From association to causation via regression

David A. Freedman
Technical Report No. 408
Statistics Department
University of California
Berkeley, Calif 94720

April 22, 1994

Talk presented at the Notre Dame Conference on
Causality in Crisis, Oct 15-17, 1993.

For nearly a century, investigators in the social sciences have used regression models to deduce cause-and-effect relationships from patterns of association. Path models and automated search procedures are more recent developments. In my view, this enterprise has not been successful. The models tend to neglect the difficulties in establishing causal relations, and the mathematical complexities tend to obscure rather than clarify the assumptions on which the analysis is based.

Formal statistical inference is, by its nature, conditional. If maintained hypotheses A, B, C, ... hold, then H can be tested against the data. However, if A, B, C, ... remain in doubt, so must inferences about H. Careful scrutiny of maintained hypotheses should therefore be a critical part of empirical work-- a principle honored more often in the breach than the observance.

This paper focuses on modeling techniques that seem to convert association into causation. The object is to clarify the differences among the various uses of regression, as well as the source of the difficulty in making causal inferences by modeling. The discussion will proceed mainly by examples, ranging from (Yule 1899) to (Spirtes, Glymour, and Scheines 1992).

1. Outline

Many treatments of regression seem to take for granted that the investigator knows the relevant variables, their causal order, and the functional form of the relationships among them; measurements of the independent variables are assumed to be without error. Indeed, Gauss developed and used regression in physical science contexts where these conditions hold, at least to a very good approximation.[1] Today, the textbook theorems that justify regression are proved on the basis of such assumptions.

In the social sciences, the situation seems quite different. Regression is used to discover relationships or to disentangle cause and effect. However, investigators have only vague ideas as to the relevant variables and their causal order; functional forms are chosen on the basis of convenience or familiarity; serious problems of measurement are often encountered.

Regression may offer useful ways of summarizing the data and making predictions. Investigators may be able to use summaries and predictions to draw substantive conclusions. However, I see no cases in which regression equations, let alone the more complex methods, have succeeded as engines for discovering causal relationships. Of course, there may be success stories that I have not found; nor does a track record of failure necessarily project into the future.

One of the first applications of regression techniques to social science is (Yule 1899). Recent examples will be found in (Spirtes, Glymour, and Scheines 1992), to be cited here as SGS. (The SGS theory is summarized in Glymour 1993, cited as CG.) SGS have attracted considerable attention in the philosophy of science, because they have developed computerized algorithms that search for path models. With their algorithms, SGS claim to make rigorous inferences of causation from association. This is a bold claim, which does not survive examination.

---

[1] Gauss was fitting orbits to astronomical observations, with least squares to estimate the elements of the orbits (Gauss 1809). (Stigler 1986, 145-46) awards priority to (Legendre 1805).

The balance of this paper is organized as follows. Section 2 discusses Yule's work. Sections 3 and 4 explain the critical idea of "exogeneity." Section 5 describes a contemporary regression model. Sections 6–10 review SGS and reanalyze some of their examples. Sections 11–12 canvass some mathematical issues. Possible responses to my critique will be found in section 13. There is a brief review of the literature in section 14, and conclusions are presented in section 15. For ease of reference, standard formulas for regression are given in an appendix. I have tried to make most of the paper accessible to non-statistical readers, particularly if they will permit the occasional undefined technical term; sections 11 and 12 are more specialized.

2. Yule's regression model for pauperism

One of the first regression models in social science was developed by Yule-- "An Investigation into the Causes of Changes in Pauperism in England, Chiefly During the Last Two Intercensal Decades."[2] In late 19th century England, poor people could be supported either inside the poor house or outside. Did provision of support outside the poor house increase the number of poor people?

To address this issue, Yule used data from the censuses of 1871, 1881, and 1891. (In England, the census is taken in years that end with 1.) He considered the periods 1871–81 and 1881–91, relating changes in the number of paupers to changes in the "outrelief ratio," that is, the ratio between the number of paupers supported outside the poor house and inside. He used regression to control for two confounders-- changes in the population and its age structure.

His equation can be written as follows:

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

Here, Δ stands for percentage difference, *Paup* for the number of paupers, *Out* for the outrelief ratio, *Pop* for population size, and *Old* for the proportion of people aged 65 and over.

Yule's unit of analysis was the "union," which seems to have been a small geographical area like a county.[3] He had four kinds of areas: rural, mixed, urban, metropolitan. He used "Ordinary Least Squares" (OLS) to estimate the coefficients from the data, with a "50 cm. Gravet" slide rule to do the arithmetic.

To be more specific, Yule estimated a separate equation for each time period (1871–81 and 1881–91) and each kind of area. There were 2 time periods and 4 kinds of areas, thus, 2 × 4 = 8 equations. Within a time period, all areas of the same kind-- for instance, all rural unions-- are governed by one equation. (By coincidence, there are 4 coefficients in each equation, and 4 kinds of areas.)

---

[2] See (Yule 1899), (Stigler 1986, 345–58), and (Desrosières 1993).

[3] There were about 600 such areas in England. A poor-law union "consisted of two more parishes combined for administrative purposes." (Stigler 1986, 346).

Yule was looking for the "Hooke's Law of Poverty." Nature ran an experiment, with lots of variation over time and geography, and Yule analyzed the results. Regression was needed to control for the confounding effects of change in population and age structure. The equations were held to show that, other things being equal, changes in the outrelief ratio create corresponding changes in the number of paupers. Indeed, if you increase the outrelief ratio by 1 percentage point but hold the other factors constant, you will increase the number of paupers by b percent, b being the coefficient of outrelief in equation (1). More qualitatively, if b is positive, welfare creates paupers.

For a moment, I turn from Yule to methodology. A regression equation like (1) is usually written as

$$(2) \quad Y = X\beta + \varepsilon.$$

In this equation, the vector Y represents the dependent variable, like pauperism; the matrix X represents the explanatory (or "independent") variables, like the outrelief ratio, population, and age structure. These are observable. The vector β represents parameters, which are not observable but may be estimated from the data: parameters are "social constants," which characterize the process that generated the data. In Yule's equation, β has four components-- the parameters a, b, c, d in equation (1). The error or "disturbance" term ε is also unobservable, and represents the impact of chance factors unrelated to X. Statistical inferences are often based on "stochastic assumptions" about ε, e.g., ε is independent of X, its components are independent and identically distributed with mean 0. For details, see the appendix below.

Three possible uses for regression equations are as follows:

(i) to summarize data, or
(ii) to predict values of the dependent variable, or
(iii) to predict the results of interventions.

Yule could certainly have summarized his data by saying that for a given time period and unions of a specific type, with certain values of the explanatory variables, the change in pauperism was about so much and so much. In other words, he could have used his equations to estimate the average value of Y, given the values of X. This use of regression may run into technical problems if there are outliers, or nonlinearities in the regression surface. However, at least in principle, there do seem to be technical fixes for such problems. Furthermore, stochastic assumptions about the disturbance term play almost no role. Therefore, like most statisticians, I believe that regression can be quite helpful in summarizing large data sets.

For prediction, there is a *ceteris paribus* assumption: the system will remain stable. Prediction is already more complicated than description. On the other hand, if you make a series of predictions and test them against data, it may be possible to show that the system is stable, or sufficiently stable for regression to be quite helpful.[4] Again, any particular use of regression to make predictions may go off the rails, but there do not seem to be essential difficulties of principle involved.

Causal inference is different, because a change in the system is contemplated: for example, there will be an intervention. Descriptive statistics tell you about the correlations that happen to hold in the data; causal models claim to tell you what will happen to Y if you change X. Indeed, regression is often used to make counter-factual inferences about the past: what would Y have been if X had been different? This use of regression to make causal inferences is the most intriguing-- and the most problematic. Difficulties are created by omitted variables, incorrect functional form, etc. Of course, if the results of causal modeling were with any frequency checked against the results of interventions, the balance of argument might be very different.[5]

For description and prediction, the numerical values of the individual coefficients fade into the background: it is the whole linear combination on the right hand side of the equation that matters. For causal inference, it is the individual coefficients that do the trick. In equation (1), for example, it is b that should tell you what happens to pauperism when the outrelief ratio is manipulated.

---

[4] (Meehl 1954) provides some well-known examples. Predictive validity is best demonstrated by making real *ex ante* forecasts in several different contexts (Ehrenberg and Bound 1993).

[5] Also see (Manski 1993).

At this remove, the flaws in Yule's argument may be apparent. For example, there seem to be some important variables missing from the equation, including variables that measure economic activity. Here is Yule's comment on the last-named factor:

> 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]

Yule seems to have used the rate of population growth-- Δ*Pop* in equation (1)-- as a proxy for economic activity, although that creates ambiguity. Other things being equal, population growth will by itself add to the number of paupers; in its role as proxy, however, population growth should reduce pauperism.

