Statistical Models for Causation

D. A. Freedman January 2005

Abstract

We review the basis for inferring causation by statistical modeling. Parameters should be stable under interventions, and so should error distributions. There are also statistical conditions on the errors. Stability is difficult to establish a priori, and the statistical conditions are equally problematic. Therefore, causal relationships are seldom to be inferred from a data set by running statistical algorithms, unless there is substantial prior knowledge about the mechanisms that generated the data. We begin with linear models (regression analysis) and then turn to graphical models, which may in principle be non-linear.

Introduction

Regression models are often used to infer causation from association. For instance, Yule (1899) showed—or tried to show—that welfare was a cause of poverty. Path models and structural equation models are later refinements of the technique. Besides Yule, examples to be discussed here include Blau and Duncan (1967) on stratification, as well as Gibson (1988) on the causes of McCarthyism. Strong assumptions are required to infer causation from association by modeling. The assumptions are of two kinds: (i) causal, and (ii) statistical. These assumptions will be formulated explicitly, with the help of response schedules in hypothetical experiments. In particular, parameters and error distributions must be stable under intervention. That will be hard to demonstrate in observational settings. Statistical conditions (like independence) are also problematic, and latent variables create further complexities. Inferring causation by regression analysis will be the primary topic. Graphical models will be considered briefly. The issues are not simple, so examining them from several perspectives may be helpful. The article ends with a review of the literature and a summary.

Regression models in social science

Legendre (1805) and Gauss (1809) developed regression to fit data on orbits of astronomical objects. The relevant variables were known from Newtonian mechanics, and so were the functional forms of the equations connecting them. Measurement could be done with great precision, and much was known about the nature of errors in the measurements and in the equations. Furthermore, there was ample opportunity for comparing predictions to reality. By the turn of the century, investigators were using regression on social science data where such conditions did not hold, even to a rough approximation. Yule (1899) was a pioneer. At the time, paupers in England were supported either inside grim Victorian institutions called "poor-houses" or outside, according to decisions made by local authorities. Did policy choices affect the number of paupers? To study this question, Yule proposed a regression equation,

$$\Delta\text{Paup} = a + b \times \Delta\text{Out} + c \times \Delta\text{Old} + d \times \Delta\text{Pop} + \text{error}. \tag{1}$$

In this equation,

Δ is percentage change over time,
Paup is the number of Paupers
Out is the out-relief ratio N/D,
$N =$ number on welfare outside the poor-house,
$D =$ number inside,
Old is the population over 65,
Pop is the population.

Data are from the English Censuses of 1871, 1881, 1891. There are two Δ's, one each for 1871–81 and 1881–91.

Relief policy was determined separately in each "union," a small geographical area like a parish. At the time, there were about 600 unions, and Yule divides them into four kinds: rural, mixed, urban, metropolitan. There are $4 \times 2 = 8$ equations, one for each type of union and time period. Yule fits each equation to data by least squares. That is, he determines a, b, c, and d by minimizing the sum of squared errors,

$$\sum \big(\Delta\text{Paup} - a - b \times \Delta\text{Out} - c \times \Delta\text{Old} - d \times \Delta\text{Pop}\big)^2.$$

The sum is taken over all unions of a given type in a given time period—which assumes, in essence, that coefficients are constant within each combination of geography and time. For example, consider the metropolitan unions. Fitting the equation to the data for 1871–81, Yule gets

$$\Delta\text{Paup} = 13.19 + 0.755\Delta\text{Out} - 0.022\Delta\text{Old} - 0.322\Delta\text{Pop} + \text{error}. \tag{2}$$

For 1881–91, his equation is

$$\Delta\text{Paup} = 1.36 + 0.324\Delta\text{Out} + 1.37\Delta\text{Old} - 0.369\Delta\text{Pop} + \text{error}. \tag{3}$$

The coefficient of ΔOut being relatively large and positive, Yule concludes that outrelief causes poverty.

Table 1 has the ratio of 1881 data to 1871 data for Pauperism, Outrelief ratio, Proportion of Old, and Population. If we subtract 100 from each entry, column 1 gives ΔPaup in equation (2). Columns 2, 3, 4 give the other variables. For Kensington (the first union in the table),

$$\Delta\text{Out} = 5 - 100 = -95, \quad \Delta\text{Old} = 104 - 100 = 4, \quad \Delta\text{Pop} = 136 - 100 = 36.$$

The predicted value for ΔPaup from (2) is therefore

$$13.19 + 0.755 \times (-95) - 0.022 \times 4 - 0.322 \times 36 = -70.$$

The actual value for ΔPaup is -73, so the error is -3. Other lines in the table are handled in a similar way. As noted above, coefficients were chosen to minimize the sum of the squared errors.

Quetelet (1835) wanted to uncover "social physics"—the laws of human behavior—by using statistical technique:

> "In giving my work the title of Social Physics, I have had no other aim than to collect, in a uniform order, the phenomena affecting man, nearly as physical science brings together the phenomena appertaining to the material world. . . . in a given state of society, resting under the

influence of certain causes, regular effects are produced, which oscillate, as it were, around a fixed mean point, without undergoing any sensible alterations. . . .

"This study. . .has too many attractions—it is connected on too many sides with every branch of science, and all the most interesting questions in philosophy—to be long without zealous observers, who will endeavour to carry it further and further, and bring it more and more to the appearance of a science."

Table 1. Pauperism, Outrelief ratio, Proportion of Old, Population. Ratio of 1881 data to 1871 data, times 100. Metropolitan Unions, England. Yule (1899, Table XIX).

	Paup	Out	Old	Pop
Kensington	27	5	104	136
Paddington	47	12	115	111
Fulham	31	21	85	174
Chelsea	64	21	81	124
St. George's	46	18	113	96
Westminster	52	27	105	91
Marylebone	81	36	100	97
St. John, Hampstead	61	39	103	141
St. Pancras	61	35	101	107
Islington	59	35	101	132
Hackney	33	22	91	150
St. Giles'	76	30	103	85
Strand	64	27	97	81
Holborn	79	33	95	93
City	79	64	113	68
Shoreditch	52	21	108	100
Bethnal Green	46	19	102	106
Whitechapel	35	6	93	93
St. George's East	37	6	98	98
Stepney	34	10	87	101
Mile End	43	15	102	113
Poplar	37	20	102	135
St. Saviour's	52	22	100	111
St. Olave's	57	32	102	110
Lambeth	57	38	99	122
Wandsworth	23	18	91	168
Camberwell	30	14	83	168
Greenwich	55	37	94	131
Lewisham	41	24	100	142
Woolwich	76	20	119	110
Croydon	38	29	101	142
West Ham	38	49	86	203

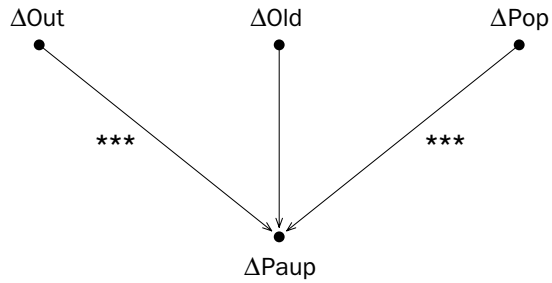
Yule is using regression to infer the social physics of poverty. But this is not so easily to be done. Confounding is one issue. According to Pigou (a leading welfare economist of Yule's era), parishes

with more efficient administrations were building poor-houses and reducing poverty. Efficiency of administration is then a confounder, influencing both the presumed cause and its effect. Economics may be another confounder. Yule occasionally tries to control for this, using the rate of population change as a proxy for economic growth. Generally, however, he pays little attention to economics. The explanation: "A good deal of time and labour was spent in making trial of this idea, but the results proved unsatisfactory, and finally the measure was abandoned altogether. [p. 253]"

The form of Yule's equation is somewhat arbitrary, and the coefficients are not consistent over time and space. This is not necessarily fatal. However, unless the coefficients have some existence apart from the data, how can they predict the results of interventions that would change the data? The distinction between parameters and estimates runs throughout statistical theory; the discussion of response schedules, below, may sharpen the point.

