Understanding Regression

James Woodward

California Institute of Technology

Although statistical techniques like regression analysis and path analysis are widely used in the biomedical, behavioral and social sciences to make causal inferences there has been surprisingly little philosophical discussion of the details of such techniques and of the conceptions of causation and explanation implicit in them. There also has been relatively little attempt to compare such techniques with various probabilistic models of causation and explanation in the philosophical literature.

In this paper I explore, for reasons of space in a very sketchy and schematic way, some issues in philosophy of science raised by regression analysis. One general conclusion I reach is that it is considerably less obvious than one might suppose that the philosophical theories alluded to above are plausibly viewed as reconstructions of regression techniques.

1. Introduction

I begin with a brief description of simple linear regression involving just one independent variable. Suppose that X and Y are variables, measurable on an interval scale, and that we have *n* pairs of observations $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$ on *X* and *Y*. In linear regression it is assumed that except for the operation of a so-called "error" or "disturbance" term, *u*; there is a definite general linear relationship between *X* and *Y*, i.e., that

(1) $y_i = \alpha + \beta x_i + u_i$ for $i = 1 \ldots n$

where α and β are fixed coefficients and u_i varies in value for different observations. I shall say more about the status of this disturbance term below, but in the simplest case it is usual to think of the term as arising as a result either of "measurement error" in *Y* (but *not* in *X*) or as the result of the operation of various other variables besides *X* which causally influence *Y* but have been left out of equation (1). One assumes that the u_i are values of a random variable *U* with a definite probability distribution and hence that *Y* is a random variable as well. However, to carry out the regression by least squares techniques one does not need to assume that *X* is a random variable.

The operation of the disturbance term u_i will result in a "spread" of values for *Y* for fixed values of *X*. In general, the regression equation for *Y* on *X* will be the equation

PSA 1988, Volume 1, pp. 255-269

https://doi.org/10.1086/psaprocbienmeetp.1988.1.192993 Published online by Cambridge University Press

which gives the path of the mean value of Y for fixed values of X. In the specific context of linear regression the problem of specifying the regression equation will reduce to the problem of estimating the values of the parameters α and β in (1) above. To do this requires the choice of an estimating procedure and certain assumptions about the distribution of the disturbance term. The usual practice is to employ the method of "least squares", i.e., to choose $\hat{\alpha}$ and $\hat{\beta}$ as estimators of α and β such that the quantity

$$(2)\quad Q = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

is minimized. Geometrically this corresponds to choosing the regression line so that it "best fits" the scatter of points (x_i, y_i), where the criterion of best fit is the minimization of the squared vertical distance of these points from the regression line.The least squares estimators obtained by minimizing Q in (2) will have desirable properties if we make the following assumptions about u_i:

3 (a) zero mean: $E(u_i)=0$ for all i,(b) common variance: $V(u_i)=\sigma^2$ for all i (c) lack of correlation between or statistical independence of error terms for $u_i \neq u_j$, (d) statistical independence of the error terms u_i and independent variable x_j. (We may think of this as automatically satisfied in the case in which X is not a random variable.)

Under assumptions (3*a* - *d*), one can show that the least squares estimators of α and β are best linear unbiased estimators.[1] These assumptions do not by themselves commit one to any definite assumptions about the probability distribution of the u_i. However, for the purposes of doing significance tests or establishing confidence intervals it is common to assume also that (3e) the u_i are normally distributed.

The problem of minimizing Q in (2) is quite straightforward: one simply takes the partial derivatives of Q with respect to $\hat{\alpha}$ and $\hat{\beta}$ and sets these equal to zero. When one does this one obtains the result that the least squares estimates for α and β are

$$(4)\quad \hat{\beta} = \frac{S_{xy}}{S_x S_x}$$

and

$$(5)\quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Here S_{xy} and S_y are respectively the sample covariance between x and y and the standard deviation of X, and $\bar{y}$ and $\bar{x}$ are the sample means for x and y.[2]

Thus given certain assumptions about the functional relationship between X and Y and the distribution of the error terms, one can derive estimates for the regression coefficients α and β in (1) from facts about the observed or directly measured values of X and Y (since S_{xy} and S_x can be inferred from the data). The result will be a linear equation relating X and Y. A non-zero value for the coefficient β will, it is hoped, under appropriate conditions reflect a structural causal connection between X and Y. Questions about the significance of the coefficient β will occupy us inconsiderably more detail below, but in general we can think of it as purporting to tell us what sort of change we may expect in the mean value of Y, if a change occurs in the level of X and if everything else is held constant. For example, in his introductory econometrics textbook (1977), Maddala regresses a variable C representing expenditures *per capita* (1958 prices) on a variable Y representing disposable

https://doi.org/10.1086/psaprocbienmeetp.1988.1.192993 Published online by Cambridge University Press

income *per capita* (also 1958 prices) for values of C and Y for the years from 1929 to 1970 and obtains the following linear equation

$$(6) \quad C = 55.432 + .8735Y$$

This seems to suggest that an average increase of $.87 in *per capita* consumer spending will be associated with each dollar increase indisposable income. Similarly, in his textbook *Data Analysis for Politics and Policy*, Edward Tufte regresses a variable Y representing percent of congressional seats in Congress won by Democrats against a variable X representing percent of the Democratic vote in 36 congressional elections from 1900 to 1972 and obtains the following relationship.

$$(7) \quad Y = -\,49.64 + 2.07X$$

Tufte suggests that "[t]his means that a one percent change in the share of the Democratic vote was typically accompanied by a change of 2.07 percent in the Democratic share of seats in Congress" [p. 68].

