CHAPTER 12

DERIVATION OF THE UNIVERSAL MODEL OF CAUSAL STATISTICS

Sections

12.1 Introduction

Causal statistics is a mathematical inquiring system which enables empirical researchers to draw causal inferences from non-experimental, quasi-experimental, and imperfectly experimental data. The universal model of causal statistics is similar to the universal model of discrete causal macromathematics. The difference is that the former allows for error in its formulation; the latter does not.

Examples of errors which can arise are (1) the omission of relevant variables from consideration, (2) specification error, (3) measurement error, (4) the dissatisfaction of the assumptions made during derivation of the basic form of the equation system, (5) sampling error, (6) stochastic error, and (7) estimation error (i.e., biasedness).

12.2 The Basic Process of Causal Inference and the Sources of Error

Say that we are investigating the causal connection between two variables, e.g., closeness of supervision and productivity. Equations (11-5) represent an exact description of all the causal relationships among all n variables relevant to this investigation. The objective of causal inference is to infer the forms of equations (11-5) from observation of empirical phenomena. Unfortunately, various types of error almost inevitably enter the inference process and, therefore, only approximations to the true form of equations (11-5) can be inferred.

The basic process of drawing causal inferences from non-experimental data will be presented in four steps. Then, these steps will be shown to be the sources of various types of error.

The process of making causal inferences from non-experimental data begins with step (1), the choice of the variables which are thought to be relevant, called the assumed relevant variables. The second step is to construct a set of causal equations with undetermined parameters, to relate the assumed relevant variables. If the equations resulting from step (2) are under identified, then we will have to adjust our choices in steps (1) and (2) so that the final equations are identified. Step (3) is the collection of data on the assumed relevant variables. Lastly, in step (4), the data is used to estimate the parameters of step (2).

In step (1) error may result from the omission of relevant variables from (i.e., the non-inclusion of relevant variables in) the assumed relevant system. If Assumption (11-1) is satisfied, no such error can occur.

Definitional error can arise from the combination of choices made in steps (1) and (2). Definitional error occurs when neither Assumptions (10-4) or (10-3) are satisfied by the combination of (a) the definition of one or more macrovariables and (b) the form of the causal equations hypothesized in step (2).

The hypothesized causal equations can, also, be a source of three types of specification error. One type of specification error occurs when it is incorrectly assumed that one variable does not cause another. The second type occurs when a causal connection between two variables is given an incorrect functional or operational form. The third type of specification error arises from the dissatisfaction of Assumption (10-5), i.e., the incorrect aggregation of causal chains into causal macrochains.

In carrying out step (3), the collection of data, we may encounter the problem of measurement error. Another potential source of error in step (3) is the dissatisfaction of assumptions used in deriving equations (11-5). In other words, it may be that, for the periods during which we are collecting data, some of the assumptions made in the derivation of the universal model of discrete causal macromathematics--like Assumptions (8-8) and (8-10)--are not satisfied. For example, the immediately prior values of the intervening variables may not be the same on each run. The dissatisfaction of any of these assumptions can lead to error. Definitional error, resulting from the dissatisfaction of Assumption (10-3), could be considered to fall under this category of error. But definitional error was, more properly, discussed in relation of steps (1) and (2). The same is true for error resulting from the omission of relevant variables.

Sampling error is another type of error which arises in the data collection step. This type of error occurs when a sample taken from the population is a less than perfect representation of the population, for the statistic of interest.

A type of error--arising in the combination (noun) of steps (3) and (4)--can be, somewhat misleadingly, called stochastic error. This error occurs when the values of a set of variables are inconsistent with the actual (i.e., true, ontological) causal connections relating the variables. Stochastic error is, in a way, a special type of error. It would not occur if there were no other errors, but typically, there are other errors. Our estimation techniques usually make it necessary for us to assume that many of these other errors cancel themselves out on the average, over a number of runs. This cancelling assumption implies that the mean behavior, indicated by the sample data, is equal to the actual (i.e., errorless) behavior. This assumption is not applicable to all types of error but to some. The average magnitude (i.e., expected value) of the stochastic error is a measure of the degree to which this assumption is invalid. The type of error discussed here appears to be stochastic, hence the name. But in actuality--based on our causal metaphysics--this error is not a result of random behavior of the universe, but simply due to other types of error.

Estimation of parameters, step (4), may result in biased estimates. Like with stochastic error, the error of bias could be eliminated, if no other type of error occurred.

12.3 Classification and Diagramatic Representation of the Variables in a Causal Universe

This section is designed to classify the variables in a causal universe according to two different schemes and show the relationship between the two schemes. In the first classification scheme some basic definitions, like "assumed relevant variables," are presented. The second scheme formulates categories of variables, like endogenous and exogenous variables.

