# 1 Statistical Models and Social Science

The social world is exquisitely complex and rich. From the improbable moment of birth, each of our lives is governed by chance and contingency. The statistical models typically used to analyze social data—and, in particular, the models considered in this book—are, in contrast, ludicrously simple. How can simple statistical models help us to understand a complex social reality? As the statistician George Box famously remarked (e.g., in Box, 1979), ''All models are wrong but some are useful'' (p. 202). Can statistical models be useful in the social sciences?

This is a book on data analysis and statistics, not on the philosophy of the social sciences. I will, therefore, address this question, and related issues, very briefly here. Nevertheless, I feel that it is useful to begin with a consideration of the role of data analysis in the larger process of social research. You need not agree with the point of view that I express in this chapter to make productive use of the statistical tools presented in the remainder of the book, but the emphasis and specific choice of methods in the text partly reflect the ideas in this chapter. You may wish to reread this material after you study the methods described in the sequel.

## 1.1 Statistical Models and Social Reality

As I said, social reality is complex: Consider how my income is ''determined.'' I am a relatively well-paid professor in the sociology department of a Canadian university. That the billiard ball of my life would fall into this particular pocket was, however, hardly predictable a half-century ago, when I was attending a science high school in New York City. My subsequent decision to study sociology at New York's City College (after several other majors), my interest in statistics (the consequence of a course taken without careful consideration in my senior year), my decision to attend graduate school in sociology at the University of Michigan (one of several more or less equally attractive possibilities), and the opportunity and desire to move to Canada (the vote to hire me at the University of Alberta was, I later learned, very close) are all events that could easily have occurred differently.

I do not mean to imply that personal histories are completely capricious, unaffected by social structures of race, ethnicity, class, gender, and so on, just that they are not *in detail* determined by these structures. That social structures—and other sorts of systematic factors—condition, limit, and encourage specific events is clear from each of the illustrations in the previous paragraph and in fact makes sense of the argument for the statistical analysis of social data presented below. To take a particularly gross example: The public high school that I attended admitted its students by competitive examination, but no young women could apply (a policy that has happily changed).

Each of these precarious occurrences clearly affected my income, as have other events—some significant, some small—too numerous and too tedious to mention, even if I were aware of them all. If, for some perverse reason, you were truly interested in my income (and, perhaps, in other matters more private), you could study my biography and through that study arrive at a detailed (if inevitably incomplete) understanding. It is clearly impossible, however, to pursue this strategy for many individuals or, more to the point, for individuals in general.

Nor is an understanding of income in general inconsequential, because income inequality is an (increasingly, as it turns out) important feature of our society. If such an understanding hinges on a literal description of the process by which each of us receives an income, then the enterprise is clearly hopeless. We might, alternatively, try to capture significant features of the process in general without attempting to predict the outcome for specific individuals. One could draw formal analogies (largely unproductively, I expect, although some have tried) to chaotic physical processes, such as the determination of weather and earthquakes.

Concrete mathematical theories purporting to describe social processes sometimes appear in the social sciences (e.g., in economics and in some areas of psychology), but they are relatively rare.[1] If a theory, like Newton's laws of motion, is mathematically concrete, then, to be sure, there are difficulties in applying and testing it; but, with some ingenuity, experiments and observations can be devised to estimate the free parameters of the theory (a gravitational constant, for example) and to assess the fit of the theory to the resulting data.

In the social sciences, *verbal* theories abound. These social theories tend to be vague, elliptical, and highly qualified. Often, they are, at least partially, a codification of ''common sense.'' I believe that vague social theories are potentially useful abstractions for understanding an intrinsically complex social reality, but how can such theories be linked empirically to that reality?

A vague social theory may lead us to expect, for example, that racial prejudice is the partial consequence of an ''authoritarian personality,'' which, in turn, is a product of rigid childrearing. Each of these terms requires elaboration and procedures of assessment or measurement. Other social theories may lead us to expect that higher levels of education should be associated with higher levels of income, perhaps because the value of labor power is enhanced by training, because occupations requiring higher levels of education are of greater functional importance, because those with higher levels of education are in relatively short supply, or because people with high educational attainment are more capable in the first place. In any event, we need to consider how to assess income and education, how to examine their relationship, and what other variables need to be included.[2]

Statistical models of the type considered in this book are grossly *simplified* descriptions of complex social reality. Imagine that we have data from a social survey of a large sample of employed individuals. Imagine further, anticipating the statistical methods described in subsequent chapters, that we regress these individuals' income on a variety of putatively relevant characteristics, such as their level of education, gender, race, region of residence, and so on. We recognize that a model of this sort will fail to account perfectly for individuals' incomes, so our model includes a ''residual,'' meant to capture the component of income unaccounted

[1]The methods for fitting nonlinear models described in Chapter 17 are sometimes appropriate to the rare theories in social science that are mathematically concrete.