There are other interpretive problems. At best, Yule has established association. Conditional on the covariates, there is a positive association between ΔPaup and ΔOut. Is this association causal? If so, which way do the causal arrows point? For instance, a parish may choose not to build poor-houses in response to a short-term increase in the number of paupers. Then pauperism is the cause and out-relief the effect. Likewise, the number of paupers in one area may well be affected by relief policy in neighboring areas. Such issues are not resolved by the data analysis. Instead, answers are assumed a priori. Although he was busily parceling out changes in pauperism—so much is due to changes in out-relief ratios, so much to changes in other variables, so much to random effects—Yule was aware of the difficulties. With one deft footnote (number 25), he withdrew all causal claims: "Strictly speaking, for 'due to' read 'associated with.' "

Figure 1. Yule's Model. Metropolitan Unions, 1871–81.



Yule's approach is strikingly modern, except there is no causal diagram with stars indicating statistical significance. Figure 1 brings him up to date. The arrow from ΔOut to ΔPaup indicates that ΔOut is included in the regression equation that explains ΔPaup. Three asterisks mark a high degree of statistical significance. The idea is that a statistically significant coefficient must differ from zero. Thus, ΔOut has a causal influence on ΔPaup. By contrast, a coefficient that lacks statistical significance is thought to be zero. If so, ΔOld would not exert a causal influence on ΔPaup.

The reasoning is seldom made explicit, and difficulties are frequently overlooked. Statistical assumptions are needed to determine significance from the data. Even if significance can be determined and the null hypothesis rejected or accepted, there is a deeper problem. To make causal inferences, it must be assumed that equations are stable under proposed interventions. Verifying such assumptions—without making the interventions—is problematic. On the other hand, if the

coefficients and error terms change when variables are manipulated, the equation has only a limited utility for predicting the results of interventions.

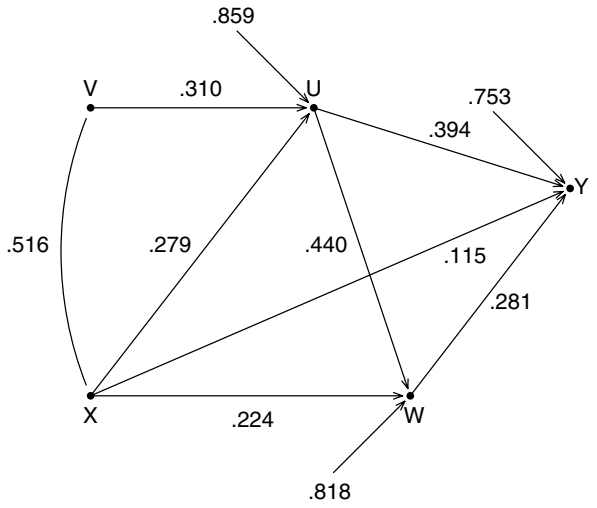
Social stratification

Blau and Duncan (1967) are thinking about the stratification process in the US. According to Marxists of the time, the US is a highly stratified society. Status is determined by family background, and transmitted through the school system. Blau and Duncan present cross-tabs (in their Chapter 2) to show that the system is far from deterministic, although family background variables do influence status. The US has a permeable social structure, with many opportunities to succeed or fail. Blau and Duncan go on to develop the path model shown in Figure 2, in order to answer questions like these:

> "how and to what degree do the circumstances of birth condition subsequent status? how does status attained (whether by ascription or achievement) at one stage of the life cycle affect the prospects for a subsequent stage?"

The five variables in the diagram are father's education and occupation, son's education, son's first job, and son's occupation. Data come from a special supplement to the March 1962 Current Population Survey. The respondents are the sons (age 20–64), who answer questions about current jobs, first jobs, and parents. There are 20,000 respondents. Education is measured on a scale from 0 to 8, where 0 means no schooling, 1 means 1–4 years of schooling, and so forth; 8 means some post-graduate education. Occupation is measured on Duncan's prestige scale from 0 to 96. The scale takes into account income, education, and raters' opinions of job prestige. Hucksters are at the bottom of the ladder, with clergy in the middle, and judges at the top.

Figure 2. Path model. Stratification, US, 1962.



V is DADS' ED
X is DADS' OCC
U is SONS' ED
W is SONS' 1st JOB
Y is SONS' OCC

How is Figure 2 to be read? The diagram unpacks to three regression equations:

$$U = aV + bX + \delta, \tag{4}$$

$$W = cU + dX + \epsilon, \tag{5}$$

$$Y = eU_i + fX + gW + \eta. \tag{6}$$

Parameters are estimated by least squares. Before regressions are run, variables are standardized to have mean 0 and variance 1. That is why no intercepts are needed, and why estimates can be computed from the correlations in Table 2.

Table 2. Correlation matrix for variables in Blau and Duncan's path model.

		Y Sons'occ	W Sons'1^{st}job	U Sons'ed	X Dads'occ	V Dads'ed
Y	Sons'occ	1.000	.541	.596	.405	.322
W	Sons'1^{st}job	.541	1.000	.538	.417	.332
U	Sons'ed	.596	.538	1.000	.438	.453
X	Dads'occ	.405	.417	.438	1.000	.516
V	Dads'ed	.322	.332	.453	.516	1.000

In Figure 2, the arrow from V to U indicates a causal link, and V is entered on the right hand side in the regression equation (4) that explains U. The path coefficient .310 next to the arrow is the estimated coefficient $\hat{a}$ of V. The number .859 on the "free arrow" that points into U is the estimated standard deviation of the error term δ in (4). The other arrows are interpreted in a similar way. The curved line joining V and X indicates association rather than causation: V and X influence each other or are influenced by some common causes, not further analyzed in the diagram. The number on the curved line is just the correlation between V and X (Table 2). There are three equations because three variables in the diagram (U, W, Y) have arrows pointing into them.

The large standard deviations in Figure 2 show the permeability of the social structure. (Since variables are standardized, it is a little theorem that the standard deviations cannot exceed 1.) Even if father's education and occupation are given, as well as respondent's education and first job, the variation in status of current job is still large. As social physics, however, the diagram leaves something to be desired. Why linearity? Why are the coefficients the same for everybody? What about variables like intelligence or motivation? And where are the mothers?

The choice of variables and arrows is up to the analyst, as are the directions in which the arrows point. Of course, some choices may fit the data less well, and some may be illogical. If the graph is "complete"—every pair of nodes joined by an arrow—the direction of the arrows is not constrained by the data (Freedman 1997 pp. 138, 142). Ordering the variables in time may reduce the number of options.

If we are trying to find laws of nature that are stable under intervention, standardizing may be a bad idea, because estimated parameters would depend on irrelevant details of the study design (see below). Generally, the intervention idea gets muddier with standardization. Are means and standard deviations held constant even though individual values are manipulated? On the other

hand, standardizing might be sensible if units are meaningful only in comparative terms (e.g., prestige points). Standardizing may also be helpful if the meaning of units changes over time (e.g., years of education), while correlations are stable. With descriptive statistics for one data set, it is really a matter of taste: do you like pounds, kilograms, or standard units? Moreover, all variables are on the same scale after standardization, which makes it easier to compare regression coefficients.

Hooke's law

According to Hooke's law, stretch is proportional to weight. If weight x is hung on a spring, the length of the spring is $a + bx + \epsilon$, provided x is not too large. (Near the elastic limit of the spring, the physics will be more complicated.) In this equation, a and b are physical constants that depend on the spring not the weights. The parameter a is the length of the spring with no load. The parameter b is the length added to the spring by each additional unit of weight. The ϵ is random measurement error, with the usual assumptions. Experimental verification is a classroom staple.

If we were to standardize, the crucial slope parameter would depend on the weights and the accuracy of the measurements. Let v be the variance of the weights used in the experiment, let σ^2 be the variance of ϵ, and let s^2 be the mean square of the deviations from the fitted regression line. The standardized regression coefficient is

$$\sqrt{\frac{\hat{b}^2 v}{\hat{b}^2 v + s^2}} \approx \sqrt{\frac{b^2 v}{b^2 v + \sigma^2}}, \tag{7}$$

as can be verified by examining the sample covariance matrix. Therefore, the standardized coefficient depends on v and σ^2, which are features of our measurement procedure not the spring.

