Reconsidering Causation

Stephen F. LeRoy

May 17, 2024

Abstract

Recent applied work in economics has displayed renewed interest in the problem of characterizing the causal relations that link economic variables. However, many discussions avoid explicit specification of what has to be true about a formal model to justify an assertion that one variable in it causes another. Such specification is supplied here. Related topics, such as determining whether correlation implies causation, or vice-versa, and when causal coefficients can be estimated using ordinary least squares or instrumental variables regressions, are discussed.

In recent years economists have displayed renewed interest in the role of causation in economic analysis (see, for example, Athey and Imbens [3] or Angrist and Pischke [1]).[1] In fields other than economics discussion has centered on causal graphs (Pearl [10]), in which one bases an account of causation on a graph, taken as given, consisting of a series of nodes representing variables connected by arrows indicating causal links. For the most part economists have not found this framework congenial (see Heckman and Pinto [7]). One can only speculate about the reasons for this, but one plausible guess is that economists are uncomfortable with the question of where the arrows come from. Economists think of models as being derivable from assumptions on preferences, endowments, market structure and the like. Arrows do not appear in such derivations.

The Cowles economists, who originated the study of structural economic models, saw the lack of clarity in discussions of causation as a serious problem. They proposed formal characterizations of causation (see especially Simon [13]), but that research tradition has

[1]University of California, Santa Barbara. leroy@ucsb.edu. Some definitions are changed here, but most of the analysis in an earlier monograph (LeRoy [9]) carries over with minor modification in the current setting.

I am indebted to Isaiah Andrews, Nancy Cartwright, Kevin Hoover, Mary Morgan, Judea Pearl and Stephen Salant for conversations and correspondence on earlier drafts of the monograph and this paper.

been mostly discontinued. As a result, in the contemporary literature we frequently see discussions of ideas related to causation without any explicit characterization of causation in terms of the formal properties of structural models.

Here this connection is supplied: causal orderings are derived from the model's formal structure without reference to intuitive ideas about causation drawn from outside the model. Further, graphical representations of causal models, in which the causal arrows are derived from the structural equations rather than specified as part of the assumed structure, are shown to play an important role in facilitating the analysis. Finally, the exercise is based on the assumed existence of a model that is not altered as part of the analysis of interventions; the contrary—a presumption that causal analysis involves changing the assumed model, with the alteration depending on the question being asked—is sometimes seen in the existing literature.

In the usage of the Cowles economists causal analysis consists of two undertakings: (1) formulating an explicit model assumed to generate the data, and (2) determining the causal links implied by that model. Only the second of these projects is considered in this paper.

1 Interventions

For the Cowles economists investigation of causation consists of analysis of interventions. An *intervention* consists of a modification of the structural equations intended to allow the analyst to determine what would happen under a given hypothetical change in the environment (Haavelmo [6]). When the cause variable is internal (that is, determined by the equations of the model) this use of a model to analyze causation involves altering the assumed model. The alteration consists of deleting the equations that determine the cause variable and substituting the assumption that the cause variable is external (that is, taken as given). Thus the alteration of the model depends on the causal question that is being asked. But coherent causal analysis using a model is possible only if, contrary to this, the model is defined independently of the contemplated intervention.

The insistence of the Cowles economists on representing interventions as modifications of structural equations led them away from a much simpler formalization of interventions using elements of the model that are already available: external variables. Representing interventions as hypothetical alterations of the values assumed to be taken on by external variables means that no change in the model is involved: internal variables are not treated as if they were external unless the structure of the model justifies doing so (that is, unless causation is implementation neutral; see the discussion below). There is no loss of generality in requiring that interventions be modeled as alterations of external variables since any conceivable intervention can be accommodated by inclusion of external "shift variables" in the model.

The first step in analyzing the effects of an intervention is to set the external variables to preassigned values. The solution to the model under these values is termed the *baseline*. Then generate an intervention by changing the assumed value of one or more of the external variables and recomputing the solution. One then determines the effect of the intervention by comparing the values taken on by the internal variables under the intervention with those under the baseline specification. In linear models the causal coefficient—the ratio between the change in the effect variable and the change in the cause variable—if uniquely defined is the same for all specifications of the values of the external variables that are consistent with the assumed intervention on the cause variable.

By designating a coefficient as an external variable rather than a constant or internal variable the analyst is allowing for interventions on that variable, and is also excluding the specification that interventions on other external variables alter the value taken on by the variable of interest. Designating as variables the coefficients of a structural equation in an otherwise linear model is perfectly acceptable, but doing so implies that the model is bilinear, not linear. These specifications are different. In a model that consists of equations characterized as linear the coefficients are interpreted as constants. Labeling the coefficient a constant implies that interventions on that constant are ruled out: we do not ask mathematicians what would happen if π were equal to a number other than 3.1416, and economists

should not be asking the analogous question about the constants of their models (or, more precisely, should interpret intervening on a constant as generating a comparison of different models, not as constituting a causal analysis of a given model).

The requirement that analysts explicitly distinguish constants from external variables and treat each consistently, even in analyzing interventions, enforces clarity about which contemplated interventions the analyst views as admissible and which are excluded from consideration.[2]

2 Structural Models

A linear *structural model* can be written as

$$y = Ay + Bx, \tag{1}$$

where y denotes the *internal variables* of the model (those determined by the model) and x denotes its *external variables* (those taken as given). Here we are taking the term "structural" to refer to the specification that external and internal variables are explicitly distinguished from one another. Both x and y are vectors. Their dimensions are unrestricted. The external variables x are assumed to be variation free, pending the stronger assumption of independence adopted below.[3] $A = \{\alpha_{ij}\}$ and $B = \{\beta_{ik}\}$ are matrices of constants. $I - A$ is assumed to be invertible to assure that the solution for y in terms of x exists and is unique (here I is the identity matrix conformable with A). A is a lower triangular matrix: $\alpha_{ij} = 0$ for $i \leq j$.[4]

[2]In the Cowles treatment of causation, and also in many recent discussions in the philosophy literature, analysts insisted that causal interpretation of a model requires a property of invariance. The meaning of invariance in the context of implementing alterations of a model's structure was never made clear despite much discussion. However, with interventions characterized as consisting of hypothetical changes in the values of external variables rather than as general structural changes, allegations of failure of invariance can only consist of assertions that terms specified as constants should instead be modeled as variables. Reminding analysts that if their models are misspecified their diagnoses of causation are likely to be wrong is hardly necessary. We see that invariance disappears as a feature of causal attributions that requires extended discussion.

