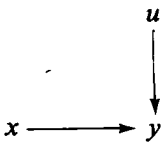
Contents

1

Preliminaries

In this book we will frequently use a notation like the following:



It may be read as "a change in x or u produces a change in y" or "y depends on x and u" or "x and u are the causes of y." In all these statements, we need to include a distinction between two sorts of causes, x and u. We intend the former to stand for some definite, explicit factor (this variable has a name) producing variation in y, identified as such in our *model* of the dependence of y on its causes. On the other hand, u stands for all other sources (possibly including many different causes) of variation in y, which are not explicitly identified in the model. It sums up all their effects and serves to account for the fact that no single cause, x, nor even a finite set of causes (a list of specific x's) is likely to explain all the observable variation in y. (The variable u has no specific name; it is just called "the disturbance.")

The letters (like x, y, and u), the arrows, and the words (like "depends on") are elements in a language we use in trying to specify

how we think the world—or, rather, that part of it we have selected for study—works. Once our ideas are sufficiently definite to help us make sense of the observations we have made or intend to make, it may be useful to formalize them in terms of a model. The preceding little arrow diagram, once we understand all the conventions for reading it, is actually a model or, if one prefers, a pictorial representation of a model. Such a representation has been found useful by many investigators as an aid in clarifying and conveying their ideas and in studying the properties of the models they want to entertain.

More broadly useful is the algebraic language of variables, constants, and functions, symbolized by letters and other notations, which are manipulated according to a highly developed grammar. In this language, our little model may be expressed as

$$y = bx + u$$

or, even more explicitly,

$$y = b_{yx}x + u$$

It is convenient, though by no means essential, to follow the rule that y, the "dependent variable" or "effect" is placed on the left-hand side of the equation while x, the "independent variable" or "cause," goes on the right-hand side. The constant, or coefficient b in the equation tells us by how much x influences y. More precisely, it says that a change of one unit in x (on whatever scale we adopt for the measurement of x) produces a change of b units in y (taking as given some scale on which we measure y). When we label b with subscripts (as in b_{yx}) the order of subscripts is significant: the first named variable (y) is the dependent variable, the second (x) the independent variable.

(**Warning:** *Although this convention will be followed throughout this book, not all authors employ subscripts to designate the dependent and independent variables.*)

The scale on which u is measured is understood to be the same as that used for y. No coefficient for u is required, for in one sense, u is merely a balancing term, the amount added to the quantity bx to satisfy the equation. (In causal terms, however, we think of y as depending on u, and not vice versa.) This statement may be clearer if we make explicit a feature of our grammar that has been left implicit

up to now. The model is understood to apply (or, to be proposed for application) to the behavior of units in some population, and the variables y, x, and u are variable quantities or "measurements" that describe those units and their behavior. [The units may be individual persons in a population of people. But they could also be groups or collectives in a population of such entities. Or they could even be the occasions in a population of occasions as, for instance, a set of elections, each election being studied as a unit in terms of its outcome (y) and being characterized by properties such as the number of candidates on the ballot (x), for example.] We may make this explicit by supplying a subscript to serve as an identifier of the unit (like the numeral on the sweater of a football player). Then the equation of our model is

$$y_i = b_{yx}x_i + u_i$$

That is, for the ith member of the population we ascertain its score or value on x, to wit x_i, multiply it by b, and add to the product an amount u_i (positive or negative). The sum is equal to y_i, or the score of the ith unit on variable y. Ordinarily we will suppress the observation subscript in the interest of compactness, and the operation of summation, for example, will be understood to apply over all members of a sample of N units drawn from the population.

It is assumed that the reader will have encountered notation quite similar to the foregoing in studying the topic of regression in a statistics course. (Such study is a prerequisite to any serious use of this book.) But what we have been discussing is not statistics. Rather, we have been discussing the form of one kind of model that a scientist might propose to represent his ideas or theory about how things work in the real world. Theory construction, model building, and statistical inference are distinct activities, sufficiently so that there is strong pressure on a scientist to specialize in one of them to the exclusion of the others. We hope in this book to hint at reasons why such specialization should not be carried too far. But we must note immediately some reasons why the last two may come to be intimately associated.