Hooke's law is an example where regression is a very useful tool. But the parameter to estimate is b, the unstandardized regression coefficient. It is the unstandardized coefficient that says how the spring will respond when the load is manipulated. If a regression coefficient is stable under interventions, standardizing it is probably not a good idea, because stability gets lost in the shuffle. That is what (7) shows. Also see Achen (1977), Blalock (1989, p. 451).

Political repression during the McCarthy era

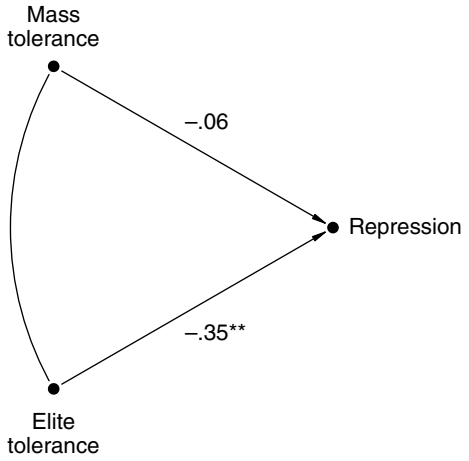
Gibson (1988) tries to determine the causes of McCarthyism in the US. Was repression due to the masses or the elites? He argues that elite intolerance is the root cause, the chief piece of evidence being a path model (Figure 3, redrawn from the paper). The dependent variable is a measure of repressive legislation in each state. The independent variables are mean tolerance scores for each state, derived from the Stouffer survey of masses and elites. The "masses" are just respondents in a probability sample of the population. "Elites" include school board presidents, commanders of the American Legion, bar association presidents, labor union leaders. Data on masses were available for 36 states; on elites, for 26 states. The two straight arrows in Figure 3 represent causal links: mass and elite tolerance affect repression. The curved double-headed arrow in Figure 3 represents an association between mass and elite tolerance scores. Each one can influence the other, or both can have some common cause. The association is not analyzed in the diagram.

Gibson computes correlations from the available data, then estimates a standardized regression equation,

$$\text{Repression} = \beta_1 \text{Mass tolerance} + \beta_2 \text{Elite tolerance} + \delta.$$

He says, "Generally, it seems that elites, not masses, were responsible for the repression of the era. . . . The beta for mass opinion is −.06; for elite opinion, it is −.35 (significant beyond .01)."

Figure 3. Path model. The causes of McCarthyism.



The paper asks an interesting question, and the data analysis has some charm too. However, as social physics, the path model is not convincing. What hypothetical intervention is contemplated? If none, how are regressions going to uncover causal relationships? Why are relationships among the variables supposed to be linear? Signs apart, for example, why does a unit increase in tolerance have the same effect on repression as a unit decrease? Are there other variables in the system? Why are the states statistically independent? Such questions are not addressed in the paper.

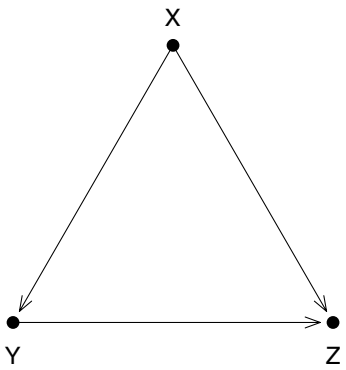
McCarthy became a force in national politics around 1950. The turning point came in 1954, with public humiliation in the Army-McCarthy hearings. Censure by the Senate followed in 1957. Gibson scores repressive legislation over the period 1945–65, long before McCarthy mattered, and long after. The Stouffer survey was done in 1954, when the McCarthy era was ending. The timetable is puzzling.

Even if such issues are set aside, and we grant the statistical model, the difference in path coefficients fails to achieve significance. Gibson finds that $\hat{\beta}_1$ is significant and $\hat{\beta}_2$ is insignificant, but that does not impose much of a constraint on $\hat{\beta}_1 - \hat{\beta}_2$. (The standard error for this difference can be computed from data generously provided in the paper.) Since $\beta_1 = \beta_2$ is a viable hypothesis, the data are not strong enough to distinguish masses from elites.

Inferring causation by regression

Path models are often thought to be rigorous statistical engines for inferring causation from association. Statistical techniques can be rigorous, given their assumptions. But the assumptions are usually imposed on the data by the analyst. This is not a rigorous process, and it is rarely made explicit. The assumptions have a causal component as well as a statistical component. It will be easier to proceed in terms of a specific example. In Figure 4, a hypothesized causal relationship between Y and Z is confounded by X. The free arrows leading into Y and Z are omitted.

Figure 4. Path model. The relationship between Y and Z is confounded by X. Free arrows leading into Y and Z are not shown.



The diagram describes two hypothetical experiments, and an observational study where the data are collected. The two experiments help to define the assumptions. Furthermore, the usual statistical analysis can be understood as an effort to determine what would happen under those assumptions *if* the experiments were done. Other interpretations of the analysis are not easily to be found. The experiments will now be described.

(i) *First hypothetical experiment*. Treatment is applied to a subject, at level x. A response Y is observed, corresponding to the level of treatment. There are two parameters, a and b, that describe the response. With no treatment, the response level for each subject will be a, up to random error. All subjects are assumed to have the same value for a. Each additional unit of treatment adds b to the response. Again, b is the same for all subjects, at all levels of x, by assumption. Thus, if treatment is applied at level x, the response Y is assumed to be

$$a + bx + \text{random error}. \tag{8}$$

For Hooke's law, x is weight and Y is length of a spring under load x. For evaluation of job training programs, x might be hours spent in training and Y might be income during a followup period.

(ii) *Second hypothetical experiment.* In the second experiment, there are two treatments and a response variable Z. There are two treatments because there are two arrows leading into Z; the treatments are labeled X and Y (Figure 4). Both treatments may be applied to a subject. There are three parameters, c, d, and e. With no treatment, the response level for each subject is taken to be c, up to random error. Each additional unit of treatment #1 adds d to the response. Likewise, each additional unit of treatment #2 adds e to the response. The constancy of parameters across subjects and levels of treatment is an assumption. If the treatments are applied at levels x and y, the response Z is assumed to be

$$c + dx + ey + \text{random error}. \tag{9}$$

Three parameters are needed because it takes three parameters to specify the linear relationship (9), namely, an intercept and two slopes. Random errors in (8) and (9) are assumed to be independent from subject to subject, with a distribution that is constant across subjects; expectations are zero and variances are finite. The errors in (9) are assumed to be independent of the errors in (8).

The observational study. When using the path model in Figure 4 to analyze data from an observational study, we assume that levels for the variable X are independent of the random errors in the two hypothetical experiments ("exogeneity"). In effect, we pretend that Nature randomized subjects to levels of X for us, which obviates the need for experimental manipulation. The exogeneity of X has a graphical representation: arrows come out of X, but no arrows lead into X.

We take the descriptions of the two experiments, including the assumptions about the response schedules and the random errors, as background information. In particular, we take it that Nature generates Y as if by substituting X into (8). Nature proceeds to generate Z as if by substituting X and Y—the same Y that has just been generated from X—into (9). In short, (8) and (9) are assumed to be the causal mechanisms that generate the observational data, namely, X, Y, and Z for each subject. The system is "recursive," in the sense that output from (8) is used as input to (9) but there is no feedback from (9) to (8).

Under these assumptions, the parameters a, b can be estimated by regression of Y on X. Likewise, c, d, e can be estimated by regression of Z on X and Y. Moreover, these regression estimates have legitimate causal interpretations. This is because causation is built into the background assumptions, via the response schedules (8) and (9). If causation were not assumed, causation would not be demonstrated by running the regressions.

One point of running the regressions is usually to separate out direct and indirect effects of X on Z. The direct effect is d in (9). If X is increased by one unit with Y held fast, then Z is expected to go up by d units. But this is shorthand for the assumed mechanism in the second experiment. Without the thought experiments described by (8) and (9), how can Y be held constant when X is manipulated? At a more basic level, how would manipulation get into the picture?