The equations for metropolitan unions are shown below, for 1871–81 and 1881–91:[6]

$$\text{(1871–81)}\quad \Delta\text{Paup} = 13.19 + 0.755\times\Delta\text{Out} - 0.322\times\Delta\text{Pop} - 0.022\times\Delta\text{Old} + \text{residual.}$$

$$\text{(1881–91)}\quad \Delta\text{Paup} = 1.36 + 0.324\times\Delta\text{Out} - 0.369\times\Delta\text{Pop} + 1.37\times\Delta\text{Old} + \text{residual.}$$

For example, one metropolitan union is Westminster. Over the period 1871–81, the percentage changes in *Out, Pop* and *Old* are –73, –9, and 5, respectively. The percentage change in *Paup* predicted from the regression equation is

$$13.19 + 0.755\times(-73) - 0.322\times(-9) - 0.022\times 5 = -39.$$

The actual percentage change in *Paup* is –48. The "residual" is

$$\text{residual} = \text{actual} - \text{predicted} = -48 - (-39) = -9.$$

---

[6] These, and the other 6 equations, are reported in Yule's Table C, p. 259. His Table XIX gives data for metropolitan unions, in the form of "percentage ratios" for 1871–81 rather than differences, apparently to avoid negative numbers. The equations were fitted to data; the numerical coefficients in the displays are estimates for the corresponding parameters in (1); the residuals are observable, but are only approximations to unobservable disturbance terms.

The coefficients in the regression equation are estimated so as to minimize the size of the residuals. (Technically, it is the sum of the squares that is minimized-- hence the term "least squares.") The linear combination of explanatory variables on the right side of the equation has therefore been optimized; but there is no guarantee that individual coefficients will make much sense.

There are some noticeable inconsistencies in Yule's coefficients, over time and across the various kinds of geography. Nor are the signs of the coefficients entirely reasonable. These inconsistencies may not by themselves be fatal, but certainly raise the question of whether the equations hold true for any well-defined population of times and places. If the coefficients do not have a life of their own-- outside Yule's particular data set-- they cannot be used to answer questions of the form, "What would happen if you change the outrelief ratio?" The coefficients may be useful for descriptive purposes, but not for causal inference or even prediction.

Moreover, there are familiar difficulties of interpretation. At best, Yule showed that changes in pauperism and the outrelief ratio were associated, even after adjusting for changes in the population and its age structure. The direction of the causal arrow, however, is by no means clear. Yule's theory is that outrelief is the cause, and pauperism the effect. That is a reasonable view. However, the opposite idea seems equally tenable-- a union that is flooded with paupers may not be able to build poor houses fast enough, and resorts to outrelief. If so, pauperism causes outrelief. Also, Governor Pete Wilson's theory may have some plausibility for 19th century England if not 20th century California: unions that provide generous outrelief attract paupers from elsewhere.[7]

---

[7] According to (Stigler 1986, 356-7), Pigou criticized Yule for ignoring "the non-quantitative facts of the situation.... It is well known that, during recent years, those unions in which out-relief has been restricted have, on the whole, enjoyed a general administration much superior to that of other unions."

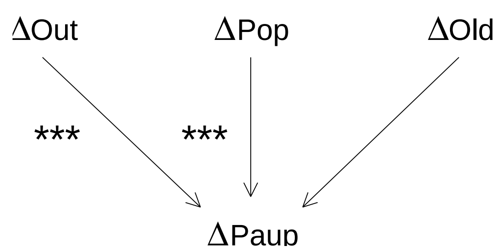
Stigler responds that "Pigou's ad hoc speculation ... could not, of course, be disproved from the data Yule used." In effect, this allows Yule to defend himself by pleading ignorance.

Yule must have been aware of these problems. After allocating the changes in pauperism to their various causes (including the residual), he withdraws all causal claims with one deft sentence:

> Strictly, for "due to" read "associated with." [footnote 25, p. 270]

Yule's paper is quite modern in spirit, with two exceptions: he did not rely on statistical significance, and he did not use a graph. Figure 1 brings him up to date.

Figure 1. Yule's model for pauperism. The figure represents equation (1) in graphical form. The asterisks denote a high degree of statistical significance.



Notes: To determine the asterisks, I recomputed Yule's regression for the metropolitan unions over the period 1871–81, using data in his Table XIX. I replicated his coefficients, as shown in the display, although roundoff error is quite large:

$$\begin{array}{llllll} \Delta\text{Paup} = & 12.884 & + 0.752\times\Delta\text{Out} & - 0.311\times\Delta\text{Pop} & + 0.056\times\Delta\text{Old} & + \text{residual.} \\ & 10.367 & 0.135 & 0.067 & 0.223 & \\ & 1.24 & 5.57 & -4.65 & 0.25 & \end{array}$$

Under the coefficients are standard errors (SEs) and t-statistics. The SE indicates the likely size of the difference between an estimated coefficient and its true value. The t-statistic is the ratio of an estimate to its SE. Generally, a t-statistic above 2 or 3 in absolute value indicates that the corresponding parameter is unlikely to be truly 0. The parameters are features of the model, and the SEs are computed on the basis of the stochastic assumptions in the model; for details, see the appendix. (Of course, Yule's model is open to serious question.) In Figure 1, the explanatory variables are correlated; such correlations are often signaled by curved, double-headed arrows; error terms are not shown either.

3. Regression estimates and conditional expectations

In the regression model (2), Y is the dependent variable, like pauperism; X represents the explanatory variables, like the outrelief ratio, population, and age structure. If all goes well, the regression equation will estimate the "conditional expectation" of Y given $X = x$, that is, the average value of Y corresponding to given values for the explanatory variables.

To clarify the definitions, consider two procedures:

Procedure #1. Select subjects with $X = x$; look at the average of their Y's.

Procedure #2. Intervene and set $X = x$ for some subjects; look at the average of their Y's.

These procedures are quite different. The first involves the data set as you find it. The second involves an intervention.

Regression does seem to let you move from selection to intervention; that is why the technique is so popular. However, regression approximates the selection procedure, rather than intervention. Nor does the statistical analysis prove that the two procedures give the same results: how could it? Instead, causal inferences are made by *assuming* that selection tells you what would happen if you were to intervene.

The phrase "X is exogenous" is often taken to mean that selecting on X will produce the same results as intervening to set the value of X-- the basic assumption in many analyses. Exogeneity also has weaker meanings, to be taken up later. The ambiguity is unfortunate, because analysts may assume exogeneity in a weak sense, and proceed as if they had established something more. It is only exogeneity in the strong sense defined above that enables you to predict the results of interventions from non-experimental data.

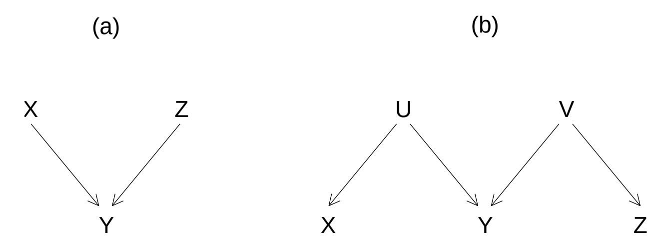
The distinction between selection and intervention is acknowledged even by the modelers:

> Formally speaking, probabilistic analysis is indeed sensitive only to covariations, so it can never distinguish genuine causal dependencies from spurious correlations.... (Pearl 1988, 396)

Such admissions-- like Yule's footnote 25-- are fatal to the enterprise. Of course, Pearl does not give up. For instance, he goes on to say that experiments just provide the opportunity to observe yet more correlations, a move he attributes to (Simon 1980).

Figure 2 is Pearl's. On the left, it seems that *X* and *Z* cause *Y*; manipulating *X* or *Z* will change *Y*. However, if only we had measured the variables *U* and *V*, we might have seen that they were the joint causes of *X*, *Y* and *Z*, as in the right hand panel. If so, manipulating *X* and *Z* will not change *Y* at all. No amount of statistical analysis on the observables-- on *X*, *Y* and *Z*-- can tell us which panel expresses the right theory. Indeed, matters can be arranged so that both theories lead to the same joint distribution for the observables.

Figure 2. After Judea Pearl (1988, 397). Causation cannot be inferred from association by using causal models.



Notes: In panel (a), *X* and *Z* are assumed to be independent. In panel (b), *U* and *V* are assumed to be independent; it may be shown in consequence that *X* and *Z* are independent. Also see (Duncan 1975, 113–27).

4. Two ideas of conditional probabilities

The distinction between the two ideas of conditioning-- selecting subjects with $X = x$, or intervening to set $X = x$-- seems fundamental. A concrete example may help, and conditional probabilities are easier to deal with than conditional expectations.