[3]The members of a collection of random variables are variation free if the domain of each does not depend on the domains or the realizations of the others. Random variables that are probabilistically independent are variation free, but not necessarily vice-versa. For example, two random variables that are distributed as bivariate normal with nonzero correlation are variation free, but not independent.

[4]The extent to which the analysis here generalizes to non-triangular A is an open question. At a minimum it would be necessary to include new symbols like $\longleftrightarrow$ and $\Longleftrightarrow$ to represent the relation between y_i and y_j when both $\alpha_{ij} \neq 0$ and $\alpha_{ji} \neq 0$. Proceeding along these lines would constitute generalizing the analysis of causation to deal with a hybrid of causation and

The *reduced form* is

$$y = (I - A)^{-1}Bx \equiv Gx. \tag{2}$$

The Cowles economists viewed the reduced form as lacking valuable information that is available in the structural form. It is difficult to extract from their discussions an account of why this information disappears under the arithmetic operations involved in going from the structural form to the reduced form. A recurrent theme has been that the structural form coefficients can be used to analyze interventions, and therefore locate causal orderings among internal variables, whereas the reduced-form coefficients cannot be used in this way. There remains the question of what it is about structural models that makes this so.

We will show that a version of the Cowles argument is correct. A definition of causation is proposed that clarifies the precise nature of the information that is lost in passing from the structural to the reduced form.

3 Causation

We begin by establishing some terminology for structural models, starting with the idea of connected variables. Two external variables are not (directly) connected. Two variables at least one of which is internal are *directly connected* if there exists a structural equation in which both appear. If the variables are x_k and y_j, an equivalent statement of the condition is that β_{jk} is nonzero; if they are y_i and y_j the corresponding condition is that either α_{ij} or α_{ji} is nonzero. Two variables are *indirectly connected* if there exists a *path*—an ordered n-tuple of variables—that goes from one to the other, where each variable is directly connected to its neighbors. Assuming that indirect paths contain at least one interior variable allows them to be distinguished from direct paths. The variables are *connected* if they are connected either directly or indirectly (or both along different paths). A *connectedness graph* is a graph displaying the variables of the model with an edge drawn between each pair of variables that are directly connected.

simultaneity.

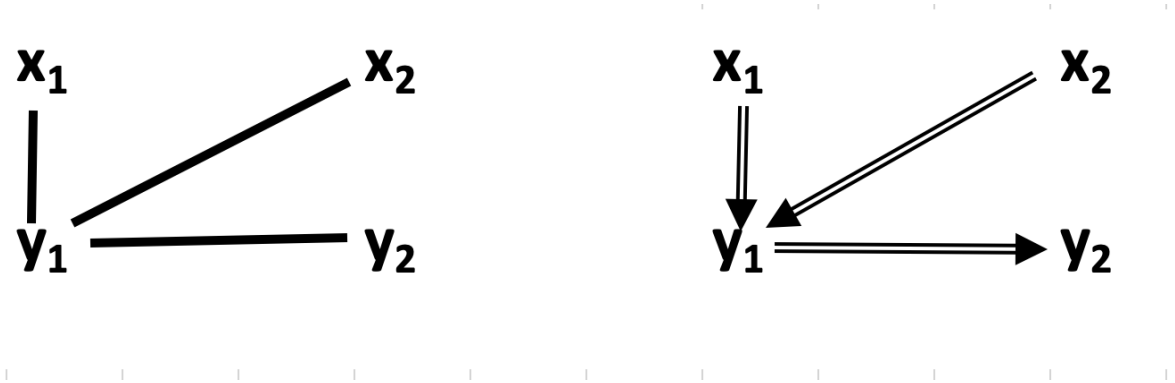

Figure 1: Connectedness and IN-Causation

As an example, consider the linear model

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{3}$$

$$y_2 = \alpha_{21}y_1. \tag{4}$$

The connectedness graph of this model is shown in the left panel of Figure 1.

We turn to the definition of causation. The *external set* of y_i, denoted $\mathcal{E}(y_i)$, is the set of external variables x_k such that $x_k \in \mathcal{E}(y_i)$ if and only if $\gamma_{ik} \neq 0$. Thus x_k is a member of $\mathcal{E}(y_i)$ if and only if x_k is one of the external determinants of the value of y_i.

Pairs of external variables (x_k, x_j) are never causally related. Pairs of variables (x_k, y_j) are *directly causally related*, with x_k directly causing y_j, if x_k and y_j are directly connected. They are *indirectly causally related* if x_k is not directly connected to y_j but is a member of the external set of y_j. Finally, they are *causally related* if they are causally related either directly or indirectly (or both, along different paths).

An internal variable y_i *directly IN-causes* another internal variable y_j if and only if (1) the two are directly connected with $\alpha_{ji} \neq 0$, and (2) for each $x_k \in \mathcal{E}(y_i)$ all paths that connect x_k to y_j pass through y_i. An internal variable y_i *indirectly IN-causes* a connected internal variable y_j if there exists a path that indirectly connects them, and (2) above is satisfied. One internal variable *IN-causes* another if it does so either directly or indirectly (or both, along different paths).