Another path-analytic objective is to determine the effect e of Y on Z. If Y is increased by one unit with X held fast, then Z is expected to go up by e units. (If $e = 0$, then manipulating Y would not affect Z, and Y does not cause Z after all.) Again, the interpretation depends on the thought experiments. Otherwise, how could Y be manipulated and X held fast?

To state the model more carefully, we would index the subjects by a subscript i in the range from 1 to n, the number of subjects. In this notation, X_i is the value of X for subject i. Similarly, Y_i and Z_i are the values of Y and Z for subject i. The level of treatment #1 is denoted by x, and $Y_{i,x}$ is the response for variable Y if treatment at level x is applied to subject i. Similarly, $Z_{i,x,y}$ is the response for variable Z if treatment #1 at level x and treatment #2 at level y are applied to subject i. The response schedules are to be interpreted causally:

- $Y_{i,x}$ is what Y_i would be if X_i were set to x by intervention.
- $Z_{i,x,y}$ is what Z_i would be if X_i were set to x and Y_i were set to y by intervention.

Counterfactual statements are even licensed about the past: $Y_{i,x}$ is what Y_i would have been, if X_i had been set to x. Similar comments apply to $Z_{i,x,y}$.

The diagram unpacks into two equations, which are more precise versions of (8) and (9), with a subscript i for subjects. Greek letters are used for the random error terms.

$$Y_{i,x} = a + bx + \delta_i. \tag{10}$$

$$Z_{i,x,y} = c + dx + ey + \epsilon_i. \tag{11}$$

The parameters a, b, c, d, e and the error terms δ_i, ϵ_i are not observed. The parameters are assumed to be the same for all subjects.

Additional assumptions, which define the statistical component of the model, are imposed on the error terms:

(i) δ_i and ϵ_i are independent of each other within each subject i.
(ii) δ_i and ϵ_i are independent across subjects.
(iii) The distribution of δ_i is constant across subjects; so is the distribution of ϵ_i. (However, δ_i and ϵ_i need not have the same distribution.)
(iv) δ_i and ϵ_i have expectation zero and finite variance.
(v) The δ's and ϵ's are independent of the X's.

The last is "exogeneity."

According to the model, Nature determines the response Y_i for subject i by substituting X_i into (10):

$$Y_i = Y_{i,X_i} = a + bX_i + \delta_i.$$

Here, X_i is the value of X for subject i, chosen for us by Nature, as if by randomization. The rest of the response schedule—the $Y_{i,x}$ for other x—is not observed, and therefore stays in the realm of counterfactual hypotheticals. After all, even in an experiment, subject i would be assigned to one level of treatment, foreclosing the possibility of observing the response at other levels.

Similarly, we observe $Z_{i,x,y}$ only for $x = X_i$ and $y = Y_i$. The response for subject i is determined by Nature, as if by substituting X_i and Y_i into (11):

$$Z_i = Z_{i,X_i,Y_i} = c + dX_i + eY_i + \epsilon_i.$$

The rest of the response schedule, $Z_{i,x,y}$ for other x and y, remains unobserved. Economists call the unobserved $Y_{i,x}$ and $Z_{i,x,y}$ "potential outcomes." The model specifies unobservable response schedules, not just regression equations. Notice too that a subject's responses are determined by levels of treatment for that subject only. Treatments applied to subject j are not relevant to subject i. The response schedules (10) and (11) represent the causal assumptions behind the path diagram.

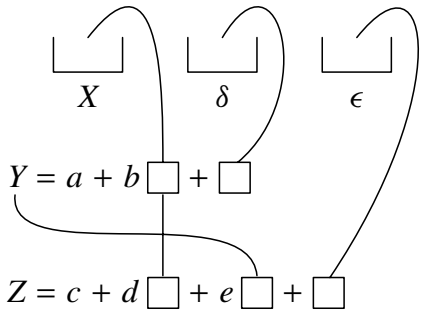
The conditional expectation of Y given $X = x$ is the average of Y for subjects with $X = x$. The formalism connects two very different ideas of conditional expectation: (i) finding subjects with $X = x$, versus (ii) an intervention that sets X to x. The first is something you can actually do with observational data. The second would require manipulation. The model is a compact way of stating the assumptions that are needed to go from observational data to causal inferences.

In econometrics and cognate fields, "structural" equations describe causal relationships. The model gives a clearer meaning to this idea, and to the idea of "stability under intervention." The parameters in Figure 3, for instance, are defined through the response schedules (8) and (9), separately from the data. These parameters are constant across subjects and levels of treatment (by assumption, of course). Parameters are the same in a regime of passive observation and in a regime of active manipulation. Similar assumptions of stability are imposed on the error distributions. In summary, regression equations are structural, with parameters that are stable under intervention, when the equations derive from response schedules like (10) and (11).

Path models do not infer causation from association. Instead, path models *assume* causation through response schedules, and—using additional statistical assumptions—estimate causal effects from observational data. The statistical assumptions (independence, expectation zero, constant variance) justify estimation by ordinary least squares. With large samples, confidence intervals

and significance tests would follow. With small samples, the errors would have to follow a normal distribution in order to justify t-tests.

Figure 5. The path diagram as a box model.



The box model in Figure 5 illustrates the statistical assumptions. Independent errors with constant distributions are represented as draws made at random with replacement from a box of potential errors (Freedman-Pisani-Purves 1998). Since the box remains the same from one draw to another, the probability distribution of one draw is the same as the distribution of any other. The distribution is constant. Furthermore, the outcome of one draw cannot affect the distribution of another. That is independence. Verifying the causal assumptions (10) and (11), which are about potential outcomes, is a daunting task. The statistical assumptions present difficulties of their own. Assessing the degree to which the modeling assumptions hold is therefore problematic. The difficulties noted earlier—in Yule on poverty, Blau and Duncan on stratification, Gibson on McCarthyism—are systemic.

Embedded in the formalism is the conditional distribution of Y, if we were to intervene and set the value of X. This conditional distribution is a counterfactual, at least when the study is observational. The conditional distribution answers the question, what would have happened if we had intervened and set X to x, rather than letting Nature take its course? The idea is best suited to experiments or hypothetical experiments.

There are also 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. Investigators may hope that regression equations are like laws of motion in classical physics. (If position and momentum are given, you can determine the future of the system and discover what would happen with different initial conditions.) Some other formalism may be needed to make this non-manipulationist account more precise.

Latent variables

There is yet another layer of complexity when the variables in the path model remain "latent"—unobserved. It is usually supposed that the manifest variables are related to the latent variables by a series of regression-like equations ("measurement models"). There are numerous assumptions about error terms, especially when likelihood techniques are used. In effect, latent variables are reconstructed by some version of factor analysis and the path model is fitted to the results. The scale of the latent variables is not usually identifiable, so variables are standardized to have mean 0 and variance 1. Some algorithms will infer the path diagram as well as the latents from the data,

but there are additional assumptions that come into play. Anderson (1984) provides a rigorous discussion of statistical inference for models with latent variables, given the requisite statistical assumptions. He does not address the connection between the models and the phenomena. Kline (1998) is a well-known text. Ullman and Bentler (2003) survey recent developments.

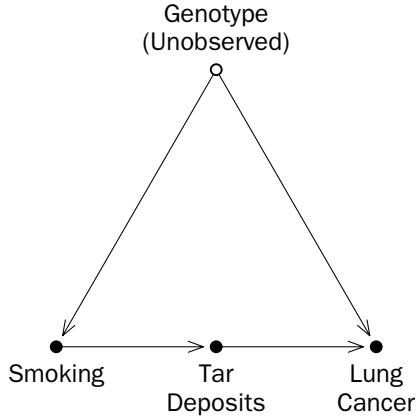
A possible conflict in terminology should be mentioned. In psychometrics and cognate fields, "structural equation modeling" (typically, path modeling with latent variables) is sometimes used for causal inference and sometimes to get parsimonious descriptions of covariance matrices. For causal inference, questions of stability are central. If no causal inferences are made, stability under intervention is hardly relevant; nor are underlying equations "structural" in the econometric sense described earlier. The statistical assumptions (independence, distributions of error terms constant across subjects, parametric models for error distributions) would remain on the table.