Statistics, in one of its several meanings, is an application of the theory of probability. Whenever, in applied work—and all empirical inquiry is "applied" in this sense—we encounter a problem that probability theory may help to solve, we turn to statistics for guidance.

There are two broad kinds of problems that demand statistical treatment in connection with scientific use of a model like the one we are discussing. One is the problem of inference from samples. Often we do not have information about all units in a population. (The population may be hypothetically infinite, so that one could never know about "all" units; or for economic reasons we do not try to observe all units in a large finite population.) Any empirical estimate we may make of the coefficient(s) in our model will therefore be subject to sampling error. Any inference about the form of our model or the values of coefficients in it that we may wish to base on observational data will be subject to uncertainty. Statistical methods are needed to contrive optimal estimators and proper tests of hypotheses, and to indicate the degree of precision in our results or the size of the risk we are taking in drawing a particular conclusion from them.

The second, not unrelated, kind of problem that raises statistical issues is the supposition that some parts of the world (not excluding the behavior of scientists themselves, when making fallible measurements) may be realistically described as behaving in a stochastic (chance, probabilistic, random) manner. If we decide to build into our models some assumption of this kind, then we shall need the aid of statistics to formulate appropriate descriptions of the probability distributions.

This last point is especially relevant at this stage in the presentation of our little model, for there is one important stipulation about it that we have not yet stated. We think of the values of u as being drawn from a probability distribution. We said before that, for the ith unit of observation, u_i is the amount added to bx_i to produce y_i. Now we are saying that u_i itself is produced by a process that can be likened to that of drawing from a large set of well-mixed chips in a bowl, each chip bearing some value of u. The description of our model is not complete until we have presented the "specification on the disturbance term," calling u the "disturbance" in the equation (for reasons best known to the econometricians who devised the nomenclature), and meaning by the "specification" of the model a statement of the assumptions made about its mathematical form and the essential stochastic properties of its disturbance.

Throughout this book, we will assume that the values of the disturbance are drawn from the same probability distribution for all units in

the population. This subsumes, in particular, the assumption of "homoskedasticity." It can easily be wrong in an empirical situation, and tests for departures from homoskedasticity are available. When the assumption is too wide of the mark, special methods (for example, transformation of variables, or weighting of regression estimators) are needed to replace the methods sketched in this book. No special attention is drawn to this assumption in the remainder of the text: but the reader must not forget it, nonetheless. Another assumption made throughout is that the mean value of the disturbance in the population is zero. The implications of assumptions about the disturbance are discussed in Chapter 11.

Frequently we shall assume explicitly that the disturbance is uncorrelated with the causal variable(s) in a model, although this assumption will be modified when the logic of the situation requires. Thus for the little model under study now, we specify that $E(xu) = 0$. (E is the sign for the expectation operator. If the reader is not familiar with its use in statistical arguments, he should look up the properties of the operator in an intermediate statistics text such as Hays, 1963, Appendix B.) The assumption that an explanatory or causal variable is uncorrelated with the disturbance must always be weighed carefully. It may be negated by the very logic of the model, as already hinted. If it is supposed, not only that y depends on x, but also that x simultaneously depends on y, it is contradictory to assume that the disturbance in the equation explaining y is uncorrelated with x. The specification $E(xu) = 0$ may also be contrary to fact, even when it is not inherently illogical. The difficulty is that we will never know enough about the facts of the case to be sure that the assumption is true—that would be tantamount to knowing everything about the causes of y. Lacking omniscience, we rely on theory to tell us if there are substantial reasons for faulting the assumption. If so, we shall have to eschew it—however convenient it may be—and consider how, if at all, we may modify our model or our observational procedures to remedy the difficulty. For *we must have this assumption in the model in some form*—though not necessarily in regard to all causal variables—*if any statistical procedures* (estimation, hypothesis testing) *are to be justified.* Here we distinguish sharply between (*1*) statistical description, involving summary measures of the joint distributions of observed variables, which may serve the useful purpose of data reduction, and (*2*) statistical methods

applied to the problem of estimating coefficients in a *structural model* (as distinct from a "statistical model") and testing hypotheses about that model. One can do a passably good job of the former without knowing much about the subject matter (witness the large number of specialists in "multivariate data analysis" who have no particular interest in any substantive field). But one cannot even get started on the latter task without a firm grasp of the relevant scientific theory, because the starting point is, precisely, the model and not the statistical methods.

