Prepared for *Evaluation Review*

On Specifying Graphical Models for Causation, and the Identification Problem

by David A. Freedman

March 2004

Abstract

This paper (which is mainly expository) sets up graphical models for causation, having a bit less than the usual complement of hypothetical counterfactuals. Assuming the invariance of error distributions may be essential for causal inference, but the errors themselves need not be invariant. Graphs can be interpreted using conditional distributions, so that we can better address connections between the mathematical framework and causality in the world. The identification problem is posed in terms of conditionals. As will be seen, causal relationships cannot be inferred from a data set by running regressions unless there is substantial prior knowledge about the mechanisms that generated the data. The idea can be made more precise in several ways. There are few successful applications of graphical models, mainly because few causal pathways can be excluded on a priori grounds. The invariance conditions themselves remain to be assessed.

1. Introduction

In this paper, I review the logical basis for inferring causation from regression equations, proceeding by example. The starting point is a simple regression; next is a path model, and then simultaneous equations (for supply and demand). After that come nonlinear graphical models. The key to making a causal inference from nonexperimental data by regression is some kind of invariance, exogeneity being a further issue. Parameters need to be invariant to interventions: this well-known condition will be stated here with a little more precision than is customary. Invariance is also needed for (i) errors or (ii) error distributions, a topic that has attracted less attention. Invariance for distributions is a weaker assumption than invariance for errors. I will focus on invariance of error distributions in stochastic models for individual behavior, eliminating the need to assume sampling from an ill-defined super-population.

With graphical models, the essential mathematical features can be formulated in terms of conditional distributions ("Markov kernels"). To make causal inferences from nonexperimental data using such techniques, the kernels need to be invariant to intervention. The number of plausible examples is at best quite limited, in part because of sampling error, in part because of measurement error, but more fundamentally because few causal pathways can be excluded on a priori grounds. The invariance condition itself remains to be assessed.

Many readers will "know" that causal mechanisms can be inferred from nonexperimental data by running regressions. I ask from such readers an unusual boon—the suspension of belief. (Suspension of disbelief is all too readily at hand, but that is another topic.) There is a complex chain of assumptions and reasoning that leads from the data via regression to causation. One objective in the present essay to is explicate this logic. Please bear with me: what seems obvious at first may become less obvious on closer consideration, and properly so.

2. A first example: simple regression

$$\mathsf{X} \longrightarrow \mathsf{Y}$$

Figure 1. Linear regression

Figure 1 is the easiest place to start. In order to make causal inferences from simple regression, it is now conventional (at least for a small group of mathematical modelers) to assume something like the setup in equation (1) below. I will try to explain the key features in the formalism, and then offer an alternative. As will become clearer, the equation makes very strong invariance assumptions, which cannot be tested from the data on X and Y.

$$Y_{i,x} = a + bx + \delta_i. \tag{1}$$

The subscript i indexes the individuals in a study, or the occasions in a repeated-measures design, and so forth. A treatment may be applied at various levels x. The expected response $a + bx$ is by assumption linear in x, with intercept a and slope b; these parameters are the same for all subjects and all levels of treatment. When treatment at level x is applied to subject i, the response $Y_{i,x}$ deviates from the expected by a “random error” or “disturbance” δ_i. This presumably reflects the impact of chance. For some readers, it may be more natural to think of $a + \delta_i$ in (1) as a random intercept; others may classify $Y_{i,x}$ as a “potential outcome:” more about that will be said later.

In this paper, as is commonplace in statistics, random variables like δ_i are functions on a probability space Ω. Informally, chance comes in when Nature chooses a point at random from Ω, which fixes the value of δ_i. The choice is made once and once only: Nature does not re-randomize if x is changed in (1). More technically, $Y_{i,x}$ is a function of x and δ_i, but δ_i does not vary with x. (The formalism is compact, which has certain advantages; on the other hand, it is easy to lose track of the ideas.)

The δ_i are assumed to be independent and identically distributed. The common “error distribution” $\mathcal{D}$ is unknown but its mean is assumed to be 0. Nothing in the equation is observable. To generate the data, Nature is assumed to choose $\{X_i : i = 1, \ldots, n\}$ independently of $\{\delta_i : i = 1, \ldots, n\}$, showing us

$$(X_i,\ Y_i),$$

where

$$Y_i = Y_{i,X_i} = a + bX_i + \delta_i$$

for $i = 1, \ldots, n$.

Notice that x could have been anything: the model features multiple parallel universes, all of which remain counterfactual hypotheticals—because, of course, we did no intervening at all. Instead, we passively observed X_i and Y_i. (If we had done the experiment, none of these interesting issues would be worth discussing.) Nature obligingly randomizes for us. She chooses X_i at random from some distribution, independently of δ_i, and sets $Y_i = a + bX_i + \delta_i$ as required by (1).

“Exogeneity” is the assumed independence between the X_i and the errors δ_i. Almost as a bookkeeping matter, your response Y_i is computed from your X_i and error term δ_i: nobody else’s X and δ get into the act, precluding interactions across subjects. According to the model, δ_i exists—incorruptible and unchanging—in all the multiple unrealized counterfactual hypothetical universes, as well as in the one real factual observed universe. This is a remarkably strong assumption: all is flux, except a, b and δ_i.

An alternative setup will be presented next—more like standard regression—to weaken the invariance assumption. We start with parameters a, b and an error distribution $\mathcal{D}$. The last is unknown, but has mean 0. Nature chooses $\{X_i : i = 1, \ldots, n\}$ at random from some n-dimensional distribution. Given the X's, the Y's are assumed to be conditionally independent, and the random errors

$$Y_i - a - bX_i$$

are assumed have common distribution $\mathcal{D}$. In other words, the Y's are built up from the X's as follows: Nature computes the linear function $a + bX_i$, then adds some noise drawn at random from $\mathcal{D}$ to get Y_i. We get to see the pairs (X_i, Y_i) for $i = 1, \ldots, n$.

In this alternative formulation, there is a fixed error distribution $\mathcal{D}$ but there are no context-free random errors: errors may be functions of treatment levels among other things. The alternative has both a causal and an associational interpretation. (i) Assuming invariance of error distributions to interventions leads to the causal interpretation. (ii) Mere insensitivity to x when we condition on $X_i = x$ gives the associational interpetation—the probability distribution of $Y_i - a - bX_i$ given $X_i = x$ is the same for all x. This can at least in principle be tested against the data; invariance to interventions cannot, unless interventions were part of the design.

The key difference between equation (1) and the alternative is this. In (1), the errors themselves are invariant: in the alternative, only the error distribution is invariant. In (1), inference is to the *numerical value* that Y_i would have had, if X_i had been set to x. In the alternative formulation, causal inference can only be to the *probability distribution* that Y_i would have had. With either setup, the inference is about specific individuals, indexed by i; inference at the level of individuals is possible because—by assumption—parameters a, b are the same for all individuals. The two formulations of invariance, with the restrictions on the X's, express different ideas of exogeneity. The second set of assumptions is weaker than the first, and seems generally more plausible.

An example to consider is Hooke's law. The stretch of a spring is proportional to the load: a is length under no load and b is stretchiness. The disturbance term would represent measurement error. We could run an experiment to determine a and b. Or, we could passively observe the behavior of springs and weights. If heavier weights are attracted to bigger errors, there are problems. Otherwise, passive observation might give the right answer. Moreover, we can with more or less power test the hypothesis that the random errors $Y_i - a - bX_i$ are independent and identically distributed. By contrast, consider the hypothesis that $Y_i - a - bX_i$ itself would have been the same if X_i had been 7 rather 3. Even in an experiment, testing that seems distinctly unpromising.

