals of Statistics* 24: 2148-2178.

Manski, C. (1995). *Identification Problems in the Social Sciences*. Harvard University Press, Cambridge, MA.

Morton, D., Saah, A., Sliberg, S., Owens, W., Roberts, M., and Saah, M. (1982). Lead absorption in children of employees in a lead related industry. *American Journal of Epidemiology*, **115**, 549-555.

Needleman, H., Geiger, S., and Frank, R. (1985). Lead and IQ scores: a reanalysis, *Science*, 227, pp. 701-704.

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

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, **82**, 669-709.

Rosenbaum, P. (1995) Observational Studies. Springer Series in Statistics, New York.

Scheines, R. (1994) Causation, indistinguishability, and regression, in Sofstat '93: Advances in Statistical Software 4, Frank Faulbaum (editor), Gustav Fischer Verlag, pp. 89-98.

Scheines, R., Boomsma, A., Hoijtink, H. (1999). Bayesian Estimation and Testing of Structural Equation Models. *Psychometrika* **64**(1): 37-52.

Spirtes, P., Glymour, C., and Scheines, R.(1993) Causation, Prediction, and Search**,** Springer-Verlag Lecture Notes in Statistics 81, New York.

Spirtes, P. (1995). Directed cyclic graphical representation of feedback models. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, Morgan Kaufmann: 491-498.

Spirtes, P., Richardson, T., Meek, C., Scheines, R., and Glymour, C. (1998). Using path diagrams as a structural equation modeling tool. *Sociological Methods and Research*, **27**, 148-181.

Spirtes, P., Meek, C., and Richardson, T. (1999). An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, Causation and Discovery*. C.Glymour and G. Cooper. Cambridge, MA, MIT Press.