CAUSAL INFERENCE

Stephen F. LeRoy

University of California, Santa Barbara; July 27, 2019
E-mail address: leroy@ucsb.edu

Contents

Preface

The study of causation has had a checkered history. Although causal inference plays a central role in all scientific work, the topic has undergone extensive analysis only in the philosophy literature. Most applied researchers view the discourses there as largely unrelated to their problems. It is hard not to agree: philosophical analysis centers on detailed examination of ordinary-language usage of causal terms, to the near-exclusion of investigating the role of causation in formal models.

In their theorizing statisticians express reluctance to engage in explicit causal attributions in the absence of controlled experiment. They point to the vagueness and lack of a rigorous foundation that attend many discussions involving causation. In contrast, in their applied work statisticians routinely use causal language. Other concepts, such as probability, that are of at least equally controversial provenance play a central role in theoretical statistics—why the willingness to think hard about the foundations of probability, but not of causation?

Sociologists have opined that correlation can be identified with causation given a sufficient degree of sample stratification, but they did so without specifying a definition of causation under which this proposition could be evaluated. Beginning 50 years ago economists did away with the problem by relabeling predictability as causation despite the existence of readily available examples of the difference between the two.

For the most part the topic has been ignored. This has had the predictable consequence that causal language is used without discipline: analysts debate whether two variables are or are not causally related without any clear shared understanding of what it means for their relation to be causal.

In an earlier literature economists had made headway by introducing the idea of interventions, and situating causal analysis as the determination of the effects of interventions on variables that appear in formal models. However, the intervention typically involved hypothesizing a change in a constant, which, unlike hypothesizing a change in an external variable, amounts to changing the model. Doing so constitutes using the causal question to define the model, as opposed to using the model, taken as defined independently of the causal question to be addressed, to investigate the causal question.

Avoiding this problem involves distinguishing a model's constants from its external and internal variables and associating interventions on internal variables with interventions on the external variables that cause them. Under this protocol the model—the map from external to internal variables—is not altered as part of the analysis of the intervention, so there is some hope of obtaining satisfactory answers.

The idea that coherent causal analysis in the context of a formal model is possible only if the model itself is defined independently of the specified intervention seems obvious. However, it is difficult to find sources in which the matter is discussed explicitly, or in which an effort is made to determine which interventions on internal variables are admissible by the above standard, and why.

In formal analyses of causal relations in linear models two problems must be distinguished: (1) what is the effect on an internal variable of an intervention on an external variable? and (2) what is the effect of an intervention on one internal variable on another? In (1) there is no ambiguity in theory about assigning magnitudes to causal effects, although identification and estimation of causal coefficients can be problematic. The fact that (1) and (2) are often not distinguished creates a presumption that (2) is not substantially different from (1), suggesting that the basic problems of causation analysis are being correctly handled in the existing literature both when the cause variable is external and when it is internal. Here it is argued that this presumption is incorrect. We find that, contrary to the presumption, analysis of causal relations among internal variables involves issues that do not appear when the cause variable is external.

Consider an example. Two internal variables (variables determined by the model) y_1 and y_2 are determined by three external variables (variables taken as given) x_1, x_2 and x_3:

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{0.1}$$

$$y_2 = \alpha_{21}y_1 + \beta_{23}x_3. \tag{0.2}$$

Here there is no doubt that the external variables x_1 and x_2 cause y_1, and also that these and x_3 cause y_2. Most analysts would interpret this model as implying that y_1 causes y_2: an alteration in y_1 of magnitude Δy_1 is induced either by an alteration of $\Delta y_1/\beta_{11}$ on x_1 or an alteration of $\Delta y_1/\beta_{12}$ on x_2, and this causes an alteration Δy_2 on y_2 of $\alpha_{21}\Delta y_1$. But in the reduced form of this model,

$$y_1 = \gamma_{11}x_1 + \gamma_{12}x_2 \tag{0.3}$$

$$y_2 = \gamma_{21}x_1 + \gamma_{22}x_2 + \gamma_{23}x_3, \tag{0.4}$$

with

$$\frac{\gamma_{11}}{\gamma_{21}} = \frac{\gamma_{12}}{\gamma_{22}} \tag{0.5}$$

it appears that y_1 does not cause y_2. This is so even though the same argument for y_1 causing y_2 would seem to apply in the model (0.3)-(0.4) (due to the fact that the coefficients in eqs. (0.3)-(0.4) conform to the restrictions implied by eqs. (0.1)-(0.2)[1]). As this example suggests, reduced forms display causal relations when the cause variable is external, but not when the cause variable is internal.

What is it about the arithmetic operations involved in solving the structural model for its reduced form that alters the causal relation between y_1 and y_2? Clearly some general characterization of causation underlies our willingness to label y_1 as causing y_2 in one model but not the other. Equally clearly this difference has something to do with the fact that in the model (0.1)-(0.2) y_1 and y_2 both appear in one of the equations, whereas in the model (0.3)-(0.4) neither equation contains both y_1 and y_2. Thus at an intuitive level the inference of causation appears to be based on the presence of both the cause variable and the effect variable in at least one of the equations of the model. If causation in formal models is to be explicated along these lines, it seems worthwhile to develop a precise formulation of this idea.

This monograph is aimed at developing a definition of causation under which analysis of interventions does not involve altering the model assumed to characterize the environment. In general the cause variable may be either external or internal, so that the definition includes as a special case the characterization of causal dependence as involving a specification of the cause variable as external.

In many, but not all, cases the formal treatment of causation proposed here agrees with informal discussions. For instance, under the definitions proposed here y_1 does in fact cause y_2 in the model (0.1)-(0.2), but does not do so in (0.3)-(0.4), agreeing with the informal analysis just discussed.

Discussion of theoretical aspects of causation is accompanied by extended analysis of examples. Earlier versions of this material were presented in Cooley-LeRoy [**7**] and LeRoy [**19**], [**20**], [**21**], [**22**], [**23**]. I am indebted to Nancy Cartwright, Christian Gilles, Fei Jia, Judea Pearl, Stephen Salant and Rish Singhania for comments on this monograph and its predecessors.

[1]That is, due to the fact that we have $\gamma_{11} = \beta_{11}$, $\gamma_{12} = \beta_{12}$, $\gamma_{23} = \beta_{23}$, $\gamma_{21} = \alpha_{21}\beta_{11}$ and $\gamma_{22} = \alpha_{21}\beta_{12}$. These imply that we have $\Delta y_2 = \alpha_{21}\Delta y_1$ regardless of whether the intervention that generates Δy_1 is on x_1 or x_2.

Part 1

Theory

CHAPTER 1

Structural Models

A linear *structural model* can be written as

$$Ay = Bx, \tag{1.1}$$

where y denotes the *internal variables* of the model (those determined by the model) and x denotes its *external variables* (those taken as given).[1] Both x and y are vectors. $A = \{\alpha_{ij}\}$ and $B = \{\beta_{ik}\}$ are matrices of constants. A is square and nonsingular, and is normalized by setting the elements of the main diagonal equal to one. The dimensions of x and y are unrestricted. Prior to Chapter 7 attention is restricted to models that are linear in variables, as indicated in eq. (1.1). In most places additive constants are deleted.

Economists associated with the Cowles Commission in the 1940s and 1950s, who first developed the analysis of structural models, distinguished the structural form of a model from its solution form,

$$y = A^{-1}Bx \equiv Gx, \tag{1.2}$$

where $G = \{\gamma_{ij}\}$. Eq. (1.2) is usually called the *reduced form.*

The variables of a model may be either observed by the analyst or unobserved. If x_j is unobserved the coefficients β_{ij} and γ_{ij} ($i = 1, ..., n$, where n is the number of internal variables) are not identified. This being so, for each j such that x_j is unobserved one of the coefficients β_{ij} or γ_{ij} can be set equal to 1 as an arbitrary choice of units. For the present we are not concerned with whether or not variables are observed, but starting in Chapter 5 where the distinction is introduced the convention just specified is adopted.

The Cowles economists viewed the structural form as conveying valuable information not contained in the reduced form. It is difficult to extract from their discussions an account of why this information disappears in going from the structural form to the reduced form, and exactly how it is connected with causation (it was usually associated with identification). In this respect a

[1] In an earlier literature the preferred terms were "endogenous" and "exogenous". However, more recently the latter term was assigned a more specialized meaning in the econometrics literature (Engle, Hendry and Richard [**10**]), so it is avoided here. See Leamer [**17**] for a discussion of the many meanings attached to the term "exogenous". Leamer took the view that exogeneity involves invariance of probability distributions, whereas here probabilistic considerations are not involved in the characterization of variables as internal or external.

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 in Chapters 2 and 4 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. Before presenting this material it is necessary to discuss some treatments of causation that are alternatives to ours.

1. Equality and Causation

Many contemporary applications of structural models, particularly those directed toward graphical analysis of causation, use an alternative specification of structural models, written as

$$y = Ay + x, \tag{1.3}$$

with reduced form

$$y = (I - A)^{-1}x. \tag{1.4}$$

Here A has zeros on the main diagonal. In eq. (1.3) the symbol $=$ is taken to denote causation, with the right-hand side variables of each equation interpreted as directly causing the left-hand side variable. Thus $=$ is an assignment operator, as in computer languages.

Characterizing structural models using eq. (1.3) rather than eq. (1.1) appears to imply that linear operations on the equations of the model are ruled out, due to the assumption that in eq. (1.3) x has no coefficient matrix. Linear operations are consistent with the absence of the coefficient matrix only if x is replaced by a vector each element of which is a function of, in general, all the x_j. This redefinition of the set of external variables changes the model.

The interpretation of $=$ as an assignment operator in eq. (1.3) appears to allow each of two internal variables to cause the other. This occurs when A is not triangular. Some analysts have accepted this implication, but others share the view expressed below that causation is inherently asymmetric. If so it follows that simultaneous determination of sets of internal variables does not fall under the rubric of causation, and therefore must be treated separately. Under the definition to be proposed causation is asymmetric whether or not the coefficient matrix A in eq. (1.1) is triangular. Further, the proposed definition of causation is consistent with existence of sets of variables that are simultaneously determined.

Under the interpretation of $=$ as an assignment operator each equation in eq. (1.3) has a distinct identity: the variables that are direct causes of y_i are all located on the right-hand side of the i-th equation. The characterization

of = as a reflexive, symmetric and transitive operator, on the other hand, implies that it is arbitrary which variable or variables appear on the left-hand side of an equation. In the formulation (1.1), in which = is interpreted as reflexive, symmetric and transitive, all the internal variables are on the left-hand side of the equations and all the external variables are on the right-hand side. This is purely a matter of convenience. The equations of the reduced form are best thought of as defining a single map from an m-dimensional space of external variables to an n-dimensional space of internal variables. With = interpreted as a reflexive, symmetric and transitive operator, writing the model as $y = Ay + x$ does not connect with causation in any obvious way.

The alteration in the meaning of = from its mathematical definition to its interpretation as representing causation has led some writers to express the view that graphical depictions of causal models, which have incorporated the altered meaning of =, are fundamentally different from their algebraic counterparts. Below we will conclude that, contrary to this, there is no reason to avoid using = with its usual mathematical meaning in analyzing causation. This is a major attraction: economic models are derived from primitives by using mathematical calculations in which = is interpreted as a reflexive, symmetric and transitive operator, as opposed to an assignment operator. Proposing to change the interpretation of = upon termination of such derivations creates more problems than it solves. With = preserving its mathematical interpretion in the analysis of causation these problems do not arise.

The objection here is not to the interpretation of the right-hand side variables as causing the left-hand side variables. The problems appear when the analyst starts from that specification and takes it to constitute the definition of causation. The procedure here, in contrast, begins with proposing a definition of causation. Then, starting with the (non-causal) representation of the model as $Ay = Bx$, that definition can be used to reparametrize the model. The reparametrization involves redefining A and B so that the model is of the form $y = Ay + Bx$, with the right-hand side variables directly causing the left-hand side variable, and with A triangular. The parametrization is always feasible if there are no simultaneously-determined variables. This procedure is discussed in the following chapter.

2. Causation Based on *"Ceteris Paribus"*

Angrist and Pischke [**2**] is one of the few recent sources in the economics literature that discusses causation in structural models explicitly and clearly (although, in our view, not correctly). Their account outlines a treatment of causation that is widespread, if not universal, in contemporary economics. If y_j appears on the right-hand side of the structural equation determining y_i, then y_j is defined to cause y_i "ceteris paribus". Here "ceteris paribus" means that other variables in the equation determining y_i, which may include both

internal and external variables, are held constant. The i, j element of A in eq. (1.3) is interpreted as giving a quantitative measure of the causal dependence of y_i on y_j, ceteris paribus. The intervention on y_j is not connected with the external variables that, according to the model, determine y_j.

The ceteris paribus definition of causation relies on the problematic characterization of equality as an asymmetric relation, as discussed in the preceding section. Interpreting the equality symbol instead as having its usual mathematical meaning, as recommended here, implies that a definition of causation based on the "ceteris paribus" condition is not admissible inasmuch as it treats the left-hand side variable differently from the right-hand side internal variables.

Another problem (or, perhaps better, another manifestation of the same problem) is that analyzing causation using the ceteris paribus condition when the "ceteris" includes internal variables implies the existence of functional relations linking purportedly external variables. As a consequence, the causal analysis is conducted using a model different from that actually proposed: holding constant the internal variable effectively redefines it to be an external variable, and one of variables originally labeled as external becomes internal. If such model respecifications are to be avoided it is necessary to disallow causal statements that are conditional on internal variables. Conditioning on external variables is admitted, since replacing an external variable by a constant does not involve redefining an internal variable as external, nor does it introduce functional relations among purportedly external variables.[2] Note here that "conditioning" is defined to consist of replacing internal variables with constants; working with conditional probability distributions, in which "conditional" has a different meaning, causes no problems.

An example will make clear the assertion that holding internal variables constant alters the model. Consider the model

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{1.5}$$

$$y_2 = \beta_{22}x_2 + \beta_{23}x_3 \tag{1.6}$$

$$y_3 = \alpha_{31}y_1 + \alpha_{32}y_2 \tag{1.7}$$

(a graph of this model is found as Example 3.3 in Chapter 3). On the received account of causation, this model implies that y_1 causes y_3, with constant α_{31}, ceteris paribus. Here ceteris paribus means that y_2 is respecified to be external.

From eq. (1.6), this respecification implies that either x_2 or x_3 must become an internal variable. Suppose that x_2 is internal. There results the model

[2]Holding constant internal variables that are functions of a single external variable causes no problem, since doing so is the same as holding constant the external variable.

$$y_1 = \beta_{11}x_1 + \beta_{12}\widehat{x}_2 \tag{1.8}$$

$$\widetilde{y}_2 = \beta_{22}\widehat{x}_2 + \beta_{23}x_3 \tag{1.9}$$

$$y_3 = \alpha_{31}y_1 + \alpha_{32}\widetilde{y}_2. \tag{1.10}$$

Here $\widehat{x}_2$ denotes the variable x_2 now redefined to be an internal variable (if instead x_3 is internal it is replaced by $\widehat{x}_3$ rather than x_2 by $\widehat{x}_2$), and $\widetilde{y}_2$ denotes the variable y_2 now defined to be external. The model (1.8)-(1.10) is perfectly acceptable, and it generates the desired conclusion that y_1 causes (by any reasonable definition) y_3 with causal coefficient α_{31}. However, the model (1.8)-(1.10) is different from the original model—eqs. (1.5)-(1.7): the model as altered has different internal and external variables. Accordingly, the variables in these models will be seen to have different causal relations. Transferring to the alternative model does not constitute an analysis of causation in the original model.

Properly viewed, the statement that one internal variable causes another "ceteris paribus" consists of the assertion that external variables that are not determinants of the cause variable, but not internal variables or external variables that are determinants of the cause variable, are held constant. Reversing the status of external and internal variables is not involved. In the remainder of this monograph the term "causation" is always taken to mean causation that is ceteris paribus in this sense, so the "ceteris paribus" proviso can be omitted.

3. Interventions

We discuss our preferred treatment of causation in most of the remainder of this monograph.

In the Cowles treatment causation is analyzed in terms of interventions. In the usage of the Cowles analysts 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 [**11**]; see also Heckman and Pinto [**13**]). Using a model in this way to analyze causation involves altering the assumed model, with the alteration depending on the causal question that is being asked.

The insistence of the Cowles economists on representing interventions as modifications of structural equations led them away from an alternative 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 in analyzing interventions, and enforces explicit specification of what is held constant under the intervention. 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.

Let us then initially 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 recompute 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.

By designating a coefficient as an external variable rather than a constant the analyst is allowing for interventions on that variable. Designating the coefficients in eq. (1.1) as variables is perfectly acceptable, but doing so implies that the model is bilinear, not linear. These specifications are different. In an equation 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.[3] Also, interventions on external variables do not affect the value of constants.

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. Here we part company from the Cowles economists, who were sometimes unclear about this distinction.[4]

In forecasting exercises the general practice is to specify probability distributions for external variables and then derive the distributions of internal variables by applying the reduced-form equations. Analyzing interventions on such models, in contrast, involves specifying particular realizations of the external variables, as noted above. Contrary to some discussions, there is no contradiction between assigning probability distributions to external

[3]Thus analyses of interventions differ from comparative statics or comparative dynamics exercises, in which changes in constants are acceptable. This is so because the purpose of the latter exercises is to compare different models, not to determine the effects of an intervention in a given model.

[4]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, failure of invariance can only mean that terms specified as constants should instead be modeled as variables. In well-specified models labeling α as a constant means that it does not get changed during the course of the analysis. Therefore α is not a candidate for intervention, and its value is not affected by interventions.
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.

variables in using a model to generate predictions and setting the realizations of these variables to determine effects of interventions. In modeling the price of some crop an analyst could specify that the price depends on the harvest, and the harvest depends on the weather. He or she could then produce a prediction by assuming a probability distribution for weather-related external variables. Equally, the analyst could analyze what the price of the crop would be if the weather were good. The former exercise is a forecast, while the latter constitutes analysis of an intervention. The same model can be used in either application.

CHAPTER 2

Causation

For any internal variable y_i one can define the *external set* of y_i, denoted $\mathcal{E}(y_i)$, as the set of external variables x_j such that the i, j element of the reduced form coefficient matrix is nonzero. $\mathcal{E}(y_i)$ is nonempty; otherwise y_i would be a constant rather than a variable. $\mathcal{E}(y_i)$ is the smallest subset of the set of external variables required to determine the solution value of y_i. In the model

$$y_1 = \beta_{11} x_1 \tag{2.1}$$

$$y_2 = \alpha_{21} y_1 + \beta_{22} x_2 \tag{2.2}$$

the external sets are $\mathcal{E}(y_1) = \{x_1\}$ and $\mathcal{E}(y_2) = \{x_1, x_2\}$.

Two distinct variables z_j and z_i (here z is a variable that may be external or internal), not both external, are *directly connected* if there exists at least one equation in which both appear. Two variables z_j and z_i are *connected* if there exists a *path* (an ordered set of variables) z_j, z_1, ..., z_n, z_i ($n \geq 0$) such that each member is directly connected to its neighbors.

An external variable $x_j \in \mathcal{E}(y_i)$ *directly causes* an internal variable y_i if x_j is directly connected to y_i. An internal variable y_j directly causes an internal variable y_i if y_j is directly connected to y_i and also $\mathcal{E}(y_j) \subset\subset \mathcal{E}(y_i)$. Here "... $\subset\subset$ ..." means "... is a proper subset of ...". This specification constitutes the *proper subset condition* for direct causation. The requirement $\mathcal{E}(y_j) \subset \mathcal{E}(y_i)$ states that every external variable that is in the external set of y_j is also in the external set of y_i, as the intuitive idea of causation suggests. The rationale for this specification is that an intervention on y_j is identified with an intervention on one or more of the variables in its external set, and causation of y_i by y_j requires that this intervention also affect y_i. This condition is a critical element of the treatment of causation here.

The requirement for a proper subset, rather than just a subset, means that there exists at least one variable that is a member of $\mathcal{E}(y_i)$ but not of $\mathcal{E}(y_j)$. This condition distinguishes causation from simultaneous determination, under which we have $\mathcal{E}(y_i) = \mathcal{E}(y_j)$. Simultaneous determination is discussed below. The proper subset condition introduces the asymmetry that, on our view, is a defining element of causation.[1]

[1]Weakening the requirement that $\mathcal{E}(y_j)$ be a proper subset of $\mathcal{E}(y_i)$ to a requirement for a (not necessarily proper) subset would imply that simultaneously-determined variables, having the same external sets, would be variables that directly cause each other, rather

Direct causation is indicated by an arrow: x_j directly causing y_i is denoted $x_j \rightarrow y_i$, and similarly when the cause variable is internal. The set of pairs of variables of which one directly causes the other is termed the *direct causal relation* (direct causation defines a relation but, being intransitive, not an ordering).

