CAUSAL INFERENCE

Stephen F. LeRoy

University of California, Santa Barbara; November 27, 2018
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 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 some relation is or is not causal without any clear shared understanding of what it means for a 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 contained 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 given, to address 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 not involved in the 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. Typically the coefficients of internal variables in structural models are identified with causation whether or not doing so can be justified in terms of the implied interventions on the associated external variables.

This book is aimed at developing the view of causation just stated.

Earlier versions of this material were presented in Cooley-LeRoy [**5**] and LeRoy [**16**], [**17**], [**18**], [**19**], [**20**].

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. They may be either observed by the analyst or unobserved. $A = \{\alpha_{ij}\}$ 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. If there are fewer external variables than internal variables some of the internal variables are necessarily determined in simultaneous blocks. On the other hand, one often uses models in which there are more external variables than internal variables. For example, one might specify that an internal variable is a function of both observed external variables and an unobserved external variable. Prior to Chapter 7 attention is restricted to models that are linear in variables, as indicated in eq. (1.1).

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

The coefficients of B and G with respect to unobserved external variables are well defined only subject to an arbitrary scaling. The scaling usually adopted is to set either β_{ij} or γ_{ij}, depending on whether one is working with the structural form or the reduced form, equal to 1 when x_j is unobserved. 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.

[1] In an earlier literature the preferred terms were "endogenous" and "exogenous". However, more recently the latter term was assigned a different meaning in the econometrics literature (Engle, Hendry and Richard [**8**]), so it is avoided here. See Leamer [**14**] 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 external variables.

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 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 connection a recurrent theme has been that the structural form coefficients can be used to analyze interventions, and therefore locate causal orderings among internal variables, whereas the reduced-form coefficients cannot be used in this way. There remains the question of what it is about structural models that makes this so.

We will show 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.

This definition appears to allow each of two internal variables to cause the other. Some analysts have accepted this implication, but others share the view expressed below that simultaneously-determined variables should be distinguished from causally ordered variables—that is, causation is inherently asymmetric. If so it follows that causation is well defined only in fully recursive models, in which case A is triangular. One would prefer to have a treatment of models in which there may be recursive blocks, but equations within such blocks are simultaneous.

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. In the philosophy literature this property is known as "modularity". In the formulation (1.1), in contrast, the characterization of $=$ as a reflexive, symmetric and transitive operator implies that it is arbitrary which variable appears on the

left-hand side of an equation. Thus we do not have modularity: 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 incorporate 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, and this is so in both equation-system-based and graph-based discussions. 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. Thus structural models may or may not contain simultaneous blocks, consistent with their having a well-defined causal structure.

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. 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 formulation 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]

An example will make this clear. 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

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

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

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

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

[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.

(by any reasonable definition) y_3 with causal parameter 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 variables and different 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 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 [**9**]; see also Heckman and Pinto [**11**]). 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. (1.3) 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 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 harvest depends on the weather, and 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.

[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 really is constant. 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.

CHAPTER 2

Causation

For any internal variable y_i one can define as the *external set* of y_i the set of external variables x_j such that the i, j element of the reduced form coefficient matrix is nonzero. The external set of y_i is denoted $\mathcal{E}(y_i)$. $\mathcal{E}(y_i)$ 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 $\mathcal{E}(y_2) = \{x_1, x_2\}$.

The members of a set of m internal variables ($2 \leq m \leq n$, where n is the number of internal variables) are *simultaneously determined* if and only if the variables have the same external sets. In models containing a set of m simultaneously-determined variables there exists a unique set of $n - m$ equations that do not include any of the simultaneously-determined variables. The remaining m equations are termed the *simultaneous block*.

A model containing a simultaneous block can always be redefined by solving out internal variables in the simultaneous block so that only one member of a set of simultaneously-determined variables appears in each equation in the block. For example, in the model

$$y_1 = \beta_{11}x_1 \tag{2.3}$$

$$y_2 = \alpha_{21}y_1 + \alpha_{23}y_3 + \beta_{22}x_2 \tag{2.4}$$

$$y_3 = \alpha_{31}y_1 + \alpha_{32}y_2 + \beta_{33}x_3 \tag{2.5}$$

y_2 and y_3 are determined in the simultaneous block (2.4)-(2.5). The coefficients of the model can be redefined so that an equivalent model is

$$y_1 = \beta_{11}x_1 \tag{2.6}$$

$$y_2 = \alpha_{21}y_1 + \beta_{22}x_2 + \beta_{23}x_3 \tag{2.7}$$

$$y_3 = \alpha_{31}y_1 + \beta_{32}x_2 + \beta_{33}x_3. \tag{2.8}$$

Here eqs. (2.7) and (2.8) still constute a simultaneous block, but each equation in the block contains only one of the simultaneously-determined variables. This procedure will play a role in deriving causal orderings in models containing simultaneous blocks.

Two variables z_j and y_i (here z is a variable that may be external or internal) are *directly connected* if there exists at least one equation in which both appear (thus two external variables are never directly connected).

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 causes y_j also causes y_i, as the intuitive idea of causation suggests. The requirement for a proper subset, rather than just a subset, assures the asymmetry that is a defining element of causation. Also, it implies that simultaneously-determined variables are never causally related.

Direct causation is indicated by an arrow: x_j directly causing y_i is denoted $x_j \to y_i$, 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* (not being transitive, direct causation defines a relation, but not an ordering). 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 $x_j \to y_1 \to ... \to y_n \to y_i$. Here $x_j \to y_1 \to ... \to y_n \to y_i$ is termed a *causal path.* The indirect causal relation among 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 along one or more causal paths, or both.[1] [2]

By the definition 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

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{2.9}$$

$$y_2 = \beta_{22}x_2 + \beta_{23}x_3 \tag{2.10}$$

$$y_3 = \beta_{33}x_3 + \beta_{34}x_4 \tag{2.11}$$

$$y_4 = \alpha_{41}y_1 + \alpha_{43}y_3. \tag{2.12}$$

Here we have $\mathcal{E}(y_2) \subset\subset \mathcal{E}(y_4)$, but no causal path connects y_2 and y_4. Therefore y_2 does not cause y_4. A graph of this model is displayed in Example 3.4.

---

[1] 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 [**24**], p. 67).

[2] The fact that variables can be causally related along multiple paths appears to create the possibility that the causal effects cancel. However, for x to affect y along each of two canceling paths would imply that x, not being an element of y's external set, does not in fact cause y. Hereafter if one variable causes another on multiple paths it is assumed that these paths do not cancel.

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, we elect not to adopt the alternative definition. In any case, the question of how to define the causal ordering from the direct causal relation appears not to involve any substantive issues. Further, it turns out that the adopted definition of causation is easier to analyze than the alternative because under the adopted definition two variables that are causally related are always connected along a causal path.

Causation is transitive. That being so, it defines a partial ordering. A model's *causal ordering* is the set of pairs x_j, y_i and y_j, y_i such that x_j or y_j causes y_i. If y_j causes y_i it follows that y_j precedes y_i in the ordering, but, as with any partial ordering, the converse is not necessarily true. 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

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{2.13}$$

$$y_2 = \alpha_{21}y_1 + \beta_{22}x_2 + \beta_{23}x_3 \tag{2.14}$$

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. Graphs are presented in Example 3.5.