So far we have considered regression with a single independent variable X. I now turn to some very brief remarks on multiple regression, which one can think of as a generalization of the procedures described above. In multiple regression one is interested in the relationship between dependent variable Y and a number of independent variables X_1 , $X_2 \ldots X_k$. In the linear case, one assumes the model

$$(8) \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdot\cdot \beta_k x_{ki} + u_i \text{ , } i=1,2 \text{ ,..., } n$$

where the y_i are the n observations one makes on Y and $x_{1i} \ldots x_{k}i$ are the corresponding observations one makes on the variables $X_1 \ldots X_k$ and u_i represents as before a disturbance term. It will be convenient to write (8) in matrix notation as

$$(9) \quad Y = X\beta' + V$$

$$\text{where } Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \bullet \\ \bullet \\ Y_n \end{bmatrix} \text{ , } \beta = [\beta_0\, \beta_1 \ldots \beta_k]$$

$$U = \begin{bmatrix} U_1 \\ \bullet \\ \bullet \\ U_n \end{bmatrix} \quad \text{and } X = \begin{bmatrix} 1 & X_{11} & X_{12} & \bullet & X_{1k} \\ 1 & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ 1 & X_{n\,1} & \bullet & \bullet & X_{n\,k} \end{bmatrix}$$

(The role of the column of 1's added to the X matrix is to provide constant multipliers for β_0 and β' is just the transpose of β .)

Using again the least squares criterion of fit, one can show that if one makes assumptions like (3*a* - *d*) above regarding the distribution of the disturbance term, the best linear unbiased estimator for the vector β is, in close analogy with (4), the vector

$$(10) \qquad \hat{\beta} = (X'X)^{-1} X'Y$$

where X' is the transpose of X and $(X'X)^{-1}$ is the inverse of $X'X$. One can think of the observations on($X_1 \dots X_k$, Y) as representing points in $k + 1$ dimensional space, to which one is fitting a k-dimensional hyperplane. The coefficient β_i on the variable x_i can be interpreted as"the hypothetical change that would occur in the dependent variable if [the variable x_i] were to change by one unit and if the other independent variables were to remain constant" (Blalock 1971, p. 479).

2. Theses

In what follows I shall argue, rather schematically, for three general claims. (a) Regression analysis does not yield lawlike generalizations but rather yields claims about causal connections obtaining in particular populations. Regression analysis is a technique for making causal inferences in circumstances in which one lacks knowledge of exceptionless general laws or systematic theory. (b) Regression analysis and other causal modeling techniques are not plausibly viewed as part of a neo-Humean program of analyzing or defining causal claims in terms of claims about regular patterns of statistical association. (c) The causal and explanatory claims embodied in regression analysis are often most plausibly interpreted as population-level claims, rather than direct claims about particular individuals in a population. In particular, the use of regression analysis does not commit one to the claim that the causal processes at work with respect to individuals in the population of interest are indeterministic or correctly described by some probabilistic theory of causality of the sort to be found in the philosophical literature. Nor does regression analysis involve the explanation of facts about individuals by subsumption under statistical generalizations in the fashion described by various philosophical models of statistical explanation, such as Hempel's IS model or Salmon's SR model. Instead the conception of explanation associated with regression analysis is deductive and involves the exhibition of patterns of counter-factual dependence. What is explained by regression analysis is facts about such population-level parameters as changes in the mean value of the dependent variable. Regression coefficients thus represent facts about the average or aggregate impact of individual causal processes at work in a population.

3. Laws

A substantial philosophical tradition connects the notion of cause and the activity of constructing causal explanations with various claims about the existence of laws of nature. While a detailed assessment of these claims must be beyond the scope of this paper, it is worth emphasizing that the notion of a law of nature does not seem to explicitly enter into the above account of regression techniques at any point. While the use of regression analysis to draw conclusions about causal connections certainly requires (as we shall see) "extra-statistical" causal or theoretical assumptions of various kinds, these are not assumptions about the obtaining or non-obtaining of natural laws. Nor is it plausible to hold that the results of such an analysis–the regression equation itself–represents a law of nature. To begin with, the results of such an analysis will be a claim about a causal connection obtaining in a particular population, and not over a wide range of populations. As Christopher Achen puts it in his monograph *Using and Interpreting Regression*, the researchers intent, in using regression analysis, is to describe, for example,

> ...the effect of the Catholic vote on the Nazi rise to power or the impact of a pre-school cultural enrichment program like Head Start on poor children's success in school. Whatever the truth in such cases, one would not characterize it as a law . Neither Catholics nor impoverished youngsters would behave the same way in other times and places (Achen 1982, p. 12).

