Experiments offer more reliable evidence on causation than observational studies, which is not to gainsay the contribution to knowledge from observation. Experiments should be analyzed as experiments, not as observational studies. A simple comparison of rates might be just the right tool, with little value added by "sophisticated" models. This article discusses current models for causation, as applied to experimental and observational data. The intention-to-treat principle and the effect of treatment on the treated will also be discussed. Flaws in per-protocol and treatment-received estimates will be demonstrated.

Keywords: causation, models, experiments, observational studies, intention-to-treat, per-protocol, treatment-received, instrumental variables, effect of treatment on the treated

Statistical Models for Causation: What Inferential Leverage Do They Provide?

DAVID A. FREEDMAN
University of California, Berkeley

The object here is to discuss some current statistical models for causation. Observational studies will be considered, with procedures for handling confounders by stratification or by making statistical adjustments. However, the starting point is experiments. Indeed, one objective of statistical modeling is to create an analogy, perhaps forced, between an observational study and an experiment; hence the focus on experiments. Some of the key philosophical issues behind the models will be mentioned, if not resolved. Section 10 gives explicit mathematical formulations for models, estimators, and estimands.

Statistical models for causation go back to Jerzy Neyman's work on agricultural experiments in the early part of the 20th century. The key paper, Neyman (1923), was in Polish. There was an extended discussion by Scheffé (1956), and an English translation by Dabrowska and Speed (1990). The model was discussed in elementary textbooks in the 1960s. See, for instance, Hodges and Lehmann (1964, section 9.4). The setup is often called "Rubin's model," due in part to Holland (1986, 1988) who cites Rubin (1974). That simply mistakes the history.

AUTHOR'S NOTE: For their many helpful comments, I would like to thank Dick Berk, David Collier, Thad Dunning, Paul Humphreys, Jon McAuliffe, Máire Ní Bhrolcháin, and an anonymous referee. Reprinted with minor changes from Evaluation Review 2006; 30: 691–713.

Neyman's model covers observational studies, in effect by assuming these studies are experiments after suitable controls have been introduced. Neyman does not require random assignment of treatments, assuming instead an urn model, which applies rather neatly to the as-if randomized natural experiments of the social and health sciences. The model is nonparametric, with a finite number of treatment levels. Holland and Rubin (among others) discuss real-valued treatment variables and parametric models, including linear causal relationships. Neyman's model and its extensions will be referred to here as the "Neyman-Holland-Rubin" model.

Turn now to the simplest sort of experiment, which has a treatment group and a control group. There is a large population of subjects (*the study population* or just *the population*). Some subjects are chosen at random and assigned to the treatment group; the remaining subjects are assigned to the control group. According to the Neyman-Holland-Rubin model, each subject has two potential responses. The first is observed if the subject is assigned to treatment. The second is observed if the subject is assigned to control. In the nature of things, both responses cannot be observed. We take the population to be large because random error in estimators will be only a minor detail in what follows. For a discussion of statistical inference in regression models, see, for instance, Freedman (2005b, 2006).

Three parameters are of interest. These parameters describe the study population:

(i) the average response, if all subjects were assigned to treatment,
(ii) the average response, if all subjects were assigned to control,
(iii) the difference between (i) and (ii).

The third parameter, called the *intention-to-treat parameter*, is perhaps the most interesting (it is sometimes called the *average causal effect* or the *average treatment effect*). This parameter represents the average effect obtained by assigning everyone in the study population to treatment, relative to the average effect obtained by assigning everyone to the control regime. Generalizing from the experimental subjects to a broader population—*external validity*—is a major concern, but beyond the scope of this article.

Given the model, it is easy to construct unbiased estimates for the three parameters. The estimates are, respectively,

(i) the average response among subjects assigned to treatment,
(ii) the average response among subjects assigned to control,
(iii) the difference between (i) and (ii).

The third is the *intention-to-treat estimator*. Although subjects are heterogeneous, the intention-to-treat estimator makes no statistical adjustments for heterogeneity. Instead, randomization is relied upon to balance the treatment and control groups, within the limits of random error. That, after all, is the whole point of doing randomized experiments. Adjustments might in the end bring no additional clarity, a topic considered below.

1. IDEAS OF CAUSATION

The idea of causation built into the Neyman-Holland-Rubin model is simple enough. If you assign the subject to treatment, there is one response. If you assign the subject to control, there is another response. Assignment is the cause, response is the effect. The model is well suited to experiments or quasi-experiments, where assignment can (at least in principle) be manipulated by the investigator. The formalism applies less well to non-manipulationist ideas of causation: the moon causes the tides, earthquakes cause property values to go down, time heals all wounds. Time is not manipulable; neither are earthquakes or the moon. Other models may be needed to handle non-manipulationist causation.

Evans (1993) has an interesting review of causal ideas in epidemiology, with many examples. In the legal context, the survey to read is Hart and Honoré (1985). Hume (1748) discusses regular succession and mentions hypothetical counterfactuals, although the latter idea is usually associated with David Lewis. Interestingly enough, Hume (section VII, part II) says that regularities and counterfactuals give equivalent definitions of causation:

> "we may define a cause to be *an object, followed by another, and where all objects similar to the first are followed by objects similar to the second*. Or in other words *where, if the first object had not been, the second never had existed*."

See Lewis (1973) or Mackie (1974) for a discussion of Hume's ideas, and other ideas of causation.

2. CLINICAL TRIALS

In real experiments, some subjects do not follow protocol: for example, a subject assigned to the treatment group may on reflection decline to be treated. That sort of person is said to *cross over* from treatment to control. (The intention-to-treat estimator focuses on assignment, which is under the control of the experimenter, not on the treatment actually received by the subjects.) Crossover is our next topic, but first, a quick look at medical studies, where some of the vocabulary may be unfamiliar. In medicine, a randomized controlled experiment with human subjects is often called a *clinical trial*. The *treatment arm* is the treatment group; the *control arm* is the control group. The experiment runs according to a detailed plan called *the protocol*. Subjects who *follow protocol* accept the regime to which they are assigned.