12.3.1 Formulation of Basic Definitions

Say that we were interested in determining the causal relationship between closeness of supervision and level of output. These two variables would then be called the variables under investigation.

For it to be theoretically possible to establish the exact causal relationship between the variables under investigation, we must obtain data on them and also on a number of other variables. These variables for which data is desired are called mathematically relevant variables. All other variables in the universe are called mathematically irrelevant variables. Mathematically irrelevant variables are those variables whose omission from the mathematical formulation will not affect the inferred causal relationships between the variables under investigation.

But, typically, we will be unable and unwilling to observe all mathematically relevant variables. So, we will use the variables which we believe to be the most important and the simplest to observe. These variables are called the assumed relevant variables (also, called the considered variables) or, as a group, they are called the assumed relevant system or considered system.

Now that these definitions are stated, we turn to a different sheme of classification of the variables in a causal universe.

12.3.2 Formulation of Categories

Figure 12-1 is a representation of the variables in a causal universe, partitioned into five categories. In discussing the figure, we will begin with the center circle, category (1), and work out way out, to category (5).

In formulating these categories, we must distinguish between the categories, called hypothesized categories, used for discussing the hypothesized causal model resulting from step (2) and the group of categories, called ontological categories, used to discuss the true causal model. We will first present the hypothesized categories.

Endogenous variables (category (1)) are those variables, among the assumed relevant variables, which are caused by themselves or other variables in the assumed relevant system. Endogenous variables are designated as Y_i's.

Exogenous variables (category (2)) are the group of all considered variables which cause endogenous variables, but which are, themselves, not caused by any endogenous or exogenous variable. Exogenous variables are designated by X_i's.

The line arrows in Figure 12-1 represent causal connections which are necessary, by definition. For example, exogenous variables must cause endogenous variables. The line arrows in the box indicate that by

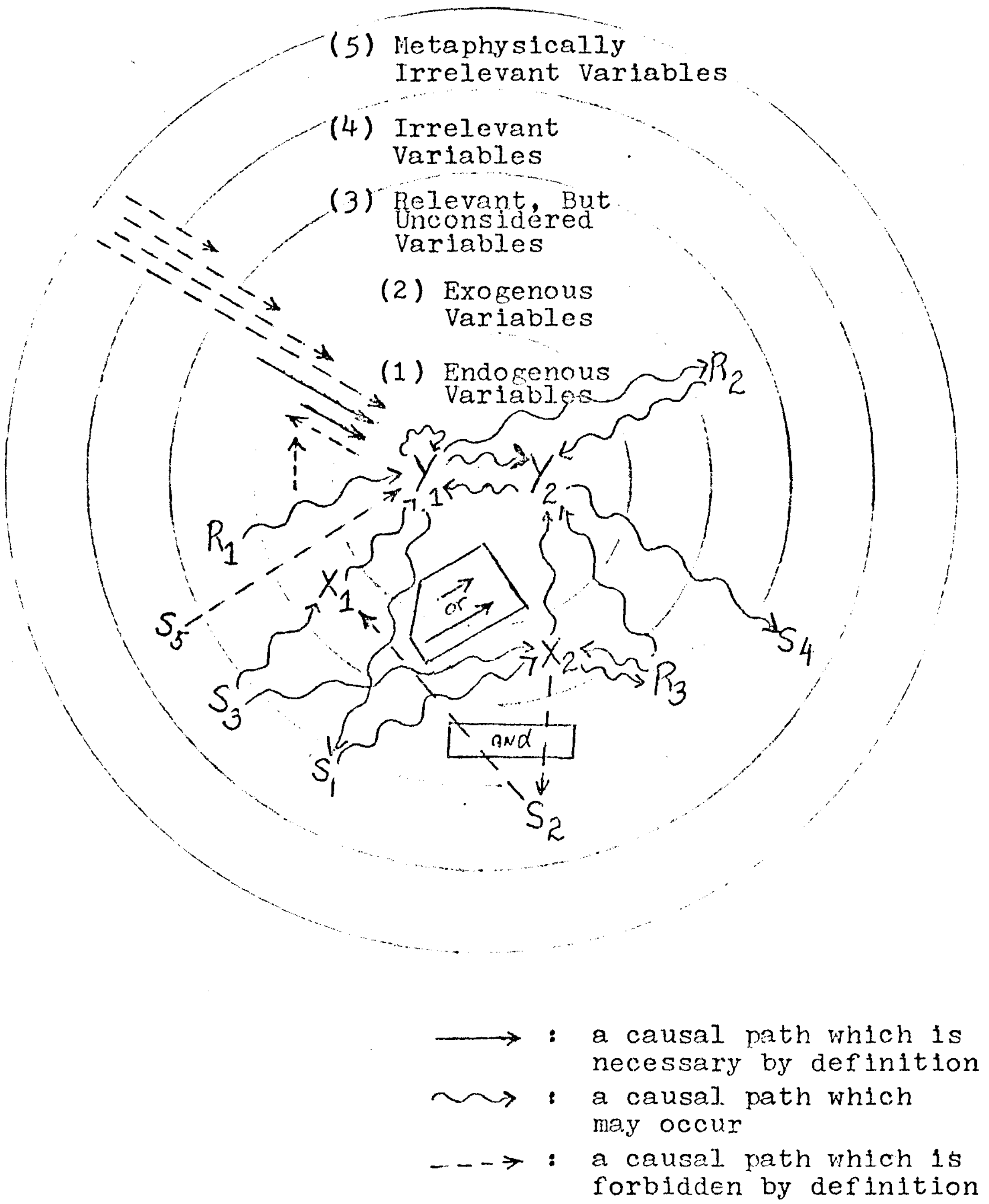
(5) Metaphysically Irrelevant Variables
(4) Irrelevant Variables
(3) Relevant, But Unconsidered Variables
(2) Exogenous Variables
(1) Endogenous Variables
R_2
Y_1
Y_2
R_1
X_1
or
S_5
S_4
X_2
R_3
S_3
and
S_1
S_2
——→ : a causal path which is necessary by definition
〰→ : a causal path which may occur
- - -→ : a causal path which is forbidden by definition