Many studies have demonstrated an association between cervical cancer and exposure to two sexually transmitted diseases-- herpes and chlamydia. Suppose we had data as shown in Table 1. The incidence rate of cervical cancer is 200 per 100,000 for women exposed to herpes and chlamydia (top left); 116 per 100,000 for women exposed to herpes but not chlamydia; and 130 per 100,000 for those exposed to herpes, the two exposure categories for chlamydia being combined. Other cells may be read in a similar way.

Table 1. Rate of cervical cancer cases per 100,000 women, by exposure to chlamydia and herpes. Data are hypothetical.

		Chlamydia		Marginal
		Yes	No	
Herpes	Yes	200	116	130
	No	180	80	87
Marginal		190	90	100

With sample data, there is a role for technical statistics in estimation and testing-- for instance to see if the rates within a row are constant across columns. However, the real question is not association but causation. Does herpes cause cervical cancer? what about chlamydia? Biotechnology might find a way to eliminate *Herpes simplex* as well as *Chlamydia trachomatis*. That would be a great relief, but would it reduce the incidence rate of cervical cancer?

To consider the issue of causality more directly, suppose that we actually know the rates for the population of interest, as shown in Table 1. Statistical testing must now fade into the background. The overall incidence rate is 100 cervical cancers per 100,000 women (Table 1, bottom right); among women exposed neither to herpes nor to chlamydia, the rate is lower-- 80 per 100,000. If cervical cancer is caused by herpes and chlamydia, eliminating the microorganisms responsible for those diseases should reduce the incidence rate of cervical cancer from 100 to 80 per 100,000. On the other hand, if the relationship is not causal, eliminating those microorganisms will have little effect on the incidence rate of the cancer.

To be more explicit, 80/100,000 has been found by selecting women who are exposed to neither herpes nor chlamydia, and computing the incidence rate of cervical cancer for that group: one interpretation of conditional probability. If we intervene and eliminate the two diseases, we want to know the rate after the intervention: that is another interpretation. The two interpretations are different, because the underlying procedures are different. Statistical analysis of the numbers in the table, however refined or complex, cannot prove that a hypothetical intervention will give the same results as selection. This may seem obvious, even banal; but if you grant the point, the causal modeling game is largely over.

What is the situation for Table 1? The story is far from certain. Current epidemiological opinion favors the idea that cervical cancer is caused by certain strains of human papilloma virus (HPV); herpes and chlamydia have no etiologic role, but serve only as markers for exposure to HPV. If that opinion is correct, wiping out herpes and chlamydia will have no impact on rates of cervical cancer.

Due in part to the rarity of cervical cancer, cohort studies do not seem to be available. (The numbers in Table 1, although hypothetical, are not unreasonable.) My point is even stronger for the real studies of the association between cervical cancer and herpes or chlamydia. Problems created by incomplete data cannot simplify the task of inferring causation from association.[8]

5. Another regression example

Rindfuss et al. (1980) propose a model to explain the process by which a woman decides how much education to get, and when to have her first child. The model illustrates many features of contemporary technique.[9] Before we take up the model, let the authors say what they were trying to do:

> The interplay between education and fertility has a significant influence on the roles women occupy, when in their life cycle they occupy these roles, and the length of time spent in these roles.... This paper explores the theoretical linkages between education and fertility.... It is found that the reciprocal relationship between education and age at first birth is dominated by the effect from education to age at first birth with only a trivial effect in the other direction. [Abstract]
>
> No factor has a greater impact on the roles women occupy than maternity. Whether a woman becomes a mother, the age at which she does so, and the timing and number of subsequent births set the conditions under which other roles are assumed.... Education is another prime factor conditioning female roles. [p. 431, footnote omitted]

---

[8] For a discussion of the epidemiology, see (Cairns 1978, Peto and zur Hausen 1986, Sherman et al. 1991, Hakama et al. 1993).

[9] I use this example because it is discussed by SGS pp. 139–40; also see CG pp. 14–15.

> The overall relationship between education and fertility has its roots at some unspecified point in adolescence, or perhaps even earlier. At this point aspirations for educational attainment as a goal in itself and for adult roles that have implications for educational attainment first emerge. The desire for education as a measure of status and ability in academic work may encourage women to select occupational goals that require a high level of educational attainment. Conversely, particular occupational or role aspirations may set standards of education that must be achieved. The obverse is true for those with either low educational or occupational goals. Also, occupational and educational aspirations are affected by a number of prior factors, such as mother's education, father's education, family income, intellectual ability, prior educational experience, race, and number of siblings. [p. 432, citations omitted]

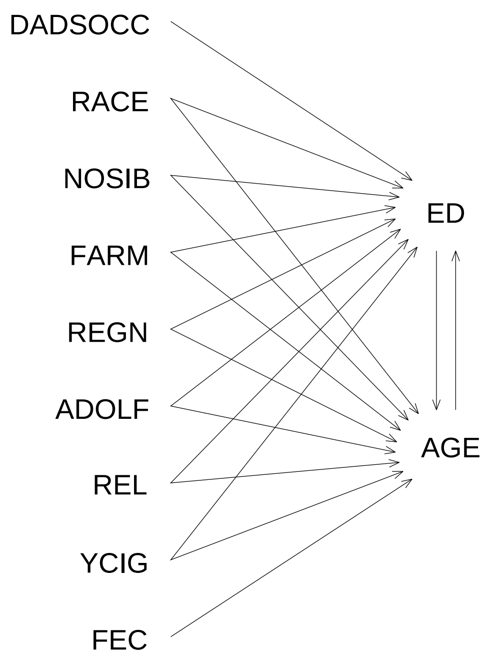
The model used by Rindfuss et al. (1980) is shown in Figure 3. The diagram corresponds to two linear equations in two unknowns, ED and AGE (variables are defined in Table 2):

(3) $\mathrm{ED} = a \times \mathrm{AGE} + A,$

(4) $\mathrm{AGE} = a' \times \mathrm{ED} + A'.$

According to the model, a women chooses her educational level and age at first birth as if by solving these two equations for the two unknowns.

Figure 3. The model in diagram form (Rindfuss et al. 1980; SGS p. 140; CG p. 15). Variables are defined in Table 2 below. Explanatory variables (DADSOCC, RACE, etc.) are correlated; error terms are not shown in the diagram.



The coefficients a and a' are "social constants," to be estimated from the data. The terms A and A' take background factors into account:

(5) $$A = A_0 + b \times \text{DADSOCC} + c_1 \times \text{RACE} + \ldots + c_7 \times \text{YCIG}$$
$$+ \text{ random error drawn from a box,}$$

(6) $$A' = A_0' + b' \times \text{FEC} + c_1' \times \text{RACE} + \ldots + c_7' \times \text{YCIG}$$
$$+ \text{ another random error drawn from a box.}$$

Again, the parameters A_0, b, c_1, ... are social constants to be estimated from the data. The random errors are assumed to have mean 0, to be statistically independent from woman to woman, and to be identically distributed. Correlations across equations (5) and (6) are permitted.

Equations (3-6) are not quite regression equations, due to the simultaneity of (3) and (4); fitting by OLS (ordinary least squares) would create "simultaneity bias." Thus, Rindfuss et al. use an estimation procedure called "two-stage least squares."[10] FEC does not come into equation (5), nor DADSOCC into equation (6). Graphically, there is no arrow from DADSOCC to AGE in Figure 3; likewise, there is no arrow from FEC to ED. These behavioral assumptions are critical to the statistical enterprise. Without them, or some similar assumptions, two-stage least squares could not be used. Technically, the system would not be "identifiable" (section 11.4).

The main empirical finding: The estimated coefficient of AGE in the first equation is not "statistically significant," i.e., the coefficient *a* in (3) could be zero. The sort of woman who drops out of school to have a child would drop out anyway.

Table 2. Variables in the model (Rindfuss et al. 1980).

The endogenous variables

ED	Respondent's education (Years of schooling completed at first marriage)
AGE	Respondent's age at first birth

The exogenous variables

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)

Notes: The data are from a probability sample of 1,766 women 35-44 years of age residing in the continental United States; the sample was restricted to ever-married women with at least one child. DADSOCC was measured on Duncan's scale, combining information on education and income; missing values were imputed at the overall mean. SGS (p. 139) give the wrong definitions for NOSIB and ADOLF.

---

[10] See, e.g., (Maddala 1992); for discussion, see (Daggett and Freedman 1985).

If looked at coldly, the argument may seem implausible. A critique can be given along the following lines.

(i) Statistical assumptions. Just why are the errors independent and identically distributed across the women? Independence may be reasonable, but heterogeneity is more plausible then homogeneity.