What happens without invariance? The answer will be obvious. If intervention changes the intercept a, the slope b, or the mean of the error distribution, the impact of the intervention becomes difficult to determine. If the variance of the error term is changed, the usual confidence intervals lose their meaning. How would any of this be possible? Suppose, for instance, that—unbeknownst to the statistician—X and Y are both the effects of a common cause operating through linear statistical laws like (1); errors are independent and normal: and Nature has randomized the common cause to have a normal distribution. The scatter diagram will look lovely, a regression line is easily fitted, and the straightforward causal interpretation will be wrong.

3. Conditionals

Let us assume (informally) that the regression in Figure 1 is causal. What the Y_i's would have been if we had intervened and set X_i to x_i—this too isn't quite mathematics, but does correspond to either of two formal objects. One object is generated by equation (1): the random variables $Y_i = a + bx_i + \delta_i$ for $i = 1, \ldots, n$. The second object is this: n independent Y's, the ith being distributed as $a + bx_i$ plus a random draw from the error distribution $\mathcal{D}$. One object is defined in terms of random variables; the other, in terms of conditional distributions. There is a similar choice for the examples presented below.

So far, I have been discussing linear statistical laws. In Figure 1, for example, if we set $X = x$, then the conditional distribution of Y is $a + bx$, plus some random noise with distribution $\mathcal{D}$. Call this conditional distribution $K_x(dy)$. On the one hand, K_x may just represent the conditional distribution of Y given $X = x$, a rather dry statistical idea. On the other hand, K_x may represent the result of a hypothetical intervention: the distribution that Y would have had, if only we had intervened and set X to x. This is the more exciting causal interpretation. Data analysis on X and Y cannot decide whether the causal interpretation is viable. Instead, to make causal inferences from a system of regression equations, causation is assumed from the beginning. As Cartwright (1989) says, "No causes in, no causes out." This view contrasts rather sharply with rhetoric that one finds elsewhere.

Of course, solid arguments for causation have been made from observational data, but fitting regressions is only one aspect of the activity (Freedman, 1999). Replication seems to be critical, with good study designs and many different kinds of evidence. Also see Freedman (1997, pp.120–21), noting the difference between conditional probabilities that arise from selection of subjects with $X = x$, and conditional probabilities arising from an intervention that sets X to x. The data structures may look the same, but the implications can be worlds apart.

4. A second example: two linear regressions

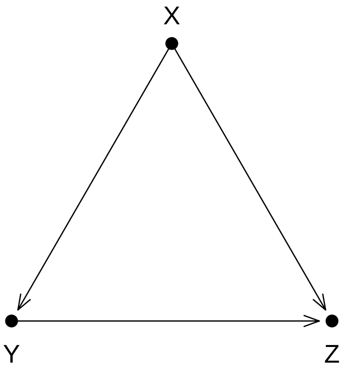

Figure 2. A linear path model.

The discussion can now be extended to path diagrams, with similar conclusions. Figure 2 involves three variables, and is a cameo version of applied statistics. If we are interested in the effect of Y on Z, then X confounds the relationship. Some adjustment is needed to avoid biased estimates, and regression is often used. The diagram unpacks into two response schedules:

(2a) $$Y_{i,x} = a + bx + \delta_i$$

$$Z_{i,x,y} = c + dx + ey + \epsilon_i. \tag{2b}$$

We assume that $\delta_1, \ldots, \delta_n,\ \epsilon_1, \ldots, \epsilon_n$ are all independent. The δ's have a common distribution $\mathcal{D}$. The ϵ's have another common distribution $\mathcal{F}$. These two distributions are unknown, but are assumed to have mean 0. Again, nothing in (2) is observable.

To generate the data, Nature chooses $\{X_i : i = 1, \ldots, n\}$ independently of $\{\delta_i, \epsilon_i : i = 1, \ldots, n\}$. We observe

$$(X_i,\ Y_i,\ Z_i)$$

for $i = 1, \ldots, n$, where

$$Y_i = Y_{i,X_i} = a + bX_i + \delta_i$$
$$Z_i = Z_{i,X_i,Y_i} = c + dX_i + eY_i + \epsilon_i.$$

Basically, this is a recursive system with two equations. The X's are "exogenous," that is, independent of the δ's and ϵ's. According to the model, Nature plugs the X's into (2a) to compute the Y's. In turn, those very X's and Y's get plugged into (2b) to generate the Z's. That is the recursive step. In other words, Y_i is computed as a linear function of X_i, with intercept a and slope b, plus the error term δ_i. Then Z_i is computed as a linear function of X_i and Y_i. The intercept is c, the coefficient on X_i is d, the coefficient on Y_i is e; at the end, the error ϵ_i is tagged on. Again, the δ's and ϵ's remain the same no matter what x's and y's go into (2); so do the parameters a, b, c, d, e. (Interactions across subjects are precluded because, for instance, subject i's response Y_i is computed from X_i and δ_i rather than X_j and δ_j.)

The proposed alternative involves not random errors but their distributions $\mathcal{D}$ and $\mathcal{F}$. These distributions are unknown but have mean 0. We still have the parameters a, b, c, d, e. To generate the data, we assume that Nature chooses $X_1, \ldots, X_n$ at random from some n-dimensional distribution. Given the X's, the Y's are assumed to be conditionally independent: Y_i is generated by computing $a + bX_i$, then adding some independent noise distributed according to $\mathcal{D}$. Given the X's and Y's, the Z's are assumed to be conditionally independent: Z_i is generated as $c + dX_i + eY_i$, with independent additive noise distributed according to $\mathcal{F}$. The exogeneity assumption is the independence between the X's and the errors.

As before, the second setup assumes less invariance than the first: it is error distributions that are invariant, not error terms; the inference is to distributions rather than specific numerical values. Either way, there are unbiased estimates for the parameters a, b, c, d, e; the error distributions $\mathcal{D}$ and $\mathcal{F}$ are identifiable: parameters and error distributions are constant in both formulations. As before, the second setup may be used to describe conditional distributions of random variables. If those conditional distributions admit a causal interpretation, then causal inferences can made from observational data. In other words, regression succeeds in determining the effect of Y on Z if we know (i) X is the confounder and (ii) the statistical relationships are linear and causal.

What can go wrong? Omitted variables are a problem, as discussed before. Assuming the wrong causal order is another issue. For example, suppose equation (2) is correct; the errors are independent and normally distributed; moreover, the exogenous variable X has been randomized to have a normal distribution. However, the unfortunate statistician regresses (i) Y on Z, then (ii) X on Y and Z. Diagnostics will indicate success: the distribution of residuals will not depend on the explanatory variables. But causal inferences will be all wrong. The list of problem areas can easily be extended beyond omitted variables and causal orderings to include functional form, stochastic specification, measurement. . . .

The issue boils down to this. Does the conditional distribution of Y given X represent mere association, or does it represent the distribution Y would have had if we had intervened and set the values of X? There is similar question for the distribution of Z given X and Y. These questions cannot be answered just by fitting the equations and doing data analysis on X, Y, and Z; additional information is needed. From this perspective, the equations are "structural" if the conditional distributions inferred from the equations tell us the likely impact of interventions, thereby allowing a causal rather than an associational interpretation. The take-home message will be clear: you cannot infer a causal relationship from a data set by running regressions—unless there is substantial prior knowledge about the mechanisms that generated the data.