If the conditions for IN-causation are satisfied causation between y_i and y_j is *implementation neutral*: an intervention on $x_k \in \mathcal{E}(y_i)$ resulting in Δy_i has the same effect on y_j—$\alpha_{ji}\Delta y_i$, assuming IN-causation is direct—for each x_k (here for any member of $\mathcal{E}(y_i)$ the assumed intervention is set to produce the same Δy_i). That being so, the causal relation between y_i and y_j is implementation neutral in the sense that the effect on y_j of an intervention resulting in Δy_i does not depend on how the intervention is implemented (that is, on which member or members of $\mathcal{E}(y_i)$ is the underlying intervention variable). Implementation-neutral causation is indicated by a double arrow: $y_i \Rightarrow y_j$. The causal graph of the model (3)-(4) is shown as the right-hand panel of Figure 1.

The assumption that A is lower triangular implies that if $i < j$ y_i may or may not directly IN-cause y_j, but y_j never IN-causes y_i. Equivalently, y_i directly IN-causes y_j only if $i < j$.

We will also write $x_k \Rightarrow y_j$, so that x_k IN-causes y_j, when x_k is in the external set of y_j. This notational specification reflects the fact that y_i plays the same role in $y_i \Rightarrow y_j$ as x_k plays in $x_k \Rightarrow y_j$.

Frequently we have pairs of directly connected internal variables y_i and y_j such that $\alpha_{ji} \neq 0$ but the condition for IN-causation that all paths connecting members of $\mathcal{E}(y_i)$ with y_j pass through y_i is not satisfied. In that case the effect on y_j of an intervention represented by Δy_i is ambiguous: different interventions consistent with a given Δy_i generate different effects on y_j. Thus causation is not implementation neutral. Members of $\mathcal{E}(y_i)$ which have causal paths that connect with y_j without passing through y_i are termed *confounders*. The intervention that defines causation is restricted by holding the confounder constant. The relevant notion of causation is called *conditional causation*, as opposed to IN-causation, which occurs when there are no confounders.[5] Conditional causation is indicated by a single arrow accompanied by | and followed by a list of the confounders.[6]

[5]It might appear that the interventions that define conditional causation are not likely to be relevant in applications. This presumption is incorrect in general. Consider a family deciding whether a prospective student should attend a high-status private university or a cheaper public university. The decision depends on the effect of the enrollment decision on the student's subsequent lifetime income. Lifetime income depends on the enrollment decision, which in turn depends on current family income. Family income also affects lifetime income directly. The family here, knowing its income, is interested in the effect of the enrollment decision on future earnings conditional on current family income. They are interested in conditional causation, not IN-causation, which in any case is undefined due to the role of family income as a confounder.

[6] Here "conditioning" has a meaning different from that in probability theory. Of course, there is no problem in considering the probability distribution of any internal variable conditional on any other. The context indicates which meaning of "con-

Figure 2: Connectedness and Conditional Causation

In the model

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{5}$$

$$y_2 = \alpha_{21}y_1 + \beta_{22}x_2 \tag{6}$$

the effect of Δy_1 on y_2 is $\alpha_{21}\Delta y_1$ if the alteration of y_1 is due to an intervention of $\Delta y_1/\beta_{11}$ on x_1, or $(\alpha_{21} + \beta_{22}/\beta_{12})\Delta y_1$ if the alteration of y_1 is due to an intervention of $\Delta y_1/\beta_{12}$ on x_2. With $\beta_{22} \neq 0$ these are different, reflecting the fact that causation of y_2 by y_1 is not implementation neutral. The variable x_2 is a confounder. Causation is conditional: $y_1 \rightarrow y_2|x_2$, with α_{21} being the coefficient of conditional causation.

In graphs pairs of variables of which one conditionally causes the other are connected by single arrows. Figure 2 shows the connectedness and causal graphs of the model just presented.

Note that the causal graph omits specification of the confounders in the case of conditional causation. That this omission is permissible reflects the fact that the set of confounders can be inferred from the pattern of arrows in the causal graph, so listing the confounders explicitly would be redundant.

We will use the language that y_i *directly causes* y_j if either y_i directly IN-causes y_j, or y_i directly causes y_j conditional on some nonempty set of confounders. Suppose that y_i directly causes, but does not IN-cause, y_j. If the confounders are held constant we have IN-causation in the model so truncated, and conditional causation in the original model.

ditioning" is intended. The term "holding constant", which is often taken to be equivalent to "conditioning", has the same double meaning.

An external variable x causes an internal variable y if and only if there exists a *causal path* (a path along which each member except the last directly causes its successor) connecting the two. This condition is satisfied if and only if $x \in \mathcal{E}(y)$, which in turn is equivalent to whether the appropriate element of the reduced-form coefficient matrix is nonzero. Thus whether or not an external variable causes an internal variable can be determined from the reduced form without reference to the structural form of a model.[7]

Whether or not one internal variable causes or directly causes another cannot be determined from the reduced form. This is so because in the structural form causal paths are involved in the determination, and these are not preserved under the arithmetic operations involved in computing the reduced form from the structural form. Most simply, internal variables are never causally related in reduced forms. These results give content to the Cowles assertion that structural models contain causal information not available from the reduced form alone.

Note here that the variables being conditioned upon are external. Conditioning on internal variables is not admissible inasmuch as setting internal variables equal to zero conflicts with the assumed dependence in the model of those variables on the members of their external sets. Suppressing this dependence by setting these variables equal to zero constitutes alteration of the model, and must be disallowed when model respecification is not intended.