The direct causal relation can be derived by checking for all pairs x_j, y_i and y_j, y_i whether the conditions for $x_j \to y_i$ and $y_j \to y_i$ (direct connectedness and the proper subset condition) are satisfied. Doing so will produce the same direct causal relation whether or not simultaneous blocks have been resolved. This is so because, first, solving out variables in simultaneous blocks does not alter the external set of any internal variable. Doing so does eliminate direct connections between pairs of internal variables both of which are members of the same simultaneous block. However, second, these pairs of variables have the same external sets by definition, so neither causes the other regardless of whether or not they are directly connected. It follows that the set of pairs of internal variables that are causally connected is not affected by solving out simultaneously-determined variables.[3]

[3]The fact that solving out variables in simultaneous blocks, as just outlined, is innocuous does not mean that the same is true for any linear operations on the equations of a model. Solving out some of the variables of a model, which can be done via linear operations,

Put more simply, the conclusion is that solving out simultaneously-determined variables in a simultaneous block does not affect the direct causal relation or, therefore, the causal ordering. Accordingly, we can assume without loss of generality that models containing simultaneous blocks have been reparametrized in the manner described. Doing so facilitates construction of the causal ordering by the recursive algorithm outlined in the following section.

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

1. Structural Forms and Reduced Forms

The location of zeros in a model's reduced form determines whether the proper subset condition is satisfied for two internal variables. However, this information is not sufficient to allow determination of a model's causal ordering or direct causal relation. As discussed above, it is possible that neither of two internal variables causes the other even if the proper subset condition is satisfied. It follows that knowledge of a model's reduced form allows verification of a necessary condition for causation, but sheds no light on whether the two variables are connected by a causal path, which is necessary and sufficient.

As regards the direct causal relation, the situation is similar. Direct connectedness is a necessary condition for direct causation, but the reduced form provides no information about whether or not two variables are directly connected. Eqs. (2.13)-(2.14) provide an example of two models (one with β_{22} unrestricted, one with $\beta_{22} = 0$) that have different direct causal relations, but the same external sets: $\mathcal{E}(y_1) = \{x_1, x_2\}$ and $\mathcal{E}(y_2) = \{x_1, x_2, x_3\}$. The reduced forms of these models have zeros in the same places, implying that the reduced form cannot distinguish between them.

These results bear out the claim of the Cowles economists that structural models, from which the direct causal relation and causal ordering are derivable, provide information about causation that is lost in the reduced form. This theme is taken up again in Chapter 4, p. 24.

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. The construction begins with the model written in the 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

---

results in a different model. The altered model has fewer internal variables, and therefore necessarily has a different direct causal relation.

of the internal variables as effects of other variables. These 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 is a member of Λ. Each of the variables appearing in each of these equations other than the new variable is designated as a direct cause of the new variable in that equation. The new variables, one per equation so identified, are then included in Λ.

The second sub-round of the second round is the same as the first sub-round except 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 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 variable.

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. The process continues until all the internal variables are members of Λ.

An example will illustrate the construction. Consider the model

$$y_1 = \beta_{11}x_1 \tag{2.15}$$

$$y_2 = \alpha_{21}y_1 + \beta_{22}x_2 \tag{2.16}$$

$$y_3 = \alpha_{32}y_2 + \beta_{33}x_3 \tag{2.17}$$

$$y_4 = \alpha_{42}y_2 + \alpha_{43}y_3 + \beta_{44}x_4. \tag{2.18}$$

The first round identifies the internal variable y_1 in eq. (2.15), and adds it to Λ. The first sub-round of the second round identifies y_2 in eq. (2.16) and adds it to Λ. The second sub-round of the second round identifies y_3 in eq. (2.17) and adds it to Λ. Finally, the third round identifies y_4 in eq. (2.18) and adds it to Λ. This completes the construction.

In this construction it is assumed that in each round and sub-round there exists at least one equation that contains exactly one new (that is, not a member of Λ) internal variable. That would not be the case if in

some round all the internal variables not yet in Λ were determined in one or more simultaneous blocks, since in that event each remaining equation might contain two or more new variables. For example, in the model (2.3)-(2.5) this occurs in the second round of the construction of the direct causal relation. This problem would render continuation of the construction impossible. To forestall this problem one reparametrizes the equations of the simultaneous block so that only one new internal variable appears in each equation of the simultaneous block, as outlined above (p. 9). In the example this involves replacing the model (2.3)-(2.5) with (2.6)-(2.8). Doing so makes possible completion of the construction.

An alternative (but equivalent) version of this algorithm would involve reparametrizing each equation in each round and sub-round so that the new internal variable in each equation appears on the left-hand side, with all other variables in that equation appearing on the right-hand side. There results a model written in the form $y = Ay + Bx$.

This appears similar to the construction criticized in Chapter 1, but the construction here has a different basis. Here we begin with the model in the form $Ay = Bx$ and derive a model of the 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 variables on the left-hand side is justified even though $=$ is interpreted in its usual mathematical sense rather than as an assignment operator. Thus the model written in the form $y = Ay+Bx$ is derived from a prior characterization of direct causation, with this characterization then applied to a model initially written as $Ay = Bx$. 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.[4]

In Chapter 1 it was noted that the model written as $y = Ay + x$ with causation generated by interpreting $=$ as an assignment operator cannot accommodate simultaneous blocks. That is not true here. The derivation of $y = Ay + Bx$ in this chapter, with right-hand side variables directly causing left-hand side variables, implies that A is triangular (and singular due to the fact that the main diagonal of A consists of zeros). This is so whether or not the model includes simultaneous blocks.

[4]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. Doing so, 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. Accordingly, it is not clear how such a specification is to be justified.

3. 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 [**26**]. 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, assumed to be written in the form $y = Ay + Bx$. 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 above.

In our usage a *causal graph* is a graph that represents the direct causal relation 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 causation. Under the alternative form of the algorithm the reparametrization results in a graph with arrows pointing from the right-hand side variables to the left-hand side variable of each equation.

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. This procedure generates a graph in which each internal variable is caused by its ancestors and 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.

The causal graph allows an easy representation of reduced-form coefficients. With each causal path linking an external variable to an internal variable that it causes is associated a path coefficient consisting of the product of the coefficients associated with the direct causal relations that generate the path. Each internal variable is linked to each of the variables in its external set by one (or more) causal path(s). If there exists only one such path connecting the two variables the reduced-form coefficient of that internal variable with respect to each external variable in its external set equals the path coefficient for that path. If there exist pairs of variables on a causal path that are linked by indirect as well as direct subpaths, the contribution of those pairs to the reduced-form coefficient equals the sum of the path coefficients along the subpaths.

Note that this characterization of causal coefficients applies without qualification only when the cause variable is external. When the cause variable is internal there may or may not exist an analogue to the reduced-form coefficient that measures the causal effect of one variable on another. The corresponding characterization when the cause variable is internal is found in Chapter 4.