Even on a rather permissive conception of lawlikeness, according to which generalizations which hold only over limited spatio-temporal intervals can count as laws, the claims that result from regression analysis are just too closely tied to particular

https://doi.org/10.1086/psaprocbienmeetp.1988.1.192993 Published online by Cambridge University Press

populations and too non-resilient in the sense of Skyrms (1980) to qualify as laws of nature. Secondly (and relatedly) a regression analysis may identify just one (or some small number of) the variables which are relevant to the dependent variable–the remaining omitted variables are represented by the catch-all error term. Here again, this sort of omission (and this strategy for dealing with omitted variables) does not seem characteristic of genuine laws of nature. Third, as recent accounts emphasize, lawlike status often seems to have something to do with integration into organized, systematic theory. Regression equations typically lack this feature–they are not procedures for the construction of theories consisting of a theoretically integrated system of lawlike generalizations analogous to a good physical theory,but rather represent in part an alternative explanatory strategy, in which one eschews the search for such a theory (perhaps on the supposition that it does not exist in the domain in which one is interested) in favor of a more piecemeal, less systematic, more data-driven investigation into the role of various causal factors in particular populations. The role of theory in such an investigation is not to supply candidates for laws of nature, but rather, as Achen (1982, p. 12-17) claims, to provide information about possibly causally relevant variables or about "causal ordering," or to support claims about causal irrelevance (important matters, as we shall see below).

Writers on explanation in history and the social sciences often claim that to identify a cause of some event is to "tacitly commit" oneself to or to "implicitly rely on" or "to invoke" some claim about the existence of a law linking this cause and its effect (cf. Hempel 1965, Gardiner 1961). Other writers, supposing that all serious explanations must involve appeal to laws and that the scientific status of a discipline depends upon whether there are laws specific to that discipline conclude from the evident absence of laws in the social sciences, that such disciplines do not contain serious explanatory theory (cf. Rosenberg 1980). Regression analysis and other causal sophisticated techniques for establishing claims about causal connections in the absence of knowledge of laws. Moreover, such claims about causal connections sometimes seem to provide (and are certainly regarded by users of the above techniques as providing) explanations. It seems arbitrary and unmotivated to maintain that the results of those techniques is not really science or not really explanatory.

4. Causes and Probabilities

Regression analysis and other causal modeling techniques are sometimes regarded as attempts to define or reduce the notion of "cause" to claims about frequencies or patterns of regular statistical association (cf. Suppes 1970, especially pp. 60-62). While the role of probabilities in regression analysis will receive more attention below, it seems implausible to regard causal modeling techniques as embodying any such wholesale reductionist strategy. The regression techniques described above are not techniques for inferring causal conclusions from purely statistical premises; they are rather techniques which allow one by making certain assumptions about causal connections ("causal assumptions," as I shall call them) to use statistical information about variances and covariances to test other causal claims. These causal assumptions are commonly described as "*a priori*" or "extra-statistical," where what this means is not that they are non-empirical or incapable of being tested, but rather that they are not inferred just from the statistical data at hand, but rather have at least in part some other rationale or justification (cf. Kendall and Stuart, 1961, Simon 1954). These extra-statistical, causal assumptions take a number of different forms. First, and perhaps most centrally, they include assumptions about which variables to include or exclude from the regression equation. As (10) shows, one can always alter the coefficient of any variable in a regression equation by the inclusion or deletion of other variables, as long as these variables exhibit a non-zero correlation with one or more of the original variables.[3] As users of causal modeling techniques know all too well, it will virtually always be possible to find such additional variables. This yields, in the context of regression analysis, a version of (or analogue to) Simpson's paradox, as recently discussed by Nancy Cartwright (1979).[4] The moral drawn in the literature on regression analysis is

https://doi.org/10.1086/psaprocbienmeetp.1988.1.192993 Published online by Cambridge University Press

essentially the moral drawn by Cartwright. One needs, among other things, extra-statistical assumptions about which are the potentially causally relevant variables if the regression analysis is going to be used to support causal claims. A good illustration of this general point is provided in Edward Tufte's (1974), in the context of an investigation into automobile fatality rates. Tufte notes that states with high rates tend, in addition to lacking inspections and having low population density, not to have been one of the original 13 states and to have seven or less letters in their names, while states with low rates tend to lack these characteristics. He writes

> While we observe many different associations between the death rate and other characteristics of the state, it is our substantive judgment, and not merely the observed association that tells us density and inspections might have something to do with the death rate and that the number of letters in the name of the state has nothing to do with it (Tufte 1974, p. 9).

A second kind of extra-statistical, causal assumption which must be made in the context of regression analysis has to do with "causal ordering." We can think of such assumptions as assumptions about causal direction and causal independence. A particularly simple illustration is furnished by the observation that the regression of *Y* on *X* will be different from the regression of *X* on *Y* (unless *X* and *Y* are perfectly correlated). If the situation is one in which we can rule out the possibility of reciprocal or "simultaneous" causation and if our interest is in explanation or the identification of causes, whether it is appropriate to regard *X* or *Y* as the independent variable will depend upon whether *X* is the sort of thing that could cause *Y* or *vice-versa*. Here again, one often cannot infer causal direction just from available statistical data–indeed one often cannot reliably infer causal direction even if one is allowed in addition to rely on such acceptably Humean information as observations about temporal order. Rather one must rely on prior, independent ideas about which variables are causally prior to others.[5]

More generally, we can say that given any body of statistical data there will be a large number of regression equations (or systems of such equations or linear causal models) which are consistent with this data.[6] To determine which causal model is the correct one we must be able to eliminate other possible candidates for such models; this will generally require rather strong "causally-committed" assumptions about, among other matters, possibly causally relevant variables and about causal direction. Which causal conclusions a researcher gets out of a body of statistical data thus depends very much on the causal assumptions he is willing to make. Given a willingness to assume certain causal claims, one may be able to test others by reference to a body of statistical data; and these assumptions may in turn be tested piecemeal by relying on other causal claims, and other statistical data, and so forth, but one never arrives at a point at which causal claims are warranted just in terms of a body of statistical data or are translatable without a remainder just into claims about patterns of statistical dependence. Rather than implementing a Humean program in which causal claims are analyzed solely in terms of claims about statistical regularities, regression and other causal modeling techniques strongly suggest that such a project cannot be carried out.