5. Simultaneous equations

Similar considerations apply to models with simultaneous equations. The invariance assumptions will be familiar to many readers. Changing pace, I will discuss hypothetical supply and demand equations for butter in the state of Wisconsin. The endogenous variables are Q and P, the quantity and price of butter. The exogenous variables in the supply equation are the agricultural wage rate W and the price H of hay. The exogenous variables in the demand equation are the prices M of margarine and B of bread (substitutes and complements). For the moment, "exogeneity" just means "externally determined." Annual data for the previous twenty years are available on the exogeneous variables, and on the quantity of Wisconsin butter sold each year as well as its price. Linearity is assumed, with the usual stochastics.

The model can be set up formally with two linear equations in two unknowns, Q and P:

(3a) Supply $$Q = a_0 + a_1 P + a_2 W + a_3 H + \delta_t,$$

(3b) Demand $$Q = b_0 + b_1 P + b_2 M + b_3 B + \epsilon_t.$$

On the right hand side, there are parameters (the a's and b's). There are also error terms (δ_t, ϵ_t) which are assumed to be independent and identically distributed for $t = 1, \ldots, 20$. The common two-dimensional "error distribution" $\mathcal{C}$ for (δ_t, ϵ_t) is unknown, but is assumed to have mean 0.

Each equation describes a thought experiment. In the first, we set P, W, H, M, B and observe how much butter comes to market: by assumption, M and B have no effect on supply, while P, W, H have additive linear effects. In the second we set P, W, H, M, B and observe how much butter is sold: W and H have no effect on demand, while P, M, B have additive linear effects. In short, we have linear supply and demand schedules. Again, the error terms themselves are invariant to all interventions, as are the parameters. Since this is a hypothetical, there is no need to worry about the EEC, NAFTA, or the economics.

A third gedanken experiment is described by taking equations (3a) and (3b) together. Any values of the exogenous variables W, H, M, B —perhaps within certain ranges—can be substituted in on the right, and the two equations solved together for the two unknowns Q and P, giving us the transacted quantity and price in a free market, denoted

(4) $$Q_{W,H,M,B} \quad \text{and} \quad P_{W,H,M,B}.$$

Since δ and ϵ turn up in the formulas for both Q and P, the random variables in (4) are correlated—barring some rare parameter combinations—with the error terms. The correlation is "simultaneity."

So far, we have three thought experiments expressing various assumptions, but no data: nothing so far is observable. We assume that Nature generates data for us by choosing W_t, H_t, M_t, B_t for $t = 1, \ldots, 20$, at random from some high-dimensional distribution, independently of the δ's and ϵ's. This independence is the exogeneity assumption, which gives the concept a more technical shape. For each t, we get to see the values of the exogenous variables

$$W_t,\ H_t,\ M_t,\ B_t,$$

and the corresponding endogenous variables computed by solving (3ab) together, namely,

$$Q_t = Q_{W_t,H_t,M_t,B_t} \quad \text{and} \quad P_t = P_{W_t,H_t,M_t,B_t}.$$

Of course, we do not get to see the parameters or the disturbance terms. A regression of Q_t on P_t and the exogenous variables leads to “simultaneity bias,” because P_t is correlated with the error term; hence two-stage least squares and related techniques. With such estimators, enough data, and the assumptions detailed above, we can (almost) recover the supply and demand schedules (3ab) from the free market data—using the exogenous variables supplied by Nature.

The other approach, sketched above for Figures 2 and 3, suggests that we start from the parameters and the error distribution $\mathcal{C}$. If we were to set P, W, H, M, B, then Nature would be assumed to choose the errors in (3) from $\mathcal{C}$: farmers would respond according to the supply equation (3a), and consumers according to the demand equation (3b). If we were to set only W, H, M, B and allow the free market to operate, then quantity and price would in this parable be computed by solving the pair of equations (3ab).

The notation for the error terms in (3) is a bit simplistic now, since these terms may be functions of W, H, M, B. Allowing the errors to be functions of P may make sense if (3a) and (3b) are considered in isolation; but if the two equations are considered together, this extra generality would lead to a morass. To generate data, we assume that Nature chooses the exogenous variables at random from some multidimensional distribution. The market quantities and prices are still computed by solving the pair of equations (3ab) for Q and P, with independent additive errors for each period drawn from $\mathcal{C}$; the usual statistical computations can still be carried out.

In this setup, it is not the error terms that are invariant, but their distribution. Of course, parameters are taken to be invariant. The exogeneity assumption is the independence of $\{W_t, H_t, M_t, B_t : t = 1, 2 \ldots\}$ and the error terms. The inference is for instance to the probability distribution of butter supply, if we were to intervene in the market by setting price as well as the exogenous variables. By contrast, with assumed invariance for the error terms themselves, the inference is to the numerical quantity of butter that would be supplied.

I have presented the second approach with a causal interpretation; an associational interpretation is also possible, although less interesting. The exposition may seem heavy-handed, because I have tried to underline the critical invariance assumptions that need to be made in order to draw causal conclusions from nonexperimental data: parameters are invariant to interventions, and so are errors or their distributions. Exogeneity is another concern. In a real example, as opposed to a butter hypothetical, real questions would have to be asked about these assumptions. Why are the equations “structural,” in the sense that the required invariance assumptions hold true?

Obviously, there is some tension here. We want to use regression to draw causal inferences from nonexperimental data. To do that, we need to know that certain parameters and certain distributions would remain invariant if we were to intervene. That invariance can seldom if ever be demonstrated by intervention. What then is the source of the knowledge? “Economic theory” seems like a natural answer, but an incomplete one. Theory has to be anchored in reality. Sooner or later, invariance needs empirical demonstration, which is easier said than done.

6. Nonlinear models: Figure 1 revisited

Graphical models can be set up with nonlinear versions of equation (1), as in Pearl (1995, 2000). The specification would be something like $Y_{i,x} = f(x, \delta_i)$, where f is a fairly general (unknown) function. The interpretation is this: if the treatment level were set to x, the response by subject i would be $Y_{i,x}$. The same questions about interventions and counterfactual hypotheticals would then have to be considered. Instead of rehashing such isues, I will indicate how to formulate the models using conditional distributions (“Markov kernels”), so that the graphs can be interpreted either distributionally or causally. In the nonlinear case, K_x—the conditional distribution of Y given that $X = x$—depends on x in some fashion more complicated than linearity with additive noise. For example, if X, Y are discrete, then K can be visualized as the matrix of conditional probabilities $P(Y = y|X = x)$. For any particular x, K_x is a row in this matrix.

Inferences will be to conditional distributions, rather than specific numerical values. There will be some interesting new questions about identifiability. And the plausibility of causal interpretations can be assessed separately, as will be shown later. I will organize most of the discussion around two examples used by Pearl (1995); also see Pearl (2000, pp.66–68 and 83–85). But first, consider Figure 1. In the nonlinear case, the exogenous variables have to be assumed independent and identically distributed in order to make sense out of the mathematics; otherwise, there are substantial extra complications, or we have to impose additional smoothness conditions on the kernel.

Assume now that (X_i, Y_i) are independent and distributed like (X, Y) for $i = 1, \ldots, n$; the conditional distribution of Y_i given $X_i = x$ is K_x, where K is an unknown Markov kernel. With a large-enough sample, the joint distribution of (X, Y) can be estimated reasonably well; so can K_x, at least for x’s that are likely to turn up in the data. If K is only a conditional probability, that is what we obtain from data analysis. If K admits a causal interpretation—by prior knowledge or assumption, not by data analysis on the X’s and Y’s—then we can make a causal inference: What would the distribution of Y_i have been, if we had intervened and set X_i to x? (Answer: K_x.)