[2]See Section 1.2.

for by the systematic part of the model, which incorporates the ''effects'' on income of education, gender, and so forth.

The residuals for our model are likely very large. Even if the residuals were small, however, we would still need to consider the relationships among our social ''theory,'' the statistical model that we have fit to the data, and the social ''reality'' that we seek to understand. Social reality, along with our methods of observation, produces the data; our theory aims to explain the data, and the model to describe them. That, I think, is the key point: Statistical models are almost always fundamentally *descriptive*.

I believe that a statistical model cannot, and is not literally meant, to capture the social process by which incomes are ''determined.'' As I argued above, individuals receive their incomes as a result of their almost unimaginably complex personal histories. No regression model, not even one including a residual, can reproduce this process: It is not as if my income is partly determined by my education, gender, race, and so on, and partly by the detailed trajectory of my life. It is, therefore, not sensible, at the level of real social processes, to relegate chance and contingency to a random term that is simply added to the systematic part of a statistical model. The unfortunate tendency to *reify* statistical models—to forget that they are descriptive summaries, not literal accounts of social processes—can only serve to discredit quantitative data analysis in the social sciences.

Nevertheless, and despite the rich chaos of individuals' lives, social theories imply a structure to income inequality. Statistical models are capable of capturing and describing that structure or at least significant aspects of it. Moreover, social research is often motivated by questions rather than by hypotheses: Has income inequality between men and women changed recently? Is there a relationship between public concern over crime and the level of crime? Data analysis can help to answer these questions, which frequently are of practical—as well as theoretical—concern. Finally, if we proceed carefully, data analysis can assist us in the discovery of social facts that initially escape our hypotheses and questions.

It is, in my view, a paradox that the statistical models that are at the heart of most modern quantitative social science are at once taken too seriously and not seriously enough by many practitioners of social science. On one hand, social scientists write about simple statistical models as if they were direct representations of the social processes that they purport to describe. On the other hand, there is frequently a failure to attend to the descriptive accuracy of these models.

As a shorthand, reference to the ''effect'' of education on income is innocuous. That the shorthand often comes to dominate the interpretation of statistical models is reflected, for example, in much of the social science literature that employs structural-equation models (once commonly termed ''causal models,'' a usage that has thankfully declined). There is, I believe, a valid sense in which income is ''affected'' by education, because the complex real process by which individuals' incomes are determined is partly conditioned by their levels of education, but—as I have argued above—one should not mistake the model for the process.[3]

Although statistical models are very simple in comparison to social reality, they typically incorporate strong claims about the descriptive pattern of data. These claims rarely reflect the

[3]There is the danger here of simply substituting one term (''conditioned by'') for another (''affected by''), but the point is deeper than that: Education affects income because the choices and constraints that partly structure individuals' lives change systematically with their level of education. Many highly paid occupations in our society are closed to individuals who lack a university education, for example. To recognize this fact, and to examine its descriptive reflection in a statistical summary, is different from claiming that a university education literally adds an increment to individuals' incomes.

substantive social theories, hypotheses, or questions that motivate the use of the statistical models, and they are very often wrong. For example, it is common in social research to assume a priori, and without reflection, that the relationship between two variables, such as income and education, is linear. Now, we may well have good reason to believe that income tends to be *higher* at higher levels of education, but there is no reason to suppose that this relationship is *linear*. Our practice of data analysis should reflect our ignorance as well as our knowledge.

A statistical model is of no practical use if it is an inaccurate description of the data, and we will, therefore, pay close attention to the descriptive accuracy of statistical models. Unhappily, the converse is not true, for a statistical model may be descriptively accurate but of little practical use; it may even be descriptively accurate but substantively misleading. We will explore these issues briefly in the next two sections, which tie the interpretation of statistical models to the manner in which data are collected.