(ii) The assumption of constant coefficients. Rindfuss et al. are assuming that the same parameters apply to all women alike, from poor blacks in the cities of the Northeast to rich whites in the suburbs of the West. Why?

(iii) Omitted variables. Surely, important variables have been omitted from the model, including two that were identified by Rindfuss et al. themselves-- aspirations and ability. Malthus thought that wealth was an important factor. Social class matters, and DADSOCC measures only one of its aspects.[11]

(iv) What about the "no arrow" assumptions, from DADSOCC to AGE and FEC to ED?

(v) Are FEC and DADSOCC exogenous?

(vi) Are the equations "structural"?

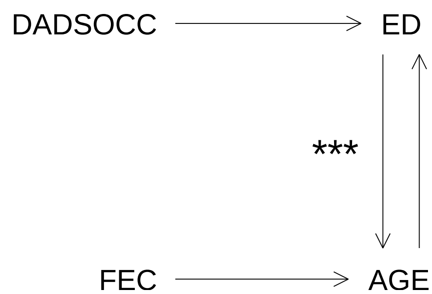
Questions (iv-vi) will be discussed in the next section, as will the idea of "structural" equations.

---

[11] The solution to the "omitted variable" problem may seem easy-- just throw some more variables into the model. The difficulties are explored in (Clogg and Haritou 1994). Also see (Freedman 1983).

5.1 A thought experiment

Figure 4. A simpler version of the model.



A simpler version of the model restricts attention to a more homogenous group of women, where the only relevant background factors are DADSOCC and FEC. To make causal inferences from the data using the model, we need to believe that the arrows are as shown in Figure 4, that DADSOCC and FEC are exogenous, and that the equations are "structural." The following thought experiment may help to define the last term, and the empirical commitments behind the words.[12]

The *gedanken* experiment involves two groups of women. In both groups, fathers are randomized to jobs, and some of the daughters are chosen at random to have a miscarriage before their first child. (The statistical terminology of randomization is dry; the *gedanken* experimentalist intervenes, for instance, to make the fathers do one job rather than another: professors are caused to work as plumbers, and taxi drivers are installed as hospital anesthetists.)

> *Group I.* Daughters are randomized to the various levels of ED, and AGE is observed as the response. (The *gedanken* experimentalist strikes again, forcing some women stay in school longer than they wish, while preventing others from continuing their education.)

> *Group II*. Daughters are randomized to the various levels of AGE, and ED is observed as the response. (More *gedanken* intervention is needed.)

---

[12] Also see (Pearl 1994ab).

The statistical model can now be translated. For the women in Group I, AGE should not depend on DADSOCC-- the "no arrow" assumption; however, AGE should depend linearly on ED. For the women in Group II, ED should not depend on FEC-- the other "no arrow" assumption; however, ED should depend linearly on DADSOCC. Rindfuss et al.'s discovery is that ED would not depend on AGE.

There is one final assumption: the equations and parameters that describe the responses of the women in the experiment must also describe the natural situation. That is what "structural" means. For instance, a woman who freely chooses her educational level and her time to bear children does so by using the same equations as a woman made to give birth at a certain age. In short, with respect to the matters at issue, life in Des Moines proceeds more or less along the same lines as life in the Gulag.

The thought experiment provides the intellectual foundation for the model, by articulating the background assumptions. These assumptions have not been subjected-- cannot be subjected-- to direct empirical proof. Nor can assumptions be validated by appealing to thought experiments that are almost unthinkable. Do the modelers have some other method in reserve? If the assumptions remain unvalidated, what is the logical status of their implications?

5.2 Exogeneity

Identifying the exogenous variables is a major problem. For example, you can obtain results quite different from those of Rindfuss et al., by using variables other than DADSOCC and FEC as "instruments."[13]

Rindfuss et al. respond that estimates made by

> instrumental variables... require strong theoretical assumptions... and can give quite different results when alternative assumptions are made... it is usually difficult to argue that behavioral variables are truly exogenous and that they affect only one of the endogenous variables but not the other. (Rindfuss et al. 1984, 981-82).

In short, results can depend quite strongly on assumptions of exogeneity, and there is no good way to justify one set of assumptions rather than another. Also see (Bartels 1991), who comments on the impact of exogeneity assumptions, and the difficulty of verification.

6. Automated searches for causality

SGS (Spirtes, Glymour, and Scheines 1992) have computerized algorithms that search for path models. Using the algorithms, SGS claim to make rigorous inferences of causation from association. Their theory is summarized in (Glymour 1993), cited as CG. For present purposes, a "path model" is a recursive system of regression equations, in which the dependent variables

---

[13] See (Hofferth and Moore 1979, Moore and Hofferth 1980). An "instrument" is an exogenous variable, used as part of the two-stage least squares estimation procedure. Some investigators may draw a terminological distinction: an "instrument" is exogenous, but does not appear as an explanatory variable in the equation being estimated. For purposes of estimation, exogenous variables are assumed to be independent of error terms; this does not suffice for causal inference (section 11). Even the independence assumption is not to be made lightly (Clogg and Haritou 1994).

from some equations are used as explanatory variables in later equations.[14]

The basic idea in path models is this: putative causes combine with parameters and random errors by multiplication and addition in order to produce their effects. I have discussed such models elsewhere, and do not believe they offer much help in deducing causation from association, because there is little evidence to support the basic assumptions (Freedman 1987). To pursue the discussion here, a slightly more explicit definition of the models may be in order.

Definition. A "path model" starts with variables at "level 0," which are exogenous in the minimal sense that they are not explained within the model. Variables "at level 1" are built up as linear combinations of level 0 variables, plus independent random errors. More generally, variables "at level k" are built up as linear combinations of variables at previous levels; again, there are additive, independent random errors. Variables at level 1, level 2, ... are "endogenous," in the sense that they are explained within the system. The path model may be presented as a "path diagram," like Figure 1, or Figure 5 below. Nodes represent variables in the model; if there are arrows from X, Y,... to Z, then X, Y,... are explanatory variables in the regression equation for Z. Nodes are often called "vertices," and the diagrams are referred to as "graphs" or "causal graphs."[15]

---

[14] The model used by Rindfuss et al. would not fall into this category, if ED and AGE really influenced each other. The SGS framework excludes reciprocal causation, by assumption; so do path models, as I define them. However, some authors extend the definition of path models to include simultaneous equation models for reciprocal causation.

[15] SGS seem to make the strong -- and quite unusual-- assumption that exogenous variables are independent of each other. That may be part of the reason why their algorithms estimate such peculiar models in Figures 5 and 6 below. There is another, even more esoteric, point. To estimate an equation, its error term need only be assumed independent of the explanatory variables. If so, error terms from different equations may be correlated; then standard procedures for computing the correlations among the variables will not apply (Freedman 1987, 112-14; Seneta 1987, 199). SGS seem to interpret correlated errors as indicating the presence of "latent variables." Such variables will be mentioned in notes to Figures 5 and 6, below.

The path model may represent mere association-- conditional dependence and independence relations. Or the model may represent causation. I will take that up later. For now, however, either interpretation suffices. Suppose the graph is "sparse"-- each equation in the model involves relatively few variables. Suppose too there are no troublesome algebraic identities among the regression coefficients; in SGS terminology, the distribution is "faithful" to its graph (SGS p. 35; CG p. 9; and see section 11.2 below). You have a sample-- many independent realizations of variables *X, Y, Z*,.... You are willing to assume the distribution conforms to a path model, but do not know which model. You do not even know which variables are at level 0, which are at level 1, and so forth.

SGS claim their algorithms are likely to find the underlying path model, or a rather similar model, and in short order. Their most convincing evidence is based on simulation experiments, where the computer generates data from a path model and the SGS algorithms try to infer the model from the data (SGS pp. 145ff, 152ff, 250ff, 320ff, 332ff); in these experiments, the algorithms do very well. Roughly speaking, the SGS algorithms are variants of "best subsets" regression, the search being over graphs rather than subsets. The data come into the SGS algorithms only through the covariance structure. The rest of the apparatus-- the diagrams, the Markov property, faithfulness, etc.-- consists of assumptions.

SGS seem to assert that their algorithms determine causality, as a matter of mathematics. Such assertions are not defensible. In the SGS formalism, causation is obtained not by mathematical proof but by mathematical assumption. If you assume that the arrows in the underlying path diagram represent causes, then the arrows found by the algorithms represent causes. If you assume that the underlying arrows represent mere associations, then the arrows found by the algorithms represent associations. Causation has to do with empirical reality, not with mathematical proofs based on axioms. The issue is not one of theorems, but of the connection between theorems and reality.