7. Technical notes

The conditional distribution of Y given X tells you the conditional probability that Y is in one set C or another, given that $X = x$. A Markov kernel K assigns a number $K_x(C)$ to pairs (x, C); the first element x of the pair is a point; the second, C, is a set. With x fixed, K_x is a probability. With C fixed, the function that sends x to $K_x(C)$ should satisfy some minimal regularity condition. Below, I will write $K_x(dy)$ as shorthand for the kernel whose value at (x, C) is $K_x(C)$, where C is any reasonable set of values for Y. Matters will be arranged so that $K_x(C)$ is the conditional probability that $Y \in C$ given $X = x$, and perhaps some other information: $K_x(C) = P(Y \in C|X = x \ldots)$.

Without further restrictions, graphical models are nonparametric, because kernels are infinite-dimensional “parameters.” Our ability to estimate such things depends on the degree of regularity that is assumed. With minimal assumptions, you may get minimal performance—but that is a topic for another day. Even in the linear case, some of the fine points about estimation have been glossed over. To estimate the model in Figure 1, we would need some variation in X and δ. To get standard errors, we would assume finite variances for the error terms. Conditions for identifiability in the simultaneous-equations setup do not need to be rehearsed here, and I have assumed a unique solution for (3). Two-stage least squares will have surprising behavior unless variances are assumed for the errors; some degree of correlation between the exogenous and endogenous variables would also be needed.

More general specifications can be assumed for the errors. For example, in (1), the δ_i may be assumed to be independent, with common variances and uniformly bounded fourth moments; then the hypothesis of a common distribution can be dropped. In (3), an ARIMA model may be assumed. And so forth. The big picture does not change, because (i) questions about invariance remain, and (ii) even an ARIMA model requires some justification.

8. More complicated examples

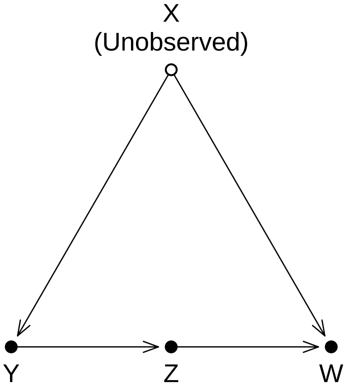

Figure 3. A graphical model.

The story behind Figure 3 will be explained below. For the moment, it is an abstract piece of mathematical art. The diagram corresponds to three kernels: $K_x(dy)$, $L_y(dz)$, and $M_{x,z}(dw)$. These kernels describe the joint distribution of the random variables shown in the diagram (X, Y, Z, W). The conditional distribution of Y given $X = x$ is K_x. The conditional distribution of Z given $X = x$ and $Y = y$ is L_y: there is no subscript x on L because—by assumption—there is no arrow from X to Z in the diagram. The conditional distribution of W given $X = x$, $Y = y$, $Z = z$ is $M_{x,z}$: there is no subscript y on M because—again by assumption—there is no arrow leading directly from Y to W in the diagram.

You can think of building up the variables X, Y, Z, W from the kernels and a base distribution μ for X, in a series of steps:

(i) Chose X at random according to $\mu(dx)$.

(ii) Given the value of X from step (i), say $X = x$, choose Y at random from $K_x(dy)$.

(iii) Given $X = x$ and $Y = y$, choose Z at random from $L_y(dz)$.

(iv) Given $X = x$, $Y = y$, and $Z = z$, choose W at random from $M_{x,z}(dw)$.

The recipe is equivalent to the graph.

By assumption, the 4-tuples (X_i, Y_i, Z_i, W_i) are independent and distributed like (X, Y, Z, W) for $i = 1, \ldots, n$. There is one more wrinkle: the circle marked "X" in the diagram is open, meaning that X is not observed. In other words, Nature hides $X_1, \ldots, X_n$ but shows us

$$Y_1, \ldots, Y_n, \; Z_1, \ldots, Z_n, \; W_1, \ldots, W_n.$$

That is our data-set.

The base distribution μ and the kernels K, L, M are unknown. However, with many observations on independent and identically distributed triplets (Y_i, Z_i, W_i), we can estimate their joint distribution reasonably well. Moreover—and this should be a little surprising—we can compute L_y from that joint distribution, as well as

$$\text{(5a)} \qquad \mathcal{M}_z(dw) = \int M_{x,z}(dw)\, \mu(dx),$$

where μ is the distribution of the unobserved confounder X. Hence we can also compute

$$\text{(5b)} \qquad \mathcal{L}_y(dw) = \int \mathcal{M}_z(dw)\, L_y(dz).$$

Here is the idea: L is computable because the relationship between Y and Z is not confounded by X. Conditional on Y, the relationship between Z and W is not confounded, so $\mathscr{M}_z$ in (5a) is computable. Then (5b) follows.

More specifically, with "P" for probability, the identity

$$P(Z \in C | Y = y) = P(Z \in C | X = x, Y = y) = L_y(C)$$

can be used to recover L from the joint distribution of Y, Z. Likewise, we can recover $\mathscr{M}$ in (5a) from the joint distribution of Y, Z, W, although the calculation is a little more intricate. Let $P_{x,y,z} = P(\bullet | X = x, Y = y, Z = z)$ be a regular conditional probability given X, Y, Z. Then

$$\begin{aligned} P(W \in D | Y = y, Z = z) &= \int P_{x,y,z}(W \in D)\, P(X \in dx | Y = y, Z = z) \\ &= \int M_{x,z}(D)\, P(X \in dx | Y = y), \end{aligned}$$

because

$$P_{x,y,z}(W \in D) = M_{x,z}(D)$$

by construction, and X is independent of Z given Y by a side-calculation. We have recovered $\int M_{x,z}(D)\, P(X \in dx | Y = y)$ from the joint distribution of Y, Z, W. Hence we can recover

$$\begin{aligned} \int\!\!\int M_{x,z}(D)\, P(X \in dx | Y = y) P(Y \in dy) &= \int M_{x,z}(D)\, \mu(dx) \\ &= \mathcal{M}_z(D), \end{aligned}$$

although the distribution μ of X remains unknown, and so does the kernel M.

These may all just be facts about conditional distributions, in which case (5) is little more than a curiosity. On the other hand, if K, L, M have causal interpretations, then $\mathcal{M}_z$ in (5a) tells you the effect of setting $Z = z$ on W, averaged over the possible X's in the population. Similarly, $\mathcal{L}_y$ in (5b) tells you the effect of Y on W: if you intervene and set Y to y, then the distribution of W will be $\mathcal{L}_y$, on the average over all X and Z in the population. (There may be exceptional null sets, which are being ignored.) How to estimate $\mathcal{M}$ and $\mathcal{L}$ in a finite sample is another question, not discussed here.

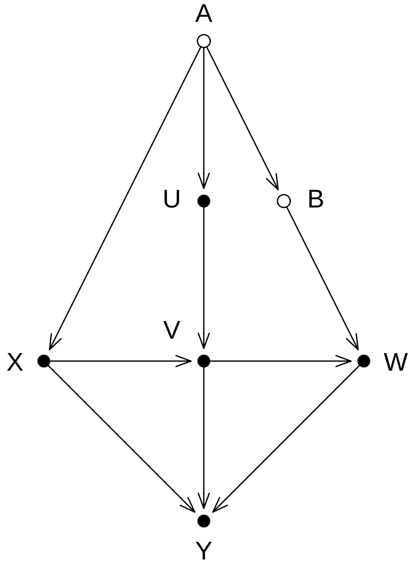

Figure 4. A graphical model: seven variables, of which five are observed.