In a clinical trial to see whether vitamin supplements prevent cancer and heart disease, subjects randomized to the treatment arm will be given vitamin supplements; subjects in the control arm will not be given the supplements. In the treatment arm, subjects who follow protocol take their vitamins; in the control arm, subjects who follow protocol do not sneak off to find vitamins. Empirical findings may be of interest. In too-brief summary, many observational studies suggest that vitamins have a strong protective effect; but the experiments go the other way. See, for instance, Virtamo et al. (2003), U. S. Preventive Services Task Force (2003), Smith and Ebrahim (2005). The conflict between observational studies and experiments is not confined to vitamins; another recent example is provided by hormone replacement therapy for post-menopausal women. Many observational studies suggest that hormone replacement therapy protects against heart disease. The experiments show that, if anything, hormones have adverse effects. See, for instance, Petitti (1998, 2002).

Why is there a conflict between the observational data and the experimental data? People who eat lots of vitamins are different from the rest of us in other ways too. Similarly, women who take hormones differ from women who do not. Some of the differences can perhaps be adjusted out by statistical modeling, but statistical adjustments are incomplete. That is why observational studies, no matter how intricate the statistical analysis, often get it wrong. And that in turn is why clinical trials are needed. For parallel examples in political science, see Arceneaux, Gerber, and Green (2006). On the other hand, most of what we know about causation in the medical and social sciences derives from observational studies. See, for instance, chapter 1 in Freedman (2005b).

3. SINGLE CROSSOVER

We return to the Neyman-Holland-Rubin model for experiments. Intention-to-treat analysis was considered above, in the context of a randomized controlled experiment with one treatment group and one control group. However, subjects may not follow protocol. In a relatively simple case, all subjects assigned to control accept the control regime. However, some subjects assigned to the treatment group decline treatment, following the control regime instead. This is *single crossover*. To avoid potential ambiguity, define "the treatment group" as the group of individuals assigned to treatment, whether they accept treatment or not; the "assigned-to-treatment group" may be clearer, and is occasionally used for emphasis. Similar comments apply to the control group. Problems due to crossover are minimized if the trial can be run "blind," so that subjects do not know which treatment they are receiving. Blinding is often difficult to achieve; here, we will be assuming that the blind is at best imperfect.

The Neyman-Holland-Rubin is easily adapted to handle single crossover, as follows. There are two kinds of subjects, "compliers" and "never-treat." A complier follows protocol. As before, a complier has two potential responses, of which only one can be observed. If we assign the complier to treatment, the response to treatment is observed; if we assign the complier to control, the response to the control regime is observed. By contrast, a never-treat subject is assumed to have only one response. This response is observed whether the subject is assigned to treatment or to control. The idea is that assignment affects response only through the regime chosen by the subject, and a never-treat subject always chooses the control regime.

The intention-to-treat analysis remains valid. That analysis provides an unbiased estimate for the intention-to-treat parameter, which may still be the parameter of greatest policy interest. But there are now other parameters, namely,

(i) the fraction of compliers in the experimental population,
(ii) the average response of compliers to treatment,
(iii) the average response of compliers to the control regime,
(iv) the difference between (ii) and (iii), which is the average effect of treatment on the compliers,
(v) the average response of never-treat subjects to the control regime.

These parameters can all be estimated. To begin with (i), the fraction of compliers in the treatment group estimates the fraction in the whole study population. After all, due to random assignment, the treatment group is a random sample from the population, and the average of a random sample is an unbiased estimator for the average of the population. Similarly for (ii): the average response to treatment of all compliers in the study population is estimated by the average response among compliers in the treatment group, namely, the subjects assigned to treatment who accept treatment. Similarly for (v): the average response of all never-treat subjects to the control regime is estimated by the average response among subjects in the treatment group who decline treatment.

What about (iii)? This parameter is estimated by solving an algebraic equation, as shown in section 10. Basically, the average response in the control group is a mix of the average response for compliers in the control condition (the unknown), and the average response for never-treat subjects (which has already been estimated). Due to random assignment, the mix of subjects in the control group has to be about the same as the mix in the treatment group. That sets up the equation, which can be solved to get an estimate for the average response of compliers, across the whole study population, to the control regime. Finally, (iv) is estimated by subtraction.

4. DOUBLE CROSSOVER

Double crossover means that some subjects assigned to treatment cross over to the control arm, while some subjects assigned to the control arm cross over to treatment. Three estimators are widely used in this setting:

(i) instrumental-variables,
(ii) per-protocol,
(iii) treatment-received.

In section 10, we pinpoint the estimands: what are these different estimators trying to estimate?

The Neyman-Holland-Rubin model can be elaborated to handle double crossover. As a preliminary matter, there are four types of subjects in the model.

Always-treat. If assigned to the treatment group, this type of subject accepts treatment. If assigned to the control group, this type of subject insists on treatment. In other words, these subjects always take treatment.

Complier. If assigned to the treatment group, this type of subject accepts treatment. If assigned to the control group, this type of subject accepts the control regime. In other words, these subjects follow instructions.

Never-treat. If assigned to the treatment group, this type of subject declines treatment and follows the control regime. If assigned to the control group, this type of subject accepts the control regime. In other words, these subjects never take treatment.

Defier. If assigned to the treatment group, this type of subject declines treatment, and follows the control regime. If assigned to the control group, this type of subject insists on treatment. In other words, these subjects do the opposite of what they are told to do.

Partial information about subject type is available from the experimental results. For instance, if you assign a subject to treatment and he takes the treatment, he is either always-treat or a complier; if he declines treatment, he is either never-treat or a defier. On the other hand, if you assign a subject to the control group and she follows the control regime, she is either a complier or never-treat; if she insists on treatment, she is either always-treat or a defier. Finer detail is not determined by the data.

According to the model, if a subject is always-treat, the same response will be observed regardless of assignment: similarly if a subject is never-treat. In other words, it is assumed that subjects respond to the regime they select, rather than to assignment. If a subject is a complier, however, there are two potential responses. The first is to treatment and the second to control. Only one of the two can be observed. Defiers also have two potential responses.

As will be seen in section 10, the instrumental-variables estimator is getting at the differential effect of treatment on compliers. An identifying assumption is needed—that there are no defiers. The estimands of the per-protocol and treatment-received estimators are complex mixtures of structural parameters, with no obvious interpretation. Therefore, the latter two estimators are not recommended.