The SGS algorithms, like many earlier statistical procedures (factor analysis, LISREL, etc.), proceed by analyzing the correlation matrix of a set of variables. I will call such methods "correlational." Sections 7-10 consider applications of the SGS algorithms to real examples. Sections 11-12 try to explain the key ideas in the SGS formalism, and indicate by mathematical example some of the intrinsic limitations. Before proceeding, however, I discuss the SGS statement of assumptions.

6.1 The SGS statement of assumptions

SGS discuss the role of assumptions in their theory several times (pp. 53-69, pp. 75-81, pp. 324-5, p. 351). However, the clearest statement can be found when SGS are trying to discredit the evidence that smoking causes lung cancer:

> effects **** cannot be predicted from **** sample conditional probabilities. [p. 302]

Readers may consult the original for context, to see whether the omitted material affects the meaning. The advantage of the quote is clarity. If the statement is generally applicable, then SGS-- like Yule and Pearl before them-- have disavowed the ability to infer causation from association.

7. The SGS examples

SGS share my pessimistic views about regression. They claim, however, that their algorithms will succeed where regression has failed:

> In the absence of very strong prior causal knowledge, multiple regression should not be used to select the variables that influence an outcome or criterion variable in data from uncontrolled studies. So far as we can tell, the popular automatic regression search procedures [like stepwise regression] should not be used at all in contexts where causal inferences are at stake. Such contexts require improved versions of algorithms like those described here to select those variables whose influence on an outcome can be reliably estimated by regression. In applications, the power of the specification searches against reasonable alternative explanations of the data is easy to determine by simulation.... [p. 257]

At first reading, SGS seems to be filled with real examples showing the successful application of their algorithms. That is an illusion. Many of the examples are based on simulation, and I set those aside.[16] The real examples are mostly to be found on pp. 132–52 and 243–256.[17]

---

[16] Simulations tell us how well the SGS algorithms do *if* the underlying statistical assumptions hold good; the assumptions are built into the computer code that generates the simulated data. When applying statistical algorithms to real data, a critical question is *whether* those assumptions hold. The simulations do not address such questions.

[17] The parallel material in CG is on pp. 13–16 and 21–23.

The main examples given in SGS are path models. But these cannot withstand scrutiny-- section 5 above, sections 8-9 below. One exception is the stratification model of Blau and Duncan (1964). SGS (pp. 142-5) and CG (pp. 21-22) seem to be quite critical of this model; their current position is almost diametrically opposite to the one in (Glymour et al. 1987, 33-9). Like SGS, I do not believe that the Blau-Duncan regressions are a satisfactory causal model. On the other hand, as descriptions of the data, the equations can tell us something important about our society (Freedman 1987, 122, 220). The discussion in SGS adds little to our understanding either of the model or of stratification.

SGS appear to use the health effects of smoking as a running example to illustrate their theory.[18] Again, there is an illusion. The causal diagrams are all hypotheticals, no contact is made with data, and no substantive conclusions are drawn. If the diagrams were proposed as real descriptions of causal mechanisms, they would be open to devastating criticism.

What about the substantive question: does smoking cause lung cancer, heart disease, and many other illnesses? SGS appear not to believe the epidemiological evidence. When they actually get down to arguing their case, they use a rather old-fashioned method-- a literature review with arguments in ordinary English (pp. 291-302). Causal models and search algorithms have disappeared.

I approve of the method if not the implementation: the summary is wrong in some places and tendentious in others. However, the review does show the complexity of the issues. To make judgments about causation, you need to consider death certificate data, necropsy data, case control and cohort studies, twin studies, dose response curves, as well as animal experiments and human experiments. The force of the epidemiological evidence-- and the SGS critique-- depends on the complex interplay among these various studies and data sets.

---

[18] See, e.g., SGS p. 18 and pp. 216-37; also see CG pp. 19-20 and 30-31.

In the end, SGS and CG do not really make bottom-line judgments on the health effects of smoking, at least so far as I can see. Their principal conclusion is methodological: nobody understood the issues.

> Neither side understood what uncontrolled studies could and could not determine about causal relations and the effects of interventions. The statisticians pretended to an understanding of causality and correlation they did not have; the epidemiologists resorted to informal and often irrelevant criteria, appeals to plausibility, and in the worst case to *ad hominem*.... While the statisticians didn't get the connections between causality and probability right, the .... 'epidemiological criteria for causality' were an intellectual disgrace, and the level of argument ... was sometimes more worthy of literary critics than scientists. [pp. 301-2]

Part of a sentence in SGS (p. 4) does seem to grant one of the major claims made by the epidemiologists, "smoking does cause lung cancer." But that only complicates the puzzle. If you don't believe the evidence, why accept the claim?

Despite SGS, the epidemiologists did have a good understanding of the issues, and made a strong case against smoking. The arguments were imperfect, and some reasonable doubts may remain. But the data, taken all in all, are compelling. The epidemiological literature on smoking is far stronger than anything I have seen in the social sciences. For a survey of the evidence, see (Cornfield et al. 1959); this paper is still worth reading. More recent data are reviewed in (International Agency for Research on Cancer 1986).

SGS elected not to use their analytical machinery on the smoking data-- a remarkable omission. When applied to the examples that SGS actually chose, the algorithms produce one small disaster after another, as will now be seen. In sum, SGS claim to have developed techniques for generating causal models; but they do not have any success stories.

8. Using the SGS search procedure

The SGS search procedures are embodied in a computer program called TETRAD. Version 2.1 of this program was kindly provided by Richard Scheines and Peter Spirtes. The BUILD module is the part of TETRAD used to discover path models with no latent variables. I ran BUILD on two examples-- Rindfuss et al. and AFQT (to be discussed in section 9).

8.1 Rindfuss et al.

To explain AGE (age at first birth) in the Rindfuss et al. example, the SGS algorithms select the variables shown in Table 3. Regression estimates for the coefficients, based on summary data in SGS, are reported in the first three columns of the table. The coefficients for ADOLF (the indicator for women from broken homes) and YCIG (an indicator for smoking by age 16) have positive signs. That is paradoxical: women from broken homes and women who smoke should be having children earlier, not later.[19] The signs should be negative, not positive. SGS do not comment on this issue.

Rindfuss et al. (1980) give standard deviations and correlations for their data; SGS (p. 139) used these statistics to compute a covariance matrix, but reversed some of the signs. The last three columns of Table 3 report regression estimates computed from the correct covariances. The problem with YCIG disappears, but the sign for ADOLF stays positive. Anyone can make a mistake entering data; ignoring paradoxical signs in a causal model is quite another matter.

---

[19] Smoking, broken homes, and early childbearing seem to be correlates of social disadvantage, and indicators of personality traits. DADSOCC and RACE are quite imperfect controls for family background; therefore, YCIG and ADOLF are likely to pick up effects of background, as well as effects of omitted personality variables. See note (iv) to Table 3. This sort of bias is discussed in section 12.2 below. Also see Clogg and Haritou (1994).

Table 3. The SGS model for age at first birth, computed using the SGS covariance matrix or the Rindfuss et al. covariance matrix. (Intercepts are not reported; OLS estimates.)

	SGS covariance			Rindfuss et al. covariance		
	$R^2 = 0.27$			$R^2 = 0.24$		
	Estimate	SE	t	Estimate	SE	t
RACE	−1.66	.30	−5.50	−1.66	.30	−5.46
REGN	−0.56	.19	−3.01	−0.63	.19	−3.35
ADOLF	1.89	.22	8.60	2.01	.22	8.98
YCIG	2.14	.25	8.63	−0.89	.25	−3.53
FEC	2.72	.28	9.70	2.77	.28	9.72
ED	0.67	.04	18.00	0.60	.04	15.72

Notes: (i) The first column in Table 3 shows parameter estimates. The second shows standard errors, or SEs, which indicate the likely size of the differences between the estimates and the true parameter values. The *t*-statistics in the third column are the ratios of estimates to SEs. Generally, a *t*-statistic above 2 or 3 in absolute value indicates that the corresponding parameter is unlikely to be truly 0. For details, see the appendix.

(ii) The parameters are features of the model, and the SEs are computed using the model. If you do not believe in the existence of the parameters apart from the data, or do not accept the statistical assumptions in the model, the SEs and *t*-statistics are likely to be meaningless. In any case, performing multiple tests-- as in a search algorithm-- complicates the interpretation of the *t*-statistics (Freedman 1983; CG p.45).

(iii) R^2 is generally interpretable as a descriptive statistic, whether or not the assumptions of the model hold true. An R^2 of 0.27 indicates that about 27% of the variance in AGE has been explained; that isn't much, and models in the social science literature often have even less explanatory power. For a critical discussion of R^2, see (Freedman and Lane 1981, 78–81).