The direct causal relation in a structural model can be derived by checking for all pairs x_j, y_i and y_j, y_i whether the conditions for $x_j \rightarrow y_i$ (direct connectedness and $x_j \in \mathcal{E}(y_i)$) and $y_j \rightarrow y_i$ (direct connectedness and $\mathcal{E}(y_j) \subset\subset \mathcal{E}(y_i)$) are satisfied.

We have that x_j *indirectly causes* y_i along the path x_j, $y_1, ..., y_n$, y_i $(n > 0)$ if we have $x_j \rightarrow y_1 \rightarrow ... \rightarrow y_n \rightarrow y_i$. Here $x_j \rightarrow y_1 \rightarrow ... \rightarrow y_n \rightarrow y_i$ is termed a *causal path* that connects x_j and y_i. The indirect causal relation between internal variables is similar: y_j indirectly causes y_i if there exists a causal path connecting y_j and y_i. One variable *causes* another if the two variables are causally connected either directly or indirectly, or both, along one or more causal paths.[2] [3]

Whenever we have $x_j \rightarrow y_i$ there exists a coefficient measuring the effect of x_j on y_i. If we have that x_j causes y_i indirectly along a unique path, then

---

than variables neither of which causes the other, as in the text. Under this specification causation would not be asymmetric (a relation is asymmetric if two variables satisfy the relation in both directions only when they are equal).

[2]Curiously, several contributors to the causation literature appear to have confused the question of whether a variable is external with the unrelated question of whether it causes an internal variable via multiple paths (see, for example, Nakamura-Steinsson [**27**], p. 67).

[3]The fact that variables can be causally related along multiple paths appears to create the possibility that the causal effects cancel. That cannot occur: for x to affect y along canceling paths would contradict the specification that x is in the external set of y.
This treatment of canceling paths has the implication that the formal treatment of causation can diverge from ordinary-language usage. An example, widely discussed in the philosophy literature, specifies that the season determines whether it rains or not. If it rains the pavement is slippery. If it does not rain sprinklers are turned on and, again, the pavement is slippery. The (candidate) internal variable consisting of whether the pavement is slippery is a constant: the pavement is always slippery. The rain does not cause whether or not the pavement is slippery.
For another example, consider a pilot who is able to steer in such a way as to offset exactly the effect of waves on the direction taken by his boat. Ordinary-language usage would have that both the waves and the steering are causes of the boat's direction. In contrast, the causation analysis proposed here would require relabeling the direction variable as a constant, that variable having by assumption no variation to explain. The causal ordering in the model that remains would consist of waves causing steering, and that alone.
It is not clear that the formal treatment of causation proposed here can readily be altered so as to correspond exactly to informal usage. This divergence between formal and informal usage is not a problem; it occurs frequently in scientific applications (consider terms from physics such as "force" and "momentum"). The important question is whether the formal treatment of causation has useful applications that justify the departure from ordinary-language usage. Whether this is so is best considered in the context of the examples discussed below.

the causal effect of x_j on y_i—the reduced-form coefficient of y_i with respect to x_j—equals the product of the coefficients of each of the causal pairs along the causal path. For example, if we have $x_1 \to y_1$ with coefficient β_{11}, and also $y_1 \to y_2$ with coefficient α_{21}, and if there are no other causal paths connecting x_1 and y_2, then the reduced-form coefficient of y_2 with respect to x_1 equals $\beta_{11}\alpha_{21}$. If x_1 and y_2 are connected along multiple paths one of which passes through y_1, then the effect of x_1 on y_2 along that path equals $\beta_{11}\alpha_{21}$, but the total effect of x_1 on y_2 requires consideration of the other paths (as in Example 3.3).

When the cause variable is internal, and the cause y_j is connected to the effect y_i along a unique path, there exists a coefficient associated with the path connecting the two, but that coefficient does not necessarily represent the effect on y_i of an intervention on y_j. That issue is taken up in Chapter 4 (to anticipate, the problem is that y_j causing y_i does not rule out the possible existence of confounding variables even if the causal path connecting y_j and y_i is unique).

Aside from the central assumption here that an intervention on an internal variable is identified with interventions on the elements of that variable's external set, the present formalization of causation is similar to that proposed in some of the Cowles analyses, especially Simon [**34**].

By the definition of causation just given, two internal variables not directly connected can satisfy $\mathcal{E}(y_j) \subset\subset \mathcal{E}(y_i)$ without y_j causing y_i. This occurs when there is no causal path connecting y_j and y_i. An example is the model

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{2.3}$$

$$y_2 = \beta_{22}x_2 + \beta_{23}x_3 \tag{2.4}$$

$$y_3 = \beta_{33}x_3 + \beta_{34}x_4 \tag{2.5}$$

$$y_4 = \alpha_{41}y_1 + \alpha_{43}y_3. \tag{2.6}$$

We have $\mathcal{E}(y_2) \subset\subset \mathcal{E}(y_4)$, but no causal path connects y_2 and y_4 (the paths $y_2 \leftarrow x_2 \to y_1 \to y_4$ and $y_2 \leftarrow x_3 \to y_3 \to y_4$ connect y_2 and y_4 but are not causal paths because the arrows reverse directions). Therefore y_2 does not cause y_4. A graph of this model is displayed in Example 3.4.

It could be argued from the fact that interventions on y_j are associated with interventions on the elements of $\mathcal{E}(y_j)$ that existence of causal paths from each of the elements of $\mathcal{E}(y_j)$ to y_i (which is implied by $\mathcal{E}(y_j) \subset \mathcal{E}(y_i)$) should justify defining y_j as a cause of y_i whenever we have $\mathcal{E}(y_j) \subset\subset \mathcal{E}(y_i)$. Then $\mathcal{E}(y_j) \subset\subset \mathcal{E}(y_i)$ would imply that y_j causes y_i regardless of whether the two are connected along a causal path. Under the revised definition y_2 would cause y_4 in the example.

However, the fact that the alternative definition would allow y_j to cause y_i even though they are not connected along any causal path may be viewed as counterintuitive. In view of this consideration we elect not to adopt the

alternative definition. In any case, this question involves the definition of causal orderings, but not the logically prior and more important question of how to define the direct causal relation, discussed above. Whether to adopt the proposed definition or the alternative appears not to involve any substantive (as opposed to semantic) issues.

Causation is asymmetric and transitive. That being so, it defines a partial ordering.[4] A model's *causal ordering* is the set of pairs z_j, y_i such that z_j causes y_i. The external variables are minimal elements of a causal ordering. The members of a causal ordering, like any partially ordered set, can be arranged in a sequence such that if z_j causes y_i it precedes y_i in the sequence. For causal orderings that are not total orderings, as is generally the case with causation, the converse is not true: z_j may precede y_i in the ordering without z_j causing y_i. When causation is not total there are many sequences that represent the ordering.

The preceding discussion applies to models that have no simultaneous blocs. Simultaneous blocs are discussed below.

The causal ordering is implied by the direct causal relation, but not vice-versa: two models with different direct causal relations can have the same causal ordering. In the model

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{2.7}$$

$$y_2 = \alpha_{21}y_1 + \beta_{22}x_2 + \beta_{23}x_3 \tag{2.8}$$

x_2 directly causes y_2 and indirectly causes y_2 via y_1. If β_{22} equals zero x_2 causes y_2 only indirectly, implying that the two versions have different direct causal relations. Despite this, the two versions have the same causal ordering, represented by x_1, x_2, y_1, x_3, y_2. Graphs of the direct causal ordering of this model are presented in Example 3.5.

The causal ordering associated with the reduced form of a structural model differs from that of the structural model (as seen in the model discussed in the preface). In a reduced form each internal variable is caused by each element of its external set, and only these: internal variables do not cause other internal variables. In a structural model, in contrast, one internal variable can cause another.

The direct causal relation, unlike the causal ordering, plays a central role in causal analysis. This is so because the direct causal relation recognizes the distinction between direct and indirect causation, unlike the causal ordering (as seen in the model just presented).

Uninterpreted examples of the definitions just presented are found in the following chapter. Interpreted examples are discussed in Chapter 10.

---

[4]Strictly, causation is not an ordering because it is irreflexive: variables do not cause themselves. Despite this we will follow established practice and use the term "ordering" in connection with causation.

Asymmetry is vacuously satisfied because, from the proper subset condition, there are no pairs of variables z_1 and z_2 that satisfy both $z_1 \to z_2$ and $z_2 \to z_1$.

1. Structural Models and Causation

The notions of direct causation and causation, where in our treatment the latter is derived from the former, clarify the Cowles assertion that causal relations between internal variables can be determined from structural models, but not reduced forms. As we have seen, one internal variable directly causes another if (1) the two are directly connected, and (2) the proper subset condition is satisfied. The locations of zeros in a model's reduced form matrix make it easy to verify from the reduced form whether the proper subset condition is satisfied. It follows from this that the proper subset condition plays no role in generating the different status of structural and reduced-form models as regards causation.

Whether or not two internal variables are directly connected can be ascertained from the structural form, but not from the reduced form. It follows that even though it is possible to determine from the reduced form whether an external variable causes an internal variable, it is impossible to determine from the reduced form whether the causation is direct or indirect or both. Internal variables are never connected in reduced forms, implying that in reduced forms they are not causally related. It is impossible to determine from the reduced form whether internal variables are causally related in the corresponding structural model. Thus the fact that structural models contain a complete characterization of which variables are directly connected is what distinguishes structural models from reduced forms.

It is seen that the individual equations play a different role in a structural model than those in a reduced form. It is true that the equations of either a structural or a reduced-form model can be combined via (reversible) arithmetic operations without altering the map from external to internal variables.[5] However, such operations do affect direct connectedness relations, and therefore alter the causal ordering of the structural model. It follows that, assuming the analysis involves causal questions, arithmetic operations (involving more than one equation) must be ruled out, as such operations change the causal order, thereby changing the model.

The equations of structural models are seen to have an individual identity that has no counterpart in the equations of reduced-form models. In reduced-form models the individual equations have no role other than to collectively define a map from an m-dimensional space of external variables to an n-dimensional space of internal variables. It is seen that an essential component of the intuitive idea of direct causation is that it involves direct connectedness. This dependence of the idea of causation on direct connectedness was noted in the preface.

[5]The difference between the interpretation of structural vs. reduced-form relations may be related to the question, much discussed in the philosophy literature (see Cartwright [**5**], for example), of whether structural equations are modular. We do not pursue this line.

The fact that structural models and reduced-form models have different causal orderings has relevance for an earlier debate between statisticians, economists and members of other disciplines about the meaning of structural equations. Statisticians and econometricians (see Haavelmo [**11**], Wermuth [**38**] and Pearl [**30**] for discussion) have taken the view that the coefficients of structural models have no clear meaning because they are not connected to the probability distribution of internal variables. It is correct that structural models define the same map from external to internal variables as reduced-form models. However, the fact that the reduced-form model has a different direct causal relation from that of the structural model means that causal information is lost in passing from the structural form to the reduced form. Specifically, moving from the structural form to the reduced form implies loss of information about the direct causal relations in which both the cause and effect variables are internal.

The fact that different structural models may have the same reduced forms, in which case they are observationally equivalent, implies that, as the Cowles economists asserted, causation between pairs of internal variables may not be testable empirically. However, we will see in Chapter 4 that implementation-neutral causation, a refinement of the notion of causation discussed here, does have testable implications.

2. Constructing the Direct Causal Relation

An easily implemented recursive algorithm, rather than inspection of pairs of variables as outlined above, can be used to derive the direct causal relation in structural models that do not have simultaneous blocs. The construction begins with the model written in the structural form $Ay = Bx$. The derivation consists of a series of rounds and sub-rounds. Each round and sub-round consists of identifying one or more of the internal variables as being directly caused by other variables. These internal variables are elements of the set Λ.

Λ initially is the empty set. In rounds after the first, Λ is defined as the set of internal variables identified as effects in earlier rounds. Λ gains members with each round and sub-round.

The first round consists of identifying the equations in which only one internal variable appears. Each of the external variables appearing in each of these equations is designated as a direct cause of the internal variable appearing in that equation. Λ is redefined to consist of the internal variables so identified. The first round has no sub-rounds.

The first sub-round of the second round begins with identification of the equations that contain exactly two internal variables, one of which was designated a member of Λ in the first round. In each of these equations the new internal variable has an external set that strictly contains the external set of the variable in Λ. Therefore the internal variable that is in Λ is designated as a direct cause of the new internal variable. Also, the external variables that

appear in the equation are labeled as direct causes of the new effect variable. Finally, the new variables, one per equation so identified, are included in Λ.

The second sub-round of the second round is the same as the first sub-round—attention is focused on equations that contain two internal variables, one of which is in Λ—with one alteration. The alteration is that Λ as redefined in the first sub-round of the second round replaces Λ as redefined in the first round (the former has more members, implying that new equations now may satisfy the criterion). As in the first sub-round, the other variables in each equation are identified as direct causes of the new internal variables. The third and subsequent sub-rounds are similar. The second round ends when none of the remaining equations meets the requirement that it contain exactly one new internal variable.[6]

The third, fourth and subsequent rounds are similar to the second round except that the identified equations consist of those containing exactly three, four or more internal variables, all but one of which are elements of Λ as defined in earlier rounds. All the variables in each equation are designated as causes of the new effect variable, and that variable is included in Λ. The process continues until all the internal variables are members of Λ.

An example will illustrate the construction. Consider the model

$$y_1 = \beta_{11}x_1 \tag{2.12}$$

$$y_2 + \alpha_{21}y_1 = \beta_{22}x_2 \tag{2.13}$$

$$y_3 + \alpha_{32}y_2 = \beta_{33}x_3 \tag{2.14}$$

$$y_4 + \alpha_{42}y_2 + \alpha_{43}y_3 = \beta_{44}x_4. \tag{2.15}$$

The first round identifies the internal variable y_1 in eq. (2.12) as an effect variable, and includes it in Λ. We have $x_1 \to y_1$. The first sub-round of the second round identifies y_2 in eq. (2.13) as an effect variable, establishes $y_1 \to y_2$ and $x_2 \to y_2$ and adds y_2 to Λ. The second sub-round of the second round identifies y_3 in eq. (2.14) as an effect variable, resulting in $y_2 \to y_3$ and $x_3 \to y_3$, and adds y_3 to Λ. Finally, the third round identifies y_4 in eq.

[6]If for some round or sub-round no equation satisfies the requirement, that round or sub-round is skipped. For example, in the model

$$y_1 = \beta_{11}x_1 \tag{2.9}$$

$$y_2 = \beta_{22}x_2 \tag{2.10}$$

$$y_3 = \alpha_{31}y_1 + \alpha_{32}y_2 + \beta_{33}x_3 \tag{2.11}$$

none of the equations has exactly two internal variables, so the second round is skipped. Models with simultaneous blocs are prone to produce settings in which no equations satisfy the requisite condition. In such models the recursive construction of the causal ordering cannot be implemented. This is why simultaneity was excluded above. In models like that just presented, which does not have simultaneous blocs, skipping rounds does not prevent conclusion of the recursive construction.

(2.15) and adds $y_2 \rightarrow y_4$, $y_3 \rightarrow y_4$ and $x_4 \rightarrow y_4$ to the direct causal relation. This completes the construction.

The calculation just described has an important implication: x is an element of $\mathcal{E}(y)$ if and only if there exists at least one causal path connecting x and y (that $x \in \mathcal{E}(y)$ is implied by existence of a causal path between x and y follows from the definition of a causal path, so the content of the assertion is that the converse, that $x \in \mathcal{E}(y)$ implies existence of a causal path connecting x and y, is also true).

From the definition of external sets, $x \in \mathcal{E}(y)$ is equivalent to the reduced-form coefficient of x in the equation for y being nonzero. These reduced-form coefficients are readily calculated from the conditional causal relation as established by the recursive algorithm just outlined. For example, in the model just set out the first round establishes that the reduced-form coefficient of y_1 with respect to x_1 is β_{11}. With $\beta_{11} \neq 0$ we have $x_1 \in \mathcal{E}(y_1)$. Similarly, the first sub-round of the second round establishes that the reduced-form coefficient of y_2 with respect to x_1 is $\beta_{11}\alpha_{21}$, implying $x_1 \in \mathcal{E}(y_2)$. The same reasoning applied to all the variables shows that the set of pairs x, y such that $x \in \mathcal{E}(y)$ coincides with the set of pairs x, y that are connected by at least one causal path.

The corresponding result for two internal variables y_1 and y_2 is that y_1 causes y_2 if and only if y_1 becomes a member of Λ in an earlier round than y_2 does, implying existence of a causal path that connects y_1 and y_2. This corresponds to Simon's [**34**] formulation, in which one variable causes another if the former is determined in a lower recursive bloc than the latter.

In implementing this algorithm it is usually convenient to reparametrize each equation in each round and sub-round so that the new internal variable in each equation appears on the left-hand side with a coefficient normalized to 1, with all the variables that directly cause that variable appearing on the right-hand side. There results a model written in the form $y = Ay + Bx$. Here A is triangular and has zeros on the main diagonal. A and B here differ from A and B in the structural form with which we began ($Ay = Bx$). The reparametrized version of the model (2.12)-(2.15) is

$$y_1 = \beta_{11}x_1 \tag{2.16}$$

$$y_2 = \alpha_{21}y_1 + \beta_{22}x_2 \tag{2.17}$$

$$y_3 = \alpha_{32}y_2 + \beta_{33}x_3 \tag{2.18}$$

$$y_4 = \alpha_{42}y_2 + \alpha_{43}y_3 + \beta_{44}x_4. \tag{2.19}$$

We will designate a model written in the form $y = Ay + Bx$, with A and B representing the coefficient matrices A and B reparametrized from the original structural model in the manner just described as the *causal form* of the model.

This appears similar to the construction criticized in Chapter 1, but here the causal form representation $y = Ay + Bx$ is derived rather than taken as a

primitive, contrary to the specification in Chapter 1: here we begin with the model in the structural form $Ay = Bx$ and derive a model of the causal form $y = Ay + Bx$, with A and B reparametrized. Under this construction taking the variables on the right-hand side of $y = Ay + Bx$ as directly causing the variable on the left-hand side of each equation is justified even though $=$ is interpreted in its usual mathematical sense rather than as an assignment operator. In contrast, the discussion in Chapter 1 involved starting with $y = Ay + x$ and taking causation as defined by the interpretation of $=$ as an assignment operator combined with an assumption that A is triangular.[7]

3. Simultaneous Equations

As noted above, if a model contains m connected internal variables that all have the same external set ($1 < m \leq n$, where n is the number of internal variables), these variables are *simultaneously determined*. For example, in the model

$$y_1 = \alpha_{12} y_2 + \beta_{11} x_1 \tag{2.20}$$

$$y_2 = \alpha_{21} y_1 + \beta_{22} x_2 \tag{2.21}$$

y_1 and y_2 are simultaneously determined. Variables that are simultaneously determined do not cause each other due to failure of the proper subset condition. We will use the symbol $\leftrightarrow$ to denote simultaneous determination despite the apparent implication of $\leftrightarrow$ that causation runs in both directions. Thus the direct causal relation in the model (2.20)-(2.21) is $x_1 \rightarrow y_1$, $x_1 \rightarrow y_2$, $x_2 \rightarrow y_1$, $x_2 \rightarrow y_2$, $y_1 \leftrightarrow y_2$.

In the reduced-form version of the model,

$$y_1 = \gamma_{11} x_1 + \gamma_{12} x_2 \tag{2.22}$$

$$y_2 = \gamma_{21} x_1 + \gamma_{22} x_2, \tag{2.23}$$

y_1 and y_2 are not simultaneously determined because they are not connected.

If m variables are simultaneously determined there exist m equations that determine these variables as functions of the external variables and the internal variables of the model other than those which are simultaneously determined. These m equations form a *simultaneous bloc*. In the preceding sections of this chapter it was assumed that the models contain no simultaneous blocs (as in most of the examples below).

[7] As noted in the preceding chapter, writing a model in the form $y = Ay + x$ has the problematic implication that performing linear operations on the equations results in a model that cannot be written in the specified format. This problem could be circumvented by including an unrestricted coefficient matrix that multiplies the vector of external variables—$y = Ay + Bx$—but still obtaining causation by interpreting $=$ as an assignment operator. Taking $=$ as an assignment operator, however, implies that causation no longer has anything to do with the proper subset condition, which is not necessarily satisfied among variables labeled as causally related under the proposed characterization of causation. Accordingly, it is not clear how such a respecification is to be justified.

The analysis presented above requires modification in the presence of simultaneous blocs. A causal path was defined above to consist of an ordered set of variables such that each is connected to its neighbors via $\rightarrow$. In the presence of simultaneously-determined variables we generalize to allow the connections to be either $\rightarrow$ or $\leftrightarrow$ (but not all $\leftrightarrow$). This change reflects the fact (or, if one prefers, assumption) that causation is communicated by paths that include simultaneously-determined variables.