5. REGRESSION MODELS FOR EXPERIMENTAL DATA

Experimental data are often analyzed by fitting regression models and their ilk. As shown in section 10, randomization does not justify such models. Experimental data should therefore be analyzed first by comparing rates or averages, according to the intention-to-treat principle. Only then should models be deployed. It must be emphasized that statistical adjustments to experimental data often depend substantially on assumptions, not just on randomization.

6. OBSERVATIONAL STUDIES

The Neyman-Holland-Rubin model extends to observational studies, where subjects assign themselves to treatment and control conditions. In a natural experiment, the investigator may be willing to assume that assignment is as-if randomized, and the model can be used without any substantive changes. (In this context, Neyman's urn model just says that the treatment group can be considered as a random sample from the study population, the remaining subjects going into control.) Usually, however, the problem of confounding must be faced: treatment and control groups differ in some obvious and not-so-obvious ways, above and beyond the difference of primary interest.

One way to deal with confounding is by stratification. The investigator may compare the treatment and control groups within relatively homogeneous categories defined by possible confounders. The assumption is that within strata, assignment to treatment or control is as-if randomized: the Neyman-Holland-Rubin model should therefore hold within strata. Take, for example, a study on the health effects of smoking. Smokers may be older than non-smokers, and more of them will be male. That would confound the relationship between smoking and heart disease. The solution would be to compare smokers and non-smokers within categories defined by gender and age, for instance, to compare men age 60–69 who are smokers with men age 60–69 who are non-smokers. Which group has the higher death rate from heart disease? (The answer will come as no surprise.)

Stratification uses up the sample with great rapidity. For this reason among others, it is quite common to handle, or try to handle, confounding variables by means of regression models and the like, including logits and probits when response variables are categorical. Difficulties in the modeling approach are well known. In brief, the models assume without warrant that effects are linear and additive on the chosen scale, with parameters that are constant across subjects and invariant to intervention. To justify the usual formulas for standard errors and significance levels—and the usual claims that regression estimates are unbiased—additional statistical assumptions are needed, for example, disturbance terms are independent across subjects and independent of explanatory variables in the equation. This further complicates the picture. (For probits and logits, similar assumptions would need to be made about latent variables in the model.) The number of successful applications is limited. For discussion, see Berk (2004), Brady and Collier (2004), Freedman (2005ab).

Regression models for causation usually describe relationships among variables. Lieberson (1985) finds little merit in such models: finer-grain analytic methods are needed for causal inference. Abbott (1997, 1998) reaches similar conclusions: statistical variables may be too thin to support detailed social-science investigations, and models will rarely give the equivalent of experimental control. Also see Sobel (2000). Hedström and Swedberg (1998) find that models should grow from our understanding of social mechanisms: regression models generally fail this test. Lieberson and Lynn (2002) suggest that using regression to mimic experimental control is the wrong paradigm for the social sciences.

7. SIMULTANEOUS-EQUATION MODELS

Section 10 considers in detail a successful but narrow application of instrumental variables—estimating the differential effect of treatment on compliers, in an experiment with one treatment group and one control group. Even with two treatment groups and a control group (so there are three groups in total rather than two), the application would be less satisfying, because linearity would no longer be automatic. Assumptions that are stronger and less plausible would be needed. In the alternative, the instrumental-variables estimator can be viewed as estimating a data-dependent mixture of structural parameters, which might (or might not) be of interest.

Of course, instrumental variables are used much more broadly in the social sciences, to deal with reciprocal causation. There are all the difficulties mentioned above in connection with single-equation methods: why this functional form and not another? Why are these variables included and those excluded? Why are the coefficients constant across subjects and invariant to intervention? What about the statistical assumptions on the disturbance terms?

With simultaneous equations, such difficulties remain. Additionally, some variables are taken to be *exogenous* (independent of the disturbance terms) and some *endogenous* (dependent on the disturbance terms). The rationale is seldom clear, because—among other things—there is seldom any very clear description of what the disturbance terms mean, or where they come from. A common formulation, that disturbance terms represent the effect of variables omitted from the equation, simply does not withstand scrutiny, especially when taken in conjunction with assumptions about exogeneity. See, for instance, Pratt and Schlaifer (1984, 1988).

There is, however, an even more fundamental question to consider. Simultaneous equations grow out of work in econometrics, where market-clearing price and quantity are fundamental. It is appealing to formalize equilibrium price and quantity as the joint solution to a pair of supply and demand equations, an idea that goes back to Alfred Marshall. Since his time, the technique has diffused outwards. Thus, in other domains, reciprocal causation between two variables is often represented by a pair of equations in those two variables, with additional "exogenous" variables entered as statistical controls. But why is this two-equation formalism appropriate? What would an equilibrium solution mean, and why would it be relevant to the substantive issue? Such questions are seldom addressed on the pages of social-science journals, and there do not seem to be any very good answers.

8. OTHER LITERATURE

Stone (1993) has a particularly elegant presentation of the Neyman-Holland-Rubin model, with a discussion of its implications for applied work. Freedman (2005ab) explains how the model gets from association to causation via regression (by making assumptions). Instrumental-variables estimators are discussed by Imbens and Angrist (1994); also see Angrist, Imbens, and Rubin (1996). There is a useful survey by Angrist and Krueger (2001). Robins (1999) demonstrates the essential ambiguity in regression-type adjustments for confounders; also see Scharfstein, Rotnitzky and Robins (1999). Heckman (2000) considers the role of potential responses in economics, and the limitations of statistical models for causation. The intention-to-treat principle goes back to Bradford Hill (1961, p. 259); for additional discussion, see Newell (1992).

9. DISCUSSION

Even for experiments, the realism of the Neyman-Holland-Rubin model may be debatable. (The moral is, do the experiment; be wary of model-based interpretations.) Heterogeneity is represented in the model because different subjects have different responses. But each subject's response is assumed to depend only on that subject's assignment: the assignment of other subjects is not material. Certain kinds of social experiments are thus precluded. In a clinical trial conducted by many cooperating physicians, patient compliance might well depend on the personality of the treating physician, so that outcomes depend not only on assignment but also on details not specified in the protocol or represented in the model. If these details matter, the model may be inadequate. The identifying restriction for the instrumental-variables estimator is troublesome: just why are there no defiers?