> With few exceptions, statistical data analysis describes the outcomes of real social processes and not the processes themselves. It is therefore important to attend to the descriptive accuracy of statistical models and to refrain from reifying them.

## 1.2 Observation and Experiment

It is common for (careful) introductory accounts of statistical methods (e.g., Freedman, Pisani, & Purves, 2007; Moore, Notz, & Fligner, 2013) to distinguish strongly between observational and experimental data. According to the standard distinction, causal inferences are justified (or, at least, more certain) in experiments, where the explanatory variables (i.e., the possible ''causes'') are under the direct control of the researcher; causal inferences are especially compelling in a randomized experiment, in which the values of explanatory variables are assigned by some chance mechanism to experimental units. In nonexperimental research, in contrast, the values of the explanatory variables are observed—not assigned—by the researcher, along with the value of the response variable (the ''effect''), and causal inferences are not justified (or, at least, are less certain). I believe that this account, although essentially correct, requires qualification and elaboration.

To fix ideas, let us consider the data summarized in Table 1.1, drawn from a paper by Greene and Shaffer (1992) on Canada's refugee determination process. This table shows the outcome of 608 cases, filed in 1990, in which refugee claimants who were turned down by the Immigration and Refugee Board asked the Federal Court of Appeal for leave to appeal the board's determination. In each case, the decision to grant or deny leave to appeal was made by a single judge. It is clear from the table that the 12 judges who heard these cases differed widely in the percentages of cases that they granted leave to appeal. Employing a standard significance test for a contingency table (a chi-square test of independence), Greene and Shaffer calculated that a relationship as strong as the one in the table will occur by chance alone about two times in 100,000. These data became the basis for a court case contesting the fairness of the Canadian refugee determination process.

If the 608 cases had been assigned at random to the judges, then the data would constitute a natural experiment, and we could unambiguously conclude that the large differences among

**Table 1.1** Percentages of Refugee Claimants in 1990 Who Were Granted or Denied Leave to Appeal a Negative Decision of the Canadian Immigration and Refugee Board, Classified by the Judge Who Heard the Case

| | *Leave Granted?* | | | |
|---|---|---|---|---|
| *Judge* | *Yes* | *No* | *Total* | *Number of cases* |
| Pratte | 9 | 91 | 100 | 57 |
| Linden | 9 | 91 | 100 | 32 |
| Stone | 12 | 88 | 100 | 43 |
| Iacobucci | 12 | 88 | 100 | 33 |
| Décary | 20 | 80 | 100 | 80 |
| Hugessen | 26 | 74 | 100 | 65 |
| Urie | 29 | 71 | 100 | 21 |
| MacGuigan | 30 | 70 | 100 | 90 |
| Heald | 30 | 70 | 100 | 46 |
| Mahoney | 34 | 66 | 100 | 44 |
| Marceau | 36 | 64 | 100 | 50 |
| Desjardins | 49 | 51 | 100 | 47 |
| All judges | 25 | 75 | 100 | 608 |

SOURCE: Adapted from Table 1 in Greene and Shaffer, "Leave to Appeal and Leave to Commence Judicial Review in Canada's Refugee-Determination System: Is the Process Fair?" *International Journal of Refugee Law*, 1992, Vol. 4, No. 1, p. 77, by permission of Oxford University Press.

the judges reflect differences in their propensities to grant leave to appeal.[4] The cases were, however, assigned to the judges not randomly but on a rotating basis, with a single judge hearing all of the cases that arrived at the court in a particular week. In defending the current refugee determination process, expert witnesses for the Crown argued that the observed differences among the judges might therefore be due to characteristics that systematically differentiated the cases that different judges happened to hear.

It is possible, in practice, to "control" statistically for such extraneous "confounding" variables as may explicitly be identified, but it is not, in principle, possible to control for *all* relevant explanatory variables, because we can never be certain that all relevant variables have been identified.[5] Nevertheless, I would argue, the data in Table 1.1 establish a prima facie case for systematic differences in the judges' propensities to grant leave to appeal to refugee claimants. Careful researchers control statistically for potentially relevant variables that they can identify; cogent critics demonstrate that an omitted confounding variable accounts for the observed association between judges and decisions or at least argue persuasively that a specific omitted variable *may* be responsible for this association—they do not simply maintain the abstract possibility that such a variable *may* exist.