One generally cannot begin the analysis of causation with an arbitrarily specified direct causal relation or causal graph. 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. In the discussion below we do in fact sometimes begin with a causal graph, but in each case that graph was derived starting with the mathematical version. The derived causal graph is then reported, but exposition of the mathematical model is deleted for brevity.

CHAPTER 3

Examples

The analysis presented in the preceding chapter is illustrated using examples. In each case the model is defined mathematically and the associated causal graph is displayed.

Example 3.1

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

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

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

the graph of which is shown in Figure 3.1. 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.

Example 3.2

The standard economist's supply-demand model is

$$y_1 = \alpha_{12}y_2 + \beta_{11}x_1 \tag{3.3}$$

$$y_2 = \alpha_{21}y_1 + \beta_{22}x_2. \tag{3.4}$$

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 block of internal variables. The block 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 either by (1) comparing variables pairwise in the model as written to determine the existence of direct causation, (2) doing the same thing with the model derived after solving out the simultaneous block (which in this case results in a model that coincides with the reduced form), or (3) solving out the simultaneous block and applying the recursive derivation. The causal graph is shown as Figure 3.2.

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

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{3.5}$$

$$y_2 = \beta_{22}x_2 + \beta_{23}x_3 \tag{3.6}$$

$$y_2 = \beta_{22}x_2 + \beta_{23}x_3 \tag{3.7}$$

(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 $\alpha_{31}\beta_{12} + \alpha_{32}\beta_{22}$. See Figure 3.3.

Example 3.4

In the model

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{3.8}$$

$$y_2 = \beta_{22}x_2 + \beta_{23}x_3 \tag{3.9}$$

$$y_3 = \beta_{33}x_3 + \beta_{34}x_4 \tag{3.10}$$

$$y_4 = \alpha_{41}y_1 + \alpha_{43}y_3 \tag{3.11}$$

y_2 does not cause y_4 because y_2 is connected to y_4 only along paths that are not causal. See Figure 3.4.

Example 3.5

In the model

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{3.12}$$

$$y_2 = \alpha_{21}y_1 + \beta_{22}x_2 + \beta_{23}x_3 \tag{3.13}$$

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.

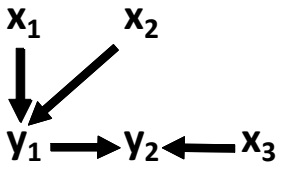

Figure 3.1

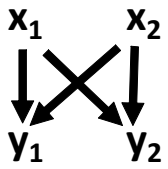

Figure 3.2

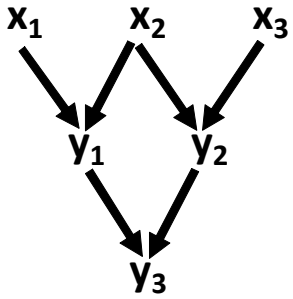

Figure 3.3

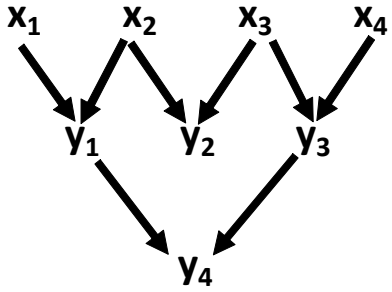

Figure 3.4

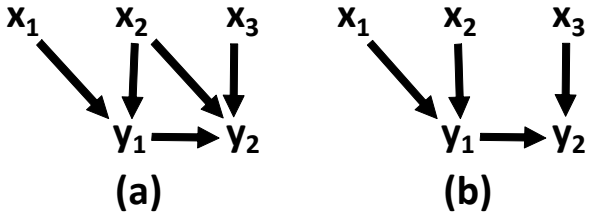

Figure 3.5

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 is 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 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 *implementation neutral.*[1] We refer to the causal relation so defined as *IN-causation.* If y_j causes y_i and the causation is implementation neutral we will write $y_j \Rightarrow y_i$. With $y_j \Rightarrow y_i$ the effect of y_j on y_i is by definition unambiguous. Thus if we have $y_j \Rightarrow y_i$ the effect of y_j on y_i is essentially the same as occurs when the cause variable is external: there exists a direct analogue of the reduced-form coefficient.

Note that under this definition causation represented by $\Rightarrow$, unlike that represented by $\rightarrow$, is not necessarily direct. This difference in definitions

[1]It appears that the first use of the term "implementation-neutral causation" was by Cartwright [**4**] in her discussion of LeRoy [**17**].

is motivated by the fact that $\rightarrow$ is used primarily in construction of causal graphs, in which restricting the use of $\rightarrow$ to direct causation is necessary if direct causation is to be distinguished from indirect causation. If one variable IN-causes another, in contrast, it does so along a unique path: if $y_j \Rightarrow y_k$ and $y_k \Rightarrow y_i$ then $y_j \Rightarrow y_i$, and the causal coefficient associated with $y_j \Rightarrow y_i$ is the product of the effect of y_j on y_k and that of y_k on y_i. Thus IN-causation $\Rightarrow$, unlike direct causation $\rightarrow$, is transitive. As with causation when the cause variable is external, if we have that $y_j \Rightarrow y_i$ causation is either direct or indirect (but not both, in contrast to the case with the direct causal relation). Further, 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 distinguish between direct and indirect IN-causation. We use $\Rightarrow$ for both.

If the cause variable 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}.

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.

Sometimes it is useful to work with graphs that depict IN-causation rather than causation, although we do not do so. 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 $z_j \Rightarrow y_i$ and there does not exist y_k such that $z_j \Rightarrow y_k \Rightarrow y_i$.

1. IN-Causation in Reduced-Form Models

It was seen in Chapter 2 that the direct causal relation and causal ordering can generally not be derived from the reduced form. However, in some simple models that derivation is possible. For example, consider the structural model

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 \tag{4.1}$$

$$y_2 = \alpha_{21}y_1 + \beta_{23}x_3, \tag{4.2}$$

which has the reduced form

$$y_1 = \gamma_{11}x_1 + \gamma_{12}x_2 \tag{4.3}$$

$$y_2 = \gamma_{21}x_1 + \gamma_{22}x_2 + \gamma_{23}x_3. \tag{4.4}$$

In such models the direct causal relation and the causal ordering coincide, and can be derived from the reduced form by exploiting the fact that the reduced form has a zero in the 1,3 position.

In this simple case one can also determine from the reduced form whether or not variables are IN-causally related. If the reduced-form coefficients satisfy

$$\frac{\gamma_{21}}{\gamma_{11}} = \frac{\gamma_{22}}{\gamma_{12}} \tag{4.5}$$

we can define the constant α_{21} by

$$\frac{\gamma_{21}}{\gamma_{11}} = \frac{\gamma_{22}}{\gamma_{12}} \equiv \alpha_{21} \tag{4.6}$$

and derive the structural model (4.1)-(4.2) from the reduced form (4.3)-(4.4). Thus the restriction (4.5) implies $y_1 \Rightarrow y_2$.