The discussion here involved one treatment group and one control group, with deterministic responses at the individual level. Several levels of treatment can be considered, and responses could have random components. Real experiments take place over an extended time period; compliance may well depend on a subject's response to the treatment or control regimes, and may not be fully observed. Some subjects will follow neither the treatment regime nor the control regime; others will drop out of the study completely. Such issues create substantial additional difficulties.

For the instrumental-variables estimator with several levels of treatment and random responses, identifying restrictions could be hard to accept. In many circumstances, the instrumental-variables estimator turns out to be estimating some data-dependent average of structural parameters, whose meaning would have to be elucidated. By contrast, for the intention-to-treat estimator, the average response among subjects assigned to a particular level of treatment remains an unbiased estimator for the mean response, if all subjects were assigned to that level of treatment.

Even in a true experiment, only intention-to-treat is an experimental comparison. That comparison is based on assignment, which is under the control of the investigator. Other analyses are observational, because it is the subjects who decide which regime to follow. That is why the intention-to-treat estimator is the most robust. The instrumental-variables estimator has its place, to estimate the differential effect of treatment on compliers. Per-protocol and treatment-received estimators should be used sparingly if at all, because the estimands have no obvious interpretation. Randomization does not justify regression models, or probits, or logits, and the list could be extended. Experimental data should therefore be analyzed first by comparing rates or averages, following the intention-to-treat principle. Such comparisons are justified because the treatment and control groups are balanced, within the limits of chance variation, by randomization. Modeling is potentially useful, as a secondary mode of analysis.

Regression models (broadly understood) are often used to draw causal inferences from observational data, potential confounders being entered as additional explanatory variables alongside the putative causal variable. The number of successful applications, however, is limited. Restrictive assumptions are needed in order to make causal inferences from observational data, and these assumptions can seldom seldom be justified. Instrumental variables may help in some circumstances, but the technique is no panacea.

Simpler analytic techniques and stronger research designs are to be preferred. Sometimes, randomized controlled experiments can be done. In other cases, natural experiments will be available, although data collection can be expensive. Convergent lines of evidence from different kinds of studies add strength to causal inferences. For this reason among others, combining qualitative and quantitative analysis may be helpful. If models are to be used, assumptions need to be discussed, and limitations of technique should be acknowledged.

In an observational study, confounding is a key difficulty. The Neyman-Holland-Rubin model does not really provide any new tools to handle confounders. In that sense, it gives no inferential leverage. On the other hand, the model brings additional clarity to the discussion of foundational issues. What is the estimand for this estimator? What would have to be assumed, in order to justify analyzing those data by that technique? The model is therefore a step forward.

10. TECHNICAL NOTES

Intention-to-treat

The intention-to-treat estimator is the average response in the assigned-to-treatment group, minus the average response in the assigned-to-control group. The estimand is the average response of the study population if all were assigned to treatment, minus the average response if all were assigned to control.

To pursue these ideas, it will be convenient to introduce some mathematical notation. We index subjects by i, running from 1 to N. If subject i is assigned to treatment, the response is T_i; if assigned to control, the response is C_i. If all subjects in the experimental population are assigned to treatment, the average response is

$$\overline{T} = \frac{1}{N}\sum_{i=1}^{N} T_i.$$

If all are assigned to control, the average response is

$$\overline{C} = \frac{1}{N}\sum_{i=1}^{N} C_i.$$

The intention-to-treat parameter is $\overline{T} - \overline{C}$, which measures the average difference that assignment to treatment would make, in the study population. These quantities are all parameters: they are computed at the level of the population, not the data. (Remember, if you see the treatment response T_i, you don't see the control response C_i.)

The estimators are the obvious ones: $\overline{T}$ is estimated by the average response of the subjects assigned to treatment; $\overline{C}$ is estimated by the average response of the subjects assigned to control; and the difference between these two sample averages estimates the intention-to-treat parameter. The estimators are unbiased, even in finite samples, because the average of a random sample is an unbiased estimator for the average of the parent population.

The version of the model described above is deterministic at the level of individuals. If you assign i to treatment, the response is T_i; if you assign i to control, the response is C_i. But two different subjects i and j may well have different responses to treatment ($T_i \neq T_j$); they may also have different responses to the control regime ($C_i \neq C_j$). Moreover, the model is easily generalized so that T_i and C_i are random variables; independence across subjects would be needed to justify the usual variance calculations. Although we do not pursue the idea here, parameters would be defined as follows:

$$\overline{T} = \frac{1}{N}\sum_{i=1}^{N} E(T_i), \quad \overline{C} = \frac{1}{N}\sum_{i=1}^{N} E(C_i).$$

A model for crossover

Let α denote the fraction of always-treat subjects in the study population. This is a parameter. We assume $T_i = C_i$ for always-treat subjects, the idea being that the response is to treatment not assignment, and the subjects in question always seek out treatment. Let A be the average response for always-treat subjects. This is another parameter: the average is taken over the totality of always-treat subjects in the experimental population.

Let β be the fraction of compliers in the study population; β is a parameter. A complier i has two potential responses, T_i if assigned to treatment and C_i if assigned to control. Let T be the average response of the compliers, if all of them are assigned to treatment. Let C be the average response of the compliers, if all of them are assigned to control. These are parameters too.

Let γ be the fraction of never-treat subjects in the study population; γ is a parameter. We assume $T_i = C_i$ for never-treat subjects: the response is to treatment not assignment, and these subjects always seek out the control regime. Let N be the average response for all the never-treat subjects in the study population. This is another parameter.