It is now possible to demonstrate the invalidity (in general) of the common practice of analyzing the causal dependence of one internal variable on another via respecifying the cause variable as external and deleting the equations that determine it. There is no problem if the cause variable IN-causes the effect variable in the original model. However, if causation is conditional the procedure involves suppressing the dependence of the effect variable on

[7] The causal graph is unambiguously implied by the structural model (and vice-versa), but not necessarily by the connectedness graph. The model

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{7}$$

$$y_2 = \alpha_{11}y_1 + \beta_{22}x_2 + \beta_{23}x_3 \tag{8}$$

implies $y_1 \to y_2|x_2$, and the model

$$y_1 = \alpha_{12}y_2 + \beta_{11}x_1 + \beta_{12}x_2 \tag{9}$$

$$y_2 = \beta_{22}x_2 + \beta_{23}x_3 \tag{10}$$

implies $y_2 \to y_1|x_2$. Thus the two models have different causal orderings. However, they have the same connectedness graph. It is seen that the causal graph, which conveys the same causal information as the structural equations, may contain strictly more causal information than the connectedness graph.

the confounders. Again, this constitutes an inadmissible alteration of the model.

4 Observed and Unobserved Variables

We have not distinguished variables according to whether they are observed by the analyst. Whether or not one variable causes or IN-causes another in a model—the subject of our discussion up to now—does not depend on whether the analyst can observe them.[8] Thus we used x and y to denote external and internal variables whether or not they are observed.

If x_k is unobserved the coefficients β_{ik} and γ_{ik} ($i = 1, ..., n$, where n is the number of internal variables) are not identified. This being so, in one of the equations in which an unobserved x_k appears—equation i, for concreteness—one can set β_{ik} equal to 1 as an arbitrary choice of units. Up to this point we have not been concerned with whether or not variables are observed, so the coefficients were not normalized to 1.

Now we are passing from determining causal orderings in prespecified structural models to testing of causal relations and estimation of causal coefficients. Causal coefficients are identified statistically only when both the cause variable and the effect variable are observed. Regressions can contain variables as explanatory variables only when these variables are observed. Consequently, in any applied work related to causation the analyst must specify which variables are observed.

We will use capital letters to denote variables observed by the analyst and lower-case letters to denote unobserved variables (and when the discussion does not depend on whether or not they are observed, as throughout the preceding sections).[9] [10]

[8] Heckman and Pinto [7]: "Issues of identification and estimation are important for making the concept of causality empirically operational, but not for defining it."

[9] This characterization does not apply to matrices A, B and G, which are assumed to consist of unobserved constants.

[10] Under the received graphical treatment of causation a dashed line is sometimes used to connect two causally related variables when one or both is unobserved. Here that notation would be unsuitable because causal links are defined independently of which variables are observed. Observability is a property of variables, not causal links, and our avoidance of dashed lines reflects this fact.

5 Probabilistically Independent External Variables

It is assumed that the external variables, not being connected by causal paths, are unconditionally independently distributed random variables (of course, they are generally correlated conditional on internal variables). The independence assumption is imposed whether or not the variables in question are observed. Independence is a very strong assumption, and most of the difficulty in determining causal orderings in applied work comes from the fact that it is usually not obvious which variables are to be taken as external, given that that specification entails the assumption that they are independent random variables (but recall that we are (1) allowing a given external variable to appear in more than one structural equation, and (2) allowing more than one external variable to appear in any given structural equation; these specifications mitigate the restrictiveness of the independence assumption).[11]

The reason the independence assumption is needed is that without that restriction we often find that the coefficients associated with causal orderings are not identified even if the cause and effect variables are observed. For example, in the model $Y_1 = \beta_{11}X_1 + x_2$, in which X_1 IN-causes Y_1, β_{11} is not identified if the external variables X_1 and x_2 are correlated.

If in a proposed model some of the variables provisionally specified as external are observed they may have nonzero sample correlations. The population counterpart of this assertion conflicts with the requirement just stated. The simplest way to respond to this problem is to interpret nonzero sample correlations as reflecting sample variation, so that in the population the correlation is assumed to be zero. All models are simplifications, and in some settings ignoring evidence of correlations among variables labeled as external may be an admissible procedure.

However, in many contexts taking that path is unacceptable, inasmuch as it amounts to assuming away the question of what causal relations underlie correlations among observed external variables. An alternative and usually preferable procedure is to assume that existence of a nonnegligible correlation between two observed variables indicates that those

[11] Investigators exhibit a strong preference for controlled experiments when they are feasible. This is so because when treatments are assigned by lotteries there is no doubt about the validity of the assumption that the treatment variable is probabilistically independent of all external variables other than the lottery outcome.

variables cannot both be external. The assumed model is a misspecification.

At a minimum, resolving the conflict between the independence assumption and nonzero correlations among variables provisionally classified as external involves respecifying one of the two correlated variables as internal.[12] Doing so makes it necessary to introduce a new external variable, presumably unobserved, and augmenting the model by including a new equation expressing the variable respecified to be internal as a function of the other of the correlated external variables and the new external variable. The operative assumption now is that the new external variable is independent of whichever of the correlated variables is assumed to be external, and also of all other external variables.[13] Thus all external variables are independent in the reformulated model. Failure to do this means that the correlations induced by explicitly modeled causal relations are conflated with unmodeled correlations among external variables.

6 Causation and Correlation

Holland [8] cited G. A. Barnard as writing "That correlation is not causation is perhaps the first thing that must be said." It is often also the last thing that is said. Repeating this mantra does not make clear what the relations are between causation and statistical measures of association. Results from the preceding analysis allow clarification of such questions.