Graphical models

Yule's equation (1) was linear: a unit increase in ΔOut is supposed to cause an increase of b units in ΔPaup, for any value of ΔOut and any values of the control variables ΔOld and ΔPop. Similarly, the Blau and Duncan equations (4-5-6) were linear, as were equations (10-11). Linearity is a restrictive assumption. Graphical techniques have been suggested for relaxing this assumption and dealing with relationships that may be non-linear. Developments can only be sketched here.

In one setup, the graph is known a priori, and the issue is to achieve control of unmeasured confounders. (Another setup, where the graph is inferred from the data, will be considered below.) Figure 6 is an example used by Pearl (1995, pp. 675-66; 2000, pp. 81–83) to illustrate his methods. The graph is to be taken as given. The arrows are assumed by Pearl to represent causation rather than mere association. The variables at the nodes are governed by a joint probability distribution. What features of this distribution correspond can be read off the graph?

Figure 6. A graphical model for smoking and lung cancer. genes are unobserved, confounding the relationship between smoking and lung cancer.



Notice that—by assumption—there is no arrow from genes to tar deposits, or from smoking to lung cancer. The first exclusion means that genes have no direct influence on tar deposits. In probabilistic terms, the implication is that

$$P(\text{Tar Deposits}|\text{Genes, Smoking}) = P(\text{Tar Deposits}|\text{Smoking})$$

The second exclusion—no arrow from smoking to lung cancer—means that smoking affects lung cancer risj only through the buildup of tar deposits, implying that

$$P(\text{Lung Cancer}|\text{Genes, Smoking, Tar Deposits}) = P(\text{Lung Cancer}|\text{Genes, Tar Deposits})$$

The probabilistic conditions are said to make the graph "Markovian." Another key point about the graph: genotype is unobserved, signalled by the open dot. The joint distribution of the observed variables—smoking, tar deposits, and lung cancer—is taken as given. However, the joint distribution of all four variables remains unknown, because genotype is unobserved.

Does smoking cause lung cancer? The relationship between smoking and lung cancer is confounded by an unobserved variable. But the causal effect of smoking on lung cancer (averaged over the various possible genotypes in the population) can be determined from the data—given the assumptions behind Figure 6. This intriguing theorem is due to Robins (1986, 1987). It was rediscovered by Pearl (1995) as well as Spirtes, Glymour and Scheines (1993).

What are the implications for applied work? To begin with, it is only by assumption that the arrows in Figure 6 represent causation. Moreover, there are three special assumptions.

(i) Genes have no direct effect on tar deposits.

(ii) Smoking has no direct effect on lung cancer.

(iii) Smoking, tar deposits, and lung cancer can be measured with good accuracy.

Pearl (2000, p. 83) acknowledges making these assumptions, but there is no support for them in the literature. (i) The lung has a mechanism—"the mucociliary escalator"—for eliminating foreign matter, including tar. This mechanism seems to be under genetic control. (Of course, clearance mechanisms can be overwhelmed by smoking.) The forbidden arrow from genes to tar deposits may have a more solid empirical basis than the permitted arrows from genes to smoking and lung cancer. Assumption (ii) is just that—an assumption. And (iii) is not plausible, especially for tar deposits in living subjects. If arrows are permitted from genes to tar deposits or from smoking to lung cancer, then the theory does not apply to Figure 6. If measurements are subject to large errors, the theory does not apply either. Other examples in Pearl (1995, 2000) are equally problematic. Graphical models cannot overcome the difficulties created by unmeasured confounders without introducing strong and artificial assumptions.

Figure 6 addresses a question with some intellectual history. Fisher's "constitutional hypothesis" explained the association between smoking and disease on the basis of a gene that caused both. This idea is refuted not by making assumptions but by empirical work. For example, Kaprio and Koskenvuo (1989) present data from their twin study. The idea is to find pairs of identical twins where one smokes and one does not. That sets up a race: who will die first, the smoker or the non-smoker? The smokers win hands down, for total mortality or death from heart disease. The genetic hypothesis is incompatible with these data.

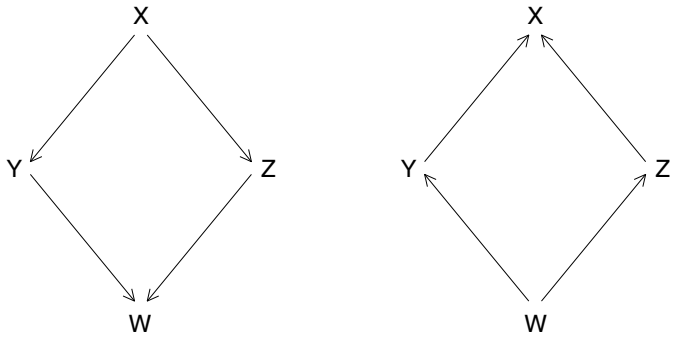
For lung cancer, the smokers win both of the races that have been run. (Why only two? Smoking-discordant twin pairs are unusual, lung cancer is a rare disease, and the population of Scandinavia is small.) Carmelli and Page (1996) have a similar analysis with a larger cohort of twins. Do not bet on Fisher. International Agency for Research on Cancer (1986) reviews the health effects of smoking and indicates the difficulties in measuring tar deposits (pp.179–98). Nakachi et al. (1993) and Shields et al. (1993) illustrate conflicts on the genetics of smoking and lung cancer.

Also see Miller et al. (2003). Other examples in Pearl (1995, 2000) are equally unconvincing on substantive grounds. Finding the mathematical consequences of assumptions matters, but connecting assumptions to reality matters even more. For additional detail, see Freedman (1997, 2004).

Inferring the graph from the data

Spirtes, Glymour, and Scheines (1993, 2000) and Pearl (1988) have algorithms for inferring causal graphs from the data, if the "faithfulness" assumption is imposed. It will be easier to explain this idea by example. Let us assume that the graphs in Figure 7 are Markovian. In the left hand panel, Y and Z will be independent given X; moreover, X and W will be independent given Y and Z. In the right hand panel, these independence relations will hold only for special values of the parameters governing the joint probability distribution of the variables X, Y, Z, W. The faithfulness condition precludes such "accidents:" the only independence relations that are permitted are independence relations that can be read off the graph. Given the faithfulness condition, there is some theory to determine which features of graphs can be recovered from the joint distributions of observables, and there are statistical algorithms to implement the theory.

Figure 7. The faithfulness condition: No accidental independence



Rather than exploring theoretical issues, it will be more helpful to consider applications. SGS (Spirtes, Glymour, and Scheines) seem to give abundant examples to show the power of their algorithms. However, many of the examples turn out to be simulations, where the computer generates the data. Assumptions are satisfied by fiat, having been programmed into the computer; questions about the real world are finessed. Many other examples relate to the health effects of smoking. These causal diagrams are hypothetical too. No contact is made with data, and no substantive conclusions are drawn.

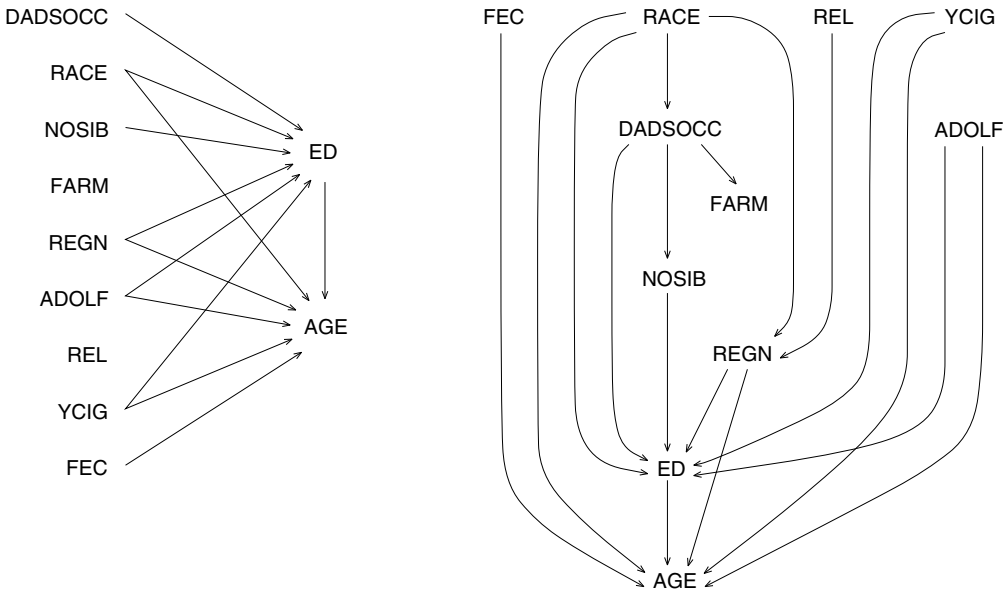
SGS do use their algorithms to analyze a number of real data sets, mainly from the social-science literature. What about those applications? Analyses were replicated for the most solid-looking cases (Freedman, 1997; Freedman and Humphreys, 1999). The examples all turned out to have the same texture; only one need be discussed here. Rindfuss et al. (1980) developed a model to explain the process by which a woman decides how much education to get, and when to have her first child. The variables in the model are defined in Table 3.