		Average response	
		when assigned to	
Group	Number	treatment	control
Always-treat	αN	A	A
Compliers	βN	T	C
Never-treat	γN	N	N
Defiers	θN	$\mathfrak{T}$	$\mathfrak{C}$

Let θ be the fraction of defiers in the study population; θ is a parameter. A defier i has two potential responses, T_i if assigned to treatment and C_i if assigned to control. Let $\mathfrak{T}$ be the average response of the defiers, if all of them are assigned to treatment. Let $\mathfrak{C}$ be the average response of the defiers, if all of them are assigned to control. These are parameters too. The notation may seem paradoxical (hence the gothic letters). For instance, defiers assigned to treatment seek out the control condition. Thus, $\mathfrak{T}$ is the average response of the defiers, if all of them are assigned to treatment—and therefore follow the control regime.

The four fractions α, β, γ, θ must add up to 1, i.e., $\alpha + \beta + \gamma + \theta = 1$. There are N subjects in the study population, so the number of always-treat subjects (for example) is αN. The structural parameters are summarized in the table.

Single crossover

Suppose that subjects assigned to control follow protocol; on the other hand, some subjects assigned to treatment accept treatment, while others seek out the control regime. We represent this state of affairs by assuming $\alpha = \theta = 0$, i.e., there are no always-treat subjects and no defiers. This assumption can be tested pretty well from the data—if either α or θ were positive, we should see crossover from control to treatment. As a consequence of the assumption, $\beta + \gamma = 1$.

To start with never-treat subjects, the estimator $\hat{\gamma}$ is the fraction of never-treat subjects in the assigned-to-treatment group. (Never-treat subjects assigned to treatment are easy to spot—they're the ones who decline treatment.) Similarly, the estimator $\hat{\mathsf{N}}$ is the average response among never-treat subjects assigned to treatment. Turn now to compliers. First, $\hat{\beta} = 1 - \hat{\gamma}$. Next, the estimator $\hat{\mathsf{T}}$ is the average response among subjects who are assigned to treatment and stay the course. (As a matter of notation, $\hat{\gamma}$ estimates γ while $\hat{\mathsf{N}}$ estimates N, and so forth.)

What about the response of compliers to the control regime? The control group is a mix of compliers and never-treat subjects. We cannot tell which is which, but we know the proportions are β and γ, sampling error apart. (Due to random assignment, the control group is a random sample from the population; and in the population, the proportions are β and γ, by definition of the parameters.) The average response of the compliers in the control group will be essentially C, just as the average response of the never-treat subjects in the control group will be close to N.

Let Y^C be the average response in the control group (this is a sample quantity). With E for expectation,

$$E(Y^C) = \beta\mathsf{C} + \gamma\mathsf{N}. \tag{1}$$

So

$$\mathsf{C} = \big(E(Y^C) - \gamma\mathsf{N}\big)/\beta. \tag{2}$$

Equation (2) suggests an estimator for C:

$$\hat{\mathsf{C}} = \big(Y^C - \hat{\gamma}\hat{\mathsf{N}}\big)/\hat{\beta}. \tag{3}$$

Here and elsewhere, we are tacitly assuming that β is positive.

To be more explicit about $\hat{\mathsf{T}}$, let Y^T be the average response in the treatment group. Then $Y^T = \hat{\beta}\hat{\mathsf{T}} + \hat{\gamma}\mathsf{N}$, so $\hat{T} = (Y^T - \hat{\gamma}\mathsf{N})/\hat{\beta}$ and

$$\hat{\mathsf{T}} - \hat{\mathsf{C}} = (Y^T - Y^C)/\hat{\beta}. \tag{4}$$

This sort of estimator is discussed by, among others, Bloom (1984), Smith, Kulik, Stromsdorfer (1984), Sommer and Zeger (1991), Imbens and Angrist (1994), Angrist, Imbens, and Rubin (1996). Also see Heckman and Robb (1985). For an example in the context of a clinical trial on mammography, see Freedman, Petitti, and Robins (2004, p. 73).

Per-protocol analysis

The per-protocol estimator is the average response of those in the assigned-to-treatment group who accept treatment, minus the average response of those in the assigned-to-control group who accept the control regime. This is an estimate of

$$\frac{\alpha\mathsf{A} + \beta\mathsf{T}}{\alpha + \beta} - \frac{\beta\mathsf{C} + \gamma\mathsf{N}}{\beta + \gamma}. \tag{5}$$

The relevance of this parameter is not obvious. The per-protocol estimator is increasingly popular, but it should not be used without careful reflection.

The logic behind (5). Let ξ be the average response among those in the treatment group who accept treatment, and let ζ be the average response among those in the control group who accept the control regime. The estimator is $\xi - \zeta$. Take ξ first. The proportions of always-treat, compliers, never-treat, and defiers in the treatment group are essentially α, β, γ, θ respectively. Only the first two types of subjects contribute to ξ. If you divide the numerator and denominator of ξ by the (large) size of the treatment group, the numerator is essentially $\alpha\mathsf{A} + \beta\mathsf{T}$, while the denominator is essentially $\alpha + \beta$. The argument for ζ is similar. We have been assuming that many subjects know which regime they are following; if the trial is blind, and few subjects can break the blind, the per-protocol analysis will be more sensible.

Treatment-received analysis

The treatment-received estimator is the average response of those who follow the treatment regime, minus the average response of those who follow the control regime. Assignment is not explicitly considered in the calculation. The estimand is

$$\frac{\alpha\lambda \mathsf{A} + \beta\lambda \mathsf{T} + \alpha \mathsf{A} + \theta \mathfrak{C}}{\alpha\lambda + \beta\lambda + \alpha + \theta} - \frac{\beta \mathsf{C} + \gamma \mathsf{N} + \gamma\lambda \mathsf{N} + \theta\lambda \mathfrak{T}}{\beta + \gamma + \gamma\lambda + \theta\lambda}, \tag{6}$$

where λ is the size of the treatment group divided by the size of the control group. Again, the relevance of the parameter is obscure. The treatment-received estimator is not recommended. The argument for (6) is like the previous one, although it is a little more complicated. In the control group, always-treat subjects and defiers contribute terms to the analog of ξ. In the treatment group, never-treat subjects and defiers contribute to the analog of ζ.

The methodological implication. As (5) and (6) show, when choosing an estimator, it is important to consider the parameter that is to be estimated.

Instrumental variables

If we allow the existence of defiers, the problem is under-identified: the structural parameters cannot all be estimated. The following *identifying restriction* is therefore often imposed.