The assumption that external variables are independent random variables implies that an internal variable y is probabilistically dependent on (that is, not probabilistically independent of) an external variable x if and only if $x \in \mathcal{E}(y)$, so that there exists a causal path that

[12] Simpson's Paradox refers to a setting in which failure to resolve correlations between external variables results in outcomes that appear counterintuitive. The supposed paradox is that it is possible that a treatment that, based on correlations, appears to be successful with both men and women taken separately may appear to be unsuccessful in a mixed population of men and women. Under the presumption that correlations necessarily represent (unconditional) causation, this appears paradoxical. The apparent paradox owes to the implicit specification that the treatment variable is external, despite being correlated with gender. The resolution is obtained by recognizing that treatment is properly modeled as internal, depending on both gender and an external shock uncorrelated with gender. A formal model incorporating this specification would specify that gender affects the outcome both directly and via the treatment variable. The causal effect of the aggregate treatment on the aggregate outcome is not implementation neutral: gender is a confounding variable in the causal relation between treatment and outcome. This implies that the correlation between aggregate treatment and aggregate outcome does not have an IN-causal interpretation. Treatment does cause the outcome conditional on gender. Therefore there is no presumption that the correlation of treatment with outcomes has the same sign as the corresponding correlations for men and women taken separately.

[13] Note the contrast with regression theory. The existence of correlation between two observed explanatory variables causes no problems in estimating coefficients in a multivariate regression. There is no conflict between this fact and the assertion here that if two variables are correlated at least one of them should be respecified to be internal.

connects x and y.[14]

The independence assumption also implies that two internal variables y_i and y_j are probabilistically dependent if and only if for some x there exists a causal path from x to y_i and also a causal path from that same x to y_j, so that the external sets of y_i and y_j overlap. Consistent with $\mathcal{E}(y_i)$ having a nonempty intersection with $\mathcal{E}(y_j)$, one of these variables may or may not cause the other. Thus absence of statistical dependence implies absence of causation, but the presence of statistical dependence between two variables does not imply that either causes the other. The mantra, construed as the assertion that neither of two correlated variables necessarily causes the other because there may exist external variables that cause both, is correct.[15]

Another related result is that if y_i IN-causes y_j then, conditional on y_i, y_j is independent of x_k for each x_k in the external set of y_i. Thus the absence of conditional correlation is consistent with the presence of causation. This implication of causation is emphasized in Glymour, Scheines and Spirtes [5].

7 Regression and Implementation Neutrality

Every internal variable y_j in a linear model can be written as

$$y_j - E(y_j) = \sum_{x_k \in \mathcal{E}(y_j)} \gamma_{jk}(x_k - E(x_k)), \tag{11}$$

where the γ_{jk} are elements of the reduced form coefficient matrix. The assumption that x_k directly causes y_j along a unique path (and that the x_k have finite second moments) implies

$$\gamma_{jk} = \frac{\text{cov}(y_j, x_k)}{\text{var}(x_k)}. \tag{12}$$

Similarly, if we have that y_i directly IN-causes y_j along a unique path (the case in which y_i

[14] This assertion must be qualified to deal with the special case in which two variables are connected along multiple causal paths that cancel.

[15] In the philosophy literature this assertion is the "principle of the common cause" (Reichenbach [11]). It is correct in our setting. Philosophers have debated whether it is true in general. See, for example, Cartwright [4] and Reiss [12].

and y_j are connected along several paths is considered in note 18) then α_{ji} satisfies

$$\alpha_{ji} = \frac{\text{cov}(y_i, y_j)}{\text{var}(y_i)}. \tag{13}$$

Here we exploit the implication of IN-causation and the assumed independence of external variables to conclude that y_j differs from $\alpha_{ji}y_i$ by a term that has zero covariance with y_i. Taking covariances with y_i, eq. (13) results. It follows that the causal coefficient coincides with the coefficient of y_i in a univariate regression of y_j on y_i.

It is important to note that the converse proposition does not obtain: existence of second moments implies that y_j can be regressed on y_i whether or not they are causally related and, if so, whether or not causation runs from y_i to y_j or vice-versa. Specifically, if we have

$$y_j = \lambda y_i + x \tag{14}$$

then λ equals $\text{cov}(y_j, y_i)/\text{var}(y_i)$ if and only if $\text{cov}(y_i, x)$ equals zero regardless of whether or not there exists a causal relation in either direction between y_i and y_j. Thus regressing one variable on another tells us nothing about the causal relation, or lack thereof, between the two.[16]

8 Regressions Associated with Causal Equations

Under the assumption that all internal variables and a specified subset of the external variables are observed a linear structural model can be written as

$$Y = AY + BX + Cx, \tag{15}$$

where we adopt notation that distinguishes between observed and unobserved external variables.[17] As above, it is assumed that A is triangular with zeros on the main diagonal.

[16] Angrist and Pischke [2], p. 128, in their critique of prevailing instruction in econometrics, made the same point: "it's hard to see how this statement [that regression errors must be assumed to be uncorrelated with regressors] promotes clear thinking about causal effects".

[17] As a special case one might assume that C is an identity matrix, so that the number of unobserved external variables equals the number of internal variables. Then unobserved external variables could be interpreted as observation errors.

The j-th row of this vector-matrix equation expresses Y_j as a function of the variables that directly cause it. In general these variables consist of internal variables Y_i (with $i < j$), observed external variables X_k and unobserved external variables x_k. It is seen that causal equations have the same format as regressions, with the explanatory variables for the dependent variable y_j being the observed variables Y and X such that $\alpha_{ji} \neq 0$ and $\beta_{jk} \neq 0$, and with the error term consisting of the unobserved external variable(s) that appear(s) in the j-th equation.

Consider an equation j in which at least one of the right-hand side variables is observed (so that either α_{ji} or β_{jk} is nonzero for at least one i or k). The question is whether or not this equation is a valid statement of causation when interpreted as a regression—that is, whether the (population counterparts of the) values taken on by the estimated regression coefficients equal the corresponding causal coefficients.