If the generic reduced form of the model just set out can be rewritten as

$$y_1 = \gamma_{11}x_1 + \gamma_{12}x_2 \tag{4.7}$$

$$y_2 = \alpha_{21}\gamma_{11}x_1 + \alpha_{21}\gamma_{12}x_2 + \gamma_{23}x_3 \tag{4.8}$$

for some α_{21}, it satisfies the restrictions (4.5) by construction, implying that the derivation of the structural form eqs. (4.1)-(4.2) is immediate. We will use the term "restricted reduced form" to refer to the version of the reduced form that incorporates the reduced-form restrictions implied by some structural model, as in eqs. (4.7)-(4.8). Thus structural models, or equivalently restricted reduced forms, can contain information about both causation and IN-causation. In this they differ from unrestricted reduced forms, which may contain information about causation, but not about IN-causation.[2]

[2]To make the same point in vector-matrix notation, note that the restricted reduced form can be written as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \beta_{11} & \beta_{12} & 0 \\ \alpha_{21}\beta_{11} & \alpha_{21}\beta_{12} & \beta_{23} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}. \tag{4.9}$$

This can be shown to coincide with the structural form (4.1)-(4.2), written in vector-matrix form as

$$\begin{bmatrix} 1 & 0 \\ -\alpha_{21} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \beta_{11} & \beta_{12} & 0 \\ 0 & 0 & \beta_{23} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \tag{4.10}$$

The possibility of encoding structural information in reduced forms 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 [**9**], Wermuth [**35**] and Pearl [**28**] 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. This statement is correct when the probability distribution of internal variables is viewed as generated by applying the generic reduced form to the external variables, the probability distribution of which is assumed. However, it is incorrect as applied to the restricted reduced form. As the above example shows, structural coefficients like α_{21} in fact appear in restricted reduced forms, and therefore can be viewed as figuring in the link between assumed distributions of external variables and the implied distributions of internal variables. Specifically, reduced forms that incorporate structural coefficients can give information about whether causal relations are or are not IN-causal.

We noted in Chapter 2 that deriving causal orderings from reduced forms in this way runs into problems with models that are even slightly more complex than (4.1)-(4.2). The problems multiply when attention is turned to IN-causation. The fact that IN-causation is a special case of causation implies that two variables are not candidates for an IN-causal relation if they are not causally related (for example, there is no possibility of finding a restriction on the nonzero elements of the reduced form (4.3)-(4.4) that will generate $y_2 \Rightarrow y_1$). Second, even if one is willing to stipulate that y_1 causes y_2 it is an open question whether there necessarily exist restrictions on reduced-form coefficients analogous to eq. (4.5) that imply $y_1 \Rightarrow y_2$. For instance, in Example 3.5 we have $y_1 \rightarrow y_2$ but not $y_1 \Rightarrow y_2$ unless in addition we have $\beta_{22} = 0$. But β_{22} is a structural coefficient, not a reduced-form coefficient, so this example does not identify a restriction on the reduced form that implies $y_1 \Rightarrow y_2$. Third, even if there exist restrictions on the reduced form that imply that y_1 causes y_2 can be strengthened to $y_1 \Rightarrow y_2$, it remains true that except in the simplest models it is generally impossible to determine causal orderings from reduced forms, as we have seen, so there is generally no way to determine whether we have $y_1 \rightarrow y_2$.

We conclude that inferring causal and IN-causal orderings from reduced forms is problematic at best. However, these problems disappear if one works directly with the structural form instead of the reduced form. Again we substantiate the Cowles economists' claims in favor of structural models over reduced forms.

2. IN-Causation in Structural Models

IN-causation is most conveniently analyzed using causal graphs. We have $y_j \Rightarrow y_i$ if and only if y_j causes y_i and all the causal paths linking

by inverting the matrix and multiplying.

members of $\mathcal{E}(y_j)$ to y_i pass through y_j.[3] 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. 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 a fixed 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 situation displayed in Example 3.4 is ruled out. In that case $\mathcal{E}(y_j) \subset\subset \mathcal{E}(y_i)$ is equivalent to y_j causing y_i. Even with this restriction, whether or not the causation is IN-causal cannot be distinguished from the reduced form. This is so because 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), which has the same reduced form as Example 3.5(a), y_1 does IN-cause y_2.

3. Conditional 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 may be available if we condition on a 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 confounding variables are members of Ψ. If so, y_j IN-causes y_i conditional on Ψ, and we will write $y_j \Rightarrow y_i|\Psi$. For any y_j and y_i such that y_j causes y_i there will usually exist 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.[4]

Hereafter "y_j IN-causes y_i" without qualification is taken to refer to unconditional IN-causation.

[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 [**36**].

[4]Sometimes there does not exist Ψ that has this property. For example, in the Thistlethwaite-Campbell [**34**] model discussed in Chapter 9 this is the case. Other instances are found in Chapter 10, Examples 1 and 4.

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 argument applies when the variables conditioned on are external. It follows that making conditional causation statements as defined here does not involve an alteration of the model. We may have $y_j \not\Rightarrow y_i$ and also $y_j \Rightarrow y_i|\Psi$ for some Ψ; these statements are different, but are not inconsistent.

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

$$y_1 = \beta_{11}x_1 + \beta_{12}x_2 + \beta_{13}x_3 \tag{4.11}$$

$$y_2 = \alpha_{21}y_1 + \beta_{23}x_3 + \beta_{24}x_4 \tag{4.12}$$

we have $y_1 \Rightarrow y_2|x_3$. If eq. (4.12) is replaced by

$$y_2 = \alpha_{21}y_1 + \beta_{21}x_1 + \beta_{23}x_3 + \beta_{24}x_4 \tag{4.13}$$

we have $y_1 \not\Rightarrow y_2|x_3$ despite the fact that the models (4.11)-(4.12) and (4.11)-(4.13) both have reduced forms that can be written as

$$y_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + \gamma_{13}x_3 \tag{4.14}$$

$$y_2 = \gamma_{21}x_1 + \gamma_{22}x_2 + \gamma_{23}x_3 + \gamma_{24}x_4. \tag{4.15}$$

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 other 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 its causal coefficient is 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 due to the arbitrary scaling of x_1. 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 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

[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

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

Only one of the examples below includes a latent variable, so we make no formal allowance for them in the notation.

G include ones to reflect the normalizations required by the fact that x is unobserved.[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 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. For example, if X IN-causes Y the causation coefficient is not identified if the association between X and the error term is unrestricted.

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. This amounts to essentially ignoring the correlation. 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 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

[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. However, in some situations setting to 1 coefficients of unobserved external variables that appear in more than one equation may be acceptable (see Chapter 10, note 1 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 assigned by lotteries there is no question that the treatment variable is statistically independent of other external variables.

example discussed below, one has a strong prior belief in the existence of such a causal link. At a minimum, resolving the misspecification of these variables as external involves respecifying one of the two correlated variables as internal.[5] Respecifying one of the variables as internal renders it necessary to introduce a new external variable. 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 block 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, either unconditional or conditional. This construction is discussed in Chapter 10, Example 1. As the Cowles economists led us to expect, models do not have strong empirical implications if no strong assumptions are imposed.