Assume that there are no defiers.

We will come to the instrumental-variables estimator shortly, but first consider the problem informally. Take the subjects assigned to treatment. Those who accept treatment are a mix of always-treat subjects and compliers; those who refuse treatment are never-treat subjects. Up to random error, the fraction who accept treatment will be $\alpha + \beta$, with an average response $(\alpha \mathsf{A} + \beta \mathsf{T})/(\alpha + \beta)$; the fraction who refuse treatment will be γ, with an average response N.

Now, take the subjects assigned to control. Those who seek out treatment are the always-treat subjects; those who accept the control regime are a mix of compliers and never-treat subjects. Up to random error, the fraction who seek out treatment will be α, with an average response A; the fraction who accept the control regime will be $\beta + \gamma$, with an average response $(\beta \mathsf{C} + \gamma \mathsf{N})/(\beta + \gamma)$.

The fractions α, β, γ are estimable. The other parameters are also estimable. (Bias and variance are discussed below; there are a few other minor technicalities, for instance, if $\alpha = 0$ then A is not identifiable.) The differential effect of treatment on compliers is $\mathsf{T} - \mathsf{C}$. The *effect of treatment on the treated* is a little ambiguous, but usually seems to mean the differential effect on subjects who would accept treatment if assigned to treatment. These are a mix of always-treat subjects and compliers, in the proportion α to β. Assignment has no effect on always-treat subjects, so the effect of treatment on the treated is $(\mathsf{T}-\mathsf{C}) \times \beta/(\alpha+\beta)$.

The usual instrumental-variables estimator, with assignment as the instrument for treatment, may be viewed as estimating the differential effect of treatment on compliers. This is an important parameter, because compliers are the only group whose behavior is influenced by assignment. (Defiers have been ruled out, by assumption.) The estimator can be written as

$$\frac{Y^T - Y^C}{X^T - X^C}, \tag{7}$$

where Y^T is the average response in the treatment group, and X^T is the fraction in the treatment group who accept treatment; similarly, Y^C is the average response in the control group, and X^C is the fraction in the control group who seek out treatment.

Before deriving the estimator (a tedious algebra exercise), we explain why it works. If we ignore random error—the experiment is a big one, so random error is the least of our problems—the fraction of always-treat subjects in the treatment group is α, and their average response is A. The fraction of compliers in the treatment group is β, and their average response is T. The fraction of never-treat subjects in the treatment group is γ, and their average response is N. Thus, $Y^T \doteq \alpha\mathsf{A}+\beta\mathsf{T}+\gamma\mathsf{N}$ and $X^T \doteq \alpha+\beta$, where $\doteq$ means nearly equal. Similarly, $Y^C \doteq \alpha\mathsf{A}+\beta\mathsf{C}+\gamma\mathsf{N}$ and $X^C \doteq \alpha$. Now, $Y^T - Y^C \doteq \beta(\mathsf{T}-\mathsf{C})$, because $\alpha\mathsf{A}+\gamma\mathsf{N}$ cancels on subtraction. Similarly, $X^T - X^C \doteq \beta$. Because β cancels on division, the ratio is essentially $\mathsf{T}-\mathsf{C}$, as required.

Why Is (7) the instrumental-variables estimator?

The equation to think about is

$$Y_i = a + bX_i + u_i, \tag{8}$$

where Y_i is the observed response and X_i is treatment received. Properties of the error term will not matter here, the object being to derive the estimator rather than determining its statistical properties.

We instrument X_i by assignment Z_i. There are two estimating equations in two unknowns, $\hat{a}$ and $\hat{b}$, namely,

$$\text{ave}(Y) = \hat{a} + \hat{b}\,\text{ave}(X), \tag{9}$$

$$\text{ave}(ZY) = \hat{a}\,\text{ave}(Z) + \hat{b}\,\text{ave}(ZX), \tag{10}$$

where "ave" is taken across all subjects. To get (9), just average (8) over all the subjects, and drop $\text{ave}(u)$; to get (10), multiply across by Z_i before averaging, and drop $\text{ave}(Zu)$ afterwards.

The system is just-identified. Solving equation (9) for $\hat{a}$ tells us that $\hat{a} = \text{ave}(Y) - \hat{b}\,\text{ave}(X)$. So (10) implies

$$\text{ave}(ZY) - \text{ave}(Z)\,\text{ave}(Y) = \hat{b}[\text{ave}(ZX) - \text{ave}(Z)\,\text{ave}(X)]. \tag{11}$$

Thus,

$$\hat{b} = \frac{\text{ave}(ZY) - \text{ave}(Z)\,\text{ave}(Y)}{\text{ave}(ZX) - \text{ave}(Z)\,\text{ave}(X)}. \tag{12}$$

Suppose there are n subjects in the assigned-to-treatment group, with average response Y^T, and the fraction who take treatment is X^T. Similarly, there are m subjects in the the assigned-to-control group; their average response is Y^C, and a fraction X^C of them take treatment. Multiply numerator and denominator of (12) by $n + m$. Now, for instance, $(n + m)\text{ave}(ZY)$ is just the sum of the responses over the assigned-to-treatment group, and so is nY^T; also $(n + m)\text{ave}(Z) = n$. After the multiplication, the numerator in (12) becomes

$$nY^T - \frac{n}{n+m}(nY^T + mY^C) = \frac{nm}{n+m}(Y^T - Y^C)$$

and the denominator becomes

$$nX^T - \frac{n}{n+m}(nX^T + mX^C) = \frac{nm}{n+m}(X^T - X^C)$$

because

$$n - \frac{n^2}{n+m} = \frac{nm}{n+m} = \frac{n}{n+m}m.$$

We get the desired formula because $nm/(n + m)$ cancels on division.

The instrumental-variables estimator (7) is a generalization of (4); see Imbens and Angrist (1994) or Angrist, Imbens, and Rubin (1996). When there is only single crossover, say from treatment to control, then the differential effect of treatment on the treated coincides with the effect on compliers: $\alpha = 0$, so $(\mathsf{T} - \mathsf{C}) \times \beta/(\alpha + \beta) = \mathsf{T} - \mathsf{C}$, which can be estimated by (7) or (4). For general information on instrumental variables, see Freedman (2005b). There is a large econometric literature that discusses the effect of treatment on the treated. One entry point is the April 1995 issue of *Journal of Business & Economic Statistics*; another is Heckman, Tobias, and Vytlacil (2001).