5. Causation and Regression

I turn now to what I take to be one of the central philosophical issues raised by regression analysis: what are the conception of causation and explanation implicit in such an analysis? There is of course a large philosophical literature on probabilistic theories of causation and on various models of statistical explanation. Philosophers of science who have discussed the matter have often assumed or suggested (usually without detailed argument) that such models capture or correspond to important features in regression analysis. For example, both Wesley Salmon (1984) and Peter Railton (1981) seem to suggest that when statistical techniques like regression analysis are used in the social

https://doi.org/10.1086/psaprocbienmeetp.1988.1.192993 Published online by Cambridge University Press

sciences to construct causal explanations, such explanations involve explaining claims about individual members of the population of interest, by subsuming them under statistical laws or generalizations, in (some rough approximation to) their SR and DNP models of explanation. To use Salmon's example, one uses statistical data about the incidence of juvenile delinquency in various subclasses of the American juvenile population (e.g., middle class boys from urban backgrounds) to explain why some particular boy, Albert, became a juvenile delinquent (Salmon 1984, pp. 36-47).

Similarly, Patrick Suppes, in his classic monograph (1970), suggests that regression analysis embodies a generalization (to continuous properties) of his well-known account of probabilistic causation, the underlying idea of which is that a cause must raise the probability of (must be positively relevant to) its effect under suitable background conditions. A related contention is made in Suppes' recent book, *Probabilistic Metaphysics*, where it is claimed to be a point in favor of a "probabilistic analysis of causality" as opposed to a "counterfactual analysis" that the former "has an extensively developed methodology and is widely used in actual science. In contrast, the counterfactual analysis does not have a developed methodology and is not used in practice, and for good reason" (1984, p. 53). Here the move from the idea that the methodology used in making causal inferences has a prominent statistical component to the idea that the notion of causality employed or assumed in this methodology is itself probabilistic in the rather special sense of Suppes' theory is quite transparent.

I think that the idea that regression analysis presupposes or embodies a probabilistic conception of causality in Suppes' sense or a model of statistical explanation in the sense of the SR model begins to look rather problematic when one looks at the details of how such analyses work. For definiteness, let us suppose that we are interested in investigating the effect of variations in exposure to sunlight (X_1) on the height (Y) of plants in a certain population. Our data consists of observations (x_{1i}, y_i) of sunlight exposure and height for n individual plants. For each plant, the relationship between its height and sunlight will be represented by the linear equation.

$$(1)\quad y_{1i} = \alpha + \beta_i x_{1i} + u_i \quad \text{for} \quad i = 1 ,..., n .$$

If (1) is taken literally, the coefficient β_1 must be regarded as constant across the entire population–it represents a fixed, linear relationship between exposure to sunlight and height which is characteristic of each individual in the population.[7] Of course individual plants with the same exposure to sunlight will differ in height, but all of this variation is taken to be due to the operation of the error term u_i. It is thus the error term, rather than the coefficient β_1, which is the source or locus of the stochastic element in (1).

Discussions of regression analysis typically suggest two general ways of thinking about what this error term represents.[8] On the one hand, it may be that the mechanism which generates particular values of Y is deterministic. In this case the error term will reflect the operation of additional unknown causes which in conjunction with X_1 are sufficient to determine the value of Y, or it may reflect the presence of measurement error in Y. On the other hand, it may be that the mechanism which produces the values of Y is itself indeterministic. In this case the error term will represent, as it were, the stochasticity that remains in Y, once we have taken account of the fixed contribution made to the value of Y by X_1.

Most discussions take the use of regression techniques to be neutral between these possibilities. In the sorts of contexts in which regression is used, it is generally uncontroversial that the spread in values of Y for a fixed X_1 results at least in part from the influence of omitted variables on Y. It is usually simply not known whether this spread reflects in addition the presence of a fundamental indeterminacy in the mechanism producing the values of Y. The use of regression techniques is thought to be justified

https://doi.org/10.1086/psaprocbienmeetp.1988.1.192993 Published online by Cambridge University Press

whether or not such indeterminacy is present, provided the appropriate assumptions about the distribution of the error term are satisfied.

What does the coefficient β_1 represent on these two possible construals of the error term? If we take the situation to be one in which the relationship between X_1 and Y is deterministic, so that the error term merely represents omitted variables and measurement error, then the coefficient β_1 appears to tell us what the change in Y would be if X_1 were to change by one unit and every other relevant variable were to be held constant. (An analogous interpretation will of course hold for each of the coefficients β_i in a multiple regression equation–in each case β_i represents the amount Y would change, given a unit change in X_i,with everything else held constant). On the other hand, if we think of the relationship between Y and X_1 as fundamentally indeterministic, then we will of course not be entitled to talk about "the" change in Y which would result from a unit change in X_1. But provided that (3*a* - *d*) are satisfied, we can think of the coefficient β_1 as telling us what average or expected change would result from a unit change in X_1. It is common to find both of these formulations (the change in Y, the average change in Y) in textbooks, which perhaps reflects the point that regression techniques are noncommittal regarding the truth or falsity of determinism. In neither case, however, does the coefficient β_1 have a straightforward interpretation which makes explicit reference to probabilities. Rather than explicitly representing a chancey or probabilistic link between X_1 and Y, the coefficient β_1 appears to tell us about a fixed linear relationship between changes in the value of X_1 and the value (or the expected value) of Y.[9]