The next example (Figure 4) is a little more complicated; again, the story behind the figure is deferred. There are two unobserved variables, A and B. The setup involves six kernels, which characterize the joint distribution of the random variables (A, B, U, X, V, W, Y) in the diagram:

$$K_a(db) = P(B \in db|A = a),$$

$$L_a(du) = P(U \in du|A = a),$$

$$M_a(dx) = P(X \in dx|A = a),$$

$$N_{u,x}(dv) = P(V \in dv|A = a, B = b, U = u, X = x),$$

$$Q_{b,v}(dw) = P(W \in dw|A = a, B = b, U = u, X = x, V = v),$$

$$R_{x,v,w}(dy) = P(Y \in dy|A = a, B = b, U = u, X = x, V = v, W = w).$$

Here, P represents "probability"; it seemed more tasteful not to have kernels labeled O or P. There is no a, b, u among the subscripts on R because there are no arrows going directly from A, B, U to Y in the diagram; similarly for the other kernels. The issue is to determine the effect of X on Y, integrating over the unobserved confounders A, B. This is feasible, because conditional on the observed U, V, W, the relationship between X and Y is not confounded. (If the kernels have causal interpretations, "effect" is meant literally; if not, figuratively.)

To fix ideas, we can go through the construction of the random variables. There is a base probability μ for A. First, choose A at random from μ. Given A, choose B, U, X independently at random from K_A, L_A, M_A, respectively. Given A, B, U, X, choose V at random from $N_{U,X}$. Given A, B, U, X, V, choose W at random from $Q_{B,V}$. Finally, given A, B, U, X, V, W, choose Y at random from $R_{X,V,W}$. The data-set consists of n independent septuples A_i, B_i, U_i, X_i, V_i, W_i, Y_i, distributed as A, B, U, X, V, W, Y—except that the A's and B's are hidden. The "parameters" are μ and the six kernels. Calculations proceed as for Figure 3. Again, the graph and the description in terms of kernels are equivalent. Details are (mercifully?) omitted.

9. Parametric nonlinear models

Similar considerations apply to parametric nonlinear models. Take the logit specification, for example. Let X_i be a p-dimensional random vector, with typical value x_i; the random variable Y_i is 0 or 1. Let β be a p-dimensional vector of parameters. For the p-dimensional data vector x, let K_x assign mass

$$e^{\beta x}\big/\big(1+e^{\beta x}\big)$$

to 1, and the remaining mass to 0. Given $X_1, \ldots, X_n$, each being a p-vector, suppose the Y_i are conditionally independent, and

$$(6) \qquad P(Y_i = 1|X_1 = x_1, \ldots, X_n = x_n) = K_{x_i}.$$

On the right hand side of (6), the subscript on K is x_i: the conditional distribution of Y for a subject depends only on that subject's x. If the $x_1, \ldots, x_n$ are reasonably spread out, we can estimate β by maximum likelihood. (With a smooth, finite-dimensional parametrization, we do not need the X_i to be independent and identically distributed.)

Of course, this model could be set up in a more strongly invariant form, like (1). Let U_i be independent (unobservable) random variables with a common logistic distribution: $P(U_i < u) = e^u/(1+e^u)$. Then

$$(7) \qquad Y_{i,x} = 1 \iff U_i < \beta x.$$

The exogeneity assumption would make the X's independent of the U's, and the observable Y_i would be Y_{i,X_i}. That is, $Y_i = 1$ if $U_i < \beta X_i$, else $Y_i = 0$.

This is all familiar territory, except perhaps for (7); so familiar that the critical question may get lost. Does K_x merely represent the conditional probability that $P(Y_i = 1|X_i = x)$, as in (6)? Or does K_x tell us what the law of Y_i would have been, if we had intervened and set X_i to x? Where would the U_i come from, and why would they be invariant if we manipulated x? Nothing in the mysteries of Euclidean geometry and likelihood statistics can possibly answer this sort of question: other kinds of information are needed.

10. Concomitants

Some variables are potentially manipulable; others ("concomitants") are not. For example, education and income may be manipulable; age, sex, race, personality, ..., are concomitants. So far, we have ignored this distinction, which is less problematic for kernels, but a difficulty for the kind of strong invariance in equation (1). However, if Y depends on a manipulable X and a concomitant W through a linear causal law with additive error, we can rewrite (1) as

$$(8) \qquad Y_{i,x} = a + bx + cW_i + \delta_i.$$

In addition to the usual assumptions on the δ's, we would have to assume independence between the δ's and the W's. In applications, defining and isolating the intervention may not be so easy, but that is a topic for another day. Also see Robins (1986, 1987).

11. The story behind figures 3 and 4

When some variables are unobserved, Pearl (1995) develops an interesting calculus to define confounding and decide which kernels or composites—see (5) for example—can be recovered from the joint distribution of the observed variables. That is a solution to the identification problem for such diagrams. He uses Figure 3 to illustrate his “back-door criterion.” The unobserved variable X is genotype; the observed variables Y, Z, W represent smoking, tar deposits in the lung, and lung cancer, respectively (Figure 5). The objective is to determine the effect of smoking on lung cancer, via (5).

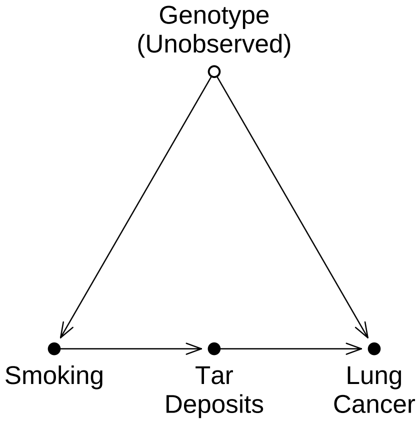

Figure 5. A graphical model for smoking and lung cancer.

Data in this example would consist of a long series of independent triplets (Y_i, Z_i, W_i), each distributed like (Y, Z, W). Pearl interprets the graph causally. The timeworn idea that subjects in a study form a random sample from some hypothetical super-population still deserves a moment of respectful silence. Moreover, there are three special assumptions in Figure 5:

(i) Genotype has no direct effect on tar deposits.

(ii) Smoking has no direct effect on lung cancer.

(iii) Tar deposits can be measured with reasonable accuracy.

There is no support for these ideas in the literature. (i) The lung has a mechanism—“the mucociliary escalator”—for eliminating foreign matter, including tar. This mechanism seems to be under genetic control. (Of course, clearance mechanisms can be overwhelmed by smoking.) The forbidden arrow from genotype to tar deposits may have a more solid empirical basis than the permitted arrows from genotype to smoking and lung cancer. Assumption (ii) is just that—an assumption. And (iii) is clearly wrong. The consequences are severe: if arrows are permitted from genotype to tar deposits or from smoking to lung cancer, or if measurements of tar are subject to error, then formula (5) does not apply. Graphical models cannot solve the problem created by an unmeasured confounder without introducing strong and artificial assumptions.

The intellectual history is worth mentioning. Fisher's "constitutional hypothesis" explained the association between smoking and disease on the basis of a gene that caused both. This idea is refuted not by making assumptions but by doing some empirical work. For example, Kaprio and Koskenvuo (1989) present data from their twin study. The idea is to find pairs of identical twins where one smokes and one does not. That sets up a race: who will die first, the smoker or the non-smoker? The smokers win hands down, for total mortality or death from heart disease. The genetic hypothesis is incompatible with these data.

For lung cancer, the smokers win two out of the two races that have been run. (Why only two? Smoking-discordant twin pairs are unusual, lung cancer is a rare disease, and the population of Scandinavia is small.) Carmelli and Page (1996) have a similar analysis with a larger cohort of twins. Do not bet on Fisher. International Agency for Research on Cancer (1986) reviews the health effects of smoking and indicates the difficulties in measuring tar deposits (pp.179–98). Nakachi et al. (1993) and Shields et al. (1993) illustrate conflicts on the genetics of smoking and lung cancer. Also see Miller et al. (2003). The lesson: finding the mathematical consequences of assumptions matters, but connecting assumptions to reality matters even more.