Figure 12-1

definition endogenous variables must be caused by one or more endogenous and/or exogenous variables. A wavy arrow represents a causal connection which may or may not occur, and a dashed arrow denotes a causal path forbidden by definition.

Relevant, but unconsidered variables (category (3)) are just what the name implies. They are relevant to the investigation and should appear in the considered system as either endogenous or exogenous variables, but they do not. As we saw in Section 12.2, relevant but unconsidered variables are one source of error in causal inference.

Irrelevant variables (category (4)) are variables which are not metaphysically irrelevant (defined in the next paragraph) but which can be omitted, from the analysis without affecting the resulting causal inferences. They may be variables (like S_3 and S_4) which have no causal effect on any of the endogenous variables or they may be well behaved (i.e., they satisfy all assumptions imposed upon them) intervening variables (like S_1) between two relevant variables. It is forbidden for S_5 to cause an endogenous variable unless S_5 acts only as an intervening variable between two relevant variables. It is also forbidden for X_2 to cause S_2 which causes X_1 because X_1 would, then, be an endogenous variable.

Metaphysically irrelevant variables (category (5)) can have no effect on variables in categories (1), (2), (3), and (4). This means that there is no way by which we can observe or know any of the metaphysically irrelevant variables. The implication of this definition is that any number of universes can be coexistent in time and space with our own and yet we would be unaware of it. If there are conscious beings in these metaphysical universes, there may or may not be any who are aware of our universe or components of it.

Using the hypothesized categories just individuated, endogenous and exogenous variables together make up the assumed relevant system. Add the relevant, but unconsidered variables (category (3)) to that and we have the mathematically relevant system. Categories (4) and (5) are composed of the mathematically irrelevant variables. Categories (1) through (4) contain the observable universe; while categories (1) through (5) contain the total universe. Note that the hypothesized categories are mutually exclusive and collectively exhaustive.

Now, what about the ontological categories? Some of the variables in the assumed relevant system may not be true endogenous or exogenous variables as they are hypothesized to be. They may actually be irrelevant variables. The ontological categories are a straightforward extension of the concept of the hypothesized categories.

The total universe of variables, i.e, categories (1) through (5), is composed of both macrovariables and fundamental variables. These variables are collectively exhaustive, but are they mutually exclusive?

The first point to be considered is that various definitions--i.e., various modes of aggregation--are properties of the mind and not properties of the universe. Therefore, it is meaningless to ask, in an ontological sense, whether or not the variables in the universe are mutually exclusive. But--from the, non-ontological, point of view of human consciousness--we can and do define variables which are to some degree overlapping.

Then, a portion of Y_2 (Figure 12-1) could simply be a portion of Y_1, by definition. This might lead to an apparent instantaneous, reciprocal causal connection between Y_1 and Y_2, which would simply be a result of their overlapping definitions. Hence, we must be careful about overlapping definitions. Nevertheless, the variables in the total universe, viewed from a human consciousness point of view, need not necessarily be mutually exclusive.