---

[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.

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). This is particularly so in treatment evaluations. 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 dependence 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

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

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

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. These equations imply 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, Example 1 includes an application of this use of the mean-independent specification.

CHAPTER 6

Regression and Correlation

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

1. Causation and Correlation

Holland [**12**] cited G. A. Barnard as writing "That correlation is not causation is perhaps the first thing that must be said."[1] It is usually 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 external variable x is probabilistically dependent on (that is, not probabilistically independent of) an internal variable y if and only if $x \in \mathcal{E}(y)$, so that there exists a causal path that connects x and y. Further, x is dependent on y_2 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.[2] 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 of this expression 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

[1]For a related analysis under the rubric of "spurious correlation" see Simon [**32**]. By "spurious correlation" Simon meant correlation where there is no causation. Here Simon can be interpreted as anticipating the idea of implementation neutrality, although his analysis differs from that found here.

[2]It was observed above that conditioning on an internal variable constitutes an alteration of the model. Conditioning there, referring to relabeling 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 by different names, but both usages of "conditioning" are well entrenched, although rarely distinguished. The same point applies to the ambiguous term "holding constant". The context will always make clear which meaning is intended.

$x_2 \to y_2 \to y_3$, 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 exist causal paths from x to y_1 and also from x to y_2, so that the external sets of y_1 and y_2 overlap. The variables y_1 and y_2 are dependent conditional on y_3 if and only if at least one of these paths does not include y_3. Thus absence of statistical dependence implies absence of causation, but the presence of statistical dependence does not imply causation. The mantra is correct.

2. Univariate 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 assumption that the x_j are independent implies that we have

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

assuming that the external variables have finite second moments.

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 (3.12)-(3.13), which has $y_1 \Rightarrow y_2|x_2$ but not $y_1 \Rightarrow y_2$. Neglecting means, 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 (3.12)-(3.13) 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, it also coincides with the population regression coefficient of y_2 on y_1.

3. Multivariate Regressions

Suppose that two internal variables are IN-causally related ($y_1 \Rightarrow y_3$) and an external variable $x \in \mathcal{E}(y_1)$ is included as an explanatory variable in the (population) regression of y_3 on y_1. In that case the population regression coefficient of x equals zero (this is the causal Markov condition, discussed below). Whether or not $x \in \mathcal{E}(y_1)$ the regression coefficient of y_3

on y_1 is unaffected by the inclusion of x. Also, if $x \notin \mathcal{E}(y_1)$ but $x \in \mathcal{E}(y_3)$ the population regression coefficient of x is nonzero and it coincides with the causal coefficient.

Assume that two internal variables are causally related, but not (unconditionally) IN-causally related: $y_1 \not\Rightarrow y_3$, and also that there exists a single confounding variable x, so that $y_1 \Rightarrow y_3|x$. Consider the bivariate regression of y_3 on y_1 and x. The coefficient of y_1 in the regression quantifies the effect of y_1 on y_3 conditional on x. It is identified, assuming that the cause variable, the effect variable and the confounding variable are observed. This is, of course, different from the unconditional effect of y_1 on y_3, which is undefined. The coefficient of x in the regression does not quantify the effect of x on y_3. This is so because, as a confounding variable, x affects y_3 both directly and indirectly, and the regression coefficient captures only the direct effect (see Chapter 10, Example 1).

Assume now that we have $y_1 \Rightarrow y_3$ and another internal variable y_2 is included with y_1 as a regressor. If in addition $y_2 \Rightarrow y_3$, with both y_1 and y_2 directly IN-causing y_3, then $\mathcal{E}(y_1)$ and $\mathcal{E}(y_2)$ are disjoint, implying that y_1 and y_2 are independent. In that case the bivariate regression coefficients of y_1 and y_2 are the same as in univariate regressions and, assuming that the relevant variables are observed, the sample regression coefficients give consistent estimates of population coefficients.

If $y_1 \Rightarrow y_3$ and $y_2 \not\Rightarrow y_3$ the regression coefficient of y_1 in a bivariate regression of y_3 on y_1 and y_2 may or may not coincide with the IN-causal coefficient. In the causal graph shown in Figure 6.1 the coefficient of y_1 is not affected by the inclusion of y_2 in the regression. This is so again because $\mathcal{E}(y_1)$ and $\mathcal{E}(y_2)$ are disjoint. In this case we have $y_2 \Rightarrow y_3|x_3$. In Figure 6.2 we still have $y_1 \Rightarrow y_3$, but now inclusion of y_2 as a regressor affects the regression coefficient of y_1. This occurs because x_3 is an element of both $\mathcal{E}(y_1)$ and $\mathcal{E}(y_2)$. It follows that y_1 and y_2 are correlated, implying that the coefficient of y_1 in the bivariate regression of y_3 on y_1 and y_2 differs from that in the univariate regression, which equals the IN-causal coefficient. Here y_2 may (as in Figure 6.1) or may not (as in Figure 6.2) conditionally IN-cause y_3.

Note that in the model of Figure 6.2 inclusion of y_2 in the regression of y_3 on y_1 introduces bias in the regression coefficient of y_1 even though y_1 IN-causes y_3.

4. Instrumental Variables

Suppose as above that one internal variable causes, but does not IN-cause, another, implying existence of a confounding variable. If the confounding variable is not observed the coefficients linking it to the cause and effect variables are not identified. However, it is still possible to analyze causation conditional on the confounding variable using instrumental variables. Doing so requires the existence of a member of the external set of

the cause variable other than the confounding variable. That variable will be the instrument. Assuming that the cause variable, the effect variable and the instrument are observed, the coefficient associated with conditional causation is identified.

In Figure 3.3 we have that y_1 IN-causes y_3 conditional on x_2. With x_2 held constant the only path connecting y_1 and y_3 originates at x_1. Therefore α_{31}, the coefficient associated with $y_1 \Rightarrow y_3|x_2$, equals the path coefficient from x_1 to y_3 divided by the path coefficient from x_1 to y_1. Using eq. (6.4) we have

$$\alpha_{31} = \frac{\beta_{31}}{\beta_{11}} = \frac{\text{cov}(x_1, y_3)/\text{var}(x_1)}{\text{cov}(x_1, y_1)/\text{var}(x_1)} = \frac{\text{cov}(x_1, y_3)}{\text{cov}(x_1, y_1)}. \tag{6.6}$$

The rightmost term in eq. (6.6) is recognized as the population counterpart of the instrumental variables estimate of α_{31}, with the instrument being x_1. Assuming that x_1, y_1 and y_3 are observed the instrumental variables regression can be implemented empirically. This is so whether or not x_2 is observed.

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.