Pearl uses Figure 4 to illustrate his "front-door criterion," calling the figure a "classical example due to Cochran," with a cite to Wainer (1989). Pearl's vision is that soil fumigants X are used to kill eelworms and improve crop yields Y for oats. The decision to apply fumigants is affected by the worm population A before the study begins, hence the arrow from A to X. The worm population is measured at baseline, after fumigation, and later in the season: the three measurements are U, V, W. The unobserved B represents "birds and other predators."

This vision is whimsical. The example originates with Cochran (1957, p.266) who had several fumigants applied under experimental control, with measurements of worm cysts and crop yield. Pearl converts this to an observational study with birds, bees, and so forth—entertaining, a teaching tool, but unreal. It might be rude to ask too many questions about Figure 4, but surely crops attract predators. Don't birds eat oat seeds? If early birds get the worms, what stops them from eating worms at baseline? In short, where have all the arrows gone?

12. Models and kernels revisited

Graphical models may lead to some interesting mathematical developments. The number of successful applications, however, is at best quite limited. Figures 4 and 5 are not atypical (there are citations to the literature, below). And it is all too tempting to forget the limitations of such methods. Given that the arrows and kernels represent causation, while variables are independent and identically distributed, we can use Pearl's framework to determine from the diagram which effects are estimable. This is a step forward. However, we cannot use the framework to answer the more basic question: Does the diagram represent the causal structure? As everyone knows, there are no formal algorithmic procedures for inferring causation from association; everyone is right.

Pearl (1995) considers only models with a causal interpretation, the latter being partly formalized; and there is new terminology that some readers may find discouraging. On the other hand, he draws a clear distinction between averaging Y's when the corresponding X is

- set to x, and
- observed to be x in the data.

That is a great advantage of his formalism.

The approach sketched here would divide the identification problem in two: (i) reconstructing kernels—viewed as ordinary conditional distributions—from partial information about joint distributions; and (ii) deciding whether these kernels bear a causal interpretation. Problem (i) can be handled entirely within the conventional probability calculus. Problem (ii) is one of the basic problems in applied statistics. Of course, kernels—especially mixtures like (5)—may not be interesting without a causal interpretation.

In sum, graphical models can be formulated using conditional distributions ("Markov kernels"), without invariance assumptions. Thus, the graphs can be interpreted either distributionally or causally. The theory governing recovery of kernels and their mixtures can be pushed through with just the distributional interpretation. That frees us to consider whether or not the kernels admit a causal interpretation. So far, however, the graphical modelers have few if any examples where the causal interpretation can be defended. Pearl generally agrees with this discussion:

> Causal analysis with graphical models does not deal with defending modeling assumptions, in much the same way that differential calculus does not deal with defending the physical validity of a differential equation that a physicist chooses to use. In fact no analysis void of experimental data can possibly defend modeling assumptions. Instead, causal analysis deals with the conclusions that logically follow from the combination of data and a given set of assumptions, just in case one is prepared to accept the latter. Thus, all causal inferences are necessarily *conditional*. These limitations are not unique to graphical models. In complex fields like the social sciences and epidemiology, there are only few (if any) real life situations where we can make enough compelling assumptions that would lead to identification of causal effects [Pearl, private communication].

13. Literature review

The model in (1) was proposed by Neyman (1923). It has been rediscovered many times since; see, for instance, Hodges and Lehmann (1964, section 9.4). The setup is often called "Rubin's model," but this simply mistakes the history: see Dabrowska and Speed (1990), with a comment by Rubin; also see Rubin (1974) and Holland (1986). Holland (1986, 1988) explains the setup with a super-population model to account for the randomness, rather than individualized error terms. These error terms are often described as the overall effects of factors omitted from the equation. But this description introduces difficulties of its own, as shown by Pratt and Schlaifer (1984, 1988). Stone (1993) presents a clear super-population model with some observed covariates and some unobserved.

Dawid (2000) objects to counterfactual inference. Counterfactual distributions may be essential to any account of causal inference by regression methods. On the other hand, as the present paper tries to show, invariant counterfactual random variables—like δ_i in equation (1)—are dispensable. In particular, with kernels, there is no need to specify the joint distribution of random variables across inconsistent hypotheticals.

There is by now an extended critical literature on linear statistical models for causation, starting perhaps with the exchange between Keynes (1939, 1940) and Tinbergen (1940). Other familiar citations in the economics literature include Liu (1960), Lucas (1976), Leamer (1978), Sims (1980), Hendry (1993), Manski (1993), Angrist, Imbens, and Rubin (1996). Heckman (2000) traces the development of econometric thought from Haavelmo and Frisch onwards, stressing the role of "structural" or "invariant" parameters, and "potential outcomes"; also see Heckman (2001ab).

According to Heckman (2000), the enduring contributions of the field are the following insights:

> . . . . that causality is a property of a model, that many models may explain the same data and that assumptions must be made to identify causal or structural models. . . . recognizing the possibility of interrelationships among causes . . . . [clarifying] the conditional nature of causal knowledge and the impossibility of a purely empirical approach to analyzing causal questions. . . . The information in any body of data is usually too weak to eliminate competing causal explanations of the same phenomenon. There is no mechanical algorithm for producing a set of "assumption free" facts or causal estimates based on those facts. [pp. 89–91]

For another discussion of causal models from an econometric perspective, see Angrist (2001). Angrist and Krueger (2001) provide a nice introduction to instrumental variables; an early application of the technique was to fit supply and demand curves for butter (Wright, 1928, p.316). Engle, Hendry, and Richard (1983) distinguish several kinds of exogeneity, with different implications for causal inference.

One of the drivers for modeling in economics and cognate fields is rational choice theory. Therefore, any discussion of empirical foundations must take into account a remarkable series of papers, initiated by Kahneman and Tversky (1974), that explores the limits of rational choice theory. These papers are collected in Kahneman, Slovic, and Tversky (1982), and in Kahneman and Tversky (2000). The heuristics and biases program has attracted its own critics (Gigerenzer, 1996). That critique is interesting and has some merit; but in the end, the experimental evidence demonstrates severe limits to the descriptive power of choice theory (Kahneman and Tversky, 1996). If people are trying to maximize expected utility, they don't do it very well. Errors are large and repetitive, go in predictable directions, and fall into recognizable categories: these are biases, not random errors. Rather than making decisions by optimization—or bounded rationality, or satisficing—people seem to use plausible heuristics that can be identified. If so, rational choice theory is generally not a good basis for justifying empirical models of behavior.

Recently, modeling issues have been much canvassed in sociology. Berk (2003) is skeptical about the possibility of inferring causation by modeling, absent a strong theoretical base. Abbott (1997) finds that variables (like income and education) are too abstract to have much explanatory power. Clogg and Haritou (1997) review various difficulties with regression, noting in particular that you can all too easily include endogenous variables as regressors. Hedström and Swedberg (1998) edited a lively collection of essays by a number of sociologists, who turn out to be quite skeptical about regression models; rational choice theory also takes its share of criticism. Goldthorpe (1998, 2001) describes several ideas of causation and corresponding methods of statistical proof, with different strengths and weaknesses. Ní Bhrolcháin (2001) has some particularly forceful examples to illustrate the limits of regression. There is an influential book by Lieberson (1985), with a followup by Lieberson and Lynn (2002); the latest in a series of papers is Sobel (2000). Meehl (1978) reports the views of an empirical psychologist; also see Meehl (1954), with data showing the advantage of using regression to make predictions—rather than experts. Meehl and Waller (2002) discuss the choice between two similar path models, viewed as reasonable approximations to some underlying causal structure, but do not reach the critical question—how to assess the adequacy of the approximation. Steiger (2001) has a critical review. There are well-known books by Cook and Campbell (1979), Shadish, Cook, and Campbell (2002). In political science, Brady and Collier (2004) compare regression methods with case studies; invariance is discussed under the rubric of causal homogeneity. Cites from other perspectives include Freedman, Rothenberg, and Sutch (1985), Oakes (1986), as well as Freedman (1985, 1987, 1991, 1995, 1999).