To see this, consider the model

$$y_1 = \alpha_{12}y_2 + \beta_{11}x_1 \tag{2.24}$$

$$y_2 = \alpha_{23}y_3 + \beta_{22}x_2 \tag{2.25}$$

$$y_3 = \alpha_{31}y_1 + \beta_{33}x_3, \tag{2.26}$$

in which the model is a simultaneous bloc ($\mathcal{E}(y_1) = \mathcal{E}(y_2) = \mathcal{E}(y_3) = \{x_1, x_2, x_3\}$; also, y_1, y_2 and y_3 are connected). Here the indirect causal relation is readily ascertained by comparing pairs of variables subject to the modification just described so as to allow paths to include simultaneously determined variables. We have that x_1 indirectly causes y_3 via the paths $x_1 \rightarrow y_1 \leftrightarrow y_3$ and $x_1 \rightarrow y_2 \leftrightarrow y_3$ (see Figure 3.6). There exists no path connecting x_1 and y_3 the members of which are all connected by $\rightarrow$, so without the modification in the definition of causal paths it would follow from the definition of a causal path that x_1 does not cause y_3 despite the fact that we have $x_1 \in \mathcal{E}(y_3)$.

Sometimes two internal variables that appear on intuitive grounds to be causally related turn out to be simultaneously determined. In an example from Imbens-Rubin [**16**], Chapter 1, each of a group of patients is characterized by a pair of potential outcomes depending on whether they are treated by drugs or surgery (the potential outcomes approach to causal analysis is discussed in Chapter 8 here). Some patients do better using drugs, others do better using surgery. The doctor knows the potential outcomes of each patient and assigns whichever treatment works better with that patient. Imbens-Rubin characterized four patients by specifying four pairs of potential outcomes, but, as is standard practice in the potential outcomes literature, did not present an explicit model. The formal model is easily specified: the potential outcomes x_0 and x_1 for each agent are external, and the treatment for each patient depends nonlinearly (it is the value of a maximum function) on both potential outcomes for that patient ($t = 1$ if surgery works better than drugs, 0 otherwise). The observed outcome is a function of both potential outcomes and the treatment (specifically, the outcome function is $y^{obs} = x_0 + t(x_1 - x_0)$, where y^{obs} is the observed potential outcome). In this model the external sets for the treatment for each patient and the observed outcome both consist of the potential outcomes for that patient. Accordingly, the treatment and the observed outcome for each patient are both caused by the potential outcomes, implying that their external sets are

the same. Therefore treatment and outcome are simultaneously determined, not causally ordered.[8]

In the presence of simultaneous blocs the analysis of this chapter prior to the present section does not completely apply in several other respects. First, the calculation of reduced-form coefficients from path coefficients is altered in the presence of simultaneously-determined variables. This is easily verified from the simplest example (eqs. (2.20)-(2.21)). Second, as noted above the recursive construction of the direct causal relation cannot be completed in the presence of simultaneously-determined variables. This occurs because at some point prior to the completion of the construction there will exist no equations with only one new internal variable, so the construction cannot be concluded. For example, this occurs in the first round in the model (2.20)-(2.21).

4. Causal Graphs

The easiest way to analyze the direct causal relation, at least with simple models, is to use graphical methods. In advocating the use of graphical methods in analyzing causation we follow the mainstream in causal analysis, notably Pearl [**28**]. However, our use of graphical methods differs from that found in the mainstream tradition. In the received analysis the causal graph is taken directly from the given structural model. When the structural model is explicitly specified the variables on the right-hand side of each equation are identified as direct causes of the left-hand side variable owing to the interpretation of = as an assignment operator. Since this procedure takes the causal ordering as given, causation itself remains undefined. We took issue with this specification in Chapter 1.

In our usage a *causal graph* is a graph that represents the direct causal relation as defined in the preceding sections: if x_j or y_j directly causes y_i ($x_j \to y_i$ or $y_j \to y_i$) the two variables are connected with $\to$ in the graph. Thus the meaning of $\to$ in the graph is the same as in the definition of direct causation. Similarly, simultaneously-determined variables are connected by $\leftrightarrow$ in the graph. Under the alternative form of the algorithm that determines the direct causal ordering from the structural form of a model, the reparametrization results in a graph with arrows pointing from the right-hand side variables of each equation to the left-hand side variable.

Determining from a causal graph whether x_j or y_j causes y_i consists of ascertaining whether there exists a causal path connecting (directly or indirectly) x_j or y_j and y_i by $\to$ or $\leftrightarrow$ (but not all $\leftrightarrow$). This procedure generates a graph in which each internal variable is caused by its ancestors and

[8]Incidentally, this example provides further motivation for associating direct causation with the proper subset condition rather than just a subset condition. Under the latter definition we would have that the treatment and the outcome cause each other, which would be bizarre.

causes its descendants. Parents and children are special cases of ancestors and descendants where the connection is achieved via a single arrow, so that causation is direct. Variables connected with $\leftrightarrow$ are presumably brothers and sisters.

A final point is that one generally cannot begin the analysis of causation with an arbitrarily specified causal graph (or, equivalently, direct causal relation). For example, consider Figure 3.1 with x_3 deleted. The resulting graph displays y_1 as causing y_2 despite the fact that it also indicates that these variables have the same external sets, implying that these variables are simultaneously determined rather than causally ordered.

CHAPTER 3

Examples

The analysis presented in the preceding chapter is illustrated using examples. In each case the associated causal graph is displayed at the end of the chapter.

Example 3.1

The simplest model in which the internal variables are causally ordered is

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{3.1}$$

$$y_2 = \alpha_{21}y_1 + \beta_{23}x_3 \tag{3.2}$$

(this is the example discussed in the preface). The two internal variables are causally ordered: y_1 directly causes y_2. The causal effect of x_1 on y_2 is indirect: the causal coefficient equals $\alpha_{21}\beta_{11}$, which is the product of the direct effect of x_1 on y_1 and the direct effect of y_1 on y_2. The other causal effects are similar.

The reduced form for this model is

$$y_1 = \gamma_{11}x_1 + \gamma_{12}x_2 \tag{3.3}$$

$$y_2 = \gamma_{21}x_1 + \gamma_{22}x_2 + \gamma_{23}x_3. \tag{3.4}$$

In the reduced form y_1 does not cause y_2 because these variables are not connected.

Example 3.2

The standard economist's supply-demand model is

$$y_1 = \alpha_{12}y_2 + \beta_{11}x_1 \tag{3.5}$$

$$y_2 = \alpha_{21}y_1 + \beta_{22}x_2, \tag{3.6}$$

discussed in the preceding chapter. Here each of two equations includes price and quantity (y_1 and y_2) and one external variable. This is the simplest model that contains a simultaneous bloc. The bloc coincides with the model, since y_1 and y_2 are simultaneously determined and are the only internal variables in the model. The model's causal graph can be derived by comparing variables pairwise to determine the existence of direct causation and simultaneous determination.

Note that y_1 does not cause or directly cause y_2, or vice-versa, even though the two are directly connected in both equations.

Example 3.3

In the model

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{3.7}$$

$$y_2 = \beta_{22}x_2 + \beta_{23}x_3 \tag{3.8}$$

$$y_3 = \alpha_{31}y_1 + \alpha_{32}y_2 \tag{3.9}$$

(eqs. (1.5)-(1.7)) the variables y_1 and y_2 have external sets neither of which is a subset of the other, and y_3 has an external set that properly contains the external sets of each of y_1 and y_2. Therefore y_1 and y_2 are neither causally related nor simultaneously determined, but each directly causes y_3. The external variable x_2 affects y_3 via two indirect paths, so the reduced-form coefficient of y_3 with respect to x_2 is $\alpha_{31}\beta_{12} + \alpha_{32}\beta_{22}$.

Example 3.4

In the model

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{3.10}$$

$$y_2 = \beta_{22}x_2 + \beta_{23}x_3 \tag{3.11}$$

$$y_3 = \beta_{33}x_3 + \beta_{34}x_4 \tag{3.12}$$

$$y_4 = \alpha_{41}y_1 + \alpha_{43}y_3 \tag{3.13}$$

y_2 does not cause y_4 because y_2 is connected to y_4 only along paths that are not causal, as noted in the preceding chapter.

Example 3.5

In the model

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{3.14}$$

$$y_2 = \alpha_{21}y_1 + \beta_{22}x_2 + \beta_{23}x_3 \tag{3.15}$$

x_2 directly causes y_2 and indirectly causes y_2 via y_1 (Figure 3.5(a)). If β_{22} equals zero x_2 does not directly cause y_2 (Figure 3.5(b)), implying that the two versions have different causal graphs. Despite this, the two versions have the same causal ordering, as discussed in the text.

Example 3.6

The internal variables y_1, y_2 and y_3 in the model

$$y_1 = \alpha_{12}y_2 + \beta_{11}x_1 \tag{3.16}$$

$$y_2 = \alpha_{23}y_3 + \beta_{22}x_2 \tag{3.17}$$

$$y_3 = \alpha_{31}y_1 + \beta_{33}x_3 \tag{3.18}$$

are determined in a single simultaneous bloc. The accompanying figure shows the causal graph.

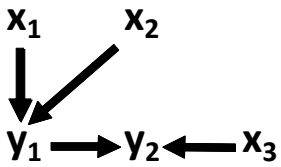

Figure 3.1

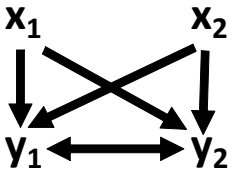

Figure 3.2

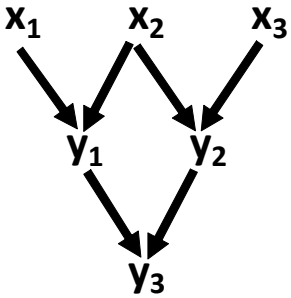

Figure 3.3

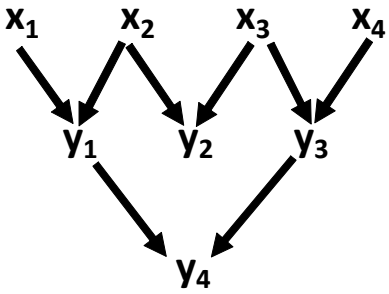

Figure 3.4

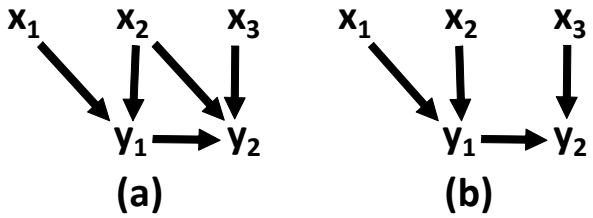

Figure 3.5

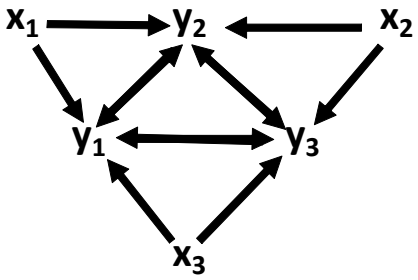

Figure 3.6

CHAPTER 4

Implementation-Neutral Causation

Chapter 2 included description of the calculation of the numerical effect of x_j on y_i when x_j causes y_i. That algorithm does not generally apply when the cause variable is internal. An intervention on an internal variable is generated by any of a set of underlying interventions on the variables in its external set; that being so, even if y_j causes y_i different interventions consistent with a given Δy_j can induce different Δy_i. In the model of Example 3.3 the intervention Δy_1 could have been caused by an intervention of $\Delta y_1/\beta_{11}$ on x_1 or $\Delta y_1/\beta_{12}$ on x_2 (or, of course, a linear combination of these). There results $\Delta y_3 = \alpha_{31}\Delta y_1$ in the first case and $\Delta y_3 = (\alpha_{31} + \alpha_{32}\beta_{22}/\beta_{12})\Delta y_1$ in the second. The question "What i·s the effect of y_1 on y_3?" does not specify which intervention produced Δy_1, leading to the conclusion that the magnitude of the causal effect of y_1 on y_3 is not well defined. Accordingly, when the cause variable is internal there is generally no analogue to the reduced-form coefficient that measures causal magnitude when the cause variable is external.

One could object against this line that in the model of Example 3.3 Δy_1 results unambiguously in an effect $\alpha_{31}\Delta y_1$ on y_3 if y_2 is held constant. We argued in Chapter 1, Section 2 that holding constant an internal variable in this way constitutes an alteration of the model by inducing a functional relation between variables specified as external (in this case x_2 and x_3). Avoiding altering the model leaves us with the conclusion that the effect of y_1 on y_3 in the model of Example 3.3 is in fact inherently ambiguous.

In other cases this ambiguity does not occur. If in addition to y_j directly causing y_i we have that all the interventions that lead to a given value of Δy_j map onto the same value of Δy_i, the effect of Δy_j on y_i does not depend on how Δy_j is implemented (that is, which element(s) of $\mathcal{E}(y_j)$ is (are) intervened upon). In that case causation is *directly implementation neutral.*[1] We refer to the causal relation so defined as *direct IN-causation.* If there exists a path connecting y_j and y_i such that each member directly IN-causes its successor, we have *indirect IN-causation.* The term IN-causation without qualification denotes either direct or indirect IN-causation. If y_j IN-causes y_i we will write $y_j \Rightarrow y_i$. Thus if we have $y_j \Rightarrow y_i$ the analysis of the effect of y_j on y_i is essentially the same as when the cause variable is external: the causal coefficient connecting y_j and y_i is directly analogous to the reduced-form

[1]It appears that the first use of the term "implementation-neutral causation" was by Cartwright [**5**] in her discussion of LeRoy [**20**].

coefficient connecting x_j and y_i. Specifically, if y_j indirectly IN-causes y_i then the associated coefficient equals the product of the coefficients associated with direct IN-causation along the path connecting y_j and y_i.

Note that, as this discussion indicates, under this definition IN-causation represented by $\Rightarrow$, unlike direct causation as represented by $\rightarrow$, is not necessarily direct. This difference in the meanings attached to $\rightarrow$ and $\Rightarrow$ is motivated by the fact that $\rightarrow$ is used to distinguish direct causation from indirect causation. If one variable IN-causes another, in contrast, it does so along a unique path: if y_j directly IN-causes y_k and y_k directly IN-causes y_i then y_j (indirectly) IN-causes y_i (thus IN-causation $\Rightarrow$, unlike direct causation $\rightarrow$, is transitive). As with causation along a given path when the cause variable is external, if we have that $y_j \Rightarrow y_i$ IN-causation is either direct or indirect. However, unlike causation, which can be both direct and indirect along different paths, IN-causation cannot be both direct and indirect. In linear models IN-causal connections, whether direct or indirect, are always associated with coefficients giving a quantitative measure of the causal effect. Accordingly, there is no need to use notation that distinguishes between direct and indirect IN-causation.

If x_j is external and x_j causes y_i we always have $x_j \Rightarrow y_i$, in view of the fact that when the cause variable is external there is no ambiguity about the intervention.

The causal relation between y_1 and y_2 in Example 3.1 is implementation neutral: the effect on y_2 of an intervention of $\Delta y_1/\beta_{11}$ on x_1 (equal to $\alpha_{21}\Delta y_1$) is the same as that of an intervention of $\Delta y_1/\beta_{12}$ on x_2. Note that, in the discussion in Chapter 1 of the ceteris paribus condition, in the model (1.5)-(1.7) y_1 does not IN-cause y_3: if the intervention inducing Δy_1 is on x_1, the effect on y_3 is different from that occurring if the intervention is on x_2. Therefore the constant α_{31} cannot be interpreted as representing IN-causation. The same observation applies to α_{32}.

If an external variable x_1 and an internal variable y_2 are connected along a unique causal path then all the causal links on the path from x_1 to y_2 are IN-causal. For instance, Example 3.1 has $x_1 \Rightarrow y_1$, $y_1 \Rightarrow y_2$ and $x_1 \Rightarrow y_2$. The IN-causal coefficients are β_{11}, α_{21} and $\beta_{11}\alpha_{21}$, respectively.

The *IN-causal ordering* consists of all the pairs $\{x_j, y_i\}$ and $\{y_j, y_i\}$ such that $x_j \Rightarrow y_i$ and $y_j \Rightarrow y_i$. IN-causation will be our primary notion of causation: if y_j causes y_i but not $y_j \Rightarrow y_i$ we do not have enough information about the intervention to characterize its effect on y_i quantitatively.[2]

[2]Sometimes it is useful to work with graphs that depict IN-causation rather than direct causation, although we do not do so in this monograph. IN-causal graphs are constructed in the same way as causal graphs: in an IN-causal graph z_j is connected to y_i with $\Rightarrow$ if z_j directly IN-causes y_i.

1. IN-Causation in Structural Models

IN-causation is most conveniently analyzed using causal graphs. We have $y_j \Rightarrow y_i$ if y_j causes y_i and all the causal paths linking members of $\mathcal{E}(y_j)$ to y_i pass through y_j.[3] [4] If y_j directly causes y_i but $y_j \not\Rightarrow y_i$ there exists at least one member of $\mathcal{E}(y_j)$ that is connected to y_i via at least one causal path that does not pass through y_j.[5] Such variables are *confounding variables*. Existence of confounding variables implies that the effect on y_i of an intervention on y_j is different under different interventions, even those generating the same Δy_j.

In Example 3.3 y_1 does not IN-cause y_3 because of the existence of a path directly connecting x_2 and y_1, and also a path connecting x_2 and y_3 that does not pass through y_1. Thus x_2 is a confounding variable in the causal relation $y_1 \rightarrow y_3$; existence of a confounding variable implies $y_1 \not\Rightarrow y_3$.

In Example 3.4 we have $y_1 \Rightarrow y_4$ despite existence of a path that links x_2 to y_4 but does not pass through y_1. However, that path, while connected, is not causal. Therefore x_2 is not a confounding variable.

IN-causal orderings cannot be deduced from the reduced form even if every pair y_j, y_i that satisfies $\mathcal{E}(y_j) \subset\subset \mathcal{E}(y_i)$ is connected by a causal path, so that the specification displayed in Example 3.4 is ruled out. This is so because by definition confounding variables connect with effect variables along multiple paths, and the reduced form does not distinguish between cause variables that are connected to effect variables along single paths and those connected along multiple paths. In Example 3.5(a) we have that y_1 does not IN-cause y_2 because of the presence of x_2 as a confounding variable. In Example 3.5(b), the reduced form of which has zeros in the same places as that of Example 3.5(a), y_1 does IN-cause y_2.

---

[3]The above representation of IN-causation in terms of graphs in which all paths from the external set of the cause variable to the effect variable pass through the cause variable is described in Woodward [**39**].

[4]The converse is not true: in the model

$$y_1 = \beta_{11} x_1 \tag{4.1}$$

$$y_2 = \alpha_{21} y_1 + \beta_{21} x_1 + \beta_{22} x_2 \tag{4.2}$$

we have that the effect of an intervention on y_1 unambiguously induces an effect of y_2 of $(\alpha_{21} + \beta_{21}/\beta_{11})\Delta y_1$. Therefore we have $y_1 \Rightarrow y_2$ despite existence of a path connecting x_1 and y_2 that does not pass through y_1. This can occur because $\mathcal{E}(y_1)$ is a singleton, implying that there is no ambiguity about the intervention. An interpreted example is found in Chapter 9. We will (without notice) sometimes ignore the case of singleton external sets and identify implementation neutrality with absence of paths that connect effect variables with elements of their external sets without passing through the cause variable.

[5]Recall that in this monograph linearity is assumed (except where noted). It will be observed below that the result just cited does not carry over to nonlinear models.

2. Conditional IN-Causation

It is useful to formulate a notion of causation that can be quantified when we have that y_j causes y_i but not $y_j \Rightarrow y_i$. Such statements are available if we condition on a nonempty proper subset Ψ of $\mathcal{E}(y_j)$, meaning that the variables in that subset are replaced by constants. These statements involve *conditional IN-causation.* Depending on Ψ, we may or may not have that all the variables that are confounders of the causal relation between y_j and y_i are members of Ψ. If so, y_j IN-causes y_i conditional on Ψ, and we will write $y_j \Rightarrow y_i|\Psi$. In linear models that do not contain simultaneous blocs, for any y_j and y_i such that y_j causes but does not IN-cause y_i there exists some Ψ such that we have $y_j \Rightarrow y_i|\Psi$. For example, this necessarily occurs if Ψ consists of all but one of the elements of $\mathcal{E}(y_j)$ and the remaining external variable connects with y_i only via paths that pass through y_j.[6] [7]

Hereafter "y_j IN-causes y_i" without qualification is taken to refer to unconditional IN-causation. Also, "y_j conditionally causes y_i" will be taken to denote conditional IN-causation for some Ψ.