The statistical assumptions made by Rindfuss et al., let alone the conditions imposed by SGS, may seem rather implausible if examined at all closely. Here, we focus on results. According to SGS,

Table 3. Variables in the model.

ED	Respondent's education (Years of schooling completed at first marriage)
AGE	Respondent's age at first birth
DADSOCC	Respondent's father's occupation
RACE	Race of respondent (Black=1, other=0)
NOSIB	Respondent's number of siblings
FARM	Farm background (coded 1 if respondent grew up on a farm, else 0)
REGN	Region where respondent grew up (South=1, other=0)
ADOLF	Broken family (coded 0 if both parents present at age 14, else 1)
REL	Religion (Catholic=1, other=0)
YCIG	Smoking (coded 1 if respondent smoked before age 16, else coded 0)
FEC	Fecundability (coded 1 if respondent had a miscarriage before first birth; else coded 0)

Figure 8. The left hand panel shows the model reported by SGS (1993, p. 140; 2000, p. 104). The right hand panel shows the whole graph produced by the SGS algorithms.



> Given the prior information that ED and AGE are not causes of the other variables, the PC algorithm (using the .05 significance level for tests) directly finds the model [in the left hand panel of Figure 8] where connections among the regressors are not pictured. [1993, p. 139; 2000, p. 103]

The main conclusion in Rindfuss et al. is that AGE does not influence ED. Apparently, the left hand panel in Figure 8 confirms this finding, which allows SGS to claim a success for their algorithms. However, the graph in the left hand panel is not the one actually produced by the algorithms. The unedited graph is shown in the right hand panel. The unedited graph says, for instance, that race and religion cause region of residence. Other peculiarities need not detain us.

The SGS algorithms are successful only if one is very selective in reading the computer output. The difficulty seems to be this. The algorithms depend on strong and artificial assumptions, which are unlikely to be satisfied in real applications. Graphical models are interesting, and may provide a natural mathematical language for discussing certain philosophical issues. But these models are unlikely to help applied workers in making valid causal inferences from observational data.

Directed Acyclic Graphs

The graphs in Figures 6-7-8 are DAGs (Directed Acyclic Graphs)—directed because each arrow points in a certain direction, acyclic because you cannot get from a node back to itself by following arrows. In particular, reciprocal causation is excluded by assumption. Interestingly, the SGS algorithms do sometimes produce graphs with cycles, showing that the algorithms are not internally consistent. For additional detail on DAGs and the SGS algorithms, see (Freedman, 1997, 2004; Freedman and Humphreys, 1996, 1999).

Literature review

There is by now an extended critical literature on statistical models, starting perhaps with the exchange between Keynes (1939, 1940) and Tinbergen (1940). Other familiar citations in the economics literature include Liu (1960), Lucas (1976), and Sims (1980). Manski (1995) returns to the under-identification problem that was posed so sharply by Liu and Sims. In brief, a priori exclusion of variables from causal equations can seldom be justified, so there will typically be more parameters than data. Manski suggests methods for bounding quantities that cannot be estimated. Sims' idea was to use simple, low-dimensional models for policy analysis, instead of complex-high dimensional ones. Leamer (1978) discusses the issues created by specification searches, as does Hendry (1993). Heckman (2000) traces the development of econometric thought from Haavelmo and Frisch onwards, stressing the role of "structural" or "invariant" parameters, and "potential outcomes." Lucas too was concerned about parameters that changed under intervention. Engle, Hendry, and Richard (1983) distinguish several kinds of exogeneity, with different implications for causal inference. Recently, some econometricians have turned to natural experiments for the evaluation of causal theories. These investigators stress the value of careful data collection and data analysis. Angrist and Krueger (2001) have a useful survey.

One of the drivers for modeling in economics and other fields is rational choice theory. Therefore, any discussion of empirical foundations must take into account a remarkable series of papers, initiated by Kahneman and Tversky (1974), that explores the limits of rational choice theory. These papers are collected in Kahneman, Slovic, and Tversky (1982), and in Kahneman and Tversky (2000). The heuristics and biases program has attracted its own critics (Gigerenzer, 1996). That

critique is interesting and has some merit. But in the end, the experimental evidence demonstrates severe limits to the power of rational choice theory (Kahneman and Tversky, 1996). If people are trying to maximize expected utility, they generally don't do it very well. Errors are large and repetitive, go in predictable directions, and fall into recognizable categories. Rather than making decisions by optimization—or bounded rationality, or satisficing—people seem to use plausible heuristics that can be identified. If so, rational choice theory is generally not a good basis for justifying empirical models of behavior. Drawing in part on the work of Kahneman and Tversky, Sen (2002) gives a far-reaching critique of rational choice theory. This theory has its place, but also leads to "serious descriptive and predictive problems."

Almost from the beginning, there were critiques of modeling in other social sciences too (Platt, 1996). Bernert (1983) reviews the historical development of causal ideas in sociology. Recently, modeling issues have been much canvassed in sociology. Abbott (1997) finds that variables like income and education are too abstract to have much explanatory power, with a broader examination of causal modeling in Abbott (1998). He finds that "an unthinking causalism today pervades our journals"; he recommends more emphasis on descriptive work and on middle-range theories. Berk (2003) is skeptical about the possibility of inferring causation by modeling, absent a strong theoretical base. Clogg and Haritou (1997) review difficulties with regression, noting that you can too easily include endogenous variables as regressors.

Goldthorpe (1998, 2000, 2001) describes several ideas of causation and corresponding methods of statistical proof, with different strengths and weaknesses. Although skeptical of regression, he finds rational choice theory to be promising. He favors use of descriptive statistics to determine social regularities, and statistical models that reflect generative processes. In his view, the manipulationist account of causation is generally inadequate for the social sciences. Hedström and Swedberg (1998) present a lively collection of essays by sociologists who are quite skeptical about regression models; rational choice theory also takes its share of criticism. There is an influential book by Lieberson (1985), with a followup by Lieberson and Lynn (2002). Ní Bhrolcháin (2001) has some particularly forceful examples to illustrate the limits of modeling. Sobel (1998) reviews the literature on social stratification, concluding that "the usual modeling strategies are in need of serious change." Also see Sobel (2000).

Meehl (1978) reports the views of an empirical psychologist. Also see Meehl (1954), with data showing the advantage of using regression to make predictions, rather than experts. Meehl and Waller (2002) discuss the choice between two similar path models, viewed as reasonable approximations to some underlying causal structure, but do not reach the critical question—how to assess the adequacy of the approximation. Steiger (2001) has a critical review of structural equation models. Larzalere and Kuhn (2004) offer a more general discussion of difficulties with causal inference by purely statistical methods. Abelson (1995) has an interesting viewpoint on the use of statistics in psychology.

There is a well-known book on the logic of causal inference, by Cook and Campbell (1979). Also see Shadish, Cook, and Campbell (2002), which has among other things a useful discussion of manipulationist vs non-manipulationist ideas of causation. In political science, Duncan (1984) is far more skeptical about modeling than Blau and Duncan (1967). Achen (1982, 1986) provides a spirited and reasoned defense of the models. Brady and Collier (2004) compare regression methods with case studies; invariance is discussed under the rubric of causal homogeneity.