(iv) According to current epidemiological opinion, smoking does have some biological effect, delaying conception by several weeks. However, the women who choose to smoke are different from the non-smokers, and have their first child almost a year earlier. This effect remains even after controlling for the measured background factors in the regression: the coefficient of YCIG is -0.89 years.

SGS report only a graphical version of their model. They say,

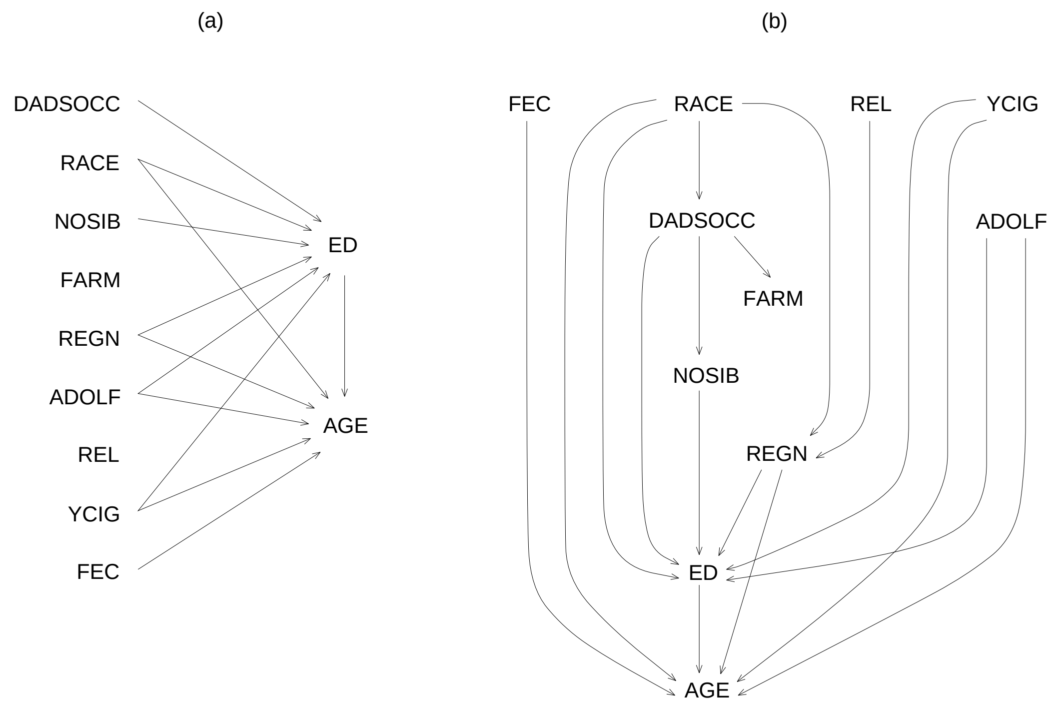
> Given the prior information that ED and AGE are not causes of the other variables, the PC algorithm (using the .05 significance level for tests) directly finds the model [in Figure 5(a)] where connections among the regressors are not pictured. (SGS p. 139; CG p.15)

However, connections among regressors can be of interest. Although TETRAD is supposed to discover the causal ordering of explanatory variables, it produces the very strange model shown in Figure 5(b). For example, the model says that race and religion cause region of residence. Comments on the sociology may be out of place, but consider the statistics. The equation is

$$\text{(7)} \qquad \text{REGN} = a + b \times \text{RACE} + c \times \text{REL} + \varepsilon.$$

REGN is a dummy variable, coded 1 for respondents who grew up in the South, 0 for others; RACE is 1 for black respondents and 0 for others; REL is 1 for Catholics, 0 for others; ε is normally distributed. In consequence, this equation forces impossible values on REGN: the left hand side is 0 or 1, the right hand side varies from $-\infty$ to $+\infty$. Now R^2 is only 0.16, so ε contributes most of the variance: equation (7) can hardly be defended as an approximation. Having dummy variables in the middle of path diagrams is a blunder. (FARM creates a similar problem; so does NOSIB, although less extreme.) In short, the SGS algorithms have produced a model that fails the most basic test-- internal consistency.

Figure 5. The left hand panel shows the model reported by SGS. The right hand panel also shows connections among the regressors, as determined by the SGS search program TETRAD.



Notes: BUILD indicates that latent variables are present, i.e., errors are correlated across equations. BUILD asks whether it should assume "causal sufficiency"; without this assumption (SGS p. 45; CG p. 4), the program output is uninformative. Therefore, I told BUILD to make the assumption; I believe that is what SGS did for the Rindfuss example. Also see (Spirtes et al. 1993, 13–15). I told BUILD that ED and AGE could not cause the remaining variables, following (SGS p. 139). However, SGS actually made the stronger assumption that (i) FEC, ED and AGE could not cause YCIG, and (ii) FEC, ED, AGE and YCIG could not cause the remaining variables. With the assumption of causal sufficiency, BUILD seems to use the PC algorithm; without the assumption, the FCI algorithm comes into play. Much of this information comes from Richard Scheines (personal communication). Data are from (Rindfuss et al. 1980), not SGS; with the SGS covariance matrix, FARM causes REGN and YCIG causes ADOLF.

9. The Armed Forces Qualification Test

SGS discuss an example based on the Armed Forces Qualification Test (AFQT).[20] The AFQT is a linear combination with fixed weights of scores on certain subtests. Some of these subtests, as well as subtests that are not part of the AFQT, are listed in Table 4. The problem is to decide which subtests go into the AFQT and which do not.

Table 4. Subtests analyzed by SGS.
Some go into the AFQT and some do not.

1.	Numerical Operations	NO
2.	Word Knowledge	WK
3.	Arithmetical Reasoning	AR
4.	Mathematical Knowledge	MK
5.	Electronics Information	EI
6.	Mechanical Comprehension	MC
7.	General Science	GS

The problem may be stated more algebraically, as follows:

$$\text{(8)}\quad \text{AFQT score} = a_1 \times \text{NO} + a_2 \times \text{WK} + \ldots + a_7 \times \text{GS} + b_1 \times \text{UN}_1 + \ldots + b_n \times \text{UN}_n,$$

where $\text{UN}_1, \ldots, \text{UN}_n$ are unobservable. Some of the a's are zero, and the challenge is to figure out which ones.

We have data on 6,224 subjects, summarized as a covariance matrix. SGS say:

> a linear multiple regression of AFQT on the other seven variables gives significant regression coefficients to all seven and thus fails to distinguish the tests that are in fact linear components of AFQT... Given the prior information that AFQT is not a cause of any of the other variables, the PC algorithm in TETRAD II correctly picks out {AR, NO, WK} as the only ... variables that can be components of AFQT.... (SGS pp. 243–4, also see CG p. 16)

---

[20] SGS p. 243, also see CG p. 16. Institutional background on the AFQT will be found in section 12.5.

Table 5. Regression of AFQT on all the observable subtests.

	Estimate	SE	t
NO	0.24	.022	10.8
WK	1.17	.029	40.5
AR	1.03	.028	36.4
MK	−0.24	.028	−8.7
EI	−0.03	.024	−1.3
MC	0.03	.024	1.3
GS	−0.13	.029	−4.6

Note: Variables were centered at their means.

To test the claims about regression, I ran AFQT on all the observable subtests. As Table 5 shows, EI and MC are related to AFQT only at the chance level. Moreover, MK and GS have negative coefficients, but psychometric practice frowns on subtests that are negatively related to overall test scores. It is a natural conjecture that NO, WK and AR go into AFQT while the other four subtests do not. Contrary to the claims of SGS, the AFQT can be handled by ordinary statistical methods.

The AFQT problem is in some ways quite easy. By definition, the "causes" or subtests combine linearly with the parameters to produce the AFQT as an "effect." Joint normality of test scores seems to follow from the procedures used to construct the tests: consequently, scores on any one subtest can be presented as a linear combination of other subtest scores, with additive random errors. Thus, critical issues in most empirical studies have disappeared.[21]

---

[21] On the other hand, unobserved variables may create serious problems (section 12.4).

9.1 TETRAD

According to SGS, given the prior information that AFQT does not cause the other variables, TETRAD correctly picks out AR, NO, and WK as the components of the AFQT.[22] Without that prior information, however, TETRAD declares AFQT to be the *cause* of these subtests, rather than the *effect*. With the prior information, TETRAD produces the strange results shown in Figure 6.[23] Now, for instance, the subtest NO may "cause" the overall test score AFQT, but it can hardly cause the other subtests AR or MK. Furthermore, there is a cycle in the figure:

MC $\rightarrow$ AR $\rightarrow$ WK $\rightarrow$ GS $\rightarrow$ MC.