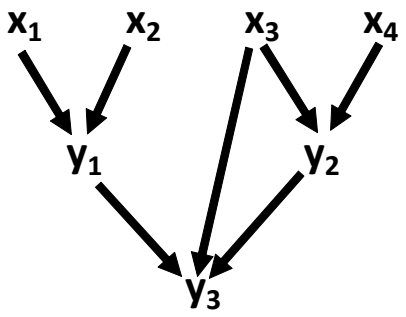

Figure 6.1

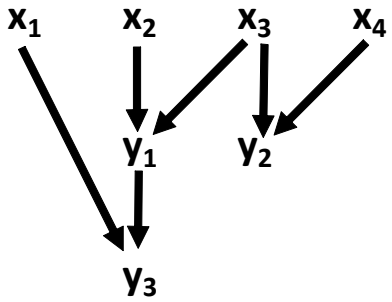

Figure 6.2

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 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.

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.

We have seen that when the map from external to internal variables is linear, reduced forms are very simple: for each pair consisting of one external and one internal variable, with the former causing the latter, there exists a coefficient that maps interventions on the external variable to their effects on the internal variable. With nonlinear models such maps are well-defined functions only when the structural equations have a solution that is unique (existence and uniqueness are guaranteed in the linear case by the assumption that A in eq. (1.1) is nonsingular, but no such simple condition carries over to nonlinear settings). Assume that the solution exists and is unique.

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 an external variable to be a member of the external set of an internal variable if that variable affects the internal variable for some set of values taken on by other elements of the external set, not necessarily for all such values. By that standard all of the members of the firing squad cause the outcome, because each member's shot determines the outcome 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. In contrast, in linear settings the effect of an intervention is never zero if an external variable causes an internal variable. 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.

Even if a nonlinear model has a solution that is unique, it is no longer sufficient to characterize the intervention by the variation it induces on one or more of the external variables: it is necessary to specify the baseline value of the external variable and its value under intervention separately. Also, it is necessary to specify the assumed values of the other variables in the external set of the cause variable. The constant measuring causation in the linear case is replaced in the nonlinear case by an internal variable that depends on all these variables.

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 $x_1 \in \mathcal{E}(y_2)$ is directly connected to y_2, and 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 causal paths from the external set of the cause variable to the effect variable pass through the cause 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

$$y_1 = \begin{cases} 1 \text{ if } \beta_{21} x_1 + x_2 \geq 0 \\ 0 \text{ otherwise} \end{cases} . \tag{7.1}$$

If y_2 is determined by

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

we have that x_1 and x_2 IN-cause y_1, with the effect of an intervention on x_1 depending on both the baseline value of x_1 and its value under intervention, and also on x_2, due to the nonlinearity. Further, y_1 IN-causes y_2 with causal coefficient α_{21}. The fact that IN-causation is representable by a single coefficient here reflects the fact that even though the model is nonlinear, eq. (7.2) is linear. This model has the causal diagram shown as Figure 3.1.

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, 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 the variables x and y. 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 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 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 of interventions on shallow parameters on internal variables are generally not defined due to failure of IN-causation. This is the Lucas [**21**] Critique. An example is found in the following section.

2. Multidate Models

Causation analysis is essentially the same in multidate models as in the single-date models studied up to now. Demonstrating this requires clarification of terminology. In many discussions involving multidate models the term "variable" is used in reference to an n-tuple or sequence (assuming discrete time) of mathematical variables indexed by 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 of the process y_t will be termed variables. Thus in multidate models we have two types of mathematical variables: elements of processes and parameters.

---

[1] In classroom lectures Lawrence Klein defined parameters as "constants that vary". We graduate students were amused; 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.

In many discussions causation is represented as being inextricably linked to time: causes are viewed as necessarily preceding effects in time. If so, static multidate models—models in which internal variables have external sets that include future-dated variables—are ruled out. Doing so is a mistake: there is no justification for a doctrinaire injunction against deterministic multidate models. 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 specify that information strictly increases over time, so that the σ-algebra at any date is a proper subset of that at later dates. 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 $y_{j.t+1} \to y_{i,t}$.

The simplest efficient markets finance model will illustrate analysis of causation in multidate models, and will indicate the consequences of specifying coefficients as parameters vs. constants. The model relates three processes: two internal processes 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 independent 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 is the reciprocal of the expected gross rate of return. 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 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. 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 $\{\delta, \alpha_{dd}, x_t, ..., x_0\}$. 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 [**21**]: effects of interventions on shallow parameters may not be well defined due to failure of IN-causation.

3. 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 [**33**], 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 [**10**] 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 variable can be written as a deterministic function of its parents, any 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, of course, 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 unobserved variables that are causal parents from those that cause a variable but are not counted among its parents. Variables characterized as errors are, of course, 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 which 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.

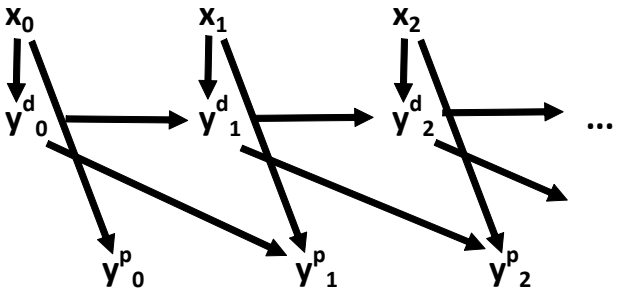

Figure 7.1

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 [**31**]. 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. As 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 [**23**] and Imbens-Rubin [**13**] 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 [**12**]). 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 [**27**] in his comments on Dawid [**7**] 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 [**12**]). 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]Note in this connection that 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 (usually called expectations, although no random variables have been defined) of agent-specific values.

In the terminilogy here 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 value of 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 [**30**]), 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. 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.[2]

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 necessary and 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 would not necessarily 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 linear regression.

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 a 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 this is not clear. These definitions are not the same: any variable x in the external set 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 effect variable pass through the treatment variable.

In the potential outcomes literature the causal coefficient is assumed to be well defined even in the presence of confounding variables, whereas here we have asserted that under our definition 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.

In many discussions in the potential outcomes literature confounding variables are identified with differential access to treatment. A variety of statistical treatments are proposed to deal with the bias supposedly introduced by differential access to treatment induced by confounding variables, assuming these to be observed. Rosenbaum and Rubin [**29**] proposed *propensity scores*, defined as statistics based on observed confounding variables that measure differences in treatment probabilities (see Athey-Imbens [**3**] for a recent discussion). It is asserted that if propensity scores are held constant then the bias induced by differential access to treatment is eliminated.

The contention that differential access to treatment necessarily induces bias 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 (except in a footnote) on aspects of treatment evaluation other than those associated with use of the potential outcomes framework.

An extended example will make clear some of the issues. This example is drawn from Thistlethwaite-Campbell [**34**]. Suppose that a group of students has taken Examination 1, with scores denoted X_1. Some time later they take another examination, 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 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. Here we are labeling the treatment as internal, pending discussion below. We have the model

$$Y_2 = \alpha_0 + \alpha_{2t} Y_t + \beta_{21} X_1 + x_2, \tag{9.1}$$

where x_2 is an unobserved error. The task is to estimate α_{2t}.