Note the stipulation that the variables being held constant are external. It was observed above that conditioning on internal variables effectively converts these to external variables, and also induces functional relations among variables characterized as external. Therefore doing so constitutes an alteration of the model. No such functional relations among external variables are induced when the variables conditioned on are external.

As with unconditional IN-causation, the existence of conditional IN-causation can be inferred from the direct causal relation, and therefore from the causal graph, but generally not from the causal ordering. In the model

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 + \beta_{13}x_3 \tag{4.3}$$

$$y_2 = \alpha_{21}y_1 + \beta_{22}x_2 + \beta_{23}x_3 + \beta_{24}x_4 \tag{4.4}$$

we have $y_1 \Rightarrow y_2|x_2, x_3$ but not $y_1 \Rightarrow y_2|x_3$, because the confounding variable x_2 is not a member of Ψ, which in the latter case consists of the set $\{x_3\}$. With $\beta_{22} = 0$ we do have $y_1 \Rightarrow y_2|x_3$ despite the facts that the reduced forms of both versions of the model can be written as

$$y_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + \gamma_{13}x_3 \tag{4.5}$$

$$y_2 = \gamma_{21}x_1 + \gamma_{22}x_2 + \gamma_{23}x_3 + \gamma_{24}x_4 \tag{4.6}$$

and the structural models have the same causal ordering.

---

[6] For an example of the failure of this assertion in nonlinear models see the Thistlethwaite-Campbell model [**37**] discussed below; for failure in models that contain simultaneous blocs, see Figure 10.2 below.

[7] If y_j causes y_i and $\mathcal{E}(y_j)$ is a singleton, then $\mathcal{E}(y_j)$ has no nonempty proper subsets, implying that no conditional causation is defined. The Thistlethwaite-Campbell [**37**] model discussed in Chapter 9 is an example in which the treatment variable has an external set that is a singleton. Other instances are found in Chapter 10.

Conditional IN-causation may have no clear interpretation. Having specified that all the variables in $\mathcal{E}(y_j)$ cause y_j, what does it mean to reverse this by holding some of these variables constant? In Example 3.3 we have $y_1 \Rightarrow y_3|x_2$ but not $y_1 \Rightarrow y_3$. With x_2 held constant the intervention is on x_1 alone, suggesting that the causal relation is between x_1 and y_3, not y_1 and y_3. Why then refer to y_1 at all?

In many cases this argument has merit. However, we will see below that in some contexts we are primarily interested in conditional causation between internal variables, not in unconditional causation between external and internal variables. First, often the task is to determine whether the causal relation between two specific variables (such as treatment and outcome) is unconditional or conditional, and not in identifying pairs of variables that are unconditionally IN-causally related. Second, in many cases the relevant external variable may be unobserved, in which case the necessity of an arbitrary normalization of its causal coefficient renders the numerical magnitude of that coefficient meaningless. For example, in Figure 3.5(a) we have $y_1 \Rightarrow y_2|x_2$, so that a change in y_2 is necessarily attributable to an intervention on x_1. If x_1 is unobserved the coefficient measuring its effect on y_2 is meaningless. In contrast, the coefficient associated with the conditional IN-causal ordering $y_1 \Rightarrow y_2|x_2$ is well defined and identified if y_1 and y_2 (but not necessarily x_2, as shown below) are observed. Several examples in which this occurs are discussed below. Third, in some models the relation between the cause and effect variables may be linear even when the equations of the model that determine the cause variable are nonlinear. In that case conditional causal relations are easier to characterize and interpret—conditional IN-causation is quantified by one causal coefficient—than unconditional causal relations (see the discussion of nonlinear models below). Again, examples are found in Chapters 7 and 10.

Part 2

Application

CHAPTER 5

Causation and Probability

In this chapter and its successors we investigate the connection between causation, correlation and regression. Doing so requires introduction of probabilities. Up to now probability distributions have not been discussed, the reason being that probability is not involved in the characterization of causal orderings, which has been our concern up to now. Henceforth it is assumed that external variables are generated according to probability distributions that are taken to be part of the specification of the model. Probability distributions of the internal variables are derived by applying the model to the assumed distribution of external variables.

1. Observed and Unobserved Variables

Also, we have not distinguished variables according to whether they are observed (except in passing). The reason again is that 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. Thus we used x and y to denote external and internal variables whether or not they are observed.

Now we are passing from determining causal orderings to empirical testing of the causal orderings in proposed models and estimation of causal coefficients. IN-causal coefficients are identified statistically only when both the cause variable and the effect variable are observed. Consequently any empirical work related to causation requires that the analyst specify which variables are observed.

We will use capital letters to denote variables observed by the analyst and lower-case letters to denote variables that are unobserved (and when the discussion does not depend on whether or not they are observed, as throughout the preceding chapters). As a simplification, all internal variables Y are assumed to be observed.[1] As noted in Chapter 1, the presence of unobserved external variables implies that the coefficient matrices B or G include ones

[1]For full generality it would be necessary to allow for the existence of unobserved internal variables ("latent variables"). In that case we would have

$$A \begin{bmatrix} Y \\ y \end{bmatrix} = B \begin{bmatrix} X \\ x \end{bmatrix}. \tag{5.1}$$

Most of the examples below do not include latent variables, so we make no formal allowance for them in the notation.

to reflect the normalizations required in the presence of unobserved external variables.[2]

2. Independent External Variables

Except as noted in the next section it is assumed that the external variables, not being connected by the equations of the model, are unconditionally independently distributed (of course, they are generally correlated conditional on internal variables[3]). Independence is a very strong assumption, and most of the difficulty in determining causal orderings comes from the fact that it is usually not obvious which variables are to be taken as external, given that that specification usually entails the assumption that they are independent random variables.[4]

The reason the independence assumption is needed is that without some restriction on the joint probability distribution of external variables we generally find that the coefficients associated with IN-causal orderings are not identified even if the relevant variables are observed. For example, if X IN-causes Y the causation coefficient is not identified if X is correlated with the error term in a regression of Y on X.

If in a proposed model some of the variables provisionally specified as external are observed they may have nonzero sample correlations, which 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 the correlation is ignored. All models are simplifications, and in some settings ignoring apparent correlations among external variables may be an admissible procedure.

However, in most contexts taking that path is unacceptable, insofar as it amounts to assuming away problems that are likely to be of first-order importance empirically. An alternative and usually preferable procedure is

---

[2]If any unobserved external variable appears in more than one structural equation some coefficients of that variable may not equal 1. Specifying all coefficients of such variables equal to 1 would imply the assumption that the external variable has the same effect quantitatively on more than one internal variable. In general this is an unlikely specification given that coefficients depend on the units in which variables are measured, which is arbitrary in the case of unobserved variables.
However, in some situations setting to 1 several coefficients of an unobserved external variable that appears in more than one equation may be acceptable (see Chapter 10, note 3 for an example).

[3]In a well-known example of a correlation induced by conditioning on an internal variable, suppose that actors are famous if they are either good looking or talented. Even if these attributes are independently distributed across the general population of actors, they will be negatively correlated conditional on an actor being famous: an untalented famous actor is necessarily good looking.

[4]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 correctness of the assumption that the treatment variable is statistically independent of all external variables other than the lottery outcome.

to assume that existence of a nonnegligible correlation between two observed variables indicates that those variables are causally related, and therefore cannot both be external. In many applications, such as the private school example discussed below, one has a strong prior belief in the existence of such a causal link. At a minimum, resolving the misspecification involves respecifying one of the two correlated variables as internal.[5] Doing so makes it necessary to introduce a new external variable, presumably unobserved. Also, it is necessary to augment 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 external, and also of all other external variables.[6] Thus all external variables are independent in the reformulated model.

There remains the question of what happens if the analyst is not willing to specify either of two observed correlated variables as external. Analytically this is not a problem: one introduces two new unobserved independently distributed external variables instead of one as above, and relabels the two observed correlated variables as internal. Then each of the observed internal variables is specified to be a linear function of both new external variables. This results in the two observed variables being treated symmetrically: they are determined in a simultaneous bloc by the two unobserved external variables. The consequence of weakening (by relaxing the assumption that one of the correlated variables is external) the specification of the model in this way is that it is more difficult to obtain IN-causation. This

[5]Simpson's Paradox refers to a setting in which the resolution of correlated variables implicitly treated as external is more involved. 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 causation, this appears paradoxical.

The apparent paradox owes to the implicit specification that the treatment variable is external. The resolution is obtained by recognizing that treatment is properly modeled as internal, depending on both gender and an external shock. A formal model incorporating this specification would specify that gender affects the outcome directly as well as via the treatment variable. Accordingly, 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 a causal interpretation. Therefore there is no presumption that it has the same sign as the corresponding correlations for men and women taken separately, which do have a causal interpretation.

[6]Note the contrast with regression theory. The existence of correlation between two explanatory variables causes no problems in estimating coefficients in a bivariate regression. That this is true is exactly the point of multiple regression. Here, in contrast, the task is not to estimate coefficients that may or may not be interpretable causally, but to establish causal orderings and estimate IN-causal coefficients when they are well defined.

construction is discussed in Chapter 10, Section 1. As the Cowles economists led us to expect, models that do not incorporate strong assumptions do not have strong empirical implications.

3. Mean-Independent External Variables

In some applications it is desirable to specify a functional dependence between one internal variable and several explanatory variables all of which are binary (so that they take on one of two possible values, usually specified as 0 and 1). This is particularly so in treatment evaluations: one wants to specify as binary the treatment variable and sometimes also the variables that cause it. If there exist at least two explanatory variables and they are binary and independent, then in linear models the dependent variable will not be binary.

To ensure that the dependent variable is at least potentially binary it is necessary to weaken the specification of independence to mean-independence. Random variable z_2 is *mean-independent* of z_1 if $E(z_2|z_1) = E(z_2)$ for all z_1. The assumption of mean-independence is weaker than full independence, but stronger than uncorrelatedness (except with normal distributions, for which all three are equivalent). Thus weakening the independence assumption to mean-independence preserves the implication of full independence that all correlations in a model's variables reflect the structural equations of the model rather than depending on uninterpreted correlations among external variables.

The result that the range of a function of two variables, one of which is mean-independent of the other, can be binary facilitates empirical investigation via regression of causal relations among binary variables. This is so because the theory of linear regression requires that unobserved explanatory variables be mean-independent of observed explanatory variables; full independence is not required. Below we will present an interpreted example involving binary variables in which one of the external variables is mean-independent of the other, but the two are not fully independent.

It is true that substituting mean-independence for the stronger assumption of full independence may be seen as conflicting with the argument made above that external variables should be free of any probabilistic interdependence whatsoever. Under this argument the proposed weakening of the independence assumption must be disallowed. The force of this argument is not to be minimized. However, we do not take this step; insisting on full independence would complicate the analysis of models involving binary external variables.

An example will make clear the implementation of the mean-independence specification. Suppose that we have

$$y = \delta + \beta x_1 + x_2, \tag{5.2}$$

with x_1 and x_2 specified to be external binary variables.[7] Let x_1 be given by

$$x_1 = \begin{cases} 1 \text{ with probability } 1/2 \\ 0 \text{ with probability } 1/2 \end{cases} . \tag{5.3}$$

The variable x_2 is assumed to be given by

$$x_2 = \begin{cases} 1-\delta \text{ with probability } \delta \\ -\delta \text{ with probability } 1-\delta \end{cases} \tag{5.4}$$

if $x_1 = 0$, and

$$x_2 = \begin{cases} 1-\delta-\beta \text{ with probability } \delta+\beta \\ -\delta-\beta \text{ with probability } 1-\delta-\beta \end{cases} \tag{5.5}$$

if $x_1 = 1$. Here x_2 is mean-independent of x_1, but not fully independent. Eq. (5.2) implies that y is binary, with the distribution

$$y = \begin{cases} 1 \text{ with probability } \delta+\beta/2 \\ 0 \text{ with probability } 1-\delta-\beta/2 \end{cases} . \tag{5.6}$$

Chapter 10, Section 1 includes an application of this use of the mean-independent specification.

[7] Assume $0 < \delta < 1$ and $0 < \delta + \beta < 1$ to ensure that the computed probabilities are admissible.

CHAPTER 6

Causation, Correlation and Regression

In this chapter the connection between causation on one hand and correlation, regression and conditional distributions on the other is discussed.

1. Causation and Correlation

Holland [**14**] 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.

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 connects x and y. Further, y_2 is dependent on x conditional on y_1 if and only if there exists a causal path that connects x and y_2 that does not include y_1.[1] To see this, consider Example 3.1, in which the only causal path connecting x_2 and y_2 passes through y_1. We have

$$F_2(y_2) = F_3((y_2 - \alpha_{21}y_1)/\beta_{23}), \tag{6.1}$$

where F_2 is the cumulative distribution of y_2 conditional on y_1, and F_3 is the cumulative distribution of x_3. The right-hand side does not include x_2, implying that x_2 and y_2 are probabilistically independent conditional on y_1. In Example 3.3, on the other hand, existence of the path $x_2 \to y_2 \to y_3$, which does not pass through y_1, implies that y_3 is dependent on x_2 conditional on y_1.

Two internal variables y_1 and y_2 are probabilistically dependent if and only if for some x there exists a causal path from x to y_1 and also a path from x to y_2, so that the external sets of y_1 and y_2 overlap. Consistent with

[1]It was observed in Chapter 1 that conditioning on an internal variable constitutes an alteration of the model. Conditioning there, referring to respecifying internal variables as external, had no connection with probabilities. Here, in contrast, we are using "conditioning" in its probability sense. Obviously taking expectations conditional on internal variables does not constitute an alteration of the model.
It would be best to designate these dissimilar operations—one involving causation, the other correlation—by different names. Regrettably, both usages of "conditioning" are well entrenched, although rarely distinguished. The same point applies to the ambiguous term "holding constant". The context here will always make clear which meaning is intended.

$\mathcal{E}(y_1)$ having a nonempty intersection with $\mathcal{E}(y_2)$, one of these variables may or may not cause the other, or they may be simultaneously determined. Thus absence of statistical dependence implies absence of causation, but the presence of statistical dependence does not imply causation. 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.[2]

2. Covariances and Regressions

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

$$y_i - E(y_i) = \sum_j \gamma_{ij}(x_j - E(x_j)), \tag{6.2}$$

where j indexes the variables in $\mathcal{E}(y_i)$, and the γ_{ij} are elements of the reduced form coefficient matrix. The assumptions that the x_j are independent and have finite second moments implies that we have

$$\gamma_{ij} = \frac{\text{cov}(y_i, x_j)}{\text{var}(x_j)}. \tag{6.3}$$

Similarly, if we have $y_j \Rightarrow y_i$ and IN-causation is direct, α_{ij} satisfies

$$\alpha_{ij} = \frac{\text{cov}(y_i, y_j)}{\text{var}(y_j)}. \tag{6.4}$$

To see this, consider the model (3.14)-(3.15), which has $y_1 \Rightarrow y_2|x_2$ but not $y_1 \Rightarrow y_2$. Simplifying by setting the means of x_1, x_2 and x_3 equal to 0, we have

$$\text{cov}(y_1, y_2) = E(y_1 y_2) = \alpha_{21}\text{var}(y_1) + \beta_{12}\beta_{22}\text{var}(x_2). \tag{6.5}$$

In the special case $\beta_{22} = 0$ the model (3.14)-(3.15) reduces to the model of Example 3.1, in which we have $y_1 \Rightarrow y_2$. From eq. (6.5), the regression coefficient of y_2 on y_1 equals α_{21} in that case. Thus when we have $y_1 \Rightarrow y_2$ not only is the causal coefficient well defined, but also it coincides with the population regression coefficient of y_2 on y_1. It follows that, assuming that y_1 and y_2 are observed, the causal coefficient can be consistently estimated by least squares.

These results have implications for the most basic regression theory. Textbook expositions emphasize that consistent regression estimators are available only if the regressor is assumed to be uncorrelated with the error. The uncorrelatedness result is true by construction for any two variables

[2]In the philosophy literature this assertion is the "principle of the common cause". It is correct in our setting. Philosophers have debated whether it is true in general. See, for example, Reiss [**31**].

(with finite second moments) regardless of the causal relation between the dependent and explanatory variables, or lack thereof.[3] Let y_3 be the error term in a regression of y_2 on y_1. We have

$$y_3 = y_2 - \frac{\mathrm{cov}(y_2, y_1)}{\mathrm{var}(y_1)} y_1, \tag{6.6}$$

from which it is immediate that $\mathrm{cov}(y_3, y_1) = 0$. It is hard to imagine what it means for an regression error to be correlated with a regressor; if it is correlated with a regressor it is not a regression error. However, for the present purpose the question is not whether a regression error is uncorrelated with the regressors, but whether the error in the causal equation is correlated with the regressors (equivalently, whether the error in the causal equation coincides with the error in the regression), which it may or may not be. This issue is discussed in the following section.

3. Regressions Based on Causal Equations

Under the reparametrization outlined on p. 18 the causal form of a structural model displays each internal variable of the model as directly caused by each member of a subset of the external and internal variables of the model (a different subset for each internal variable, of course). Depending on the model, the causal coefficients may represent either unconditional or conditional direct IN-causation.

Consider an equation of the causal form in which at least one of the right-hand side variables is observed. That equation is a candidate to be viewed as a regression. In that candidate regression the effect variable is the dependent variable and the observed causal variables are the explanatory variables. It is assumed that all the observed variables that appear on the right-hand side of the causal equation, and only these (pending discussion below), are entered as explanatory variables. The error term in the candidate regression is the counterpart of the unobserved variables in the causal equation.

Whether or not the candidate regression is interpretable as a regression depends on the error term—that is, the unobserved variables (which are external; recall that we are assuming that all internal variables are observed)—that appear in the causal equation for Y_n. The critical determinant is whether the union of the external sets of these variables is or is not disjoint from the union of the external sets of the observed variables in that equation. If so, the error in the candidate regression is independent of the explanatory variables. That being so, the coefficients in the univariate or multivariate regression of Y_n on the explanatory variables coincide

[3]This point was made by Angrist and Pischke [**3**], p. 128 in their critique of prevailing instruction in econometrics. These authors did not indicate what conclusion could be drawn from this observation, contenting themselves with the statement that "it's hard to see how this statement [that errors must be assumed to be uncorrelated with regressors] promotes clear thinking about causal effects."

with the causal coefficients, so the candidate regression is in fact a regression (mean-independence of the error from the explanatory variables is the defining attribute of a regression). This regression is termed the *associated regression* for that causal equation. Least squares provides consistent estimates of the causal coefficients in each causal equation that has an associated regression.

The causal coefficients of the observed variables in the associated regression may calibrate either unconditional or conditional causation. For an example of the latter, consider the causal model

$$Y_1 = x_1 + \beta_{12} x_2 \tag{6.7}$$

$$Y_2 = x_2 + x_3 \tag{6.8}$$

$$Y_3 = \alpha_{31} Y_1 + \alpha_{32} Y_2 + x_4, \tag{6.9}$$

which coincides with Example 3.3 except for inclusion of the unobserved external variable x_4 in eq. (6.9). Neither Y_1 nor Y_2 IN-causes Y_3 unconditionally, but each does IN-cause Y_3 conditional on x_2. Here a bivariate regression of Y_3 on Y_1 and Y_2 is the associated regression for the causal equation (6.9), implying that the regression coefficients of Y_1 and Y_2 coincide with the conditional causal coefficients α_{31} and α_{32}.

The causal coefficients that appear in a given equation of a causal model determine the magnitude of direct causation only. The same variable that directly causes an effect variable may also cause it indirectly along a different path; these indirect links are not represented in the causal regression. Examples in which regression coefficients measure the direct effect of a cause variable on an effect variable, but not its indirect effect, are found in Chapter 10.

On the other hand, the union of the external sets of the explanatory variables may include one or more of the unobserved variables in the regression. If so, the unobserved term in the causal equation is not mean-independent of the explanatory variables in the candidate regression. Equivalently, the unobserved term in the causal relation does not coincide with the corresponding regression error, which by construction is mean-independent of the explanatory variables. Also equivalently, the regression coefficients do not coincide with the causal coefficients, implying that least-squares estimation does not produce consistent estimates of causal coefficients. In that case the causal equation under discussion does not have an associated regression. Estimation techniques other than least squares, such as instrumental variables or regression discontinuity (both discussed below), are needed to resolve this problem.

For an example in which the candidate regression is not an associated regression, suppose that eq. (6.9) is replaced by

$$Y_3 = \alpha_{31} Y_1 + \alpha_{32} Y_2 + \beta_{32} x_2 + x_4, \tag{6.10}$$

so that x_2 directly causes all three internal variables. Here the candidate regression involves the same explanatory variables, Y_1 and Y_2, as under eq. (6.9). However, it 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. Accordingly, the regression coefficients of Y_1 and Y_2 differ from α_{31} and α_{32}, the coefficients measuring the effects of Y_1 and Y_2 on Y_3 conditional on x_2.