[4]Even so, this inference is not reasonably construed as a representation of the *cognitive process* by which judges arrive at their determinations. Following the argument in the previous section, it is unlikely that we could ever trace out that process in detail; it is quite possible, for example, that a specific judge would make different decisions faced with the same case on different occasions.

[5]See the further discussion of the refugee data in Section 22.3.1.

What makes an omitted variable "relevant" in this context?[6]

1. The omitted variable must influence the response. For example, if the gender of the refugee applicant has no impact on the judges' decisions, then it is irrelevant to control statistically for gender.
2. The omitted variable must be related as well to the explanatory variable that is the focus of the research. Even if the judges' decisions are influenced by the gender of the applicants, the relationship between outcome and judge will be unchanged by controlling for gender (e.g., by looking separately at male and female applicants) unless the gender of the applicants is also related to judges—that is, unless the different judges heard cases with substantially different proportions of male and female applicants.

The strength of randomized experimentation derives from the second point: If cases were randomly assigned to judges, then there would be no systematic tendency for them to hear cases with differing proportions of men and women—or, for that matter, with systematic differences of *any* kind.

It is, however, misleading to conclude that causal inferences are completely unambiguous in experimental research, even within the bounds of statistical uncertainty (expressed, for example, in the $p$-value of a statistical test). Although we can unambiguously ascribe an observed difference to an experimental *manipulation*, we cannot unambiguously identify that manipulation with the explanatory variable that is the focus of our research.

In a randomized drug study, for example, in which patients are prescribed a new drug or an inactive placebo, we may establish with virtual certainty that there was greater average improvement among those receiving the drug, but we cannot be sure that this difference is due (or solely due) to the putative active ingredient in the drug. Perhaps the experimenters inadvertently conveyed their enthusiasm for the drug to the patients who received it, influencing the patients' responses, or perhaps the bitter taste of the drug subconsciously convinced these patients of its potency.

Experimenters try to rule out alternative interpretations of this kind by following careful experimental practices, such as "double-blind" delivery of treatments (neither the subject nor the experimenter knows whether the subject is administered the drug or the placebo), and by holding constant potentially influential characteristics deemed to be extraneous to the research (the taste, color, shape, etc., of the drug and placebo are carefully matched). One can never be certain, however, that *all* relevant variables are held constant in this manner. Although the degree of certainty achieved is typically much greater in a randomized experiment than in an observational study, the distinction is less clear-cut than it at first appears.

> Causal inferences are most certain—if not completely definitive—in randomized experiments, but observational data can also be reasonably marshaled as evidence of causation. Good experimental practice seeks to avoid confounding experimentally manipulated explanatory variables with other variables that can influence the response variable. Sound analysis of observational data seeks to control statistically for potentially confounding variables.

[6]These points are developed more formally in Sections 6.3 and 9.7.

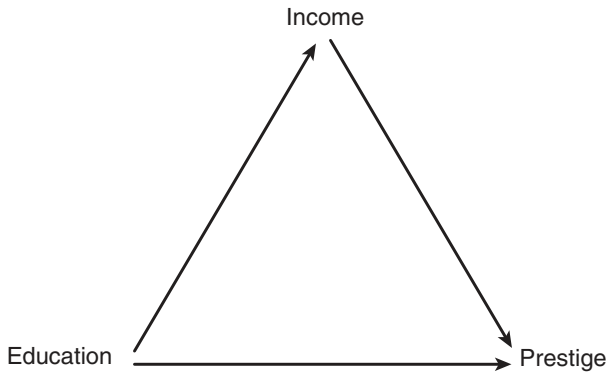


**Figure 1.1** Simple "causal model" relating education, income, and prestige of occupations. Education is a common prior cause of both income and prestige; income intervenes causally between education and prestige.

In subsequent chapters, we will have occasion to examine observational data on the prestige, educational level, and income level of occupations. It will materialize that occupations with higher levels of education tend to have higher prestige and that occupations with higher levels of income also tend to have higher prestige. The income and educational levels of occupations are themselves positively related. As a consequence, when education is controlled statistically, the relationship between prestige and income grows smaller; likewise, when income is controlled, the relationship between prestige and education grows smaller. In neither case, however, does the relationship disappear.