Why is an identifying restriction needed?

In the model for crossover, there are 9 free parameters: α, β, γ, A, T, C, N, $\mathfrak{T}$, $\mathfrak{C}$. In the treatment group, you see the fraction that accept treatment, and their average response, as well as the average response among those who decline: that is 3 pieces of information. You get another 3 pieces of information from the controls. That imposes 6 linear constraints on the 9 parameters. The argument is informal, but sound. Eliminating defiers (by assumption) eliminates 3 parameters, and makes the system just-identified.

Bias and variance

In a formula like (3), the estimators $\hat{\beta}$, $\hat{\gamma}$, and $\hat{\mathsf{N}}$ are unbiased in a strict technical sense: $E(\hat{\beta}) = \beta$, $E(\hat{\gamma}) = \gamma$, and $E(\hat{\mathsf{N}}) = \mathsf{N}$, even in small samples: the last is because, given the mix of never-treat subjects and compliers in each arm of the experiment, the conditional expectation of $\hat{\mathsf{N}}$ equals N. In more detail, the number of never-treat subjects in the treatment arm is random. Given that this number is n, say, the never-treat subjects in the treatment arm constitute a random sample of size n from the totality of never-treat subjects in the study population: this is a consequence of random assignment. The expected value of the average response is therefore N. Hence, the conditional expectation of $\hat{\mathsf{N}}$ is N. Finally, the unconditional expectation must be N too.

In the same way, the numerator and denominator of $\hat{\mathsf{C}}$ are unbiased estimates for the numerator and denominator of (2). However, $\hat{\mathsf{C}}$ itself is biased (*ratio-estimator bias*), due to the division in (3): division is a nonlinear operation. With large samples, the bias will be trivial. There is similar bias in the treatment-received, per-protocol, and instrumental-variables estimators. Thus, for example, $\mathsf{T} - \mathsf{C}$ is estimable, up to a trivial amount of ratio-estimator bias.

Bias and variance for complex estimators like the instrumental-variables estimator can be worked out, to a good approximation, using the *delta method*. In effect, nonlinear statistics are approximated by simpler linear statistics. The error is a quadratic function of the data, which accounts for the bias (van der Vaart, 1998). In the econometric literature, the bias in the instrumental-variables estimator is called *small-sample bias*. The intention-to-treat estimator is unbiased, even with small samples.

Regression models

Suppose the response is quantitative (otherwise, we switch to logits and probits). Let Z_i be the assignment variable: $Z_i = 1$ if subject i is assigned to treatment, and $Z_i = 0$ if i is assigned to control. The response variable is

$Y_i = Z_i T_i + (1 - Z_i)C_i$, which is observable. For instance, if subject i is assigned to treatment, then T_i is observed and $Y_i = 1 \times T_i + (1-1) \times C_i = T_i$: the unobserved C_i drops out of the formula.

Experimental (and non-experimental) data are often analyzed using a regression model of the form

$$Y_i = a + bZ_i + W_i\beta + \epsilon_i, \tag{13}$$

where W_i is a vector of control variables for subject i, while a, b, and β are parameters (if W_i is $1 \times p$, then β is $p \times 1$). The effect of treatment is measured by b. The disturbances ϵ_i would be assumed independent across subjects, with expectation 0 and constant variance. The Z_i and W_i would also need to be independent of the disturbances (this is the exogeneity assumption).

Randomization guarantees that the Z_i are independent of the W_i and ϵ_i. But why are W_i and ϵ_i independent? Why are the ϵ_i independent across subjects, with expectation 0 and constant variance? Replacing the indicator Z_i for assignment by an indicator X_i for treatment received makes the model less secure: why is choice of treatment independent of the disturbance term? With observational data, such questions are even thornier. Of course, there are models with assumptions that are more general and harder to fathom. But that only postpones the reckoning. More-complicated questions can in turn be asked about more-complicated models.

Estimating the average causal effect by regression

If there is only one level of treatment, and control, then $\hat{b}$ in (13) estimates the average causal effect—at least when the sample is large. With more levels of treatment, or smaller samples, regression estimates are subject to bias. Even with one level of treatment, standard errors computed by the usual procedures can be quite misleading.

Mistakes to avoid

Randomization guarantees that the assignment variable Z is statistically independent of the covariates W. That, however, does not translate to exact orthogonality on the sample data. If it did, nobody would bother adjusting, because adjustment would make no difference to estimated treatment effects. According to the Neyman-Holland Rubin model, the multiple regression estimator is conditionally biased. Indeed, given the assignment variable, the response is deterministic. Unconditionally, with suitable regularity conditions—and only two possible values for the assignment variable—the bias goes to 0 as sample size increases. Asymptotic variance may be decreased by modeling, or increased. The usual formula for asymptotic variance may be severely biased. For details, see Freedman (2006).

Summary

With models, it is easy to lose track of three essential points: (i) results depend on assumptions, (ii) changing the assumptions in apparently innocuous ways can lead to drastic changes in conclusions, and (iii) familiarity with a model's name is no guarantee of the model's truth. Under the circumstances, it may be the assumptions behind the model that provide the leverage, not the data fed into the model. This is a danger with experiments, and even more so with observational studies.

REFERENCES

Abbott, A. 1997. Of time and space: The contemporary relevance of the Chicago school. *Social Forces* 75: 1149–82.

Abbott, A. 1998. The causal devolution. *Sociological Methods and Research* 27: 148–81.

Angrist, J. D and G. W. Imbens. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62: 467–75.

Angrist, J. D, G. W. Imbens, and D. B. Rubin 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91: 444–72.

Angrist, J. D. and A. B. Krueger. 2001. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Persepctives* 15: 69–85.

Arceneaux, K., A. S. Gerber, and D. P. Green. 2006. Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Political Analysis* 14: 37–62.