The first question is whether the model just specified is linear affine. It appears from eq. (9.1) that this is so, and the designation of such models as linear is frequently encountered in the treatment evaluation literature (Lee-Lemieux [**15**], p. 286, for example). The reduced form for this model is

$$Y_t = \begin{cases} 1 \text{ if } X_1 \geq 0 \\ 0 \text{ otherwise} \end{cases} \tag{9.2}$$

$$Y_2 = \begin{cases} \alpha_0 + \alpha_{2t} + \beta_{21} X_1 + x_2 \text{ if } X_1 \geq 0 \\ \alpha_0 + \beta_{21} X_1 + x_2 \text{ otherwise} \end{cases}, \tag{9.3}$$

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 was 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. (9.2) here). If Y_t is clearly specified as an internal variable, then eq. (9.2) must be included along with eq. (9.1) as part of the structural version of the model. The revised structural form of the model, now consisting of eqs. (9.1) and (9.2), is nonlinear, like the reduced form.

In discussing 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 neither presumption is admissible. Y_t does not IN-cause Y_2, either unconditionally or conditional on any variable of the model, so α_{2t} does not have a causal interpretation.[1] The effect of X_1 on Y_2, on the other hand, is well defined (eq. (9.3)). The relation between X_1 and Y_2 being nonlinear, the causal effect cannot be associated with a single coefficient that multiplies the intervention on the cause variable. 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. 44 above). The effect on Y_2 of an intervention on X_1 is

$$\begin{cases} \alpha_{2t} + \beta_{21}(X_1^i - X_1^b) \text{ if } X_1^b < 0 \text{ and } X_1^i \geq 0, \\ -\alpha_{2t} + \beta_{21}(X_1^i - X_1^b) \text{ if } X_1^b \geq 0 \text{ and } X_1^i < 0, \\ \beta_{21}(X_1^i - X_1^b) \text{ otherwise.} \end{cases} \tag{9.4}$$

This expression incorporates both the direct effect of X_1 on Y_2, which is linear, and the indirect effect via Y_t, which is nonlinear.[2]

---

[1] Here the external set of the treatment variable consists of a single variable that also affects the outcome variable directly. Conditioning on that variable does not define conditional causation because Ψ is not a proper subset of $\mathcal{E}(Y_t)$ (the two coincide).

[2] In the potential outcomes framework the date-2 outcome would be represented by two variables—Y_2^1 and Y_2^0—given by

$$Y_2^1 = \alpha_0 + \alpha_{2t} + \beta_{21}X_1 + x_2 \tag{9.5}$$

$$Y_2^0 = \alpha_0 + \beta_{21}X_1 + x_2, \tag{9.6}$$

according to whether a student did or did not participate in the special instruction. Here Y_2^1 is observed only when $X_1 \geq 0$, and Y_2^0 is observed only when $X_1 < 0$.

This construction involves hypothesizing an intervention on an internal variable, the treatment, without connecting the intervention to the external variable, X_1, that according to the model determines treatment. Doing so constitutes an alteration of the model. The nature of the alteration is not clear; at a minimum it involves deleting eq. (9.2).

Also, characterizing Y_2^1 as unobserved when $X_1 < 0$ represents a radical departure from the usual meaning of "unobserved": if $X_1 < 0$, Y_2^1 is undefined according to the model, not unobserved. This exemplifies the observation in the preceding chapter that potential outcomes are undefined in the absence of IN-causation.

1. Population and Sample

As has been observed, the assumption that external variables are independently distributed plays an important role in causal analysis. Here the concern is with the assumed nature of the population. Another element of model specification, distinct from that just described, involves characterizing the independence, or lack thereof, of draws for the members of the sample upon which empirical work is based. As regards the latter task, two specifications are prominent in empirical work: that draws are independent and that they are dependent but random. A random sample has the property that each observation is drawn from the same population. If samples are identical and independent they are random, but the converse is not necessarily true.

An example will make clear the distinction between independent samples and random samples. Suppose, following Neyman [**25**] (analysts credit Neyman for creating the field of treatment evaluation in this paper), that a farmer wishes to determine whether fertilizer A or fertilizer B improves the crop yield more. The assumed model of the population is

$$Y = \beta X + x, \tag{9.7}$$

where Y is the crop yield, X is a binary variable indicating whether fertilizer A or fertilizer B is applied, and x is an unobserved error term. It is assumed that X and x are independently distributed.

The farmer has two identical plots of land. If he or she assigns fertilizer use via independent coin tosses there is a probability of 50% that the same fertilizer will be applied to both plots. In that event the farmer learns nothing from the experiment about which fertilizer is better. A better strategy would be to choose the fertilizer to be applied to plot 1 via a coin toss, and then assign the other fertilizer to plot 2. By so doing the farmer is certain to get some information about which fertilizer is better. This treatment assignment is random, meaning that the probabilities of each fertilizer being assigned to one plot is the same as that for the other, but the draws are not independent.

In any treatment evaluation exercise the analyst has two decisions to make: whether to assume that the population external variables are independently distributed and, assuming that he or she controls the treatment assignment, whether to use independent draws in the sample. Neither decision is related to the other. We saw above that in some settings it is appropriate to weaken the assumption of independence in the population external variables, and we see here that in some settings it is appropriate to reject treatment assignment based on independent draws.

In many treatment evaluation exercises, particularly those employing the potential outcomes approach, the distinction between the assumed distibutions of population variables—independent vs. mean-independent, for

example—and the assumed nature of the sample—independent vs. random, in the above example—is not clearly drawn. This occurs because the population and the sample are not distinguished, as was observed above.

2. Regression Discontinuity

Analysts evaluating treatment effectiveness are increasingly relying on statistical procedures involving regression discontinuity to estimate causal coefficients (see Lee-Lemieux [**15**] 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.

The model set out in the preceding section is that of Thistlethwaite-Campbell, who are credited with introducing regression discontinuity in their paper [**34**] discussing that model. There is no doubt that regression discontinuity is an important idea, but Thistlethwaite-Campbell's model is not the vehicle that makes clear why it makes sense to ignore data away from the discontinuity.

Leave aside the question of whether α_{2t} can be interpreted as quantifying a causal relation. That coefficient can be estimated using the bivariate regression (9.1). This 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 parameter 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.

We are interested in models in which Y_1 and Y_2 are both observed, and we have $Y_1 \to Y_2$ but not $Y_1 \Rightarrow Y_2$. Then there necessarily exists at least one confounding variable; suppose that there exists just one. We wish to estimate the 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

[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 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.

links the confounding variable to the cause variable or the path linking it to the effect variable contains a 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 the forcing variable now IN-causing the outcome variable, the causal coefficient is well defined and can be estimated by ordinary 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 four examples of the preceding analysis are discussed. The variables in the models discussed 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 [**6**].

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 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 earnings 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 earnings 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. 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 earnings. 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 earnings, 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 earnings, 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 earnings 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 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 earnings. 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). 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 61.7. 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 earnings of school choice conditional on affluence, not the unconditional effect.

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 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 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 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.

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.[1]

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.

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

[1] 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 these 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.

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 earnings. 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 earnings 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 earnings 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).[2] 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 linear 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 identified due to the assumption that y_a is not observed.