There is an extended literature on graphical models for causation. Greenland, Pearl and

Robins (1999) give a clear account in the context of epidemiology. Lauritzen (1996, 2001) has a careful treatment of the mathematics. These authors do not recognize the difficulties in applying the methods to real problems. Strong claims are made for non-linear methods that elicit the model from the data and control for unobserved confounders (SGS, 1993, 2000; Pearl, 2000). However, the track record is not encouraging (Freedman, 1997, 2004; Freedman and Humphreys, 1996, 1999). Cites from other perspectives include Oakes (1986), Pearl (1995), McKim and Turner (1997), as well as Freedman (1985, 1987, 1991, 1995, 1999).

The statistical model for causation was proposed by Neyman (1923). It has been rediscovered many times since: see, for instance, Hodges and Lehmann (1964, section 9.4). The setup is often called "Rubin's model," but that simply mistakes the history. See the comments by Dabrowska and Speed on their translation of Neyman (1923), with a response by Rubin; compare to Rubin (1974) and Holland (1986). Holland (1986, 1988) explains the setup with a super-population model to account for the randomness, rather than individualized error terms. Error terms are often described as the overall effects of factors omitted from the equation. But this description introduces difficulties of its own, as shown by Pratt and Schlaifer (1984, 1988). Stone (1993) presents a super-population model with some observed covariates and some unobserved. Formal extensions to observational studies—in effect, assuming these studies are experiments after suitable controls have been introduced—are discussed by Holland and Rubin among others.

Conclusion

Causal inferences can be drawn from non-experimental data. However, no mechanical rules can be laid down for the activity. Since Hume, that is almost a truism. Instead, causal inference seems to require an enormous investment of skill, intelligence, and hard work. Many convergent lines of evidence must be developed. Natural variation needs to be identified and exploited. Data must be collected. Confounders need to be considered. Alternative explanations have to be exhaustively tested. Before anything else, the right question needs to be framed. Naturally, there is a desire to substitute intellectual capital for labor. That is why investigators try to base causal inference on statistical models. The technology is relatively easy to use, and promises to open a wide variety of questions to the research effort. However, the appearance of methodological rigor can be deceptive. The models themselves demand critical scrutiny. Mathematical equations are used to adjust for confounding and other sources of bias. These equations may appear formidably precise, but they typically derive from many somewhat arbitrary choices. Which variables to enter in the regression? What functional form to use? What assumptions to make about parameters and error terms? These choices are seldom dictated either by data or prior scientific knowledge. That is why judgment is so critical, the opportunity for error so large, and the number of successful applications so limited.

Author's footnote

Richard Berk, Persi Diaconis, Michael Finkelstein, Paul Humphreys, Roger Purves, and Philip Stark made useful comments. This paper is based on Freedman (1987, 1991, 1997, 1999, 2004, 2005ab). Figure 1 is redrawn from Freedman (19970; figure 2, from Blau and Duncan (1967); figure 3, from Gibson (1988), also see Freedman (1991); figures 7 and 8, from Freedman (1997). Figure 8 appeared in Freedman and Humphreys (1996).

References

Abelson, R. (1995). *Statistics as Principled Argument*. Lawrence Erlbaum Associates, Hillsdale, N. J.

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.

Achen, C. (1977). Measuring Representation: Perils of the Correlation Coefficient. *American Journal of Political Science* 21: 805–15.

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

Achen, C. (1986). *The Statistical Analysis of Quasi-Experiments*. University of California Press, Berkeley.

Anderson, T. W. (1984). Estimating Linear Statistical Relationships. *Annals of Statistics* 12: 1–45

Angrist, J. D. and Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives* 19: 2–16.

Berk, R. A. (2003). *Regression Analysis: A Constructive Critique*. Newbury Park, CA: Sage Publications.

Bernert, C. (1983). The Career of Causal Analysis in American Sociology. *British Journal of Sociology* 34: 230–54.

Blalock, H. M. (1989). The Real and Unrealized Contributions of Quantitative Sociology. *American Sociological Review* 54: 447–60.

Blau, P. M. and Duncan O. D. (1967). *The American Occupational Structure*. Wiley. Reissued by the Free Press, 1978. Data collection described on page 13, coding of education on pages 165–66, coding of status on pages 115–27, correlations and path diagram on pages 169–70.

Brady, H. and Collier, D. (2004), eds. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman & Littlefield Publishers, Inc. To appear.

Carmelli, D. and Page, W. F. (1996). 24-year Mortality in Smoking-Discordant World War II U. S. Male Veteran Twins. *International Journal of Epidemiology* 25: 554–59.

Clogg, C. C. and Haritou, A. (1997). The Regression Method of Causal Inference and a Dilemma Confronting this Method. In *Causality in Crisis*, ed. by V. McKim and S. Turner. University of Notre Dame Press, pp. 83–112.

Cook T. D. and Campbell D. T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings.* Boston: Houghton Mifflin.

Duncan, O. D. (1984). *Notes on Social Measurement*. New York: Russell Sage.

Engle, R. F., Hendry, D. F., and Richard, J. F. (1983). Exogeneity. *Econometrica* 51: 277–304.

Freedman, D. A. (1985). Statistics and the Scientific Method. In *Cohort Analysis in Social Research: Beyond the Identification Problem,* ed. by W. M. Mason and S. E. Fienberg. New York: Springer-Verlag, pp. 343–90 (with discussion).

Freedman, D. A. (1987). As Others See Us: A Case Study in Path Analysis. *Journal of Educational Statistics*, 12, 101–223 (with discussion). Reprinted in *The Role of Models in Nonexperimental Social Science*, ed. by J. Shaffer. Washington, D.C.: AERA/ASA, 1992, pp. 3–125.

Freedman, D. A. (1991). Statistical Models and Shoe Leather. In *Sociological Methodology 1991*, ed. by Peter Marsden. Washington, D.C.: American Sociological Association, Chapter 10 (with discussion).

Freedman, D. A. (1995). Some Issues in the Foundation of Statistics. *Foundations of Science* 1: 19–83 (with discussion). Reprinted in *Some Issues in the Foundation of Statistics*, ed. by B. C. van Fraassen (1997). Dordrecht: Kluwer, pp. 19–83 (with discussion).

Freedman, D. A. (1997). From Association to Causation via Regression. In *Causality in Crisis?* ed. by V. McKim and S. Turner. South Bend: University of Notre Dame Press, pp. 113–82 (with discussion). Reprinted in *Advances in Applied Mathematics* 18: 59–110.

Freedman, D. A. (1999). From Association to Causation: Some Remarks on the History of Statistics. *Statistical Science* 14: 243–58. Reprinted in *Journal de la Société Française de Statistique* 140 (1999) 5–32, and in *Stochastic Musings: Perspectives from the Pioneers of the Late 20th Century*, ed. by J. Panaretos (2003), Lawrence Erlbaum, pp. 45–71.

Freedman, D. A. (2004). On Specifying Graphical Models for Causation, and the Identification Problem. *Evaluation Review* 26: 267–93. Reprinted in J. Stock and D. Wright, eds. *The Rothenberg Festschrift*. Cambridge University Press.

Freedman, D. A. (2005a). Linear statistical models for causation: A critical review. In B Everitt and D Howell, eds. *Wiley Encyclopedia of Statistics in Behavioral Science*.

Freedman, D. A. (2005b). *Statistical Models: Theory and Practice*. To appear.

Freedman, D. A. and Humphreys, P. (1996). The grand leap. *British Journal for the Philosophy of Science* 47: 113–23.

Freedman, D. A. and Humphreys, P. (1999). Are There Algorithms that Discover Causal Structure? *Synthese* 121: 29–54.

Freedman, D. A., Pisani, R., and Purves, R. A. (1998). *Statistics*. 3rd ed. New York: W. W. Norton, Inc.

Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium.* Perthes et Besser, Hamburg. Reprinted in 1963 by Dover, New York.