The significance of this point becomes clearer when we consider in more detail the role that assumptions like (3*a* - *d*) concerning the distribution of the error term and its relation to the independent variable play in regression analysis. Intuitively, these assumptions function so as to insure that when we look at the association between Y and X_1, the other causal factors which we have left out of the equation will not, on the average, operate in such a way as to yield a systematically misleading estimate of the effect of X_1 on Y. Similarly, the role of assumption (3e) is to make possible significance testing and the construction of confidence intervals. In general the role of the entire set of assumptions (3*a* - *e*) about the distribution of the error term is one of specifying conditions which if met insure us that we may reliably estimate the value of β_1 from observations on Y and X_1, and not the role of supporting a probabilistic analysis of the causal relation between Y and X_1. Suppose, for example, that in the above case soil quality (X_2), which has been left out of the analysis, also usually affects plant height and that, furthermore, above average exposure to sunlight is systematically correlated with above average soil quality. Since the error term will, in part, reflect the operation of the omitted variable X_2, the error term will now be correlated with the included variable X_1 in violation of one or more of the assumptions (3*a* - *e*). In consequence our estimate of the coefficient β_1 will be biased–in this case it will exaggerate the influence of exposure to sunlight on plant height.[10] To obtain an unbiased estimate for β_1 we need to bring X_2 and perhaps additional variables as well explicitly into the regression equation.

As noted above, even if we think that the underlying causal story about plant growth is ultimately a fully deterministic one, so that deterministic processes generate each value of the disturbance term, and deterministic generalizations link the height of each plant to facts about its exposure to sunlight and the various other causes represented by the error term operative in that particular case, we can still use regression analysis as a basis for conducting a causal investigation into the effects of sunlight on plant height, as long as assumptions like (3*a* - *e*) are satisfied. Conversely, suppose that the relationship between plant height and sunlight is in the case of individual plants an irreducibly indeterministic one. This fact will not by itself guard against the above difficulty of specification error (i.e., error in the estimate for β_1) if other causes correlated with sunlight are operative. To rule out this possibility, one still needs, in addition, assumptions like (3a -e). This illustrates again how the real work done by the assumptions about the distribution of the

https://doi.org/10.1086/psaprocbienmeetp.1988.1.192993 Published online by Cambridge University Press

error term has to do with their evidential role and is quite independent of issues about whether (1) presupposes a probabilistic analysis of causation.[11]

6. Explanation and Regression

I noted above that traditional models of statistical explanation like Hempel's IS model or Salmon's SR model are clearly meant to apply to explanations of why a dependent variable takes the values it does for particular individuals in a population. By contrast, regression analyses (and other causal modeling techniques) do not purport to explain the behavior of particular individuals. Such techniques are more plausibly viewed as explaining aggregate or population level facts such as facts about changes in the mean value of dependent variable of interest.

One reason for thinking this is that it seems to best fit the conception of explanation employed in the literature on the subject. One common measure adopted for the assessment of a regression equation is the proportion of variance in the dependent variable which is "explained" by the independent variable. The basic idea is typically developed (focusing for simplicity on the case in which there is just one independent variable) as follows: let $\hat{y}_i = \alpha + \beta x_i$. Let y_i be, as before, the actual value of Y associated with x_i and let $\bar{y}$ be the mean for all values of Y. Then one can readily show that

$$(11) \quad \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

The first quantity on the right is the sum of the squares of the deviations of the actual values of Y from the values predicted by the regression line. It is, as Herbert Blalock writes in a typical passage, "unexplained since it indicates the amount of error in prediction" (1979, p. 107). By contrast, the second quantity on the right represents the relative improvement in one's ability to predict obtained by using x_i to predict y_i instead of $\bar{y}$. This quantity is referred to as the "explained" sum of squares. The quantity

$$(12) \quad \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

measures the ratio of the "explained variance" or sum of squares to the total variance in Y. This is readily shown to be equal to the square of the correlation coefficient r^2_{xy} and is a measure of spread about the regression line. The closer the actual values(y_i) of Y to the predicted values $\hat{y}_i$ (and the closer the value of r_{xy} to 1), the better the explanation which the independent variable in the equation is said to provide.

It seems clear that the underlying conception of explanation here involves the idea that what is explained is what can be deduced or predicted via the use of the regression equation. To the best of my knowledge, there is no suggestion in any of the literature on regression analysis that values y_i of Y which lie off the regression line are to be thought of as explained by the regression equation–rather the extent to which those values differ from the predicted values $\hat{y}_i$ is precisely the extent to which they are regarded as "unexplained." The model of explanation at work here is simply not one in which particular values of Y for particular individuals are regarded as explained according to some model of statistical

https://doi.org/10.1086/psaprocbienmeetp.1988.1.192993 Published online by Cambridge University Press

explanation by subsumption under a statistical generalization or by reference to some causal factor (or propensity) which operates in a probabilistic fashion.