Most simply, if any of the observed explanatory variables directly IN-causes the dependent variables along a unique path the regression coefficient coincides with the corresponding causal coefficient. This was shown in the preceding section.[18]

The more interesting question is under what conditions the same is true for all the regression coefficients in the equation, whether or not they IN-cause the dependent variable. This occurs if the error is independent of the explanatory variables (here we elide the distinction between independence and mean-independence). This depends on whether the union of the external sets of the observed explanatory variables is or is not disjoint from the union of the unobserved variables in equation j. If these sets are disjoint it follows from the assumed probabilistic independence of external variables that the error in the candidate regression is independent of the explanatory variables. In that case the regression coefficients of the observed explanatory variables coincide with the direct causal coefficients, and this is so

Under the additional restriction $B = 0$, so there are no observed external variables, the model could represent a vector autoregression, following a change of notation. We do not pursue these lines.

[18] In the model

$$Y_1 = x_1 + x_2 \tag{16}$$

$$Y_2 = \alpha_{21} Y_1 + x_3 \tag{17}$$

$$Y_3 = \alpha_{31} Y_1 + \alpha_{32} Y_2 + x_4 \tag{18}$$

Y_1 IN-causes Y_3, and the two variables are connected along two paths, one direct and one passing through Y_2. The regression coefficient in a bivariate regression of Y_3 on Y_1, equal to the term in parentheses in $\Delta Y_3 = (\alpha_{31} + \alpha_{32}\alpha_{21})\Delta Y_1$, is seen to coincide with the coefficient of IN-causation. It depends on the coefficients on both paths.

whether or not causation is implementation neutral. The candidate regression is in fact a causal regression.

This regression can be termed the *associated regression* for that causal equation. Depending on the model none, some or all of the causal equations have associated regressions. Least squares provides consistent estimates of the causal coefficients in each causal equation that has an associated regression. This is so whether the causation is implementation neutral or conditional.

If the two sets just described have a nonempty intersection the error covaries with at least one of the explanatory variables, implying that for at least one of the explanatory variables the regression coefficient differs from the corresponding causal coefficient. It remains true that the regression coefficient of any explanatory variable that IN-causes the dependent variable along a unique path does coincide with the corresponding causal coefficient, as seen above (IN-causation of some explanatory variables is consistent with failure of the disjointness condition if equation j includes variables other than the cause variable as explanatory variables).

The same point can be stated differently. In the presence of conditional causation the confounder(s) may (but does not necessarily, as in the first example below) induce correlation between some or all of the explanatory variables and the error. If so regression coefficients associated with explanatory variables for which the correlation is nonzero do not coincide with the associated causal coefficients.

In sum, it follows that regression gives a consistent estimate of the causal coefficient associated with a particular observed explanatory variable either when causation is implementation neutral along a unique path, or for all explanatory variables in an equation when the disjointness condition is satisfied. In the presence of confounders the regression produces a consistent estimate of a conditional causal coefficient if the confounding path or paths pass through other explanatory variables. This is so because in that case the error term will still be independent of the explanatory variables. However, if instead any of the confounding paths passes through the error term the regression coefficients will not coincide with

conditional causal coefficients.

Two examples will make these results clearer. Consider the causal model

$$Y_1 = x_1 + x_2 \tag{19}$$

$$Y_2 = \beta_{22}x_2 + x_3 \tag{20}$$

$$Y_3 = \alpha_{31}Y_1 + \alpha_{32}Y_2 + x_4. \tag{21}$$

Neither Y_1 nor Y_2 IN-causes Y_3, but each does cause Y_3 conditional on x_2. Here a multivariate regression of Y_3 on Y_1 and Y_2 is the associated regression for the causal equation (21), implying that the regression coefficients of Y_1 and Y_2 coincide with the conditional causal coefficients α_{31} and α_{32}. The status of x_2 as a confounder for $Y_1 \to Y_3$ and $Y_2 \to Y_3$ does not result in inconsistency in the estimates of α_{31} and α_{32}. This is so because for Y_1 (Y_2) the confounding path from x_2 to Y_3 goes through Y_2 (Y_1), not through the error.

For an example in which the candidate regression is not an associated regression, consider

$$Y_1 = x_1 + x_2 \tag{22}$$

$$Y_2 = \beta_{22}x_2 + x_3 \tag{23}$$

$$Y_3 = \alpha_{31}Y_1 + \alpha_{32}Y_2 + \beta_{32}x_2 + x_4 \tag{24}$$

Here the candidate regression for the third equation involves the same observed explanatory variables, Y_1 and Y_2, as above. We have $Y_1 \to Y_3|x_2$ and $Y_2 \to Y_3|x_2$, also as above. The third equation is not an associated regression due to the fact that x_2 is in the external sets of Y_1 and Y_2, and is also a component of the regression error. The least-squares coefficient estimates of α_{31} and α_{32} are inconsistent.

A regression that omits one or more of the observed explanatory variables that appear in the causal equation will generally not produce coefficients that coincide with the corresponding causal coefficients. For example, in the model just discussed a univariate regression of Y_3 on Y_2, so that Y_1 is omitted, will not produce the coefficient α_{32}. Similarly, if the regres-

sion is specified to include explanatory variables that do not appear in the causal equation, least-squares estimates of causal coefficients will generally be inconsistent.

It was common practice two generations ago to estimate macroeconometric models by running regressions of each internal variable on external variables and other internal variables. Routinely these estimated coefficients were interpreted as measuring causation. More recently the same practice has been adopted in estimating structural vector autoregressions. It was presumed that doing empirical implementations in this way is acceptable as long as the model is correctly specified. Here we see that this presumption is in general incorrect: even if the data were in fact generated by the assumed model, estimated regression coefficients provide consistent estimators of causal coefficients only if each causal equation has an associated regression. Even then regression coefficients may represent either IN-neutral causation or conditional causation, as we have seen.