Gibson, J. L. (1988). Political Intolerance and Political Repression during the McCarthy Red Scare. *APSR* 82: 511–29.

Gigerenzer, G. (1996). On Narrow Norms and Vague Heuristics. *Psychological Review* 103: 592–96.

Goldthorpe, J. H. (1998). *Causation, Statistics and Sociology*. Twenty-ninth Geary Lecture, Nuffield College, Oxford. Published by the Economic and Social Research Institute, Dublin, Ireland.

Goldthorpe, J. H. (2000). *On Sociology: Numbers, Narratives, and Integration of Research and Theory*. Oxford University Press.

Goldthorpe, J. H. (2001). Causation, Statistics, and Sociology. *European Sociological Review* 17: 1–20.

Greenland, S., Pearl, J., and Robins, J. (1999). Causal diagrams for epidemiologic research. *Epidemiology* 10: 37–48.

Heckman, J. J. (2000). Causal Parameters And Policy Analysis In Economics: A Twentieth Century Retrospective. *The Quarterly Journal of Economics* CVX: 45–97.

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

Hendry, D. F. (1993). *Econometrics—Alchemy or Science?* Oxford: Blackwell.

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

Holland, P. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association* 8: 945–60.

Holland, P. (1988). Causal Inference, Path Analysis, and Recursive Structural Equation Models. In *Sociological Methodology 1988*, ed. by C. Clogg. Washington, D.C.: American Sociological Association, Chapter 13.

International Agency for Research on Cancer (1986). *Tobacco smoking*. Lyon, France: IARC, Monograph 38.

Kahneman, D., Slovic, P., and Tversky, A., eds. (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Kahneman, D. and Tversky, A. (1974). Judgment under Uncertainty: Heuristics and Bias. *Science* 185: 1124–31.

Kahneman, D. and Tversky, A. (1996). On the Reality of Cognitive Illusions. *Psychological Review* 103: 582–91.

Kahneman, D. and Tversky, A., eds. (2000). *Choices, Values, and Frames*. Cambridge University Press.

Kaprio, J. and Koskenvuo, M. (1989). Twins, Smoking and Mortality: A 12-year Prospective Study of Smoking-Discordant Twin Pairs. *Social Science and Medicine* 29: 1083–9.

Keynes, J. M. (1939). Professor Tinbergen's Method. *The Economic Journal* 49: 558–70.

Keynes, J. M. (1940). Comment on Tinbergen's Response. *The Economic Journal* 50: 154–56.

Kline, R. B. (1998). Principles and Practice of Structural Equation Modeling. New York: Guilford.

Larzalere, R. E. and Kuhn, B. R. (2004). The Intervention Selection Bias: An Underrecognized Confound in Intervention Research. *Psychological Bulletin* 130: 289–303.

Lauritzen, S. (1996). *Graphical models*. Oxford: Clarendon Press.

Leamer, E. (1978). *Specification Searches*. New York: John Wiley.

Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes.* Courcier, Paris. Reprinted in 1959 by Dover, New York.

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

Lieberson, S. and Lynn, F. B. (2002). Barking up the Wrong Branch: Alternative to the Current Model of Sociological science. *Annual Review of Sociology* 28: 1–19.

Liu, T. C. (1960). Under-Identification, Structural Estimation, and Forecasting. *Econometrica* 28: 855–65.

Lucas, R. E. Jr. (1976). Econometric Policy Evaluation: A Critique. In K. Brunner and A. Meltzer (eds.), *The Phillips Curve and Labor Markets*, vol. 1 of the Carnegie-Rochester Conferences on Public Policy, supplementary series to the Journal of Monetary Economics, North-Holland, Amsterdam, pp. 19–64. (With discussion.)

Manski, C. F. (1995). *Identification Problems in the Social Sciences.* Harvard University Press.

McKim, V. and Turner, S., eds. (1997). *Causality in Crisis? Proceedings of the Notre Dame Conference on Causality.* University of Notre Dame Press.

Meehl, P. E. (1954). *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis: University of Minnesota Press.

Meehl, P. E. (1978). Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology. *Journal of Consulting and Clinical Psychology* 46: 806–34.

Meehl, P. E. and Waller N. G. (2002). The Path Analysis Controversy: A New Statistical Approach to Strong Appraisal of Verisimilitude. *Psychological Methods* 7: 283–337 (with discussion).

Miller, D. P., Neuberg, D., De Vivo, I., Wain, J. C. et al. (2003). Smoking and the Risk of Lung Cancer: Susceptibility with GSTP1 Polymorphisms. *Epidemiology* 14: 545–51.

Nakachi, K., Ima, K., Hayashi, S.-I., and Kawajiri, K. (1993). Polymorphisms of the CYP1A1 and Glutathione S-Transferase Genes Associated with Susceptibility to Lung Cancer in relation to Cigarette Dose in a Japanese Population. *Cancer Research* 53: 2994–99.

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. English translation by D. Dabrowska and T. Speed (1990) *Statistical Science* 5: 463–80 (with discussion).

Ní Bhrolcháin, M. (2001). Divorce Effects and Causality in the Social Sciences. *European Sociological Review* 17: 33–57.

Oakes, M. W. (1990). *Statistical Inference*, Chestnut Hill, MA: Epidemiology Resources Inc.

Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika* 82: 669–710 (with discussion).

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Publishers, Inc.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Platt, J. (1996). *A History of Sociological Research Methods in America*. Cambridge University Press.

Pratt, J. and Schlaifer, R. (1984). On the Nature and Discovery of Structure. *Journal of the American Statistical Association* 79: 9–21.

Pratt, J. and Schlaifer, R. (1988). On the Interpretation and Observation of Laws. *Journal of Econometrics* 39: 23–52.

Quetelet, A. (1835). *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale.* Bachelier, Paris.

Rindfuss, R. R., Bumpass, L., and St. John, C. (1980). Education and Fertility: Implications for the Roles Women Occupy. *American Sociological Review* 45: 431–47.

Robins, J. M. (1987). A Graphical Approach to the Identification and Estimation of Causal Parameters in Mortality Studies with Sustained Exposure Periods. *Journal of Chronic Diseases* 40, Supplement 2: 139S–161S.

Robins, J. M. (1986). A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect. *Mathematical Modelling* 7: 1393–1512.

Rubin, D. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66: 688–701.

Sen, A. K. (2002). *Rationality and Freedom*. Harvard University Press.

Shadish W. R., Cook T. D., and Campbell D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Shields, P. G., Caporaso, N. E., Falk, K. T. , Sugimura, H. et al. (1993). Lung Cancer, Race and a CYP1A1 Genetic Polymorphism. *Cancer Epidemiology, Biomarkers and Prevention* 2: 481–5.

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica* 48: 1–47.

Sobel, M. E. (2000). Causal Inference in the Social Sciences. *Journal of the American Statistical Association* 95: 647–51.

Sobel, M. E. (1998). Causal Inference in Statistical Models of the Process of Socioeconomic Achievement—A Case Study. *Sociological Methods & Research* 27: 318–348

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. Springer Lecture Notes in Statistics, no. 81, Springer-Verlag, New York. 2nd edn. 2000, MIT Press.

Steiger, J. H. (2001). Driving Fast in Reverse. *Journal of the American Statistical Association* 96: 331–38.

Stone, R. (1993). The Assumptions on which Causal Inferences Rest. *Journal of the Royal Statistical Society* Series B 55: 455–66.

Tinbergen, J. (1940). Reply to Keynes. *The Economic Journal* 50: 141–54.

Ullman, J. B. and Bentler, P. M. (2003). Structural equation modeling. In I. B. Weiner, J. A. Schinka, and W. F. Velicer, Eds. *Handbook of Psychology. Volume 2: Research Methods in Psychology* pp. 607–634. Hoboken, NJ: Wiley.

Yule, G. U. (1899). An Investigation into the Causes of Changes in Pauperism in England, Chiefly During the Last Two Intercensal Decades. *Journal of the Royal Statistical Society* 62: 249–95.

January 2005
Technical Report No. 651
Statistics Department
University of California
Berkeley, CA 94720-3860
www.stat.berkeley.edu/~census/651.pdf