As a measure of explanatory power, (12) has certain well-known limitations, which I have discussed elsewhere (cf. Woodward 1987). These limitations have to do with the fact that (12) is relativised to the variance of the dependent variable *Y* in an undesirable way and do not provide one with a good reason for rejecting the general claim that the conception of explanation embodied in regression analysis is in part deductive and that what is explained is facts about population-level parameters, rather than facts about individuals. If one wishes to reject (12) as a measure of explanatory power, the alternative construal which I think is most naturally suggested by the literature on regression analysis, is to simply take what is explained to be the mean value of *Y* or perhaps changes in the mean value of *Y* with changes in the values of *X*. On this sort of construal one thinks of a regression analysis as purporting to explain, e.g., a certain increase in the mean rate of lung cancer in a certain population in terms of an increase in *per capita* cigarette consumption or an increase in the mean murder rate for a group of states in terms of an increase in the unemployment rate.

An additional reason for adopting this sort of construal of what is explained in regression analysis is the obvious point that there is no direct and simple inference from general population-level causal claims to causal claims about particular individuals. Suppose, for the sake of definiteness that we regress a variable *Y* representing the *per capita* incidence of lung cancer in various regions of the United States on a variable X_1 representing the concentration of certain pollutants in the air of that region and obtain a positive coefficient β_1 differing significantly from zero. Suppose also that we are satisfied that assumptions of form 3*a* - *d* are satisfied and that the relationship between X_1 and *Y* is genuinely a causal one. Presumably underlying this relationship are facts about causal processes going on in individuals (particular people in various parts of the country are caused to get lung cancer at varying rates by these pollutants) and also facts about the causal processes that would go on in these individuals if they were exposed to more or less pollutants. Nonetheless, the regression analysis will not tell us why some particular individual developed lung cancer or necessarily identify the causal factors actually relevant to this outcome.

Even given the results of the regression analysis and the fact that some individual Jones was exposed to these pollutants and has developed lung cancer, we are not entitled to infer that this exposure actually caused Jones' lung cancer–it may be that Jones' cancer was entirely caused by exposure to some other carcinogen. To show that the pollutants in question were the cause of Jones' lung cancer, one must produce additional considerations–e.g., one must show that no other possible causes of lung cancer were present or that some characteristic *modus operandi* which such other causes exhibit when they produce lung cancer was absent. This seems to me to provide an additional reason for doubting that regression analysis explains such facts as why particular individuals get lung cancer or become juvenile delinquents or directly describes the individual causal processes at work when some particular person develops these conditions. It is instead more plausible to conclude that regression analysis captures average or aggregate features of the operation of many individual causal processes in a population–that such an analysis tells us about, for example, the aggregate causal impact of various levels of air pollution on lung cancer in different areas of the United States.

I have argued above that the explanations provided in regression analysis are deductive in structure. This is reassuring since many explanations elsewhere in science also possess a deductive structure. I now want to suggest, again rather schematically, that the results of regression analyses possess another feature often associated with good explanation elsewhere in science.[12] Contrary to what Patrick Suppes claims in the remark quoted above insofar as a regression equation is understood as furnishing an explanation and as reflecting a non-spurious causal relationship, it should be understood as embodying a

https://doi.org/10.1086/psaprocbienmeetp.1988.1.192993 Published online by Cambridge University Press

counterfactual claim. An equation of form (1) or (8) tells us how the mean value of the dependent variable would change if, contrary to present fact, the average value of the independent variable(s) were to change in various ways. That is to say, a regression equation reflecting a non-spurious causal relationship claims to provide one with information about a (population-specific) pattern of counterfactual dependency obtaining between the mean values of the independent and dependent variables. If one claims, on the basis of regression analysis,that an increase of some amount (or some part of the increase) in the incidence of lung cancer among women is due to (and explained by) an increase in their average cigarette smoking, then one must also believe that if American women were to smoke less this would change the incidence of lung cancer among them in the way represented by the relevant regression equation.

The fact that when a regression equation is taken to reflect a genuine causal relationship, it will carry with it this sort of counterfactual commitment is reflected in, among other things, the use of such techniques in connection with public policy issues. Thus when Tufte, in an example referred to above, regresses a variable representing statewide automobile fatalities on a variable representing quality of state automobile inspections, he is explicitly interested in the question of whether if all states were required to make inspections (as is not now the case), this would be an effective way of reducing fatalities. It is precisely because he concludes that there is a causal connection between these two variables, and because he takes this claim to have counterfactual import, that he concludes that requiring universal inspections would reduce fatalities. It is because he thinks that the relationship between fatalities and the number of letters in a state's name is non-causal and spurious that he does not think that changing states' names would be an effective strategy for promoting automobile safety. Similar points can be made about the use of regression analysis and other causal modeling techniques to assess the deterrent effect of the death penalty or the effect on educational attainment of pre-school enrichment programs such as Head Start. Similarly, if the examples of regression equations (6) and (7) above are understood as having causal or explanatory import, they should also be understood as counterfactual claims about how consumer spending would change if contrary to fact, disposable income would change in various ways or about how Democratic representation in Congress would change if the Democratic popular vote were to change in various ways. Insofar as one's interest in regression techniques goes beyond the purely descriptive (and represents an interest in explanation and in establishing causal connections) a central point of the whole exercise is the successful identification of patterns of counterfactual dependency.[13]

Notes

[1] Suppose that we are attempting to estimate the value of a parameter ϑ and we have a sample of n observations $y_1 \ldots y_n$. An estimator t for ϑ will be a function $t\ (y_1 \ldots y_n)$– itself a random variable–of these observations and will be unbiased for ϑ if its expected value, $E(t) = \vartheta$. Linear estimators are those for which t is a linear function, that is, $t=c_1 y_1+c_2 y_2+ \ldots +c_n y_n$, where $c_1 \ldots c_n$ are constants. The estimator among the class of unbiased, linear estimators which has minimum variance will be the best linear unbiased estimator (BLUE).