In summary, we have proposed a model,

$$y = b_{yx}x + u$$

and stated a specification on its disturbance term, $E(xu) = 0$. Without mentioning it before, we have also been assuming that $E(x) = 0$, which is simply a convention as to the location of the origin on the scale of the independent variable. It follows at once that

$$E(y) = b_{yx}E(x) + E(u) = 0.$$

Now, each of the variables in our model has a variance, and it is convenient to adopt the notation,

$$\sigma_{yy} = E(y^2)$$
$$\sigma_{xx} = E(x^2)$$
$$\sigma_{uu} = E(u^2)$$

for the variances (writing σ_{yy}, for example, in place of the usual σ_y^2). There are also three covariances,

$$\sigma_{yx} = E(yx)$$
$$\sigma_{yu} = E(yu)$$
$$\sigma_{xu} = E(xu) = 0$$

The disappearance of the last of these covariances is merely a restatement of the original specification on the disturbance term. To evaluate σ_{yu} we multiply the equation by u and take expectations, finding

$$E(yu) = b_{yx}E(xu) + E(uu)$$

so that $\sigma_{yu} = \sigma_{uu}$, in view of the fact that $E(xu) = 0$. Let us multiply through the equation of our model by y, obtaining,

$$y^2 = b_{yx}xy + yu$$

We take expectations

$$E(y^2) = b_{yx}E(xy) + E(yu)$$

and thereby find that we can write the variance of y as

$$\sigma_{yy} = b_{yx}\sigma_{xy} + \sigma_{uu}$$

since $\sigma_{yu} = \sigma_{uu}$ as already noted. Let us next multiply through by x and take expectations. We find

$$E(xy) = b_{yx}E(x^2) + E(xu)$$

or

$$\sigma_{xy} = b_{yx}\sigma_{xx}$$

Substituting this result into the expression for the variance of y, we obtain

$$\sigma_{yy} = b_{yx}^2\sigma_{xx} + \sigma_{uu}$$

(The same result is obtained upon squaring both sides of $y = b_{yx}x + u$ and taking expectations.)

The three symbols on the right-hand side stand for the basic parameters of this model as it applies in a well-defined *population*:

—the structural coefficient b_{yx},
—the variance of the exogenous variable x,
—the variance of the disturbance u.

The variance in the dependent variable is traceable to these three distinct sources.

The expression just obtained for the covariance of the two observable variables is a suggestive one, for we can immediately rewrite it as

$$b_{yx} = \frac{\sigma_{xy}}{\sigma_{xx}}$$

We see that if we knew σ_{xy} and σ_{xx} we could calculate the value of the structural coefficient. We do not and, in general, cannot know these quantities exactly. But we can estimate them, or their ratio, from data

pertaining to a *sample* of the population to which the model applies. How to use this sample information in a correct and efficient manner is a topic studied in the statistical theory of estimation. In this book, we will draw upon a few important results from that theory, but will not try to demonstrate those results.

Exercise. *Our model*

$$y = bx + u$$

could be solved for x, *to read*

$$x = \frac{1}{b}y - \frac{1}{b}u$$

Let $1/b$ *be renamed* c *and* $-u/b$ *be called* v. *Someone could, therefore, assert that our model is equally well written*

$$x = cy + v$$

But on the assumption that our original model is true (including the specification on the disturbance term) show that the disturbance is not *uncorrelated with the variable on the right-hand side, that is,* $E(yv) \neq 0$. *Show also that we cannot solve for* c *using the same kind of formula developed for the original model, that is,* $c \neq \sigma_{xy}/\sigma_{yy}$. *How do you square this result with the well-known fact in statistics, that there are two regressions,* Y *on* X *and* X *on* Y?

FURTHER READING

The statistical methods of simple and multiple regression are well presented in Snedecor and Cochran (1967, Chaps. 6, 7, and 13). A judicious discussion of issues raised in the use of the regression model to represent a causal relationship is given by Rao and Miller (1971, Chap. 1).