In the preceding discussion it was assumed that the explanatory variables in the regression coincide with the observed variables in the causal equation. If, contrary to this, the regression is specified to include variables that do not appear in the causal equation, or does not include variables that do appear in the causal equation, the candidate regression is disqualified as an associated regression. Accordingly, least-squares will generate inconsistent estimates of causal coefficients. For example, assume that eq. (6.9) is replaced with

$$Y_3 = \alpha_{31} Y_1 + x_4, \tag{6.11}$$

so that the regression of Y_3 on Y_1 alone now is the associated regression for the causal equation (6.11) (Figure 6.1): the single regression coefficient coincides with the causal coefficient α_{31}. The fact that Y_1 and Y_2 have overlapping external sets implies that they are correlated. It follows that the regression coefficient of Y_3 on Y_1 in the bivariate regression that also includes Y_2 as an explanatory variable differs from the corresponding coefficient in the univariate regression of Y_3 on Y_1. Similarly, the regression coefficient of Y_2, being generally nonzero, is not interpretable as a causal coefficient. Analysis of the opposite case, in which the candidate regression deletes an observed variable that appears in the causal equation, is similar.

If the causal equation that determines Y_n has an associated regression, and if that regression contains a single observed explanatory variable, that variable IN-causes Y_n. However, we have seen that if there exist two or more internal explanatory variables in the associated regression it is possible that neither of these IN-causes Y_n, consistent with existence of an associated regression. An important implication is that coefficients associated with conditional causation can sometimes be estimated consistently by least squares even when unconditional causation fails. The model presented above exemplifies this.

4. Instrumental Variables

In the model

$$Y_1 = \alpha_{11} X_1 + \beta_{12} x_2 \tag{6.12}$$

$$Y_2 = \alpha_{21} Y_1 + x_2 + x_3 \tag{6.13}$$

we have $Y_1 \Rightarrow Y_2 | x_2$, with causal coefficient α_{21}. However, the regression of Y_2 on Y_1 is not an associated regression for the causal relation (6.13). This

is so because the unobserved term in eq. (6.13) is not mean-independent of Y_1, in view of the fact that x_2 is a determinant of Y_1, from eq. (6.12). Therefore the coefficient in the least-squares regression of Y_2 on Y_1 does not equal α_{21}. We conclude that least-squares regression does not produce a consistent estimator of the coefficient associated with $Y_1 \Rightarrow Y_2|x_2$.

However, α_{21} can be estimated, implying that the effect of Y_1 on Y_2 conditional on x_2 can be evaluated. From the fact that a reduced-form coefficient equals the product of the coefficients associated with the direct causal relations along the causal path, we have

$$\alpha_{21} = \frac{\gamma_{21}}{\beta_{11}} = \frac{\text{cov}(X_1, Y_2)/\text{var}(X_1)}{\text{cov}(X_1, Y_1)/\text{var}(X_1)} = \frac{\text{cov}(X_1, Y_2)}{\text{cov}(X_1, Y_1)}. \tag{6.14}$$

The rightmost term in eq. (6.14) is recognized as the population counterpart of the instrumental variables estimate 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_2 and Y_1, is assumed to be unobserved.

As this example indicates, a valid instrument must be an observed element of the external set of the cause variable, and must be related to the effect variable only through paths that pass through the cause variable.

See Chapter 10, Example 2 for this use of instrumental variables estimation.

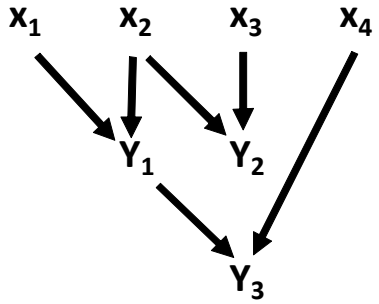

Figure 6.1

CHAPTER 7

Extensions

Our focus so far has been on static linear models. As seen in preceding and succeeding chapters, extensive results are available for that case. Frequently, however, more general settings are required. In this chapter we briefly discuss what happens under more general specifications.

1. Nonlinear Models

Sometimes the logic of the model being analyzed forces a nonlinear specification. A common example is a model of execution by firing squad, in which case the victim dies if any of the executioners hits his target. This causation is most easily modeled using a maximum function, which is inherently nonlinear. It is worthwhile discussing, if only in passing, to what extent the analysis of linear models applies in the nonlinear case.

With nonlinear models maps from external to internal variables are well defined only when the structural equations have a solution that is unique (in the linear case existence and uniqueness are guaranteed by the assumption that A in eq. (1.1) is nonsingular, but no such simple condition carries over to nonlinear settings). Therefore it is necessary to assume separately that models are such that the solution exists and is unique. We make this assumption throughout.

Even if the model has a unique solution, in nonlinear settings the effect on an internal variable of an intervention on an external variable generally depends on the assumed values of other external variables. Further, in analyzing interventions in nonlinear models it is necessary to specify the baseline value of each external variable and its value under intervention individually—the difference between the two is not sufficient to determine the effect of the intervention, contrary to the case in linear models. Accordingly, the constant that measures causation in the linear case (assuming implementation neutrality) is replaced in the nonlinear case by an internal variable that depends on all the relevant external variables.

In a nonlinear setting the definition of external sets must be modified to allow for the altered form of the dependence of effects on causes. We define the external set of y_i as the set of external variables each of which affects y_i for some, but not necessarily all, values of the other members of the set. By that standard all the members of the firing squad cause the outcome, because each member's shot determines whether the vectim is killed in the event that all the other members' shots miss the victim. This definition implies that

some specifications of the baseline and the intervention result in a causal effect of zero: a member's shot does not affect the outcome if any of the other members hits his target. In contrast, in linear settings if an external variable causes an internal variable the effect of an intervention is never zero. The definition just presented may appear arbitrary by the standards of common usage, but it preserves the interpretation of the external set as the smallest subset of the external variables that allows a complete determination of the effect variable.

We noted in Chapter 5 that in problems in which the treatment variable and also the external variables that cause the treatment variable are binary it is possible to preserve linearity, although at the cost of weakening the assumption that the external variables are independent. If instead the function determining the treatment variable is nonlinear, then the treatment variable may be binary even though the external variables are fully independent and may or may not be binary.

Given the modification specified above in the definition of external sets, we can carry over from the linear case the definition of the direct causal relation: x_1 directly causes y_2 if $x_1 \in \mathcal{E}(y_2)$ and x_1 is directly connected to y_2. Correspondingly, y_1 directly causes y_2 if y_1 is directly connected to y_2 and $\mathcal{E}(y_1) \subset\subset \mathcal{E}(y_2)$. Similarly, the definition of IN-causation is unchanged; a variable IN-causes an internal variable if it causes that variable, and if also all interventions consistent with a given change in the cause variable generate the same change in the effect variable.

Frequently it is convenient to specify models in which the treatment, being binary, is generated by a nonlinear function, but the relation between treatment and outcome is linear and IN-causal. For example, we can specify that y_1 is generated by a nonlinear equation such as

$$y_1 = \begin{cases} 1 \text{ if } \beta_{11}x_1 + \beta_{12}x_2 \geq 0 \\ 0 \text{ otherwise} \end{cases} . \tag{7.1}$$

We have that x_1 and x_2 IN-cause y_1, with the effect of an intervention on x_1 depending separately on both the baseline value of x_1 and its value under intervention (as opposed to depending on only the difference between the two, as in the linear case), and also on x_2, due to the nonlinearity of eq. (7.1). Assume that y_2 is determined by

$$y_2 = \alpha_{21}y_1 + \beta_{23}x_3. \tag{7.2}$$

In the model (7.1)-(7.2) we have $y_1 \Rightarrow y_2$, with the causal coefficient α_{21}. The fact that IN-causation of y_2 by y_1 is representable by a single coefficient here reflects the fact that even though the model as a whole is nonlinear, eq. (7.2) is linear. This model has the causal graph shown as Figure 3.1, coinciding with the causal graph that represents the corresponding linear model.

In Chapter 4 it was noted that in linear models if we have $y_1 \rightarrow y_2$ and $\mathcal{E}(y_1)$ is a singleton, then y_1 IN-causes y_2 even if there exists a path connecting the single variable constituting $\mathcal{E}(y_1)$ and y_2 that does not pass through y_1. This result does not necessarily carry over to nonlinear models. The reason is that in linear models the map from y_1 to y_2 can be inverted, implying that when $\mathcal{E}(y_1)$ is a singleton specifying Δy_1 is equivalent to specifying the intervention itself. It follows that the effect on y_2 of the intervention on y_1 is unambiguous even though it involves a path that does not pass through y_1. In the nonlinear case, however, $y_1 \rightarrow y_2$ may not be invertible even when $\mathcal{E}(y_1)$ is a singleton, implying that specifying Δy_1 may not give enough information about the intervention to determine its effect on y_2. An example in which this occurs is discussed in Chapter 9.

2. Parameters

In Chapter 1 we referred to models of the form $Ay = Bx$ as linear, implying that the coefficient matrices A and B were characterized as constants. If so, the members of A and B are not subject to intervention. But coefficients can also be variables in the mathematical sense, so we need terminology that distinguishes coefficients that are specified to be mathematical variables from those specified to be constants. Coefficients that are variables in the mathematical sense are termed *parameters*.[1] Models that take the form (1.1) in which the elements of A and B are treated as parameters are bilinear in variables, not linear. It follows that much of the discussion above of nonlinear models applies when coefficients are specified as parameters rather than constants.

It is essential that model-builders specify whether coefficients are to be interpreted as constants or parameters. If parameters, they must be designated as external or internal, just as with other variables, and model specification must include the map from external parameters to internal parameters as well as that from parameters and external variables to internal variables.[2] The reduced form of a nonlinear model, if it exists (meaning if the model's solution exists and is unique), consists of functions mapping external variables, including external parameters, to internal variables and internal parameters. Sometimes the terms "deep parameters" and "shallow parameters" are adopted in place of external and internal parameters. Thus the model consists of functions relating shallow parameters to deep

[1] In classroom lectures Lawrence Klein defined parameters as "constants that vary". We graduate students were amused, as Klein intended; with hindsight we should have been puzzled.

[2] Some settings incorporate the specification that constants are linked by functional relations. For example, if external variables are related by mean-independence rather than full independence the constants that describe their respective distributions are linked by functional restrictions (as with β and δ in Chapter 5, Section 3). Another example is found in the following section. These constants need be relabeled as parameters only if interventions on them, or affecting them, are introduced.

parameters, and also functions relating parameters to variables that are not parameters. External sets include deep parameters as well as external variables that are not parameters. Otherwise the analysis of causation is the same whether or not the model includes parameters.

The effects on internal variables of interventions on shallow parameters are generally not defined due to failure of IN-causation. This is the Lucas [**24**] Critique. An example is discussed in the following section.

3. Multidate Models

Causation analysis is essentially the same in multidate models as in the single-date models studied up to now. Demonstrating this requires extending the terminology. In many discussions involving multidate models the term "variable" is used in reference to an n-tuple or sequence of mathematical variables indexed by time (assuming discrete time), and also to the elements of the sequence. To avoid this ambiguity we will call such an n-tuple or sequence a *process,* and reserve the term "variable" for single mathematical variables. If y is a process, then, the elements y_t of the process are termed variables. Thus in multidate models we have two types of mathematical variables: elements of processes and parameters.

In many discussions causation is represented as being inextricably linked to time: causes are viewed as necessarily preceding effects in time. This identification dates at least to Hume [**15**]. If so, static multidate models—models in which internal variables have external sets that include future-dated variables—are ruled out. Insisting on this point is a mistake: there is no justification for a doctrinaire injunction against multidate models in which all information is revealed at a single initial date. Analysts typically represent revelation of information by assuming that processes are measurable with respect to a filtration. The definition of a filtration allows the σ-algebra that represents currently available information to be the same at every date. Such models allow internal variables to depend on future-date external variables. Under this treatment multidate models without gradual revelation of information are legitimate (if usually unrealistic), a special case of dynamic models. Thus there is no intrinsic link between causation and time.

Most multidate models involving causation specify that information strictly increases over time, so that the σ-algebra representing information at any date is a proper subset of each σ-algebra at a later date. In such models the dating convention is usually that the time subscript of each variable is the earliest date at which that variable is measurable. It follows that causes precede or are contemporaneous with effects: the external set of $y_{j,t+1}$ contains external variables that are not measurable at t, so we cannot have $y_{j.t+1} \rightarrow y_{i,t}$.

The simplest efficient markets finance model will illustrate analysis of causation in multidate models, and will indicate the consequences of specifying coefficients as constants vs. parameters. The model relates three processes: two internal processes consisting of dividends y^d and stock prices y^p, and an uninterpreted external process x. The variables constituting these processes are denoted y_t^d, y_t^p and x_t. Suppose that dividends follow a first-order autoregression:

$$y_0^d = x_0 \tag{7.3}$$

$$y_{t+1}^d = \alpha_{dd} y_t^d + x_{t+1},\ t = 0, 1, ... \tag{7.4}$$

($|\alpha_{dd}| < 1$), where x is a process consisting of independently and identically distributed random shocks. Stock prices obey

$$y_t^p = \delta E_t(y_{t+1}^p + y_{t+1}^d), \tag{7.5}$$

so that δ, the discount rate, is the reciprocal of the expected gross rate of return, assumed to be the same at every date. Here E_t denotes mathematical expectation conditional on information available at t. For now the coefficients α_{dd} and δ are specified to be constants, not parameters. Solving the model (and ruling out bubbles) results in

$$y_t^p = \alpha_{pd}\ y_t^d, \tag{7.6}$$

where α_{pd} satisfies

$$\alpha_{pd} = \frac{\delta \alpha_{dd}}{1 - \delta \alpha_{dd}}, \tag{7.7}$$

by an easy calculation. The causal graph for this model for dates 0, 1 and 2 is shown in Figure 7.1.

The same model can be written either as (7.3)-(7.4)-(7.5), with coefficients δ and α_{dd}, or as (7.3)-(7.4)-(7.6), with coefficients α_{dd} and α_{pd}. These models are equivalent due to the fact that the coefficients satisfy eq. (7.7). The equivalence between these two parametrizations depends on the specification of the coefficients as constants rather than parameters. The external set of y_t^d is $\{x_t, ..., x_0\}$. The external set of y_t^p is the same, implying that y_t^d and y_t^p are simultaneously determined.

Suppose now that δ, α_{dd} and α_{pd} are specified as parameters rather than constants, and assume that α_{dd} and δ are external, implying that α_{pd} is internal. This specification corresponds to the usual presentation of this model: agents' rate of time preference and the autocorrelation coefficient of dividends determine the equilibrium dividend yield. Then the external set for α_{pd} is $\{\delta, \alpha_{dd}\}$, and the external sets for y_t^d and p_t^d are $\{\delta, \alpha_{dd}, x_t, ..., x_0\}$, so that, again, y_t^d and p_t^d are simultaneously determined.

Consider the effect of an intervention on the deep parameter δ from δ^b in the baseline to δ^i under intervention. The effect on α_{pd} is given by

$$\alpha^i_{pd} - \alpha^b_{pd} = \frac{\delta^i \alpha_{dd}}{1 - \delta^i \alpha_{dd}} - \frac{\delta^b \alpha_{dd}}{1 - \delta^b \alpha_{dd}}, \tag{7.8}$$

from eq. (7.7). As discussed in the preceding section, from the nonlinearity of eq. (7.7) the left-hand side cannot be written as a function of $\delta^i - \delta^b$; the values of these variables must be specified separately. Also, the effect on α_{pd} of the intervention on δ depends on α_{dd}, which is not involved in the intervention. The analysis of the intervention on the elements of the process y^p is similar.

From eq. (7.7) there is no ambiguity about hypothesizing interventions in δ or α_{dd} on the internal parameter α_{pd} or any of the internal variables. However, the effect of α_{pd} on the internal variables is ambiguous due to failure of IN-causation: an intervention on α_{pd} could reflect an intervention on either δ or on α_{dd}, and these have different effects on the internal variables. This is the Lucas Critique [**24**] referred to in the preceding section: unconditional effects of interventions on shallow parameters may not be well defined due to failure of IN-causation. The causal graph for this model with coefficients interpreted as parameters rather than constants is shown in Figure 7.2(a).

One might prefer to solve out shallow parameters. In the present model the only shallow parameter is α_{pd}. The causal graph with α_{pd} solved out is shown in Figure 7.2(b). Although formally the causal graphs differ according to whether or not shallow parameters are solved out, substantively the two models are the same.

4. The Causal Markov Condition

An important tool that has been used in modeling networks is the causal Markov condition, which makes possible empirical testing of causal orderings. The causal Markov condition, as formulated by Spirtes, Glymour and Scheines [**36**], for example, states that every variable of a model is probabilistically independent of all variables other than its descendants and parents, given its parents.

The status of the causal Markov condition is ambiguous. In places it is treated as a derivable implication of the other assumptions defining a model. Alternatively, it is treated as an axiom separate from other assumptions specifying the structure of the model. Or it may be regarded as part of the definition of a Bayesian network; this presumption usually involves sidestepping the question of whether a causal graph is a Bayesian network. Finally, it is sometimes treated as a substantive proposition that can be evaluated on philosophical grounds (see Hausman and Woodward [**12**] for extended discussion).

The most obvious problem here is that, from elementary probability theory, two random variables are always independent conditional on one of them. It follows from the fact that any internal variable is a deterministic function of its parents that we can certainly delete "and parents" from the definition of the causal Markov condition. This point was noted by Hausman and Woodward. A slightly less obvious point is that, again because any internal variable can be written as a deterministic function of its parents, any internal variable is independent of all variables, including its descendants, conditional on its parents. It follows that under that reading the causal Markov condition as just stated is valid, but trivially so.

These points depend on the definition adopted in this monograph of parents as the set of all variables that directly cause the variable in question. Shocks, being random variables, are included in the set of parents of the variables they cause. In treatments of causation one often sees discussions that presume that error terms are not causal parents. However, no guidance is given as to the basis for distinguishing variables that are causal parents from those that cause a variable but are not counted among its parents. Variables characterized as errors are unobserved, but there is no apparent justification for denying their status as causal parents for this reason: the definition of causal orderings presented above does not depend on whether variables are observed. Hausman and Woodward explicitly posited existence of causal variables that are not included in the model under consideration and therefore do not qualify as parents. Presumably these appear as variables in some unspecified meta-model. It is not explained what purpose is served by introducing this complication.

There exist propositions similar to the causal Markov condition as formulated above that are correct and nontrivial, and are easily derived in the framework set out here. We set forth one such proposition: if we have $y_1 \Rightarrow y_2$, then y_2 is independent of any ancestor of y_1, conditional on y_1. This follows from the result from Section 1 of the preceding chapter; if y_3 is an ancestor of y_1 and is correlated with y_2 conditional on y_1, then there exists a path connecting y_3 and y_2 that does not pass through y_1. If so, any member of $\mathcal{E}(y_3)$ is a confounding variable, implying $y_1 \not\Rightarrow y_2$.

The proposition just stated has a partial converse: if $y_j \rightarrow y_k \rightarrow y_i$ and y_j is independent of y_i conditional on y_k, then we have $y_j \Rightarrow y_i$. The fact that we have $y_j \rightarrow y_k \rightarrow y_i$ implies that there exist paths connecting y_j and y_i. The fact that y_j and y_i are independent conditional on y_k implies that all causal paths connecting y_j and y_i pass through y_k. This is the definition of IN-causation.

Existence of such theoretical results implies that, as part of a compound hypothesis, IN-causation is testable. The availability of a partial converse suggests that in some settings the test may have high power.