[2] Here

$$S_x = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}}, \text{ and } S_{xy} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

where $\bar{x}$ and $\bar{y}$ are the sample means for X and Y.

https://doi.org/10.1086/psaprocbienmeetp.1988.1.192993 Published online by Cambridge University Press

[3] For example in the case in which we have a regression equation with two independent variables X_1 and X_2 and a dependent variable Y.

i.e., $Y=\beta_1 X_1+\beta_2 X_2+U$

the estimator $\hat{\beta}_1$ for the partial regression coefficient β_1 will be

$$\hat{\beta}_1 = \frac{S_{x2}^2 \cdot S_{yx1} - S_{yx2}S_{yx1x2}}{S_{x1}^2 S_{x2}^2 - (S_{yx1x2})^2}$$

where S_{xy} is, as before, the covariance between x and y. Clearly the value of the regression coefficient will depend in part on the covariance Sx_1x_2, which will reflect the correlation between X_1 and X_2. Note also that if X_1 and X_2 are not correlated, so that $Sx_1x_2=0$ then

$$\hat{\beta}_1 = \frac{S_{yx1}}{S_{x1}^2}$$

which is just the estimate for the regression coefficient in the one variable case–cf. (4). If X_1 and X_2 are not correlated, one can just ignore one variable in estimating a value for the coefficient for the other variable, but one cannot do this if the variables are correlated.

[4] Simpson's paradox, which is arguably not a paradox at all but simply a problem for any purely probabilistic definition of "cause," involves the observation that any relation of statistical association between two variables in a population can be reversed in the subpopulations characterized by reference to some third variable correlated with both of the original variables.

[5] For examples, and further discussion, see e.g., Simon (1953).

[6] This is a common-place observation in the literature on causal modeling–see, for example, Simon (1954), Asher (1983), Achen (1982), and Blalock (1979). Indeed, in many cases there will literally be millions of linear causal models which entail observed facts about correlations in the data. Developing systematic procedures (where possible) for generating such alternatives and efficient and systematic criteria for evaluating them is perhaps the central methodological problem posed by the use of causal modeling techniques–for pertinent discussion see especially Glymour, Scheines, Spirtes, and Kelly (1987).

[7] I am concerned here with the question of what (1) claims, and not with whether this claim about the constancy of the regression coefficient across the entire population is plausible. In fact, it seems clear that this claim of constancy is not plausible in many of the contexts in which regression techniques are used. The value of the regression coefficients will instead be different across different subpopulations of the total population for which the equation is estimated. The coefficient in an equation like (1) will thus represent an average over the coefficients for these subpopulations–a case of what is called an "interaction effect." I would reject the suggestion that in this sort of case, (1) should be rejected as telling us nothing of causal interest, although I lack the space to defend this view here.

[8] See, for example, Johnston (1972), p. 11.

[9]There is one rather special circumstance in which it is sometimes argued that the coefficient β_1 has a straightforward interpretation in terms of probabilities. Suppose that Y and X_1 are both dichotomous variables–that is, restricted to the values 1 and 0 which we may associate with their occurrence or non-occurrence. Then it is easy to show that the coefficient β_1, can be interpreted as the difference between the probability that Y occurs, given that X_1 occurs and the probability that Y occurs, given that X_1 does not occur. Several writers (e.g., Humphreys, 1985) appeal to this fact in claiming that there is a close connection between linear causal models and probabilistic analyses of causation of the sort found in the philosophical literature. This suggested connection strikes me as misleading. In ordinary regression analysis, while one or more of the independent variables can be dichotomous, it is essential that the dependent variables be measurable on an interval scale. When a linear model is used with a dichotomous dependent variable, several key assumptions underlying the use of ordinary least squares estimation techniques will be violated and the resulting parameter estimates will be biased. Indeed, if one assumes a linear model, one may get predicted values for the dependent variable which are negative or greater than one, and which are thus not interpretable as probabilities. For this reason, the preferred approach with a dichotomous dependent variable is to assume one of a variety of nonlinear functional forms and to estimate these by means of maximum likelihood techniques. Thus causal models with dichotomous dependent variables are not appropriately treated as "special cases" of linear regression models at all. Moreover, quite apart from this, even if it were appropriate to regard the use of a model with dichotomous variables as a special case of a linear model, the proposed probabilistic interpretation doesn't seem to generalize to yield a natural interpretation in terms of probabilities for the regression coefficient for models with variables measureable on an interval scale and such models are, after all, the standard or paradigmatic cases of regression models. In general, I think that the discontinuities and disanalogies between standard philosophical accounts of causation, which are framed in terms of dichotomous or discrete-valued variables, and linear causal models, which are generally taken to apply to variables which are measureable on an interval scale, are much more striking than the continuities. From the perspective of the literature on linear causal models, the tendency of philosophical discussion to focus so exclusively on the dichotomous or discrete-valued case looks myopic and not well motivated.