9 Instrumental Variables

The preceding sections involved only the estimators generated by ordinary least squares. We have seen that in the presence of confounders these estimators may be inconsistent. In the same settings instrumental variables that produce consistent estimates of conditional causal coefficients may be available. In the model

$$Y_1 = \beta_{11} X_1 + x_2 + x_3 \tag{25}$$

$$Y_2 = \alpha_{21} Y_1 + \beta_{22} x_2 + x_4 \tag{26}$$

we have $Y_1 \to Y_2 | x_2$ with conditional causal coefficient α_{21}. The least-squares regression of Y_2 on Y_1 is not an associated regression for the causal eq. (26), implying that the regression coefficient of Y_2 on Y_1 does not equal α_{21}. This is so because x_2, a component of the error in eq. (26), is a determinant of Y_1, from eq. (25).

Despite this, α_{21} can be estimated consistently, implying that the effect of Y_1 on Y_2

conditional on x_2 can be evaluated (subject, of course, to sample variation). From the fact that a reduced-form coefficient equals the product of the coefficients associated with the direct causal relations along a causal path (assuming that path is unique, as the path from X_1 to Y_2 is in the example), we have

$$\alpha_{21} = \frac{\gamma_{21}}{\gamma_{11}} = \frac{\text{cov}(X_1, Y_2)/\text{var}(X_1)}{\text{cov}(X_1, Y_1)/\text{var}(X_1)} = \frac{\text{cov}(X_1, Y_2)}{\text{cov}(X_1, Y_1)}. \tag{27}$$

The rightmost term in eq. (27) is recognized as the (population counterpart of the) instrumental variables estimator of α_{21}, with the instrument being X_1. Because X_1, Y_1 and Y_2 are observed the instrumental variables regression can be implemented empirically. This is so even though x_2, the variable that confounds the unconditional IN-causal relation between Y_1 and Y_2, is assumed to be unobserved.

10 Nonlinearity and Causation

The discussion so far has been restricted to linear models. To what extent does it apply in the presence of nonlinearities? If the question "What is the effect of x on y?" is read as "What multiplicative constant reflects the causal relation between x and y?" the setting is linear by definition. This is so because if Δx and Δy are causally related by some function $f(...)$, so that $\Delta y = f(...)\Delta x$ (where Δx represents the difference between intervention and baseline, as above), then setting $f(...)$ equal to the constant β (as follows if the effect of x on y is to be well-defined without specifying the value taken on by x or any other variable) results in a difference equation that integrates to the linear equation $y = \beta x$.

It follows that introducing nonlinearities implies that characterizing causal relations between variables involves operations more complicated than multiplying the assumed change in the cause variable by a constant. An example will make this clear. In the model (25)-(26)

replace eq. (25) by the nonlinear equation

$$Y_1 = \begin{cases} 1 & \text{if } X_1 x_2 x_3 \geq 0 \\ 0 & \text{otherwise} \end{cases} . \tag{28}$$

Characterizing the causal relation between X_1 (for example) and Y_1 consists of solving for the baseline and intervention values of Y_1 as functions of the baseline values of X_1, x_2 and x_3, and the intervention value of X_1. The baseline value of Y_1 is given by

$$Y_1^b = \begin{cases} 1 & \text{if } X_1^b x_2^b x_3^b \geq 0 \\ 0 & \text{otherwise} \end{cases} . \tag{29}$$

The intervention value Y_1^i of Y_1 is the same except that X_1^b is replaced by X_1^i. The effect of the X_1 intervention on Y_1 is $Y_1^i - Y_1^b$.

The above expressions are the counterpart in a nonlinear model of $\Delta y = \beta \Delta x$ in a linear model. The causal constant β in the linear version is replaced by a function that, in general, has as arguments not only the baseline and intervention values of the cause variable (and not just their difference, as is sufficient in the linear case) but also other variables. It is seen that in nonlinear models there may be no way to attach meaning to the causal dependence of one variable on another short of specifying at the same time how the effect variable depends on the assumed values of a collection of variables other than that arbitrarily labeled as the cause variable.

These observations have an immediate implication in the analysis of treatment evaluation. The fact that the treatment variable is usually assumed to take on one of a finite number of values—frequently two, represented by 0 and 1—implies that the function representing the determination of the treatment—generally referred to as the policy function—is necessarily nonlinear (no linear function which has a nontrivial domain consisting of independent random variables has a range {0,1}).

It turns out that the nonlinearity of the policy function causes no problems in estimating the effect of the treatment on the outcome if the latter function is linear. To see this

observe that in the model (28)-(26) under any baseline specification and any intervention the term $\Delta Y_1^i - \Delta Y_1^b$ takes on one of the values {1,-1} (0 is disallowed because in that case the purported intervention has no effect on the cause variable, Y_1). For nonzero values of $\Delta Y_1^i - \Delta Y_1^b$ the linearity of eq. (26) implies that the causal coefficient of Y_1 on Y_2 conditional on x_2 unambiguously equals α_{21}; despite the nonlinearity of eq. (28) α_{21} can be estimated consistently.

The general result is that if a model contains some equations that are linear and some that are nonlinear the analysis of this paper applies to causal coefficients that appear in those equations that are linear. Analysis of causation in equations that are nonlinear requires separate treatment.