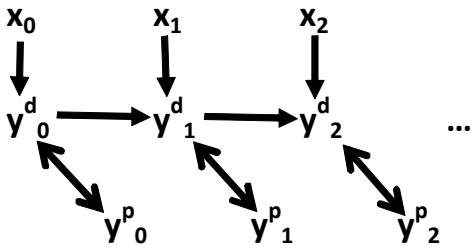

Figure 7.1

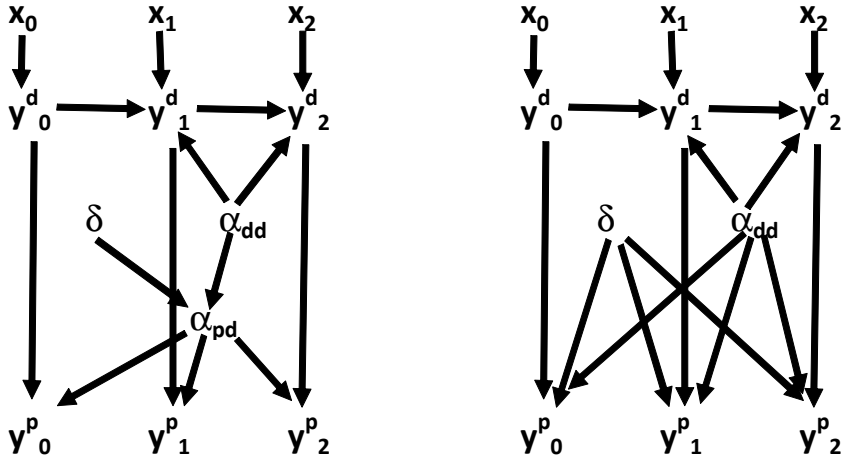

Figure 7.2

CHAPTER 8

Potential Outcomes

The treatment of causation proposed in this monograph is essentially a refinement of the approach of the Cowles economists, particularly Simon [**34**]. A central feature of the Cowles account is that it is based on explicit specification of a formal model consisting of observed and unobserved variables that are linked by equations. This model defines the population; the goal of statistical analysis is to estimate the model's coefficients and to test hypotheses about them based on a sample of draws from the population. In most (but not all) Cowles treatments, internal and external variables are distinguished, so that there is no doubt about which variables the model is intended to explain and which are taken as given. A clear distinction is drawn between the population and the sample. Thus the statistics are defined as functions of the observations in the sample, and are taken as estimators of the underlying population coefficients.

This seems basic and completely noncontroversial, but it is not so. Treatments of causation in sociology and a variety of other disciplines have taken a different path, one involving "potential outcomes" (see Morgan-Winship [**26**] and Imbens-Rubin [**16**] for recent expositions). With the increasing interest in treatment evaluation, the potential outcomes approach has been widely adopted in economics in recent years. In contrast, econometricians and macroeconomists who connect with the Cowles tradition appear to be a dying breed.

The central idea of the potential outcomes approach is that it explicitly specifies treatment outcomes for both the case when the treatment is applied to an individual and when it is not. Thus for agent i we have Y_i^1 if the treatment is applied and Y_i^0 if it is not (note that we are departing from the notation defined above, instead following the notation of the potential outcomes literature by using Y and T to denote the outcome and treatment variables).

The fact that Y_i^1 and Y_i^0 cannot both be observed has been taken to constitute the central problem of causal analysis (Holland [**14**]). Here we have an immediate consequence of the failure to distinguish beween populations and samples. If one has an underlying model of the population that is accurate to a reasonable extent, then the unobserved outcomes can be determined with reasonable accuracy. It is exactly the point of specifying formal models that doing so makes possible such identifications.

Pearl [**29**] in his comments on Dawid [**9**] made this point explicitly and clearly. Assume that we have observations of 2, 3 and 6 for mass, acceleration and force, in conformity with Newton's law. But suppose instead that mass had been 4 instead of 2, and that force is not directly observed. This does not pose any deep existential problem for causal analysis—Newton's law allows us to be confident that force is 12. Thus to the extent that the model is accurate, the value for the unobserved potential outcome implied by the model will be accurate.

The problem, of course, is that we can rarely be as confident of the underlying model as this example implies (although even Newton's laws were revised with the advent of relativity). The obvious fact that potential outcomes cannot all be observed should indeed be viewed as a critical problem with causal analysis as conducted under the potential outcomes approach. However, this is so only because the essential step of specifying a model that describes an underlying population and evaluating its accuracy, and distinguishing that model from a sample consisting of draws from the population, has been skipped (this is especially clear in Holland [**14**]). The fundamental problem of causal analysis is not that some potential outcomes are unobserved; the problem is that it is usually difficult to come up with convincing rationales for specifications of which variables can be assumed external.

The potential outcomes approach deletes the distinction, central to the approach taken here, between the population, characterized as a theoretical construct, and the sample.[1] The data, rather than being viewed as draws from a population, are themselves designated as the population, or as a subset of a larger set of agents designated as the population. The observed values of outcome variables Y_i^{obs} $(= T_i \cdot Y_i^1 + (1 - T_i) \cdot Y_i^0)$, where T_i is a binary variable designating treatment assignment, are random variables. However, this is so not because Y_i^1 and Y_i^0 are random variables, but because treatment assignment, represented by the binary variable T_i, is taken to be a random variable. That this is so is not clear from the potential outcomes literature because, although the range of T_i as $\{0, 1\}$ is clearly specified, the provenance of T_i is not characterized (if T_i is the value of a function, what is the domain of that function?). However, the frequent application in the potential outcomes literature of notation associated with mathematical expectation to Y_i^{obs} and related variables allows for no other interpretation.

[1]Observe in this connection that in the notation above we followed the potential outcomes specification, which associates what are termed the population outcomes with the agent-specific values Y_i^0 and Y_i^1. In the potential outcomes usage, potential outcomes that are not agent-subscripted denote averages (often called expectations) over individuals, not variables in a theoretical model as here.

In the terminology of the potential outcomes approach agent-specific variables are associated with the sample, not the population. Nothing in the potential-outcomes framework corresponds to the population as defined here.

1. Characterizing Potential Outcomes

It is not clear whether in adopting the notation Y_i^0 and Y_i^1 proponents of the potential outcomes approach intend the specification that the outcome for agent i depends on no variables other than the treatment variable for agent i. The notation, which displays the dependence of Y_i on T_i but not that on other causal variables, supports this interpretation. This exclusion of explanatory variables other than T leads to obvious difficulties. It is problematic to specify, for example, that a patient's outcome depends on whether he or she is treated for the disease, but not on whether he or she has the disease.

Another piece of textual evidence points in the same direction: proponents of the potential outcomes approach emphasize the importance of the Stable Unit Treatment Value assumption (Rubin [**33**]), which requires (1) that the outcome for agent i does not depend on the treatment status of other agents, and (2) that the treatment status for agent i determines a unique outcome for agent i.[2] The fact that the SUTV assumption is viewed as underlying the specification of treatment outcomes as Y_i^0 and Y_i^1 suggests that proponents of potential outcomes regard it as essential to exclude all variables other than the i-th agent's treatment as determinants of the i-th agent's outcome, as implied by the notation.

If the notation Y^0 and Y^1 is meant to specify that outcome Y depends only on treatment T we have that the external sets of Y and T are the same: any external variable that causes T also causes Y, and vice-versa. In that case T does not cause Y under the definition adopted here, due to failure of the proper subset condition. This analysis, supposedly of causation, is seen to replace causation with simultaneous determination.

To avoid this outcome we assume that the structural equation expressing the functional relation between Y and T contains an additional term:

$$Y = \alpha_{YT} T + \text{ term}. \tag{8.1}$$

If $\mathcal{E}$(term) (meaning the union of the external sets of all the variables in "term") contains at least one variable not in $\mathcal{E}(T)$, then we have $T \to Y$. Assuming that T is a binary variable and that an intervention on T does not affect "term", we have from eq. (8.1) that the potential outcomes Y^1 and Y^0 are given by

[2]There does not seem to be any substantive reason for excluding dependence of outcomes on other agents' treatments, or other potential causal variables, in this way. It is worth observing that this assumption excludes a class of models that is of central importance in applied work. For example, it is altogether reasonable to specify that the i-th agent's probability of incurring a disease depends not only on whether he was vaccinated, but also on whether others in his community were vaccinated.

$$Y^1 = \alpha_{YT} + \text{term} \tag{8.2}$$

$$Y^0 = \text{term}, \tag{8.3}$$

where "term" is the same in eqs. (8.2) and (8.3). It follows that α_{YT} measures the effect of the treatment, $Y^1 - Y^0$.

Since an intervention on T is attributable to interventions on any of the variables in $\mathcal{E}(T)$, the condition that the intervention on T does not affect "term" is guaranteed to be satisfied only if $\mathcal{E}(T)$ and $\mathcal{E}(\text{term})$ are disjoint. Disjointness of these sets is a sufficient condition for IN-causation of Y by T. An external variable that appeared in both $\mathcal{E}(T)$ and $\mathcal{E}(\text{term})$ would be a confounding variable, the presence of which would render the effect on Y of an intervention on T ambiguous. This failure of IN-causation would imply that the effect of the intervention on Y, being undefined, would not equal α_{YT}.

It is seen that, translated into the terminology set out here, potential outcomes are well defined only when either T and Y are simultaneously determined or T IN-causes Y. Thus the assumption that Y^0 and Y^1 are well defined implies that if the relation between T and Y is causal, it is IN-causal: there are no confounding variables (as defined here), and therefore there is no difficulty in estimating the causal coefficient by least-squares regression.[3]

2. Confounding Variables

As would be expected from the foregoing discussion, the analysis of confounding variables in the potential outcomes literature differs from that outlined in this monograph. The definitions of confounding variables are different: under the usage employed here a confounding variable is an external variable that causes the treatment variable and also the outcome variable via a causal path that does not pass through the cause variable. As just argued, the assumption that the potential outcomes are well defined implies that there are no confounding variables.

In the potential outcomes literature, in contrast, it appears that confounding variables are defined as variables that are correlated with both the treatment variable and the outcome variable, although again this is not clear. These definitions are not the same: any variable x in the external set

[3]It follows that if potential outcomes are well defined the effect of treatment on the treated is the same as on the untreated. This result is most easily verified from an example. Consider a model the causal graph of which is Figure 3.1. Here y_1, the treatment, is binary, and the underlying model may be linear or nonlinear. Intervention on the treatment consists of changing the assumed value of x_1 or x_2, holding constant the other, so as to change y_1 from 0 to 1, or vice-versa. The (absolute value of the) difference between the potential outcomes is the same regardless of whether $y_1 = 0$ is the baseline and $y_1 = 1$ is the intervention, or vice-versa. It follows that the effect of altering an agent's treatment status from treated to untreated is the negative of the effect of the opposite alteration. This is so despite the fact that, in general, altering x_1 or x_2 changes the probability of an agent's being treated.

of the treatment variable is correlated with both the treatment variable and the outcome variable. Despite these correlations being nonzero, x is not a confounding variable under our definition if all the paths connecting x with the outcome variable pass through the treatment variable.

In the potential outcomes literature it is assumed that the causal coefficient is well defined even in the presence of confounding variables, even though it cannot be identified with the regression coefficient of outcome on cause. In contrast, here we have asserted that in the presence of confounding variables unconditional IN-causal coefficients are undefined. As we have seen, we have conditional IN-causation if the conditioning set includes the confounding variables. If so the coefficient quantifying conditional IN-causation is well defined. For example, in Example 3.3 we have $y_1 \not\Rightarrow y_3$, due to the presence of x_2 as a confounding variable, but also $y_1 \Rightarrow y_3|x_2$. Therefore the coefficient quantifying the effect of y_1 on y_3 holding constant x_2 is well defined. In the potential outcomes literature no distinction is drawn between unconditional and conditional causation.

A variety of statistical treatments are proposed to deal with the inconsistency supposedly introduced by differential access to treatment induced by confounding variables, assuming these to be observed. Rosenbaum and Rubin [**32**] proposed *propensity scores*, defined as statistics based on observed confounding variables that measure differences in treatment probabilities (see Athey-Imbens [**4**] for a recent discussion). It is asserted that if propensity scores are held constant then the inconsistency induced by differential access to treatment is eliminated.

The contention that differential access to treatment necessarily induces inconsistency is incorrect: as just noted, existence of variables affecting the likelihood of treatment is consistent with IN-causation of outcomes by treatments as long as these variables influence outcomes only via paths that include the treatment. Having assumed that Y_i^0 and Y_i^1 are well defined, proponents of potential outcomes are excluding causal paths to the outcome variable that do not include the treatment, thus guaranteeing IN-causation regardless of the possible existence of differential access to treatment.

CHAPTER 9

Treatment Evaluation

In recent years a large literature has come into existence specializing causation analysis to the task of determining the effect of a treatment variable on an outcome variable. Most, but not all, of the papers in the treatment evaluation literature use the potential outcomes terminology, discussed in the preceding chapter. Here we avoid repetition by focusing on aspects of treatment evaluation other than those associated with use of the potential outcomes framework.

The task is to evaluate the effect of a treatment on a population of agents. The evaluation is based on analysis of a sample from the population of agents for which data are available. Each member of the sample is represented by an observed variable Z_t (which at this point may or may not be external, as the notation indicates) that characterizes the level of the treatment undergone by that member and an observed internal variable Y_o that measures the outcome for that member. The model assumed to generate the data includes these variables and generally also other variables that may or may not be observed.

To determine the effect of the treatment on the outcome the simplest assumption is that the outcome depends linearly on the treatment:

$$Y_o = \beta_{ot} Z_t + x, \tag{9.1}$$

where x is an unobserved external noise term. The variables are agent-specific; the subscript denoting agents is deleted. If Z_t is assumed to be external, we have $Z_t \ (\equiv X_t) \Rightarrow Y_o$, with β_{ot} measuring the strength of causation. The least-squares estimate $\widehat{\beta}_{ot}$ obtained by regressing Y_o on X_t in the sample of agents is a consistent estimator of the population coefficient β_{ot}.

If treatments are assigned randomly (or if for any reason the analyst is willing to assume that Z_t is independent of all the other external variables) the assumption that Z_t is external is admissible. Usually, however, in treatment evaluations the analyst believes that $Z_t \ (\equiv Y_t)$ is internal, implying that Y_t does not necessarily IN-cause Y_o. That being so, Y_t may not be independent of x in eq. (9.1). Consequently, the least-squares coefficient $\widehat{\alpha}_{ot}$ (replacing $\widehat{\beta}_{ot}$) does not necessarily consistently estimate α_{ot} (replacing β_{ot}). With Y_t designating an internal variable, the analyst must include in the model an equation or equations that determine it. The causal relation

in the revised model between Y_t and Y_o depends on whether Y_t IN-causes Y_o, either unconditionally or conditional on other variables of the model.

1. Supplemental Instruction

The analysis of treatment evaluation when the treatment variable is internal is best conducted in the context of an example. This example is drawn from Thistlethwaite-Campbell [**37**] (it is presented here rather than in Chapter 10 because it is used again in the following section). Suppose that a group of students has taken Examination 1, with scores denoted X_1. Later they take Examination 2, receiving scores Y_2. X_1 and Y_2 are presumed to be correlated; students who do well on Examination 1 are likely also to do well on Examination 2.

After Examination 1 students with scores X_1 that are above a cutoff, normalized at 0, are given special supplemental instruction that is not available to weaker students. The problem is to estimate the effect of the special instruction on Y_2. To that end, define a treatment dummy Y_t as equal to 1 if $X_1 \geq 0$, 0 otherwise. We have the model

$$Y_2 = \alpha_{2t} Y_t + \beta_{21} X_1 + x_2, \tag{9.2}$$

where x_2 is an unobserved error.

It appears from eq. (9.2) that the model is linear, and the designation of such models as linear is frequently encountered in the treatment evaluation literature (Lee-Lemieux [**18**], p. 286, for example). The reduced form for this model is

$$Y_t = \begin{cases} 1 \text{ if } X_1 \geq 0 \\ 0 \text{ otherwise} \end{cases} \tag{9.3}$$

$$Y_2 = \begin{cases} \alpha_{2t} + \beta_{21} X_1 + x_2 \text{ if } X_1 \geq 0 \\ \beta_{21} X_1 + x_2 \text{ otherwise} \end{cases}, \tag{9.4}$$

which is not linear affine. The appearance of linearity in the structural form and nonlinearity in the reduced form is surprising. It reflects the practice, frequently encountered in the treatment evaluation literature, of suppressing explicit recognition of some functional relations among variables in structural models. In the present example this is done by treating Y_t as if it were a new external variable, as distinguished from representing it as the dependent variable in an explicitly stated equation of the model (eq. (9.3) here). If Y_t is clearly specified as an internal variable, then eq. (9.3) must be included along with eq. (9.2) as part of the structural version of the model. The revised structural form of the model, now consisting of eqs. (9.2) and (9.3), is nonlinear, like the reduced form.

The relation between X_1 and Y_2 being nonlinear, the causal effect of X_1 on Y_2 cannot be associated with a single coefficient that multiplies the intervention on X_1. Instead, the baseline value of X_1 and the value of

X_1 under intervention, denoted X_1^b and X_1^i, respectively, must be specified separately (see p. 47 above). The effect on Y_2 of an intervention on X_1 is given by

$$\Delta Y_2 = \begin{cases} \alpha_{2t} + \beta_{21}\Delta X_1 \text{ if } X_1^b < 0 \text{ and } X_1^i \geq 0, \\ -\alpha_{2t} + \beta_{21}\Delta X_1 \text{ if } X_1^b \geq 0 \text{ and } X_1^i < 0, \\ \beta_{21}\Delta X_1 \text{ otherwise.} \end{cases} \tag{9.5}$$

In discussions of such models in the treatment evaluation literature it is presumed that the question "What is the effect of Y_t on Y_2?" has an unambiguous answer, and also that the answer is α_{2t}. In the model just set out an intervention on Y_t is induced by an intervention on X_1, and X_1 causes Y_2 via both a direct and an indirect path. However, we do not have $Y_t \Rightarrow Y_2$ due to existence of a path directly connecting X_1 with Y_2.[1] It follows that the effect on Y_2 of an intervention on Y_t cannot be determined without further information about the underlying intervention on X_1. In particular, the coefficient α_{2t} cannot be characterized as measuring causation, either unconditionally or conditionally.[2]

2. Regression Discontinuity

Analysts evaluating treatment effectiveness are increasingly relying on statistical procedures involving regression discontinuity to estimate causal coefficients (see Lee-Lemieux [**18**] for a survey). The idea is that if samples are restricted to data that are near (but, of course, on both sides of) a discontinuity, then the causal effect of an intervention on the forcing variable can be estimated more accurately than if observations are included in which the forcing variable is not near the discontinuity.

It is difficult to find in the treatment evaluation literature explicit attempts to justify this claim. It is suggested that if the individuals being sampled are similar with respect to the forcing variable they are likely also to be similar with respect to unobserved causal variables. This is held to mitigate inconsistency attributed to differential access to treatment, and

[1] It was noted in Chapter 7 that in linear models if the treatment variable has an external set consisting of a singleton, then the treatment variable IN-causes the effect variable despite the existence of a direct path connecting the variable constituting the external set with the effect variable. It was also observed that in nonlinear models this result may not obtain if the map from X_1 to Y_t is not invertible. That is the situation here.
Note also that in the example under discussion there does not exist Ψ such that $Y_t \Rightarrow Y_2|\Psi$; the fact that X_1 is the only member of $\mathcal{E}(Y_t)$ implies that holding constant X_1 effectively converts Y_t to a constant, rendering the model meaningless. This also was discussed in Chapter 7.

[2] In the television quiz show "Jeopardy" Alex Trebek provides the answer and the contestant responds with the corresponding question. If the answer is "$\alpha_{2t}\Delta Y_t$" the corresponding question would seem to be "Given an assumed baseline value X_1^b for X_1, what is the effect on Y_o of an intervention that sets X_1 equal to X_1^i?" The result in the text implies that there is no pair X_1^b, X_1^i that generates the requisite question (other than the trivial $X_1^b = X_1^i$).

also other unspecified biases. This assertion is problematic on several levels. Most simply, we showed in the preceding chapter that differential access to treatment does not necessarily interfere with IN-causation. Further, there is no clear basis for the assertion that individuals who are similar with respect to one variable are also similar with respect to others. Finally, restricting samples to similar individuals reduces sample variation, which in general makes estimation less accurate, not more.

Regression discontinuity is most easily discussed in the context of an example. The model is that of Thistlethwaite-Campbell set out in the preceding section. These authors are credited with introducing regression discontinuity in the paper under discussion. There is no doubt that regression discontinuity is an important idea, but we will see that Thistlethwaite-Campbell's model is not the vehicle that makes clear why it makes sense to ignore data away from the discontinuity.

Despite the fact that the coefficient α_{2t} does not have a causal interpretation, it can still be estimated using the bivariate linear regression (9.2). This exercise is essentially the same problem as that of constructing a forecast of the dependent variable at a designated value of the independent variable. As is well known, if the objective function is to minimize mean-square error the best forecast is the regression value of the dependent variable at the designated value of the independent variable. It is easily shown that this forecast gives more weight to observations near the forecast point.[3] In this sense, it is appropriate to place greater weight on observations near the discontinuity, as recommended in the regression discontinuity literature, in constructing the forecast.

However, this argument does not justify deleting any observations. Least squares regressions place equal weight on all observations in constructing coefficient estimates; deleting data strictly increases mean-square forecast errors. Thus the essential features of a model involving regression discontinuity in which efficient estimation involves ignoring data far from the discontinuity point are not found in the Thistlethwaite-Campbell model. We will have to look elsewhere for justification for regression discontinuity.