12.4 The Universal Model of Causal Statistics

In our attempts to determine equations (11-5), we select the assumed relevant variables; we assume (i.e., hypothesize) the form of the causal equations connecting the variables; we collect data; and we apply the data to the hypothesized equations. Typically, in the last step, we find that the causing side of an equation is not equal to the caused side, because of the errors encountered. It is valuable to realize the inevitability of such error and build it into the system of equations. The universal model of causal statistics is the mathematical result of specifically allowing for error in our estimation of equations (11-5). In this section we will derive the universal model of causal statistics.

Say that the true equation relating Z_1, Z_2, and Z_3 is:

$$Z_1 \leftarrow \beta_{11}Z_1+\beta_{21}Z_2+\beta_{31}Z_3+\alpha_{11}Z_1Z_2Z_3 \qquad (12\text{-}1)$$

If we make a specification error and omit the last term from the assumed equation, we must add an error term, ϵ_1, to the equation to compensate for this error, i.e., to maintain equality between the causing and caused sides. In this case ϵ_1 is numerically equal to $\alpha_{11}Z_1Z_2Z_3$. Or if, instead of making this specification error, we had omitted Z_3 from the assumed relevant system, we could compensate with either one or two error terms as follows:

$$Z_1 \leftarrow \beta_{11}Z_1+\beta_{21}Z_2+\alpha_{11}Z_1Z_2+\epsilon_1\text{, or} \qquad (12\text{-}2)$$

$$Z_1 \leftarrow \beta_{11}Z_1+\beta_{21}Z_2+\epsilon_{11}+\alpha_{11}Z_1Z_2\epsilon_{12} \qquad (12\text{-}3)$$

In equation (12-2), ϵ_1 is numerically equal to $\beta_{31}Z_3 + \alpha_{11}Z_1Z_2Z_3 - \alpha_{11}Z_1Z_2$. In equation (12-3), $\epsilon_{11} = \beta_{31}Z_3$ and

$\epsilon_{12} = Z_3$.

The use of one error term will always be able to do the job, but sometimes it may be convenient to use more than one. Hence, each equation in the universal model of causal statistics will be fitted with an error vector, $\bar{\epsilon}_i$, rather than a single error term.

To obtain the universal model of causal statistics, we will insert $\bar{\epsilon}_i$'s into equations (11-5).

$$\begin{aligned}
U_1\left[Z_1(t)\right] \longleftarrow U_1^*\Big[&t-t_{01},\ \overline{Z_1(t-\tau_{11})},\\
&\overline{Z_2(t-\tau_{21})},\ \ldots,\ \overline{Z_n(t-\tau_{n1})},\ Z_1(t_{01}),\ \bar{\epsilon}_1\Big];\\
U_2\left[Z_2(t)\right] \longleftarrow U_2^*\Big[&t-t_{02},\ \overline{Z_1(t-\tau_{12})},\\
&\overline{Z_2(t-\tau_{22})},\ \ldots,\ \overline{Z_n(t-\tau_{n2})},\ Z_2(t_{02}),\ \bar{\epsilon}_2\Big];\\
&\vdots\\
U_n\left[Z_n(t)\right] \longleftarrow U_n^*\Big[&t-t_{0n},\ \overline{Z_1(t-\tau_{1n})},\\
&\overline{Z_2(t-\tau_{2n})},\ \ldots,\ \overline{Z_n(t-\tau_{nn})},\\
&Z_n(t_{0n}),\ \bar{\epsilon}_n\Big].
\end{aligned} \qquad (12\text{-}4)$$

Since this is the final and most important model, the U's are individuated to show that they are different from each other. Otherwise, the notation in equations (12-4) is the same as that in equations (11-5). To be strictly correct all of the notation, other than t, should be different. The purpose of this would be to indicate that the operators, variables, number of relevant variables, and time lags in equations (12-4) are subject to error; whereas, in equations (11-5) they are not. To minimize the notational proliferation (i.e., notational pollution), I have resisted the urge to differentiate the notation between equations (11-5) and (12-4). Thus, it is incumbent upon the reader and/or user (when causal statistics is operationalized) to remember this difference between equations (11-5) and (12-4).

In obtaining the universal model of causal statistics, we have attained our objectives. We first forwarded a causal theory of the operation of the universe. Then a mathematical description of this theory was formulated. A derivation followed, with the end product being equations (12-4).

The universal model of causal statistics is a formulation, based upon our causal metaphysics, which can be used to represent the causal connections between any number of discrete macrovariables. This model can be considered to be a generalized template which fits any form of causal relationships. It is the duty of the empirical researcher to specialize this template, i.e., to select the assumed relevant variables, to hypothesize the specific mathematical form of the causal connections, to obtain the data, and to estimate the parameters.

Without the work contained in this dissertation, we would not know the form of the generalized model, from which we should begin our specialization. Nor would we know the axioms and assumptions upon which our specialized formulation is founded.