How are we to understand the pattern of statistical associations among the three variables? It is helpful in this context to entertain an informal "causal model" for the data, as in Figure 1.1. That is, the educational level of occupations influences (potentially) both their income level and their prestige, while income potentially influences prestige. The association between prestige and income is "spurious" (i.e., not causal) to the degree that it is a consequence of the mutual dependence of these two variables on education; the reduction in this association when education is controlled represents the removal of the spurious component. In contrast, the causal relationship between education and prestige is partly mediated by the "intervening variable" income; the reduction in this association when income is controlled represents the articulation of an "indirect" effect of education on prestige (i.e., through income).

In the former case, we partly *explain away* the association between income and prestige: Part of the relationship is "really" due to education. In the latter case, we partly *explain* the association between education and prestige: Part of the relationship is mediated by income.

> In analyzing observational data, it is important to distinguish between a variable that is a common prior cause of an explanatory and response variable and a variable that intervenes causally between the two.

Causal interpretation of observational data is always risky, especially—as here—when the data are cross-sectional (i.e., collected at one point in time) rather than longitudinal (where the data

are collected over time). Nevertheless, it is usually impossible, impractical, or immoral to collect experimental data in the social sciences, and longitudinal data are often hard to come by.[7] Moreover, the essential difficulty of causal interpretation in nonexperimental investigations—due to potentially confounding variables that are left uncontrolled—applies to longitudinal as well as to cross-sectional observational data.[8]

The notion of ''cause'' and its relationship to statistical data analysis are notoriously difficult ideas. A relatively strict view requires an experimentally manipulable explanatory variable, at least one that is manipulable in principle.[9] This is a particularly sticky point because, in social science, many explanatory variables are intrinsically not subject to direct manipulation, even in principle. Thus, for example, according to the strict view, gender cannot be considered a cause of income, even if it can be shown (perhaps after controlling for other determinants of income) that men and women systematically differ in their incomes, because an individual's gender cannot be changed.[10]

I believe that treating nonmanipulable explanatory variables, such as gender, as potential causes is, at the very least, a useful shorthand. Men earn higher incomes than women be*cause* women are (by one account) concentrated into lower paying jobs, work fewer hours, are directly discriminated against, and so on (see, e.g., Ornstein, 1983). Explanations of this sort are perfectly reasonable and are subject to statistical examination; the sense of ''cause'' here may be weaker than the narrow one, but it is nevertheless useful.

> It is overly restrictive to limit the notion of statistical causation to explanatory variables that are manipulated experimentally, to explanatory variables that are manipulable in principle, or to data that are collected over time.

## 1.3 Populations and Samples

Statistical inference is typically introduced in the context of random sampling from an identifiable population. There are good reasons for stressing this interpretation of inference—not the least of which are its relative concreteness and clarity—but the application of statistical inference is, at least arguably, much broader, and it is certainly broader in practice.

Take, for example, a prototypical experiment, in which subjects are assigned values of the explanatory variables at random: Inferences may properly be made to the hypothetical

[7]Experiments with human beings also frequently distort the processes that they purport to study: Although it might well be possible, for example, to recruit judges to an experimental study of judicial decision making, the artificiality of the situation could easily affect their simulated decisions. Even if the study entailed real judicial judgments, the mere act of observation might influence the judges' decisions—they might become more careful, for example.

[8]We will take up the analysis of longitudinal data in Chapters 23 and 24 on mixed-effects models.

[9]For clear presentations of this point of view, see, for example, Holland (1986) and Berk (2004).

[10]This statement is, of course, arguable: There are historically many instances in which individuals have changed their gender, for example by disguise, not to mention surgery. Despite some fuzziness, however, I believe that the essential point—that some explanatory variables are not (normally) subject to manipulation—is valid. A more subtle point is that in certain circumstances, we could imagine experimentally manipulating the *apparent* gender of an individual, for example, on a job application.

population of random rearrangements of the subjects, even when these subjects are not sampled from some larger population. If, for example, we find a highly ''statistically significant'' difference between two experimental groups of subjects in a randomized experiment, then we can be sure, with practical certainty, that the difference was due to the experimental manipulation. The rub here is that our interest almost surely extends *beyond* this specific group of subjects to some larger—often ill-defined—population.

Even when subjects in an experimental or observational investigation are literally sampled at random from a real population, we usually wish to generalize beyond that population. There are exceptions—election polling comes immediately to mind—but our interest is seldom confined to the population that is directly sampled. This point is perhaps clearest when *no* sampling is involved—that is, when we have data on every individual in a real population.