We are interested in models in which Y_1 and Y_2 are both observed, and we have $Y_1 \rightarrow Y_2$. If causation here is implementation neutral the regression coefficient is optimally estimated by least squares. If $Y_1 \nRightarrow Y_2$ there necessarily exists at least one confounding variable; suppose that there exists just one. We wish to estimate the IN-causal effect of Y_1 on Y_2 conditional on the confounding variable. The essential feature of the algorithm that we require is that either the path that links the confounding variable to the cause variable or the path linking it to the effect variable contains a

[3]For example, suppose that the analyst has two data points, $(x, y) = (1, y_1)$ and $(x, y) = (2, y_2)$, and wishes to forecast y at $x = 0$. The regression consists of the line that passes through the two data points, and the mean-square error minimizing forecast consists of the intercept of this line. This is easily seen to be $2y_1 - y_2$, which attaches greater weight to y_1 than to y_2.

discontinuity. If so, restricting the data on the forcing variable—the confounding variable—to values near the discontinuity has the effect of (almost) disconnecting the path that does not have a discontinuity. The result is that the forcing variable is no longer a confounding variable. With Y_1 now IN-causing the outcome variable, the causal coefficient is well defined and equal to the population regression coefficient. Accordingly, it can be estimated consistently by least squares.

An interpreted model in which regression discontinuity estimation is applied is discussed in Chapter 10, Example 3.

CHAPTER 10

Interpreted Examples

In this chapter five examples of the preceding analysis are discussed. The variables in the models analyzed in this chapter have economic interpretations; accordingly, we augment the notation by including mnemonics as subscripts. For example, in the model of Section 1 Y_i represents income (internal and observed) and x_a is a dummy for family affluence (external and unobserved). As above, we will use Z and z to denote variables not yet labeled as internal or external.

1. Private vs. Public Universities

The first example consists of a simplified version of an exercise discussed by Angrist-Pischke [**2**], which itself is a simplified version of a model developed by Dale-Krueger [**8**].

Suppose that we are interested in determining the effect on subsequent income Y_i of a student's attending a private university rather than a public university. The simplest procedure is to run a regression of income on a dummy variable Z_p representing attendance at a private university. That exercise typically results in a high number. However, Angrist-Pischke observed that there is a strong possibility of what they called an omitted variables bias here: students that attend private universities typically come from more affluent families than those who attend public universities, and this difference may affect lifetime income in ways not related to the differential effect of private university attendance. Thus family affluence is a confounding variable, the existence of which biases upward the estimated coefficient of the attendance variable.

The established practice, followed by Angrist-Pischke, is to correct for this omitted variables bias by controlling for family affluence, which is done by including a proxy for family affluence in the regression. It was presumed that including a confounding variable in a regression effectively holds that variable constant, so that it is no longer a confounding variable. Following Dale-Krueger, Angrist-Pischke proposed using the set of universities to which each student applied as a proxy for family affluence. The idea was that students from affluent families would be more likely to apply to private universities instead of, or in addition to, public universities. They defined the dummy Z_a as 1 for students who applied to more private than public universities, and as 0 for those who did not. Under the established procedure, including Z_a along with Z_p as an explanatory variable for Y_i was held

to eliminate the omitted variables bias. The resulting regression coefficient of Y_i on Z_p, being free of omitted variables bias due to the presence of Z_a in the regression, would, it was believed, provide an accurate estimate of the effect on Y_i of attendance at a private university.

In Angrist-Pischke's discussion it is stated that the roles of Z_p and Z_a are symmetric: either can be the causal variable of primary interest, with the other as the confounding variable. Thus one can reverse the roles of the causal and confounding variables so as to determine the effect of Z_a on Y_i holding constant Z_p. The bivariate regression of Y_i on Z_a and Z_p is represented as providing a good estimate of the causal effect of each explanatory variable on Y_i given the other.

Angrist-Pischke presented a simple example of this calculation using made-up data. Five former students have subsequent income shown in the second column of Table 1. They have different values for Z_p and Z_a, shown in the third and fourth columns. Table 2 reports the coefficients in a bivariate regression of Y_i on Z_p and Z_a and univariate regressions on each of these separately.[1] The coefficient of Z_p is lower in the bivariate regression than in the univariate regression. Angrist-Pischke interpreted this difference as confirming the conjecture that failure to control for family affluence in the real-world counterpart of the univariate regression of Y_i on Z_p leads to an upward-biased estimate of the effect of private university attendance on subsequent income. Angrist-Pischke's discussion implied that the coefficients in regression 1 provide good estimates of the effects of Z_p and Z_a on Y_i: each of the regression coefficients measures the effect of the associated explanatory variable on income, ceteris paribus.

Several aspects of this chain of reasoning are of interest. First, the analysis of causation proceeds without benefit of any explicit specification of which variables are external and which, besides income, are internal. The interpretation of Z_p and Z_a as both being external conflicts with the rationale—that Z_a causes both Z_p and Y_i (or Z_p causes both Z_a and Y_i)—for including Z_a in the regression in the first place. Second, the verdict that the best estimate of the effect of private university education on income is that given by the bivariate regression amounts to an assertion that one correlation provides a better estimate of causation than another. Here, it would seem, we are inferring causation directly from correlation, a practice that in many other contexts is suspect. Third, the fact that the two explanatory variables are treated symmetrically in the preferred regression implies that the status of a variable as a cause or a confounder depends on how the analyst proposes to interpret the model, not on the model itself.

We present an alternative discussion of the example based on IN-causal analysis. The fact that the bivariate regression produces different coefficients

[1]We are taking the population regression coefficients as coinciding with their sample counterparts. This is an acceptable simplification because we are not interested in sample variation here.

from the univariate regressions implies that the dummies are correlated. In any case, the existence of a positive correlation is clear from Table 1: the expectation of Z_p conditional on $Z_a = 1$ is 2/3, whereas that expectation conditional on $Z_a = 0$ is 1/2. This in turn suggests that Z_a and Z_p are linked by a (or several) functional relation(s). Any such links should be brought into the model; ignoring them will be seen to induce error in the interpretation of estimated coefficients.

The simplest specification is that one of the dummies is internal and one is external, with the internal dummy specified as a function of the external dummy and a new unobserved external variable representing idiosyncratic shocks to the internal dummy. First, assume that Z_a is external (and therefore relabeled X_a), as would be implied by the assumption that family affluence is a direct determinant of both private university attendance and lifetime income. We include in the model a regression expressing Z_p (now relabeled Y_p) as a function of X_a and an unobserved external error x_p (regession 4 in the Table 2). In the example the estimated regression is

$$Y_p = 0.5 + 0.167X_a + x_p, \tag{10.1}$$

where X_a is given by

$$X_a = \begin{cases} 1 \text{ with probability } 3/5 \\ 0 \text{ with probability } 2/5 \end{cases} . \tag{10.2}$$

The error x_p is specified as

$$x_p = \begin{cases} 1/2 \text{ with probability } 1/2 \\ -1/2 \text{ with probability } 1/2 \end{cases} \tag{10.3}$$

if $X_a = 0$, and

$$x_p = \begin{cases} 1/3 \text{ with probability } 2/3 \\ -2/3 \text{ with probability } 1/3 \end{cases} \tag{10.4}$$

if $X_a = 1$. This specification implies that Y_p takes on value 0 with probability 2/5 and 1 with probability 3/5, as in the data in Table 1. Note here that x_p is mean-independent of X_a, although not independent, as discussed above. Finally, the model also includes an equation reflecting the dependence of Y_i on its parents:

$$Y_i = 10Y_p + 60X_a + x_i. \tag{10.5}$$

Figure 10.1(a) gives a causal graph of the resulting model. From regression 1 affluence X_a affects Y_i directly, with coefficient 60. It also has an indirect effect via Y_p of 1.67, equal to the product of the coefficient of Y_p with respect to X_a (0.167) and the coefficient of Y_i with respect to Y_p (10). The total effect of X_a on Y_i is 61.67. We see that the univariate regression in

regression 3 gives the correct causal coefficient, which includes both direct and indirect effects, whereas the coefficient of X_a in regression 1 gives only the direct effect.

The unconditional effect of Y_p on Y_i is not well defined due to the failure of IN-causation: an intervention ΔY_p could have been generated either by an intervention ΔY_p on x_p, resulting in $\Delta Y_i = 10\Delta Y_p$, or an intervention equal to $(10+60/0.167)\Delta Y_p = 370\Delta Y_p$ if the intervention is on X_a. However, the effect of Y_p on Y_i conditional on $X_a, 10\Delta Y_p$, is well defined because in that case the intervention, being on x_p alone, is unambiguous. Thus one must be careful to distinguish the unconditional effect of Y_p on Y_i, which is undefined, from the effect of Y_p on Y_i conditional on X_a, which is well defined and is equal to the coefficient of the arrow connecting Y_p and Y_i in the graph.

Here we have another example in which unconditional IN-causation fails, but conditional causation exists, and is of primary interest. A family considering private versus public universities knows whether it is affluent or not, implying that its decision is based on the effect on income of school choice conditional on affluence, not the unconditional effect. It is worth noting that the model incorporates the assumption that the effect on income of university choice conditional on affluence is the same for both values of the affluence variable.[2]

We see that characterizing Z_a as external and Z_p as internal results in an asymmetric treatment of their coefficients in the bivariate regression. The coefficient of X_a reflects its direct effect on Y_i, not the total effect; the coefficient of Y_p reflects conditional causation, not unconditional causation. This contrasts with the received analysis, which as noted above would interpret both coefficients in the bivariate regression as measuring presumably the same type of causation conditional on the value of the other. We have that X_a IN-causes Y_i, with coefficient equal to that of the univariate regression of Y_i on Z_a, while Z_p does not unconditionally IN-cause Y_i, implying that neither the coefficient of Z_p in the univariate regression nor that in the bivariate regression has an unconditional IN-causal interpretation.

Now consider the case in which Z_p rather than Z_a is external (Figure 10.1(b)). To be sure, in the context of the assumed setting it is not easy to motivate this specification: the fact that a student attends a private university does not increase family affluence (just the opposite). Thus in the current context it would be acceptable to rule out this specification a priori. However, in most cases (in the following section, for example) it is not

[2]This example makes clear the importance of the distinction, emphasized in several places above, between the two meanings assigned to "conditional": the fact that the effect on Y_i of an intervention on Y_p conditional on X_p is the same for both values of X_p implies that the conditional expectation of the effect of Y_p on Y_i coincides with the unconditional expectation. Here "conditional" is used in its probabilistic sense. The observation that the unconditional effect of Y_p on Y_i is undefined uses "conditional" to mean holding constant X_p. There is no inconsistency.

obvious which variables are external, so it is worthwhile working through the reversed case here. As will be clear, the interpretation of multiple regression coefficients as measuring causation is analogous in the two cases, but with the interpretation of the variables' regression coefficients reversed.

If Z_p is external (hence is relabeled X_p) a regression generating the effect of X_p on Z_a (relabeled y_a) must be included in the model. This equation, regression 5 in Table 2, turns out to be the same as regression (10.1) (regression 4), with Z_p and Z_a reversed (reflecting the fact that the example has the special property that a plot of (Z_p, Z_a) pairs is symmetric around the 45-degree line). The univariate regression 2 shows that the effect of X_p on Y_i is 20. This consists of a direct effect of 10 and an indirect effect through Y_a of 10 (equal to the product of 0.167 and 60). The effect of Y_a on Y_i is not well defined due to failure of IN-causation. The effect of Y_a on Y_i holding constant X_p is 60, and the effect of X_p on Y_i is $60 + 6 \cdot 10 = 120$.

In Chapter 5 we considered the case when the analyst is unwilling to assume that either of two observed variables is external. It is easy to accommodate this specification, but it was noted that weakening the specification of the model in this way results in fewer implications for IN-causation and coefficient identification. It is worthwhile showing this in the context of the present model. Suppose that the binary proxies are both specified as internal, resulting in the notation Y_p and Y_a. These are specified to be linear functions of unobserved independently distributed external variables x_1 and x_2. The variances of Y_p and Y_a and their covariance can be used to parametrize the variances of x_1 and x_2 and one of the constants linking x_1 and x_2 with Y_p and Y_a. The other coefficients in the determination of Y_p and Y_a are normalized at 1, reflecting the fact that x_1 and x_2 are not observed.[3]

The causal graph for this model is shown in Figure 10.2. The binary proxies are simultaneously determined, and each causes Y_i. None of the internal variables are IN-causally related, unconditionally or conditionally, implying that these causal effects cannot be quantified. Once again we see the correctness of the Cowles emphasis on the need for strong theoretical restrictions—in this case the specification that at least one of the observed variables is external—if models are to generate testable implications.

---

[3] We have

$$Y_p = x_1 + \beta_{p2} x_2 \tag{10.6}$$

$$Y_a = x_1 + x_2. \tag{10.7}$$

Here any three of the four coefficients relating Y_p and Y_a to x_1 and x_2 could have been normalized at 1. This reflects the fact that many joint distributions of x_1 and x_2 would generate the same distribution of Y_p and Y_a. The choice among these is arbitrary.
From eqs. (10.6) and (10.7) one can calculate the variances of x_1 and x_2 and the constant β_{p2} from the variances of Y_p and Y_a and their covariance.

2. Effect of Military Service on Income

Angrist's [**1**] paper evaluating the effects of military service on lifetime income provides another setting in which the analysis proposed here can be implemented. As will be seen, the analysis here differs from that by Angrist.

The starting point in Angrist's discussion is the relation

$$Y_i = \beta_{iv} X_v + x_i, \tag{10.8}$$

stating that lifetime income Y_i depends on veteran status X_v and an unobserved error term x_i. In eq. (10.8) we have that X_v IN-causes Y_i, with β_{iv} measuring the magnitude of the effect. The simplest version of this model would have X_v and x_i external and probabilistically independent, implying that β_{iv} can be estimated by ordinary least squares. The difficulty is that eq. (10.8) is likely to be a misspecification. To the extent that veteran status is correlated with such unobserved variables as ability to earn a high income in civilian employment, which in turn may be a component of x_i, we may have a confounding problem. As a result, the coefficient β_{iv} may be interpretable as an IN-causal coefficient only due to a misspecification.

Angrist's solution was to use a measure X_e of eligibility for conscription as an instrument in estimating β_{iv}. X_e was specified to consist of the number associated with each agent under the draft lottery in the Viet Nam war. Whether or not an agent is likely to be drafted based on his lottery number is correlated with whether or not he served in the military—the treatment—but, arguably, not with other determinants of lifetime income. This, Angrist suggested, establishes the suitability of X_e as an instrument in estimating β_{iv}, interpreted as a causal coefficient.

This justification for draft eligibility as an instrument in estimating the coefficient Angrist associated with the effect of veteran status on income seems persuasive, but the informal treatment of the correlation between X_v and x_i is problematic. Investigating this difficulty involves dispensing with the purely verbal treatment of draft eligibility and income ability in favor of working with a model that incorporates these variables explicitly.

Let z_a represent an agent's ability to earn a high income in civilian employment. The new variables X_e and z_a are not part of the original formal model, consisting of eq. (10.8). We now expand that model to incorporate them, and use the augmented model to deconstruct the correlation between Z_v and z_i. Note that the notation change from x to z indicates that we are provisionally relaxing the specification that these variables are external.

The problem is to specify which variables are external in the expanded model. There are two possibilities. First, consider what Angrist characterized as the simplest specification: agents in military service accumulate human capital at a different rate from those in civilian employment, resulting in different future incomes when they compete in civilian job markets against nonveterans. This requires relabeling z_a as y_a. We also relabel X_v

as Y_v in recognition that veteran status is now specified to depend on both X_e and a new external variable x_v, unobserved. Under this respecification the augmented model can be written

$$y_a = \alpha_{av} Y_v + x_a \tag{10.9}$$

$$Y_v = \begin{cases} 1 \text{ if } \beta_{ve} X_e + x_v \geq 0 \\ 0 \text{ if } \beta_{ve} X_e + x_v < 0 \end{cases}, \tag{10.10}$$

$$Y_i = \alpha_{iv} Y_v + \alpha_{ia} y_a + x_i. \tag{10.11}$$

The external variables X_e, x_i, x_a and x_v are assumed to be distributed independently. Note that the model here is nonlinear, because of the form of eq. (10.10).[4] The causal graph of the model just specified is shown as Figure 10.3(a). As can be verified from Figure 10.3(a), Y_v IN-causes Y_i, so a least-squares regression of Y_i on Y_v produces a consistent estimate of the relevant causal coefficient; there is no need for an instrumental variables estimator. Since Y_v affects Y_i both directly and indirectly through y_a, the relevant causal coefficient is $\alpha_{iv} + \alpha_{ia}\alpha_{av}$, and that is the constant consistently estimated in a univariate regression of Y_i on Y_v. The constants α_{iv}, α_{ia} and α_{av} that quantify the breakdown of the total effect of Y_v on Y_i into the direct effect and the indirect effect are not separately identified due to the assumption that y_a is not observed.

Instead of having veteran status IN-causing income ability, we could reverse the causation and specify that earnings ability IN-causes veteran status, so that agents are more or less likely to join the armed forces according to their income ability in civilian employment. A model that reflects this respecification is the following:

$$Y_v = \begin{cases} 1 \text{ if } \beta_{ve} X_e + \beta_{va} x_a + x_v \geq 0 \\ 0 \text{ if } \beta_{ve} X_e + \beta_{va} x_a + x_v < 0 \end{cases} \tag{10.12}$$

$$Y_i = \alpha_{iv} Y_v + \beta_{ia} x_a + x_i. \tag{10.13}$$

As Figure 10.3(b) indicates, in this setting Y_v does not IN-cause Y_i due to the presence of the confounding variable x_a. Thus α_{iv} cannot be interpreted as a coefficient measuring unconditional IN-causation. However, Y_v does IN-cause Y_i conditional on x_a. The coefficient α_{iv} associated with this causal relation is consistently estimated by instrumental variables taking X_e as an instrument. The role of instrumental variables estimation of coefficients associated with conditional causation when the confounding variable

[4]Also, the model incorporates the unobserved internal variable y_a; in the discussion above it was assumed that all internal variables are observed. Formal treatment of this would require adopting the minor generalization of the causal notation set out in footnote 1 in Chapter 5.

is not observed was discussed above. Thus if z_a is taken to be external Angrist was correct in asserting that the coefficient associated with the causal relation between Y_i and Y_v is consistently estimated by instrumental variables, provided it is understood that the relevant notion of causation here is conditional causation rather than unconditional causation.

3. Regression Discontinuity

Here we present a model in which a coefficient measuring conditional causation can be estimated via regression discontinuity despite failure of unconditional IN-causation. The example is loosely based on McCrary-Royer [**25**]. The model here is drastically simplified relative to theirs; our purpose is to illustrate how regression discontinuity works, not to present an adequate empirical analysis. Unlike McCrary-Royer we explicitly specify how the confounding variable invalidates ordinary least squares estimates of causal coefficients if the regression discontinuity is not exploited.

We are interested in how maternal education affects various measures of infant health. If one were willing to assume that the former IN-causes the latter this link could be estimated directly using ordinary least squares. However, it is possible that IN-causation fails due to the presence of a confounding variable. As discussed in Chapter 9, an estimation involving regression discontinuity can disconnect the path that is continuous, thereby reversing the status of the variable that confounds the causal relation in the absence of the regression continuity estimation. With maternal education now IN-causing infant mortality, an ordinary least squares regression is justified.

We present a version of McCrary-Royer's model that illustrates this:

$$Y_b = f(X_d, x_1) \tag{10.14}$$

$$y_a = \beta_{ad} X_d + x_2 \tag{10.15}$$

$$Y_e = X_s - Y_b \tag{10.16}$$

$$Y_h = \alpha_{he} Y_e + a_{ha} y_a. \tag{10.17}$$

Eq. (10.14) connects the age at which mothers begin their education, Y_b, to the month and day they were born, X_d. As the notation suggests, this equation is nonlinear (see the discussion below). Eq. (10.15) connects y_a, family affluence, with X_d. Eq. (10.16) says that the extent of mothers' education equals the difference between X_s, the mother's age when education stops, and Y_b. Eq. (10.17) connects infant health Y_h to mother's education and family affluence. Here x_1 and x_2 are uninterpreted errors.

The external sets of this model are as follows:

$$\mathcal{E}(Y_b) = \{x_1, X_d\} \tag{10.18}$$

$$\mathcal{E}(y_a) = \{x_2, X_d\} \tag{10.19}$$

$$\mathcal{E}(Y_e) = \{x_1, X_s, X_d\} \tag{10.20}$$

$$\mathcal{E}(Y_h) = \{x_1, x_2, X_s, X_d\}. \tag{10.21}$$

The causal graph for the model is shown in Figure 10.4.