In principle, such cycles were excluded by prior assumption, as well they might be. Subtests should not cause themselves, even indirectly. To sum up:

(i) ordinary least squares techniques pick out NO, AR, and WK for the probable components of the AFQT, just as TETRAD does;

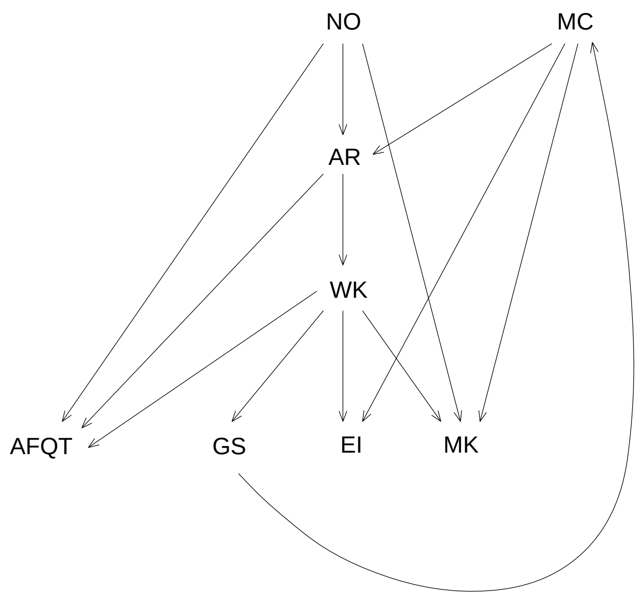
(ii) TETRAD produces the curious model in Figure 6.

---

[22] SGS p. 243, also see CG p. 16.

[23] The program output is given in (Spirtes et al. 1993, 10–11).

Figure 6. AFQT and its subtests arranged in causal order by the SGS search program TETRAD.



Notes. I believe SGS used BUILD, with the assumption of causal sufficiency, on pp. 243–44 for the AFQT example. Also see (Spirtes et al. 1993, 8–11). The program indicates there are latent variables, i.e., correlations in the errors.

10. Foreign investment and political oppression

As noted in section 7, SGS are quite pessimistic about typical social-science applications of regression. While I agree with the bottom line, their specific objections seem misplaced. One example is enough to make the point. Timberlake and Williams (1984) offer a regression model to explain political exclusion (PO) in terms of foreign investment (FI), energy development (EN) and civil liberties (CV). High values of PO correspond to authoritarian regimes that exclude most citizens from political participation; high values of CV indicate few civil liberties. Data come from 72 countries. Correlations among the Timberlake-Williams variables are shown in Table 6.

Table 6. The Timberlake and Williams data. Correlation matrix for political oppression (PO), foreign investment (FI), energy development (EN), and civil liberties (CV). Source: SGS p. 249.

	PO	FI	EN	CV
PO	1.000	−.175	−.480	.868
FI	−.175	1.000	.330	−.391
EN	−.480	.330	1.000	−.430
CV	.868	−.391	−.430	1.000

The equation proposed by Timberlake and Williams is

(9) $$\text{PO} = a + b \times \text{FI} + c \times \text{EN} + d \times \text{CV} + \text{error}.$$

Empirical results are shown in the first three columns of Table 7. The estimated coefficient of FI is significantly positive, and is interpreted as measuring the effect of foreign investment on political exclusion (Timberlake and Williams 1984, 143).

Table 7. The Timberlake and Williams model. Political exclusion (PO) is regressed on foreign investment (FI), energy development (EN), and civil liberties (CV). The first three columns show results for the observed correlation matrix (Table 6). The last three columns show what happens when r(PO,FI) is set to 0.

	$R^2 = .81$			$R^2 = .93$		
	Estimate	SE	t	Estimate	SE	t
FI	.23	.059	3.9	.44	.036	12
EN	−.18	.060	−2.9	−.22	.037	−6
CV	.88	.061	14.4	.95	.038	25

Note: The coefficients reported by SGS on p. 249 are not standardized and therefore do not match the correlation matrix. Coefficients in Table 7 are standardized, that is, computed from variables standardized to have mean 0 and variance 1.

SGS discuss this example (pp. 248–50), suggesting that Timberlake and Williams have confused cause and effect. The alternative causal sequence is not spelled out. Presumably, the idea is that dictators 'cause' foreign investment in the sense that investors think dictatorial regimes offer greater stability, etc.

The main step in the SGS statistical argument comes down to this: the correlation of −.175 between political exclusion and foreign investment is at the chance level. The calculation rides on two assumptions: (i) the 72 countries in the data set are a random sample from some much larger set of countries, and (ii) the variables follow a multivariate normal distribution. These time-honored but madcap assumptions are not stated explicitly by SGS, let alone justified. (Of course, the assumptions behind the statistics in Timberlake-Williams might seem equally antic.)

However, for the sake of argument, let us grant SGS their assumptions. On that basis, the standard error for the correlation in question is about $1/\sqrt{72} \approx .12$. I change the suspect correlation coefficient from its observed value of –.175 to the new value of 0, a difference of about 1.5 SEs. I then recompute the model (last three columns in Table 7). The results are even better for Timberlake and Williams: the estimated coefficients are bigger and more significant; the signs stay the same; and R^2 moves closer to 1.[24]

I will not defend the model any further. Measurement problems are extreme, and the list of omitted variables very long. SGS may well be right, that cause and effect have been confused. But the demonstration is peculiar. The correlation matrix cannot show that FI, EN and CV cause PO-- the fatal flaw in the Timberlake-Williams model. (Of course, Timberlake and Williams are not alone in this respect.) Nor can the matrix show that FI, EN and CV do not cause PO-- the corresponding flaw in SGS. Indeed, it is trivial to construct four variables labelled FI, EN, CV and PO, such that FI, EN and CV do cause PO; but sample correlation matrices will look rather like the one in Table 6. This only sharpens the basic question. What do any of these calculations tell us about the world outside the computer?

---

[24] The new matrix is still positive definite, so it is a legitimate correlation matrix. Section 12.1 discusses the connection between the Timberlake-Williams model and the faithfulness assumption. Also see (Cartwright 1989, 79-84).

11. Some mathematical issues

Sections 11 and 12 address by mathematical example two questions:

> (i) To what extent can correlational methods recover an underlying path diagram?
>
> (ii) When can the arrows in the diagram be interpreted as indicating causation, rather than conditional independence and dependence?

The examples will indicate how SGS use the "faithfulness" assumption to help them answer such questions. Issues of identifiability and consistency will be discussed, and methodological contributions in SGS will be delineated. Sections 11 and 12 are more technical than previous material; readers can skip to section 13 without losing the thread of the argument.

The focus is on linear models. Suppose you have a covariance matrix that describes certain variables. Assume these variables are jointly normal, with mean 0; that avoids all questions of linearity, etc., and all problems created by having only finite amounts of data. However, the statistical procedures I am considering-- like the SGS algorithms-- will operate on that covariance matrix, and on nothing else. Such procedures may be called "correlational."

Path models were defined in section 6. Briefly, you start with variables at level 0; variables at level k are linear combinations of variables at lower levels, plus independent random errors. In a path diagram, nodes represent variables. There is an arrow from X to Y if X is used as an explanatory variable in the equation for Y.

Exogeneity is a critical concept. As indicated before, the term is used in at least three senses. The weakest definition is purely mechanical: exogenous variables are not explained within the model, but are supplied to the model. Variables at level 0 in a path model are exogenous in this minimal sense. A more restrictive definition: exogenous variables are statistically independent of the error terms in the equations. The third idea is the one that is relevant to causal inference: X is exogenous if selecting subjects with $X = x$ gives the same results as intervening to set $X = x$.

There are tests for exogeneity in the literature, as well as model specification tests. However, these have limited relevance to causal inference. For example, (Hausman 1978) assumes that certain variables are known *a priori* to be exogenous, and then tests whether other variables are exogenous; he interprets exogeneity as orthogonality to disturbance terms. He also has a test that detects correlation between errors from equations in a path model. White (1980 and 1982) focuses on similar issues--for instance, testing whether the variables have a jointly normal distribution.

Another reference in the econometric literature is (Engle, Hendry, and Richard 1983). These authors distinguish several kinds of exogeneity; "strict" exogeneity means independence of variables and error terms, but only "super" exogeneity permits estimating the effects of interventions. Examples are given to illustrate the definitions (Engle, Hendry, and Richard 1983, 287-94). There is further discussion in (Leamer 1985).