[10]To be a bit more precise, suppose that the "true" regression equation is

(13) $Y=\beta_1 X_1+\beta_2 X_2+U$

but that one mistakenly omits the relevant variable X_2 and instead estimates

(14) $Y=\beta_1 X_1+V$

The least squares estimator for β_1 from (14)–the omitted variable equation (em is

$$\hat{\beta}_1 = \frac{\Sigma_{yx1}}{\Sigma_{x1}^{2}}$$

(where the values of y and x_1 have now been "standardized" by subtracting $\bar{y}$ and $\bar{x}$)
Then one can show that

$$\mathrm{E}\,(\hat{\beta}_1)=\beta_1+\beta_2 b_{12}$$

where b_{12} is the regression coefficient of the omitted variable X_2 on the included variable X_1. Thus the bias in the estimated value of β_1 resulting from the omission of X_2 will be the true coefficient of X_2 multiplied by the regression coefficient of X_2 on X_1.

https://doi.org/10.1086/psaprocbienmeetp.1988.1.192993 Published online by Cambridge University Press

[11] For an additional argument that, although probabilities are relevant as evidence for causal claims, it does not follow that the correct analysis of causality involves probabilistic elements, see David Papineau's interesting recent paper (1985). Some of Papineau's arguments echo arguments made here, but I would dissent from a number of the details of his discussion.

[12] For arguments that good explanations in science are often deductive in structure and exhibit patterns of counterfactual dependency, see Woodward (1979) and Woodward (forthcoming).

[13] But what warrants or grounds the assertion of such counterfactuals? How can one justifiably arrive at a conclusion which has counterfactual import on the basis of data about actual frequencies? The answer to this question, as I hope my discussion above suggests, is that in carrying out a regression analysis one relies on other assumptions as well as frequency data. These are the extra-statistical "causal" assumptions concerning possibly causally relevant variables, causal ordering, and so forth described above. These assumptions, since they are causally loaded themselves have counterfactual import and it is in virtue of making them, that we are able to draw conclusions which have counterfactual import via regression analysis. If we did not make any such assumptions,and attempted to make inferences by relying only on data about actual frequencies and nothing more, we would not be entitled to draw conclusions that had causal or counterfactual import.

References

Achen, C. (1982). *Interpreting and Using Regression.* Beverly Hills: Sage Publications.

Ascher, H. (1983). *Causal Modeling*. Beverly Hills: Sage Publications.

Blalock, H., ed. (1971). *Causal Models in the Social Sciences.* New York: Aldine Publishing Company.

________. (1979). *Social Statistics.* New York: McGraw-Hill.

Cartwright, N. (1979). "Causal Laws and Effective Strategies." *Nous* 13: 419-437.

Gardiner, P. (1961). *The Nature of Historical Explanation.* Oxford: Oxford University Press.

Glymour, C., Scheines, R., Spirtes, P., and Kelly, K. (1987). *Discovering Causal Structure.* Florida: Academic Press, Inc.

Hempel, C. (1965). *Aspects of Scientific Explanation.* Illinois: The Free Press.

________. (1968). "Maximal Specificity and Lawlikeness In Probabilistic Explanation." *Philosophy of Science* 35: 116-133.

Humphreys, P. (1985). "Quantitative Probabalistic Causality and Structural Scientific Realism." *PSA 1984*, Volume 2. Edited by Peter Asquith and Philip Kitcher. East Lansing, MI: Philosophy of Science Association.

Johnston, J. (1972). *Econometric Methods.* New York: McGraw-Hill.

Kendall, M. A. and Stuart, A. (1961). *The Advanced Theory of Statistics.* Volume 2, London: Griffin.

https://doi.org/10.1086/psaprocbienmeetp.1988.1.192993 Published online by Cambridge University Press

Maddala, G.S. (1977). *Econometrics.* New York: McGraw-Hill.

Papineau, D. (1985). "Probabilities and Causes." *The Journal of Philosophy* 82: 57-74.

Railton, P. (1978). "A Deductive-Monological Model of Probabilistic Explanation." *Philosophy of Science* 45: 20-226.

_______. (1981). "Probability, Explanation, and Information." *Synthese* 48: 233-256.

Rosenberg, A. (1980). *Sociobiology and the Preemption of Social Science.* Baltimore: The Johns Hopkins University Press.

Salmon, W., ed. (1971). *Statistical Explanation and statistical Relevance.* Pittsburgh: University of Pittsburgh Press.

________. (1984) *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

Simon, H. (1953). "Causal Ordering and Identifiability." In *Studies in Econometric Method.* Edited by Hood and Koopmans. New York: Wiley.

_______. (1954). "Spurious Correlation: A Causal Interpretation." *Journal of American Statistical Association* 49: 467-492.

Suppes, P. (1970). *A Probabilistic Theory of Causality.* Amsterdam: North Holland Publishing Co.

_______. (1984). *Probabilistic Metaphysics.* Oxford: Basil Blackwell.

Tufte, E. (1974). *Data Analysis For Politics and Policy.* New Jersey: Prentice-Hall.

Woodward, J.F. (1979). "Scientific Explanation." *The British Journal for the Philosophy of Science* 30: 41-67.

__________. (1987). "On An Information-Theoretic Model of Explanation." *Philosophy of Science* 54:21-44.

___________. (forthcoming). "The Causal Mechanical Model of Explanation." In a forthcoming volume of *Minnesota Studies in the Philosophy of Science.*

https://doi.org/10.1086/psaprocbienmeetp.1988.1.192993 Published online by Cambridge University Press