There is an extended literature on graphical models for causation. Greenland, Pearl and Robins (1999) give a clear account in the context of epidemiology. Lauritzen (1996, 2001) has a careful treatment of the mathematics. These authors do not recognize the difficulties in applying the methods to real problems. Equation (5) is a special case of the "g-computation algorithm" due to Robins (1986, 1987); also see Gill and Robins (2001), Pearl (1995, 2000), or Spirtes, Glymour and Scheines (1993). Robins (1995) explains—all too briefly—how to state Pearl's results as theorems about conditionals. For critical reviews of graphical models (with responses and further citations) see Freedman (1997), Humphreys (1997), Humphreys and Freedman (1996, 1999): among other things, these papers discuss various applications proposed by the modelers. Woodward (1997, 1999) stresses the role of invariance. Freedman and Stark (1999) show that different models for the correlation of outcomes across counterfactual scenarios can have markedly different consequences in the legal context. Scharfstein, Rotnitzky, and Robins (1999) demonstrate a large range of uncertainty in estimates, due to incomplete specifications; also see Robins (1999).

Acknowledgments

Over the years, I learned a great dealt about statistics from Tom Rothenberg; it is a pleasure to acknowledge the debt. I would also like to thank some other friends for many helpful conversations the topics of this paper: Dick Berk, Paul Holland, Paul Humphreys, Máire Ní Bhrolcháin, Judea Pearl, Jamie Robins, and Philip Stark. At the risk of the obvious, thanking people does not imply they agree with my opinions; nor does this caveat imply disagreement.

References

Abbott, A. (1997): "Of Time and Space: The Contemporary Relevance of the Chicago School," *Social Forces*, 75, 1149–82.

Angrist, J. D. (2001): "Estimation of Limited Dependent Variable Models with Binary Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics*, 19, 2–16.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996): "Identification of Causal Effects using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–72.

Angrist, J. D. and Krueger, A. K. (2001): "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Business and Economic Statistics*, 19, 2–16.

Berk, R. A. (2003): *Regression Analysis: A Constructive Critique*. Newbury Park, CA: Sage Publications.

Brady, H. and Collier, D. (2003), eds.: *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman & Littlefield Publishers, Inc. To appear.

Carmelli, D. and Page, W. F. (1996): "24-year Mortality in Smoking-Discordant World War II U. S. Male Veteran Twins," *International Journal of Epidemiology*, 25, 554–59.

Cartwright, N. (1989): *Nature's Capacities and their Measurement*. Oxford: Clarendon Press.

Clogg, C. C. and Haritou, A. (1997): “The Regression Method of Causal Inference and a Dilemma Confronting this Method,” in *Causality in Crisis*, ed. by V. McKim and S. Turner. University of Notre Dame Press, pp. 83–112.

Cochran, W. G. (1957): “Analysis of Covariance: Its Nature and Uses,” *Biometrics*, 13, 261–81.

Cook T. D., Campbell D. T. (1979). *Quasi-experimentation: design & analysis issues for field settings.* Boston: Houghton Mifflin.

Dawid, A. P. (2000): “Causal Inference Without Counterfactuals,” *Journal of the American Statistical Association*, 95, 407–48.

Engle, R. F., Hendry, D. F., and Richard, J. F. (1983): “Exogeneity,” *Econometrica*, 51, 277–304.

Freedman, D. A. (1999). “From Association to Causation: Some Remarks on the History of Statistics,” *Statistical Science*, 14, 243–58.

Freedman, D. A. (1997): “From Association to Causation via Regression,” in *Causality in Crisis?* ed. by V. McKim and S. Turner. South Bend: University of Notre Dame Press, pp. 113–82 (with discussion).

Freedman, D. A. (1995): “Some Issues in the Foundation of Statistics,” *Foundations of Science*, 1, 19–83 (with discussion). Reprinted in *Some Issues in the Foundation of Statistics*, ed. by B. van Fraasen. Dordrecht: Kluwer, pp. 19–83 (with discussion).

Freedman, D. A. (1991): “Statistical Models and Shoe Leather,” in *Sociological Methodology 1991*, ed. by Peter Marsden. Washington, D.C.: American Sociological Association, Chapter 10 (with discussion).

Freedman, D. A. (1987): “As Others See Us: A Case Study in Path Analysis,” *Journal of Educational Statistics*, 12, 101–223 (with discussion). Reprinted in *The Role of Models in Nonexperimental Social Science*, ed. by J. Shaffer. Washington, D.C.: AERA/ASA, 1992, pp. 3–125.

Freedman, D. A. (1985): “Statistics and the Scientific Method,” in *Cohort Analysis in Social Research: Beyond the Identification Problem*, ed. by W. M. Mason and S. E. Fienberg. New York: Springer-Verlag, pp. 343–90 (with discussion).

Freedman, D., Rothenberg, T., and Sutch, R. (1983): “On Energy Policy Models,” *Journal of Business and Economic Statistics*, 1, 24–36 (with discussion).

Freedman, D. A. and Stark, P. B. (1999): “The Swine Flu Vaccine and Guillain-Barré Syndrome: A Case Study in Relative Risk and Specific Causation,” *Evaluation Review*, 23, 619–47.

Gigerenzer, G. (1996): “On Narrow Norms and Vague Heuristics,” *Psychological Review*, 103, 592–96.

Gill, R. D. and Robins, J. M. (2001): “Causal Inference for Complex Longitudinal Data: The Continuous Case,” *Annals of Statistics*, in press.

Goldthorpe, J. H. (2001): “Causation, Statistics, and Sociology,” *European Sociological Review*, 17, 1–20.

Goldthorpe, J. H. (2000): *On Sociology: Numbers, Narratives, and Integration of Research and Theory*. Oxford University Press.

Goldthorpe, J. H. (1998): *Causation, Statistics and Sociology*. Twenty-ninth Geary Lecture, Nuffield College, Oxford. Published by the Economic and Social Research Institute, Dublin, Ireland.

Greenland, S., Pearl, J., and Robins, J. (1999): "Causal Diagrams for Epidemiologic Research," *Epidemiology*, 10, 37–48.

Heckman, J. J. (2001a): "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy*, 109, 673–748.

Heckman, J. J. (2001b): "Econometrics and Empirical Economics," *Journal of Econometrics*, 100, 3–5.

Heckman, J. J. (2000): "Causal Parameters And Policy Analysis In Economics: A Twentieth Century Retrospective," *The Quarterly Journal of Economics*, CVX, 45–97.

Hedström, P. and Swedberg, R., eds. (1998): *Social Mechanisms*. Cambridge University Press.

Hendry, D. F. (1993): *Econometrics—Alchemy or Science?* Oxford: Blackwell.

Hodges, J. L., Jr. and Lehmann, E. (1964): *Basic Concepts of Probability and Statistics.* San Francisco: Holden-Day.