11 Latent Variables and Identification

Up to now it has been assumed that causal models do not contain latent variables (internal variables that are not observed). Here we consider what happens when that very strong restriction is relaxed. As would be expected, doing so has the result that identification issues come to the foreground. This is most easily demonstrated using an example. Consider the model

$$Y_1 = x_1 + x_2 \tag{30}$$

$$y_2 = \alpha_{21} Y_1 + x_3 \tag{31}$$

$$Y_3 = \alpha_{31} Y_1 + \alpha_{32} y_2 + \beta_{32} x_2, \tag{32}$$

so that y_2 is a latent variable. The fact that there are 3 internal variables implies that there are 3 pairs of internal variables that in general may or may not be causally related, with causal coefficients that may or may not be identified.

The presumption implicit in this assertion is that the problems of determining causation and identifying causal coefficients are different and must be determined separately (the Cowles economists, like many more recent analysts of causation, largely missed this distinc-

tion, essential as it is).[19] Nonexistence of a causal coefficient can occur if either two variables are not causally related or the relevant causal coefficient exists but is not identified. In the above model the causal relations linking the members of the pairs of internal variables are $Y_1 \Rightarrow y_2$, $Y_1 \rightarrow Y_3|x_2$ and $y_2 \rightarrow Y_3|x_2$. These causal relations are the same as would obtain if y_2 were observed (it was noted above that causal orderings do not depend on which variables are observed). The constants associated with the pairs defining the causal ordering on the internal variables are α_{21}, α_{31} and $\alpha_{31} + \alpha_{32}\alpha_{21}$, respectively. The first two of these are not identified. The same is true of α_{32}, the constant directly connecting y_2 and Y_3. The constant $\alpha_{31} + \alpha_{32}\alpha_{21}$ that gives the effect of Y_1 on y_3 (given x_2) making allowance for the fact that causation occurs both directly and indirectly is identified.

The conclusion of the exercise just discussed is that even though assumptions about the observability (or lack thereof) of internal variables have no direct implications for the causal ordering, such restrictions do have implications for the identifiability of causal coefficients. As one would expect, the fewer variables are observable the larger is the set of unidentified causal coefficients. Further, there is no obvious connection between the implementation neutrality of causation and the identifiability of causal coefficients. In particular, as the example illustrates, implementation neutrality is neither necessary nor sufficient for identifiability.

This is not good news: specifying that all internal variables are observed is a restriction that one would like to avoid, but in many settings it is far from clear how this is to be achieved.

12 Summary

The work of the Cowles economists on simultaneous equations models was well received and was regarded as an important development at the time. Several of the major contributors were awarded Nobel prizes for this work. It now appears that, despite this apparent success, development based on the Cowles paradigm was terminated at an intermediate point. The

[19] As regards the more recent causation literature, this omission is yet another consequence of vagueness in specifying the data-generating model.

reasons for this abrupt shift in direction, not having been much discussed in the literature, are not apparent. Causal attributions are now based largely on natural experiments, meaning events that can credibly be characterized as external shocks. This is a positive development, but it has been combined with a reluctance to specify data-generating models that clearly distinguish between external and internal variables. Heckman and his coauthors in particular have repeatedly pointed out that the absence of explicit specification of models has resulted in lack of clarity in discussions related to causal issues. As a result of avoiding many of the central points at issue, proponents of potential outcomes and their critics have spent their time debating such semantic issues as whether controlled experiments are or are not the gold standard in causal analysis.

We do better by resuming work along the lines the Cowles economists laid out. Doing so involves addressing issues for which the Cowles treatment was at best indirect. In this paper we began with supplying a definition of what it means for one variable in an explicitly-specified model to cause another. This involved attaching a specific meaning to the assertion that one variable is connected to another, a topic dealt with only implicitly by Simon [13]. The definition of connectedness led here in turn to definitions of the notions of IN-causation, conditional causation and confounding variables. It was noted that these topics can be defined and analyzed without reference to assumptions about observability or assignments of probability distributions to external variables, a fact that guided the determination of the order in which the respective issues were discussed above. The discussion then turned to identification and estimation issues, for which observability and probabilities are central. A major topic was discussion of what has to be true for statistical measures like correlation and regression coefficients to be interpretable causally; these topics were treated only in passing by the Cowles economists.

Until recently most economists took the view that discussion of issues related to causation could safely be left to philosophers. The practice now is to take causation issues seriously, but the fact remains that it is not easy to find discussions where concerns about exogeneity and causation are dealt with explicitly and clearly. The view here is that adopting the Cowles

analytical framework, appropriately updated, will point us in a better direction.

References

[1] Joshua D. Angrist and Jorn-Steffen Pischke. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press, Princeton and Oxford, 2015.

[2] Joshua D. Angrist and Jorn-Steffen Pischke. Undergraduate econometrics instruction: Through our classes, darkly. *Journal of Economic Perspectives*, 31:125–144, 2017.

[3] Susan Athey and Guido W. Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31:3–32, 2017.

[4] Nancy Cartwright. *Hunting Causes and Using Them*. Cambridge University Press, Cambridge, 2007.

[5] Clark Glymour, Richard Scheines, and Peter Spirtes. *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science and Statistical Modeling*. Academic Press, Orlando, Florida, 1987.

[6] Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1–12, 1943.

[7] James Heckman and Rodrigo Pinto. Econometric causality: The central role of thought experiments. *University of Chicago*, 2024.

[8] P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960, 1986.

[9] Stephen F. LeRoy. *Causal Inference in Economic Models*. Cambridge Scholars, Newcastle upon Tyne, 2020.

[10] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, 2000.

[11] Hans Reichenbach. *The Direction of Time*. University of California Press, Berkeley, 1956.

[12] Julian Reiss. Time series, nonsense correlations and the principle of the common cause. In Federica Russo and Jon Williamson, editors, *Causality and Probability in the Sciences*. College Publications, 2007.

[13] Herbert A. Simon. Causal ordering and identifiability. In William C. Hood and Tjalling C. Koopmans, editors, *Studies in Econometric Method*. John Wiley and Sons, Inc., 1953.