Inspection of the graph shows that Y_e does not IN-cause Y_h. The confounding variable is X_d, which causes Y_e through a path including Y_b, and causes Y_h through a path that does not include Y_e. Following McCrary-Royer, the age at which a student begins education depends discontinuously on her date of birth (schools enroll students only when they have passed their sixth birthday, for example) by some date certain, such as December 1). Then the causal effect of X_d on Y_b is discontinuous: students born shortly before December 1 are younger (they just recently turned 6 on December 1) when they begin school than those born shortly after December 1 (who are almost 7). If their age at the date they stop their education, X_s, does not depend on birth date, then students born shortly before December 1 are more educated on average than those born after December 1.

However, the effect of X_d on y_a may reasonably be taken to be linear, and therefore continuous, implying that if the sample is restricted to students with dates of birth near December 1, the effect of X_d on y_a is negligibly small. The arrow connecting X_d to y_a can be deleted, although the arrow connecting X_d to Y_b remains due to the discontinuity. That done, X_d is no longer a confounding variable, and we have $Y_e \Rightarrow Y_h$, implying the validity of least-squares estimation.

Specifying an explicit account of how the regression discontinuity estimation works makes possible an evaluation of the conditions for adequacy of a regression discontinuity argument. The argument just summarized hinges on the implication of the model that X_d is an empirically important cause of y_a, since otherwise there is no motive to use regression discontinuity to break this link. It is, however, difficult to see any reason to specify that X_d causes y_a. On the contrary, it seems more reasonable instead to specify a model in which y_a (if external, or one of the external variables that cause it otherwise) is a major confounding variable: family income clearly strongly influences both Y_e and Y_h, the latter through paths that do not include Y_e. The regression discontinuity argument just summarized, being based on designating X_d as the forcing variable, does nothing to address the bias induced if y_a is the confounding variable.

The point is that regression discontinuity procedures can remedy failures of IN-causation if the forcing variable, in this case X_d, is also the confounding variable. Regression discontinuity arguments are therefore persuasive only to the extent that analysts can motivate the assumption that the forcing variable and the confounding variable coincide.

4. Granger Causation

Much has been written about the relation, or lack thereof, between causation and Granger-causation. These discussions—including Cooley-LeRoy [**7**]—are unsatisfactory because no precise definition of causation—as distinguished from Granger-causation—is offered and, in particular, no distinction is drawn between causation and IN-causation. It may be worthwhile to state how the above analysis of causation bears on this topic.

In the case of two stochastic processes z_1 and z_2, z_1 is *strictly exogenous* with respect to z_2 if z_1 is a function only of its own past values and an unobserved external process. The process z_2 *Granger-causes* z_1 if the optimal predictions of future values of z_1 based on past values of z_1 alone can be improved upon by including lagged values of z_2 as explanatory variables. It is easily shown that if z_1 is strictly exogenous with respect to z_2 then z_2 does not Granger-cause z_1. The contrapositive of this is that if z_2 Granger-causes z_1 then z_1 is not strictly exogenous. Thus Granger-causation is a test of strict exogeneity, in the sense that acceptance of Granger-causation implies rejection of strict exogeneity. The converse is that if z_2 does not Granger-cause z_1, then z_1 is strictly exogenous with respect to z_2. This is not generally true (see Cooley-LeRoy [**7**]). Therefore acceptance of Granger non-causation does not imply strict exogeneity.

These results are of interest to the extent that strict exogeneity can be connected to causation or IN-causation. We investigate this in the context of an example. Consider a two-equation linear autoregressive model determining the date-t values Y_{1t} and Y_{2t} of Y_1 and Y_2, both observed, as functions of each other, their own once-lagged values, and unobserved external errors x_{1t} and x_{2t}. The errors are assumed independent cross-sectionally and over time (the consequences of assuming that errors are correlated cross-sectionally are discussed in the following section). Y_1 and Y_2 can be interpreted as the monetary instrument—the money stock or the federal funds rate—and real GDP, respectively, although this interpretation is not necessary for the argument. In this model Y_1 is strictly exogenous if the errors in the equation for Y_2 do not feed back into the equation for Y_1, either currently or with a lag. This condition is satisfied if the coefficients of current and lagged Y_2 in the equation for Y_1 equal zero.

Under strict exogeneity of Y_1 the structural form of the bivariate model just described is

$$Y_{1t} = \alpha_{111}Y_{1,t-1} + x_{1t} \tag{10.22}$$

$$Y_{2t} = \alpha_{210}Y_{1t} + \alpha_{211}Y_{1,t-1} + \alpha_{221}Y_{2,t-1} + x_{2t}. \tag{10.23}$$

Here $\alpha_{ij\lambda}$ $(i, j \in \{1, 2\}; \lambda \in \{0, 1\})$ denotes the coefficient of Y_{it} with respect to $Y_{j,t-\lambda}$. The causal graph for this model for dates $t-1$ and t is shown in Figure 10.5. We have $Y_{1t} \to Y_{2t}$ because the external set of Y_{1t} consists of the errors $x_{1\tau}, \tau \leq t$, which is strictly contained in the external set of Y_{2t},

which consists of all the errors $x_{1\tau}$, $x_{2\tau}$, $\tau \leq t$. Also, Y_{1t} is directly connected to Y_{2t}, so that both conditions for direct causation are satisfied. However, we do not have $Y_{1t} \Rightarrow Y_{2t}$ in the equation system as written: the lagged error in the Y_1 equation that affects Y_{2t} via Y_{1t} also affects Y_{2t} via $Y_{1,t-1}$ and $Y_{2,t-1}$, so it is a confounding variable.[5] It follows that strict exogeneity of Y_1 does not imply that α_{210} can be interpreted as quantifying the effect of Y_{1t} on Y_{2t}. We have $Y_{1t} \Rightarrow Y_{2t}$, so that α_{210} does represent the causal effect of Y_{1t} on Y_{2t}, under the additional restriction that α_{221} equals 0.

Thus neither Granger noncausation nor the stronger assumption of strict exogeneity of Y_{1t} is sufficient to establish unconditional IN-causation. If α_{221} is nonzero the question "What is the effect on Y_{2t} of an intervention that brings about ΔY_{1t}?" does not have an unambiguous answer: different interventions consistent with a given change in Y_{1t} have different effects on Y_{2t}.

Whether or not we have $\alpha_{221} = 0$ it can be asserted that strict exogeneity implies that Y_{1t} IN-causes Y_{2t} conditional on all the lagged errors in Y_{1t}. This is, of course, a very strong restriction.

5. Vector Autoregressions

In a major paper Sims [**35**] expressed the view that the assumptions used to identify causal relations in many macroeconomic models were not credible. He argued that forecasting and policy analysis are best undertaken in the framework of models that avoid identifying assumptions supposedly (but typically not actually) drawn from economic theory. As regards policy analysis, this contention proved controversial. It is worthwhile examining the argument underlying atheoretical policy analysis.

The favored analytical tool is the vector autoregression, in which a vector of observed internal variables is expressed as a linear function of their lagged values and a vector of serially independent unobserved external variables. In the two-variable case with only once-lagged internal variables as explanatory variables we have

$$Y_{1t} = \alpha_{111} Y_{1,t-1} + \alpha_{121} Y_{2,t-1} + x_{1t} \tag{10.24}$$

$$Y_{2t} = \alpha_{211} Y_{1,t-1} + \alpha_{221} Y_{2,t-1} + x_{2t}. \tag{10.25}$$

Here the external variables x_{1t} and x_{2t} are correlated, implying that some of the causal relations of the model are contained in this uninterpreted correlation. It follows that the coefficients $\alpha_{ij\lambda}$ in (10.24)-(10.25) are not interpretable as measuring unconditional IN-causation (this would be true even

[5]Specifically, if the intervention ΔY_{1t} is caused by an intervention on x_{1t} the effect on Y_{2t} is $\alpha_{210}\Delta Y_{1t}$. If the intervention is on $x_{1,t-1}$ its effect on Y_{2t} is $\alpha_{210}(1 + \alpha_{221}/\alpha_{111})\Delta Y_{1t}$. Here one path from $x_{1,t-1}$ to Y_{2t} passes through Y_{1t} and the other does not. Thus $x_{1,.t-1}$ is a confounding variable in the causal relation between Y_{1t} and Y_{2t}. The same applies for the other lagged terms.

if x_{1t} and x_{2t} were uncorrelated because the lagged variables do not IN-cause the contemporaneous variables).

Standard practice involves transforming the model so that the error terms are independent. In the two-variable case this involves redefining one of the external variables as the residual in the regression of one of the external variables on the other. This is, of course, the same operation that was undertaken in the example of Section 1, where we began with two observed and purportedly external variables that were correlated. Suppose that x_{2t} is replaced by x'_{2t}, defined by

$$x'_{2t} \equiv x_{2t} - \alpha_{210} x_{1t}, \tag{10.26}$$

where α_{210} is the regression coefficient of x_{2t} on x_{1t}, so that x_{1t} and x'_{2t} are uncorrelated. We have

$$Y_{1t} = \alpha_{111} Y_{1,t-1} + \alpha_{121} Y_{2,t-1} + x_{1t} \tag{10.27}$$

$$Y_{2t} = \alpha_{210} Y_{1t} + \alpha_{211} Y_{1,t-1} + \alpha_{221} Y_{2,t-1} + x'_{2t}, \tag{10.28}$$

redefining α_{211} and α_{221}. The causal graph of eqs. (10.27)-(10.28) for dates $t-2$, $t-1$ and t is shown in Figure 10.6. In the model (10.27)-(10.28) the lagged internal variables cause the current internal variables, but, as noted above, do not IN-cause them. This is so because the external variables dated $t-2$ and earlier are confounding variables. Thus we have, for example, $Y_{1,t-1} \Rightarrow Y_{1t} | (\mathcal{E}(Y_{1,t-2}) \cap \mathcal{E}(Y_{2,t-2}))$. This condition, of course, is extremely strong.

Multivariate vector autoregressions, which generalize the bivariate vector autoregression just discussed, are written in vector-matrix form as

$$Y_t = A_1 Y_{t-1} + x_t, \tag{10.29}$$

the analogue of eq. (10.24)-(10.25). Generally more lagged terms are included, but there is no need to do so here. The external vector variable x'_t has covariance matrix V. The multivariate analogue of the model representation under the alteration just described for the two-variable case, eqs. (10.27)-(10.28), is

$$Y_t = A_0 Y_t + A_1 Y_{t-1} + x'_t, \tag{10.30}$$

with A_0 being a triangular matrix and x'_t having a diagonal covariance matrix (here A_1 is redefined). This is the Choleski decomposition. Note that eq. (10.30) has the same form as the causal form of a theoretical model, discussed above. Written in the form of eq. (10.30) the vector autoregression can be used to compute the impulse response functions. These, being the causal coefficients associated with independently-distributed external variables, reflect the IN-causal relation between internal variables and current and lagged external variables. The impulse response functions depend on A_0 and A_1, which are consistently estimated by least-squares regression due to

the fact that the error in each equation is uncorrelated with the explanatory variables in that regression.

Sims [**35**] characterized the Choleski decomposition as a harmless normalization, similar to the normalizations routinely used to convert underidentified models to identified models (for example, setting the coefficients of unobserved external variables equal to 1, as done above). In fact the two have nothing in common. Both the models (10.24)-(10.25) and (10.27)-(10.28) are identified, implying that application of the Choleski decomposition constitutes a substantive change in the model, not a normalization. This is obvious from the fact that the joint distribution of the external variables in eqs. (10.24)-(10.25) differs from that in eqs. (10.27)-(10.28) (unlike the joint distributions of Y_1 and Y_2, which are the same in the two models).

The fact that A_0 is triangular implies that the model is assumed to have no simultaneous equations, and also that the internal variables are causally ordered: $x_1 \rightarrow x_2 \rightarrow ... \rightarrow x_n$. Thus in formulating a vector autoregression intended for use in policy analysis the analyst must commit to which of the date-t internal variables cause which others. In the two-equation example it was assumed that we have $y_1 \rightarrow y_2$; we could equally well have assumed instead $y_2 \rightarrow y_1$. Doing so would generate different impulse-response functions. It follows that analysts who are unwilling to impose any restrictions on causal relations involving lagged terms, but are willing to exclude simultaneously-determined variables and to commit to a particular causal ordering of the date-t internal variables, will find vector autoregressions a useful analytical vehicle. This seems to be, at best, a narrow specification (however, some analysts, such as Christiano, Eichenbaum and Evans [**6**], are willing to assume triangularity explicitly; like Sims, they appear to regard this as a minor regularity condition). This dependence of policy analysis based on vector autoregressions on strong restrictions raises questions about about Sims' claim that causal analysis using vector autoregressions allows the analyst to avoid the "incredible" identifying restrictions employed under other procedures.

student #	earnings	Z_p	Z_a
1	110	1	1
2	100	1	1
3	110	0	1
4	60	1	0
5	30	0	0

Table 1

regression	dependent variable	Z_p	Z_a
1	Y_i	10	60
2	Y_i	20	
3	Y_i		61.67
4	Y_p		0.167
5	Y_a	0.167	

Table 2

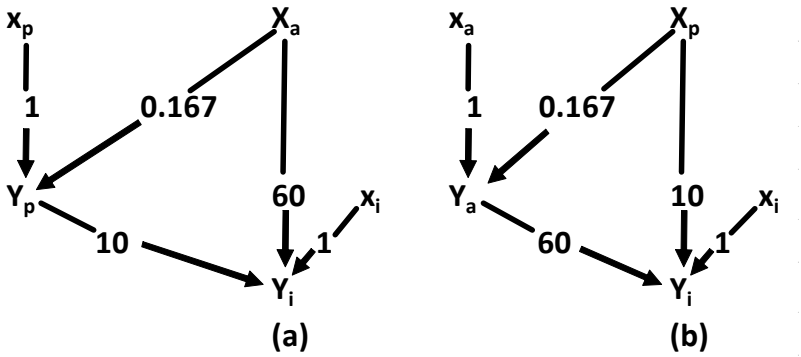

Figure 10.1

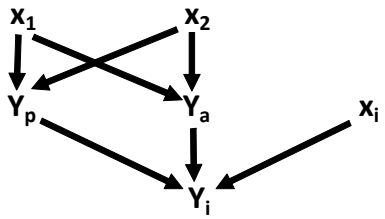

Figure 10.2

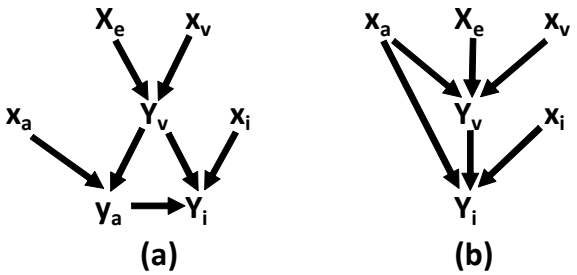

Figure 10.3

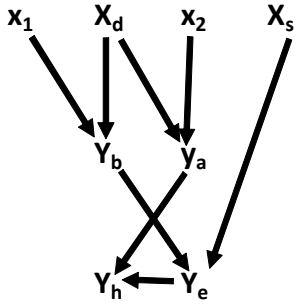

Figure 10.4

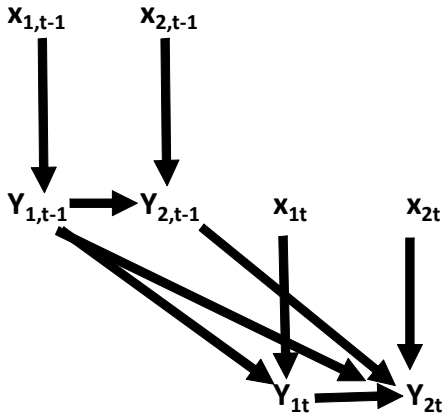

Figure 10.5

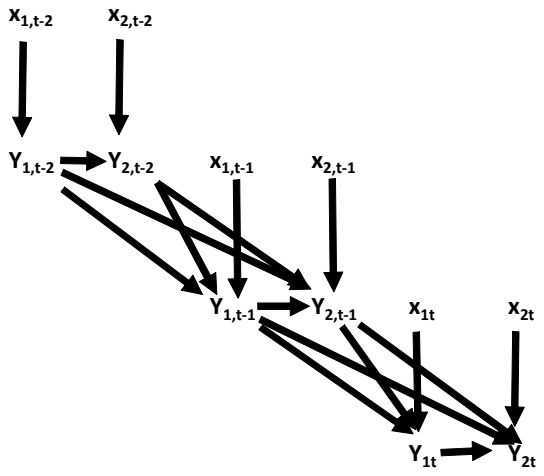

Figure 10.6

CHAPTER 11

Conclusion

Our purpose in this monograph has been to determine how to analyze causation in the context of a formal model consisting of a set of bloc-recursive equations. A preliminary step involved thinking carefully about what it means to analyze causation in the context of a model. We have taken the view that doing so implies that causal interventions be modeled as changes in the model's external variables. Failing to make this connection, we argued, would constitute implicitly altering the model, which is different from applying the model as specified. This requirement seems innocuous. However, we have seen that the developments involved in implementing the requirement take the analysis in new directions, leading to analyses that are substantially different from those generated by methods now in general use.

Bibliography

[1] Joshua D. Angrist. Lifetime earnings and the vietnam era draft lottery. *American Economic Review*, 80:313–336, 1990.

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

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

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

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

[6] Lawrence J. Christiano, Martin Eichenbaum, and Charles Evans. The effects of monetary policy shocks: Evidence from the flow of funds. *Review of Economics and Statistics*, 78:16–34, 1996.

[7] Thomas F. Cooley and Stephen F. LeRoy. Atheoretical macroeconometrics: A critique. *Journal of Monetary Economics*, 16:283–308, 1985.

[8] Stacy Berg Dale and Alan B. Krueger. Estimating the payoff to attending a more selective college. *Quarterly Journal of Economics*, 107:1491–1527, 2002.

[9] A. P. Dawid. Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95:407–424, 2000.

[10] Robert F. Engle, David F. Hendry, and Jean-Francois Richard. Exogeneity. *Econometrica*, 51:277–304, 1983.

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

[12] Daniel M. Hausman and James Woodward. Independence, invariance and the causal Markov condition. *British Journal of the Philosophy of Science*, 50:521–583, 1999.

[13] James Heckman and Rodrigo Pinto. Causal analysis after Haavelmo. *National Bureau of Economic Research*, 2013.

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

[15] David Hume. *A Treatise on Human Nature*. Clarendon Press, Oxford, 2007.

[16] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[17] Edward E. Leamer. Vector autoregressions for causal inference? volume 22. Carnegie-Rochester Conference Series on Public Policy, 1985.

[18] David S. Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48:281–355, 2010.

[19] Stephen F. LeRoy. Causal orderings. In Kevin D. Hoover, editor, *Macroeconometrics: Developments, Tensions and Prospects*. Kluwer Academic Publishers, 1995.

[20] Stephen F. LeRoy. Causality in economics. In Federica Russo and Jon Williamson, editors, *Causality and Probability in the Sciences*. College Publications, 2007.

[21] Stephen F. LeRoy. Implementation neutrality and causation. *Economics and Philosophy*, 32:121–142, 2016.

[22] Stephen F. LeRoy. Implementation-neutral causation in structural models. *Contemporary Economics*, 2018.

[23] Stephen F. LeRoy. Implementation neutrality and treatment evaluation. *Economics and Philosophy*, 34:45–52, 2018.

[24] Robert E. Lucas. Econometric policy evaluation: A critique. In Karl Brunner and Allan H. Meltzer, editors, *Carnegie-Rochester Conference Series on Public Policy*. North-Holland, 1976.

[25] Justin McCrary and Heather Royer. The effect of female education on fertility and infant health: Evidence from school entry policies using exact date of birth. *American Economic Review*, 101(1):158–195, 2011.

[26] Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Second Edition*. Cambridge University Press, New York, 2015.

[27] Emi Nakamura and Jon Steinsson. Identification in macroeconomics. *Journal of Economic Perspectives*, 32:59–86, 2018.

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

[29] Judea Pearl. Comment. *Journal of the American Statistical Association*, 95:428–431, 2000.

[30] Judea Pearl. Trygve Haavelmo and the emergence of causal calculus. *Econometric Theory*, 31:152–179, 2015.

[31] 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.

[32] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[33] Donald B. Rubin. Discussion. *Journal of the American Statistical Association*, 75:591–593, 1980.

[34] 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.

[35] Christopher Sims. Macroeconomics and reality. *Econometrica*, 48:1–48, 1980.

[36] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction and Search*. Springer-Verlag, New York, 1993.

[37] Donald L. Thistlethwaite and Donald T. Campbell. Regression discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51:309–317, 1960.

[38] Nanny Wermuth. On block-recursive regression equations. *Brazilian Journal of Probability and Statistics*, 6:1–56, 1992.

[39] James Woodward. Causation with a human face. In H. Price and R. Corry, editors, *Causation, Physics and the Constitution of Reality*. Oxford University Press, 2007.

Index