Suppose, for example, that we examine data on population density and crime rates for all large U.S. cities and find only a weak association between the two variables. Suppose further that a standard test of statistical significance indicates that this association is so weak that it easily could have been the product of ''chance.''[11] Is there any sense in which this information is interpretable? After all, we have before us data on the *entire* population of large U.S. cities at a particular historical juncture.

Because our interest inheres not directly—at least not exclusively—in these *specific* cities but in the complex social processes by which density and crime are determined, we can reasonably imagine a different outcome. Were we to replay history conceptually, we would not observe precisely the same crime rates and population density statistics, dependent as these are on a myriad of contingent and chancy events; indeed, if the ambit of our conceptual replay of history is sufficiently broad, then the identities of the cities themselves might change. (Imagine, for example, that Henry Hudson had not survived his trip to the New World or, if he survived it, that the capital of the United States had remained in New York. Less momentously, imagine that Fred Smith had not gotten drunk and killed a friend in a brawl, reducing the number of homicides in New York by one.) It is, in this context, reasonable to draw statistical inferences to the process that produced the currently existing populations of cities. Similar considerations arise in the analysis of historical statistics, for example, of time-series data.[12]

Much interesting data in the social sciences—and elsewhere—are collected haphazardly. The data constitute neither a sample drawn at random from a larger population nor a coherently defined population. Experimental randomization provides a basis for making statistical inferences to the population of rearrangements of a haphazardly selected group of subjects, but that is in itself cold comfort. For example, an educational experiment is conducted with students recruited from a school that is conveniently available. We are interested in drawing conclusions about the efficacy of teaching methods for students in general, however, not just for the students who participated in the study.

Haphazard data are also employed in many observational studies—for example, volunteers are recruited from among university students to study the association between eating disorders and overexercise. Once more, our interest transcends this specific group of volunteers.

To rule out haphazardly collected data would be a terrible waste; it is, instead, prudent to be careful and critical in the interpretation of the data. We should try, for example, to satisfy ourselves that our haphazard group does not differ in presumably important ways from the larger

[11] Cf. the critical discussion of crime and population density in Freedman (1975).

[12] See Chapter 16 for a discussion of regression analysis with time-series data.

population of interest, or to control statistically for variables thought to be relevant to the phenomena under study.

Statistical inference can speak to the *internal stability* of patterns in haphazardly collected data and—most clearly in experimental data—to causation. *Generalization* from haphazardly collected data to a broader population, however, is inherently a matter of judgment.

> Randomization and good sampling design are desirable in social research, but they are not prerequisites for drawing statistical inferences. Even when randomization or random sampling is employed, we typically want to generalize beyond the strict bounds of statistical inference.

## Exercise

**Exercise 1.1.** Imagine that students in an introductory statistics course complete 20 assignments during two semesters. Each assignment is worth 1% of a student's final grade, and students get credit for assignments that are turned in on time and that show reasonable effort. The instructor of the course is interested in whether doing the homework contributes to learning, and (anticipating material to be taken up in Chapters 5 and 6), she observes a linear, moderately strong, and highly statistically significant relationship between the students' grades on the final exam in the course and the number of homework assignments that they completed. For concreteness, imagine that for each additional assignment completed, the students' grades on average were 1.5 higher (so that, e.g., students completing all of the assignments on average scored 30 points higher on the exam than those who completed none of the assignments).

(a) Can this result be taken as evidence that completing homework assignments *causes* higher grades on the final exam? Why or why not?
(b) Is it possible to design an experimental study that could provide more convincing evidence that completing homework assignments causes higher exam grades? If not, why not? If so, how might such an experiment be designed?
(c) Is it possible to marshal stronger observational evidence that completing homework assignments causes higher exam grades? If not, why not? If so, how?

## Summary

- With few exceptions, statistical data analysis describes the outcomes of real social processes and not the processes themselves. It is therefore important to attend to the descriptive accuracy of statistical models and to refrain from reifying them.
- Causal inferences are most certain—if not completely definitive—in randomized experiments, but observational data can also be reasonably marshaled as evidence of causation. Good experimental practice seeks to avoid confounding experimentally manipulated explanatory variables with other variables that can influence the response variable.