Holland, P. (1988): "Causal Inference, Path Analysis, and Recursive Structural Equation Models," in *Sociological Methodology 1988*, ed. by C. Clogg. Washington, D.C.: American Sociological Association, Chapter 13.

Holland, P. (1986): "Statistics and Causal Inference," *Journal of the American Statistical Association*, 8, 945–60.

Humphreys, P. (1997): "A Critical Appraisal of Causal Discovery Algorithms, in *Causality in Crisis?* ed. by V. McKim and S. Turner. South Bend: University of Notre Dame Press, pp. 249–63 (with discussion).

Humphreys, P. and Freedman, D. A. (1999): "Are There Algorithms That Discover Causal Structure?" *Synthese*, 121, 29–54.

Humphreys, P. and Freedman, D. A. (1996): "The Grand Leap," *British Journal for the Philosophy of Science*, 47, 113–23.

International Agency for Research on Cancer (1986): *Tobacco Smoking* Lyon, France: IARC, Monograph 38.

Kahneman, D., Slovic, P., and Tversky, A., eds. (1982): *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Kahneman, D. and Tversky, A., eds. (2000): *Choices, Values, and Frames*. Cambridge University Press.

Kahneman, D. and Tversky, A. (1996). "On the Reality of Cognitive Illusions," *Psychological Review*, 103, 582–91.

Kahneman, D. and Tversky, A. (1974): "Judgment under Uncertainty: Heuristics and Bias," *Science*, 185, 1124–31.

Kaprio, J. and Koskenvuo, M. (1989): "Twins, Smoking and Mortality: A 12-Year Prospective Study of Smoking-Discordant Twin Pairs," *Social Science and Medicine*, 29, 1083–9.

Keynes, J. M. (1939): "Professor Tinbergen's Method," *The Economic Journal*, 49, 558–70.

Keynes, J. M. (1940): "Comment on Tinbergen's Response," *The Economic Journal*, 50, 154–56.

Lauritzen, S. (1996): *Graphical Models*. Oxford: Clarendon Press.

Lauritzen, S. (2001): “Causal Inference in Graphical Models,” in *Complex Stochastic Systems*, ed. by O. E. Barndorff-Nielsen, D. R. Cox, and C. Klüppelberg. Boca Raton, FL: Chapman & Hall/CRC, pp. 63–108.

Leamer, E. (1978): *Specification Searches*. New York: John Wiley.

Lieberson, S. (1985): *Making it Count.* Berkeley: University of California Press.

Lieberson, S. and Lynn, F. B. (2002): “Barking Up the Wrong Branch: Alternative to the Current Model of Sociological Science,” *Annual Review of Sociology*, 28, 1–19.

Meehl, P. E. and Waller N. G. (2002): “The Path Analysis Controversy: A New Statistical Approach to Strong Appraisal of Verisimilitude,” *Psychological Methods,* 7, 283–337 (with discussion).

Meehl, P. E. (1978): “Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology,” *Journal of Consulting and Clinical Psychology*, 46, 806–34.

Meehl, P. E. (1954): *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis: University of Minnesota Press.

Miller, D. P., Neuberg, D., De Vivo, I., Wain J. C., et al. (2003). “Smoking and the Risk of Lung Cancer: Susceptibility with GSTP1 Polymorphisms. *Epidemiology* 14, 545–51.

Nakachi, K., Ima, K., Hayashi, S.-I. and Kawajiri, K. (1993): “Polymorphisms of the CYP1A1 and Glutathione S-Transferase Genes Associated with Susceptibility to Lung Cancer in Relation to Cigarette Dose in a Japanese Population,” *Cancer Research*, 53, 2994–99.

Neyman, J. (1923): “Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes,” *Roczniki Nauk Rolniczki* 10: 1–51, in Polish. English translation by D. Dabrowska and T. Speed (1990), *Statistical Science*, 5, 463–80 (with discussion).

Ní Bhrolcháin, M. (2001): “Divorce Effects and Causality in the Social Sciences,” *European Sociological Review*, 17, 33–57.

Oakes, M. (1986): *Statistical Inference*. Chestnut Hill, MA: Epidemiology Resources Inc.

Pearl, J. (2000): *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J. (1995): “Causal Diagrams for Empirical Research,” *Biometrika*, 82, 669–710 (with discussion).

Pratt, J. and Schlaifer, R. (1984): “On the Nature and Discovery of Structure,” *Journal of the American Statistical Association*, 79, 9–21.

Pratt, J. and Schlaifer, R. (1988): “On the Interpretation and Observation of Laws,” *Journal of Econometrics*, 39, 23–52.

Robins, J. M. (1999): “Association, Causation, and Marginal Structural Models,” *Synthese*, 121, 151–79.

Robins, J. M. (1995): “Discussion,” *Biometrika*, 82, 695–8.

Robins, J. M. (1987): “A Graphical Approach to the Identification and Estimation of Causal Parameters in Mortality Studies with Sustained Exposure Periods,” *Journal of Chronic Diseases* 40, Supplement 2, 139S–161S.

Robins, J. M. (1986): “A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect,” *Mathematical Modelling*, 7, 1393–1512.

Rubin, D. (1974): “Estimating Causal Effects of Treatments in Randomized and Nonrandomized studies,” *Journal of Educational Psychology*, 66, 688–701.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999): “Adjusting for Non-Ignorable Drop-Out using Semiparametric Non-Response Models, *Journal of the American Statistical Association*, 94, 1096–1146 (with discussion).

Shadish W. R., Cook T. D., and Campbell D. T. (2002): *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Shields, P. G., Caporaso, N. E., Falk, K. T., Sugimura, H., Trivers, G. E., Trump, B. P., Hoover, R. N., Weston A. and Harris, C. C. (1993): “Lung Cancer, Race and a CYP1A1 Genetic Polymorphism,” *Cancer Epidemiology, Biomarkers and Prevention*, 2, 481–5.

Sims, C. A. (1980): “Macroeconomics and Reality,” *Econometrica*, 48, 1–47.

Sobel, M. E. (2000): “Causal Inference in the Social Sciences,” *Journal of the American Statistical Association*, 95, 647–51.

Spirtes, P., Glymour, C., and Scheines, R. (1993): *Causation, Prediction, and Search*. Springer Lecture Notes in Statistics, no. 81, New York: Springer-Verlag. 2nd edn (2000), Cambridge, Mass.: MIT Press.

Steiger, J. H. (2001): “Driving Fast in Reverse,” *Journal of the American Statistical Association*, 96, 331–38.

Stone, R. (1993): “The Assumptions on Which Causal Inferences Rest,” *Journal of the Royal Statistical Society*, Series B, 55, 455–66.

Tinbergen, J. (1940): “Reply to Keynes,” *The Economic Journal*, 50, 141–54.

Wainer, H. (1989): “Eelworms, Bullet Holes, and Geraldine Ferraro: Some Problems with Statistical Adjustment and Some Solutions,” *Journal of Educational Statistics*, 14, 121–40 (with discussion). Reprinted in *The Role of Models in Nonexperimental Social Science*, ed. by J. Shaffer. Washington, D.C.: AERA/ASA, 1992, pp. 129–207.

Woodward, J. (1997): “Causal Models, Probabilities, and Invariance,” in *Causality in Crisis?* ed. by In V. McKim and S. Turner. South Bend: University of Notre Dame Press, pp. 265–315 (with discussion).

Woodward, J. (1999): “Causal Interpretation in Systems of Equations,” *Synthese*, 121, 199–247.

Wright, P. G. (1928): *The Tariff on Animal and Vegetable Oils*. New York: MacMillan.

Technical Report No. 601
Department of Statistics
U.C. Berkeley, CA 94720