Berk, R. A. 2004. *Regression Analysis: A Constructive Critique*. Thousand Oaks, CA: Sage Publications.

Bloom, H. S. 1984. Accounting for no-shows in experimental evaluation designs. *Evaluation Review* 8: 225–46.

Brady, H. E. and D. Collier, eds. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, Maryland: Rowman & Littlefield Publishers, Inc.

Dabrowska, D. and T. P. Speed. 1990. On the application of probability theory to agricultural experiments. Essay on principles. English translation of Neyman (1923). *Statistical Science* 5: 463–80 (with discussion).

Evans, A. S. 1993. *Causation and Disease: A Chronological Journey*. New York: Plenum

Freedman, D. A. 2005a. Linear statistical models for causation: A critical review. In *Encyclopedia of Statistics in Behavioral Science*, ed. by B. S. Everitt and D. C. Howell. Chichester, U.K.: John Wiley & Sons.

Freedman, D. A. 2005b. *Statistical Models: Theory and Practice*. New York: Cambridge University Press.

Freedman, D. A. 2006. On regression adjustments to experimental data. Technical report, Statistics Department, U.C. Berkeley.
http://www.stat.berkeley.edu/users/census/neyreg.pdf

Freedman, D. A., D. B. Petitti, and J. M. Robins. 2004. On the efficacy of screening for breast cancer. *International Journal of Epidemiology*, 33: 43–73 (with discussion). Correspondence, pp. 1404–6.

Hart, H. L. A. and A. M. Honoré. 1985. *Causation in the Law*. 2nd ed. Oxford: Oxford University Press.

Heckman, J. J. 2000. Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics* 115: 45–97.

Heckman, J. and R. Robb (1985). Alternative methods for estimating the impact of interventions. In *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman and B. Singer. New York: Cambridge University Press, pp. 156–245.

Heckman, J., J. L. Tobias, and E. Vytlacil. 2001. Four parameters of interest in the evaluation of social programs. *Southern Economic Journal* 68: 210–223.

Hedström, P. and Swedberg, R. eds. 1998. *Social Mechanisms*. Cambridge: Cambridge University Press.

Hill, A. B. 1961. *Principles of Medical Statistics*. 7th ed. London: The Lancet.

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 8: 945–70 (with discussion).

Hodges, J. L., Jr. and Lehmann, E. 1964. *Basic Concepts of Probability and Statistics*. Holden-Day, San Francisco.

Holland, P. W. 1988. Causal inference, path analysis, and recursive structural equation models. In *Sociological Methodology 1988*, Washington, D. C.: American Sociological Association, chapter 13, ed. by C. Clogg.

Hume, D. 1748. *Philosophical Essays Concerning Human Understanding*. London: A. Millar. Retitled *An Enquiry Concerning Human Understanding* in 1758. Widely reprinted, e.g., by Oxford University Press, 2005, ed. by T. L. Beauchamp.

Imbens, G. and J. Angrist. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62: 467–75.

Lewis, D. 1973. Causation. *Journal of Philosophy* 70: 556–67.

Lieberson, S. 1985. *Making it Count*. Berkeley: University of California Press.

Lieberson, S. and F. B. Lynn. 2002. Barking up the wrong branch: Alternatives to the current model of sociological science. *Annual Review of Sociology* 28: 1–19.

Mackie, J. 1974. *The Cement of the Universe*. Oxford: Oxford University Press. Corrected edition reissued in 2002.

Newell, D. J. 1992. Intention-to-treat analysis: Implications for quantitative and qualitative research. *International Journal of Epidemiology* 21: 837–41.

Neyman, J. 1923. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* 10: 1–51, in Polish.

Petitti, D. B. 1998. Hormone replacement therapy and heart disease prevention: Experimentation trumps observation. *Journal of the American Medical Association* 280: 650–52.

Petitti, D. B. 2002. Hormone replacement therapy for prevention. *Journal of the American Medical Association* 288: 99–101.

Pratt, J. W. and R. Schlaifer. 1984. On the nature and discovery of structure. *Journal of the American Statistical Association* 79: 9–33 (with discussion).

Pratt, J.W. and R. Schlaifer. 1988. On the interpretation and observation of laws. *Journal of Econometrics* 39: 23–52.

Robins, J. M. 1999. Association, causation, and marginal structural models. *Synthese* 121: 151–79.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66: 688–701.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins. 1999. Adjusting for non-ignorable drop-out using semiparametric non-response models. *Journal of the American Statistical Association* 94: 1096-1146.

Scheffé, H. 1956. Alternative models for the analysis of variance. *Annals of Mathematical Statistics* 27: 251–71.

Smith, D. A., J. Kulik, and E. W. Stromsdorfer. 1984. The economic impact of the downriver community conference economic readjustment activity program: Choosing between retraining and job search placement strategies. In *Displaced Workers: Implications for Educational and Training Institutions*, ed. by K Hollenbeck, F. Pratzner, and H. Rosen. Columbus, Ohio: National Center for Research in Vocatoinal Education.

Smith, G. D. and S. Ebrahim. 2005. Folate supplementation and cardiovascular disease. *Lancet* 366: 1679–81.

Sobel, M. E. 2000. Causal inference in the social sciences. *Journal of the American Statistical Association* 95: 647–51.

Sommer, A. and S. L. Zeger. 1991. On estimating efficacy from clinical trials. *Statistics in Medicine* 10: 45–52.

Stone, R. 1993. The assumptions on which causal inferences rest. *Journal of the Royal Statistical Society* Series B 55: 455–66.

U. S. Preventive Services Task Force (2003). Routine vitamin supplementation to prevent cancer and cardiovascular disease: Recommendations and rationale *Annals of Internal Medicine* 139: 51–55

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

Virtamo, J., P. Pietinen, J. K. Huttunen, et al. 2003. Incidence of cancer and mortality following alpha-tocopherol and beta-carotene supplementation: A postintervention follow-up. *Journal of the American Medical Association* 290: 476–85.

David A. Freedman is professor of statistics and mathematics at the University of California, Berkeley. His research interests are in the foundation of statistics, modeling, and policy analysis. He has published numerous articles and several books, including a standard introductory text with Robert Pisani and Roger Purves.