Instead of having veteran status IN-causing earnings 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 earnings ability in civilian employment. A model that reflects this respecification is the following:

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

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

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 is not observable 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 [**22**]. 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

[2]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.

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:

$$Y_b = \beta_{bd} X_d + x_1 \tag{10.14}$$

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

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

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

Eq. (10.14) connects the age at which mothers begin their education, Y_b, to the month and day they were born, X_d. 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 structural model (10.14)-(10.17) is written in the form $y = Ay + Bx$, so that right-hand side variables directly cause left-hand side variables. 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 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) 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 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 stays 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 the role of regression discontinuity is 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 only 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 is the confounding variable. It would be well for them to address this task explicitly.

4. Granger Causation

Much has been written about the relation, or lack thereof, between causation and Granger-causation. These discussions—including Cooley-LeRoy [**5**]—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 [**5**]). Therefore acceptance of Granger non-causation does not imply strict exogeneity, although it is consistent with 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 expressing the values of the money stock $M = \{M_t\}$ and gross domestic product $Y = \{Y_t\}$ (in this section we use Y to denote GDP, not to represent a general internal variable as elsewhere), both observed, as linear functions of each other, their once-lagged values, and unobserved external errors $x_M = \{x_{Mt}\}$ and $x_Y = \{x_{Yt}\}$. The errors are assumed independent cross-sectionally and over time. In this model the money stock is strictly exogenous if the errors in the equation for income do not feed back into the equation for the money stock, either currently or with a lag. This condition is satisfied if the coefficients of current and lagged income in the equation for money equal zero.

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

$$M_t = \alpha_{MM1} M_{t-1} + x_{Mt} \tag{10.22}$$

$$Y_t = \alpha_{YM} M_t + \alpha_{YM1} M_{t-1} + \alpha_{YY1} Y_{t-1} + x_{Yt}. \tag{10.23}$$

Here we have $M_t \to Y_t$ because the external set of M_t consists of the errors $x_{M\tau}, \tau \leq t$, whereas the external set of Y_t consists of all the errors $x_{M\tau}$, $x_{Y\tau}$, $\tau \leq t$; thus the proper subset condition is satisfied, and M_t is directly connected to Y_t. However, we do not have $M_t \Rightarrow Y_t$ in the equation system as written: the lagged errors in the M equation that affect Y_t via M_t also affect Y_t via M_{t-1} and Y_{t-1}, so they are confounding variables.[3] It follows that strict exogeneity of M does not imply that α_{YM} can be interpreted as quantifying the effect of M on Y. We have $M_t \Rightarrow Y_t$, so that α_{YM} does represent the causal effect of M_t on Y_t, under the additional restriction that α_{YM1} equals 0. Further, with α_{YM1} unrestricted there exists no set of variables Ψ such that we have $M_t \Rightarrow Y_t|\Psi$, since all the elements of $\mathcal{E}(M_t)$ are confounding variables.

Thus neither Granger noncausation nor the stronger assumption of strict exogeneity of M is sufficient to establish IN-causation, either unconditional

[3]Specifically, if the intervention ΔM_t is caused by an intervention on x_{Mt} the effect on Y_t is $\alpha_{YM}\Delta M_t$. If the intervention is on $x_{M,t-1}$ its effect on Y_t is $(\alpha_{YM} + \alpha_{YM1}/\alpha_{MM1})\Delta M_t$. Thus $x_{M,.t-1}$ is a confounding variable in the causal relation between M_t and Y_t. The same applies for the other lagged terms.

or conditional. As in the static case, if α_{YM1} is nonzero the question "What is the effect on Y_t of an intervention that brings about ΔM_t?" does not have an unambiguous answer: different interventions consistent with a given change in M_t have different effects on Y_t.

One could make a case for M IN-causing Y conditional on lagged M and Y. However, as argued above, conditioning on internal variables in this way induces functional relations among the purportedly external variables—the errors—that conflict with their assumed independence. It was suggested that these functional relations constitute an alteration of the model, so that we are no longer investigating the causal relation between M and Y in the model that was specified.

Whether or not have α_{YM1} and α_{YY1} equal to 0, it can be asserted that strict exogeneity implies that M_t IN-causes Y_t conditional on all the lagged errors in money. That assertion is acceptable, these terms being external. However, this conditional causation should not be confused with the unconditional causation that was discussed above: the associated intervention is specified differently in the two cases.

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		61.7
4	Y_p		0.167
5	Y_a	0.167	

Table 2

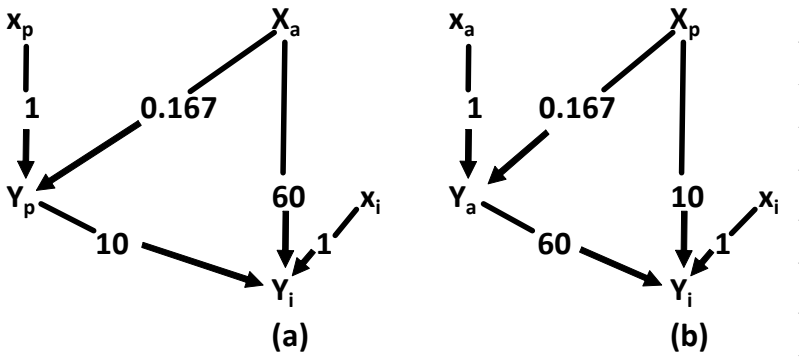

Figure 10.1

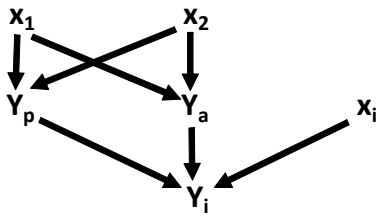

Figure 10.2

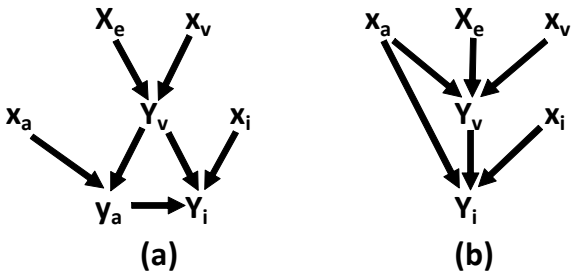

Figure 10.3

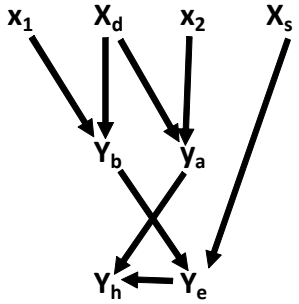

Figure 10.4

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[17] Stephen F. LeRoy. Causality in economics. Centre for Philosophy of Natural and Social Science, London School of Economics, 2004.

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

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

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

[21] 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.

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

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

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

[25] Jerzy Neyman. On the application of probability theory to agricultural experiments. *Statistical Science*, 5:465–472, 1990.

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

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

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

[29] 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.

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

[31] 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.

[32] Herbert A. Simon. Spurious correlation: A causal interpretation. *Journal of the American Statistical Association*, 49:467–479, 1954.

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

[34] 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.

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

[36] 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