11.1 The basic statistical problem

Suppose you have n random variables with a jointly normal distribution; all the variables have mean 0, and you know the covariance matrix, which is positive definite. You wish to present this covariance matrix as a path model. In a sense, nothing is easier. Simply order the variables, arbitrarily, as $X_1, X_2, \ldots, X_n$. By successively applying regression, we can find coefficients a_{ij} and error terms ε_i, such that $X_1, \varepsilon_2, \ldots, \varepsilon_n$ are all independent with mean 0, and equation (10) holds.

$$
\begin{aligned}
X_2 &= a_{21}X_1 + \varepsilon_2 \\
X_3 &= a_{31}X_1 + a_{32}X_2 + \varepsilon_3 \\
&\vdots \\
X_n &= a_{n1}X_1 + \ldots + a_{n,n-1}X_{n-1} + \varepsilon_n
\end{aligned}
\tag{10}
$$

Then X_1 is presented as exogenous and the "cause" of X_2; next, X_1 and X_2 "cause" X_3; and so forth. In short, there are many ways to present a covariance matrix as a path diagram; few if any will be relevant for causal inference.[25]

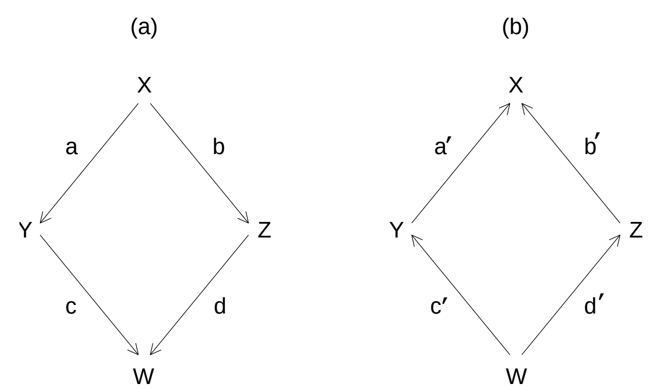
[25] For the construction in (10), simply choose a_{21} so $E\{X_2 | X_1\} = a_{21}X_1$; choose a_{31} and a_{32} so $E\{X_3 | X_1, X_2\} = a_{31}X_1 + a_{32}X_2$; and so forth. For details, see the appendix below.

Since the ordering of the variables in (10) is arbitrary, fitting such equations or drawing path diagrams cannot determine which variables are causes and which are effects. In particular, X_1 may be exogenous in the sense that it is statistically independent of disturbance terms; that by itself does not suffice to estimate the results of manipulating X_1, since we cannot tell whether X_1 is a cause or an effect.

11.2 The faithfulness assumption

How can you single out one path diagram from the many that correspond to a given covariance matrix? At this point, SGS seem to use the "faithfulness" assumption; this assumption is also used to handle confounding, as discussed in section 12.1 below. Basically, a covariance matrix is faithful to a diagram provided conditional dependencies and independencies are determined by the presence or absence of arrows in the diagram, rather than specific numerical values of parameters.

Figure 7. If two path diagrams have the same covariance matrix, correlational methods cannot tell them apart; the faithfulness assumption is made to rule out such problems. The lower case letters on the arrows denote "path coefficients," that is, standardized regression coefficients.



By way of example, Figure 7 shows two path diagrams. On the left, *X* causes *W* through the intervening variables *Y* and *Z*; on the right, the flow of causality is reversed.[26] The lower case letters on the arrows stand for "path coefficients," that is, standardized regression coefficients. How could SGS distinguish between the two theories in the figure? Their idea seems to be as follows:

> In the left hand diagram, *Y* and *Z* are conditionally independent given *X*; on the right, however, *Y* and *Z* are conditionally dependent given *X*.

Another contrast:

> In the left hand diagram, *Y* and *Z* are conditionally dependent given *W*; on the right, however, *Y* and *Z* are conditionally independent given *W*.

Therefore, the pattern of conditional dependence and independence identifies the diagram. (In both diagrams, *X* and *W* are conditionally independent given *Y* and *Z*.)

This idea works for many path diagrams, but fails for others. Indeed, the path coefficients can be chosen so the pattern of conditional dependence and independence is the same in the two diagrams. Even worse, both diagrams can give rise to the same covariance matrix-- so correlational methods cannot tell which is right. SGS make the "faithfulness assumption" in order to rule out such indeterminacies. (The workings of the assumption will be explained below.)

However, that only moves the difficulty to another place. Faithfulness is hardly an empirical fact; it is an assumption about unobservables, made to rule out situations that cannot be handled by correlational methods. The SGS analytical program can now be stated rather simply. If the arrows in a path diagram represent causation not association, and if the path diagram can be estimated from data, then SGS can indeed infer causation from association.

---

[26] In this section, I use "cause" in its ordinary (perhaps undefinable) sense. However, the technical point -- about the possibility of estimating path diagrams from covariance matrices -- still holds if the arrows are interpreted as merely representing association. "Causation" is then colorful shorthand (perhaps too colorful) for a certain kind of covariation.

The balance of section 11.2 provides technical backup; readers can skip to section 11.3. The left hand panel in Figure 7 is described by

(11) $Y = aX + \delta_1, \quad Z = bX + \delta_2, \quad W = cY + dZ + \delta_3.$

In this equation, X, δ_1, δ_2, δ_3 are independent and normal, with mean 0; X, Y, Z, W all have variance 1. The covariance matrix of X, Y, Z, W can be computed from the four parameters a, b, c, d as shown in (12).

(12)

	X	Y	Z	W
X	1	a	b	$ac+bd$
Y	a	1	ab	$c+abd$
Z	b	ab	1	$d+abc$
W	$ac+bd$	$c+abd$	$d+abc$	1

It is a little theorem, which follows by a tedious calculation from (48) in the appendix below, that

(13) $\mathrm{cov}(X, W \mid Y, Z) = 0.$

This is an example of a conditional independence relation forced by a graph; (13) holds whatever the path coefficients in Figure 7 may be.

The diagram on the left in Figure 7 is reversible provided

(14) $\mathrm{cov}(Y, Z \mid W) = 0.$

By (48) below, equation (14) is equivalent to

(15) $\mathrm{cov}(Y, Z) = \mathrm{cov}(Y, W) \times \mathrm{cov}(Z, W).$

By (12), this means

(16) $ab = (c + abd)(d + abc).$

Rearranging (16) gives the quadratic equation

(17) $cd(ab)^2 - (1-c^2-d^2)ab + cd = 0.$

One solution to (17) is

$$(18)\qquad ab = \frac{1-c^2-d^2-\sqrt{(1-c^2-d^2)^2-4c^2d^2}}{2cd}.$$

I chose *a, c, d* more or less at random, getting .1925, .2873 and .1245, respectively.[27] I computed *b* from (18), getting .2063. This choice forces the conditional independence relation (14), and violates the faithfulness assumption: conditional independence comes from the parameter values, not the presence or absence of arrows.

Given the values for the four parameters *a, b, c, d*, the covariance matrix (12) can be evaluated as

$$(19)\qquad \begin{pmatrix} 1.0000 & 0.1925 & 0.2063 & 0.0810 \\ 0.1925 & 1.0000 & 0.0397 & 0.2922 \\ 0.2063 & 0.0397 & 1.0000 & 0.1359 \\ 0.0810 & 0.2922 & 0.1359 & 1.0000 \end{pmatrix}$$

The path coefficients in the right hand panel of Figure 7 are easily computed from (19):

the path coefficient from W to Y is $c' = \text{cov}(Y,W) = .2922$;
the path coefficient from W to Z is $d' = \text{cov}(Z,W) = .1359$;
the path coefficients from Y and Z to X are obtained by multiple regression, as $a' = .1846$ and $b' = .1990$.

With these choices, faithfulness does not hold, and (19) can be represented by either diagram in Figure 7. (For details on multiple regression, see the appendix.) In effect, the faithfulness assumption precludes certain algebraic identities among the parameters, like (16). Since parameters are not observable, the faithfulness assumption is not subject to direct empirical tests based on finite amounts of data.

[27] There was a bit of luck here, because some values for *a, c, d* will not produce correlation matrices.

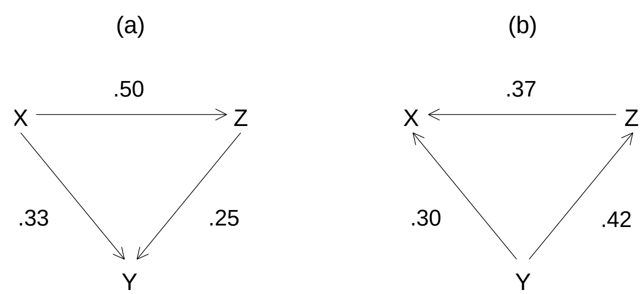
11.3 Complete graphs

Even if the covariance matrix is faithful to a graph, however, problems of indeterminacy remain-- particularly if the graph is "complete" in the sense that every pair of vertices is joined by an arrow. Figure 8 illustrates this indeterminacy. The same covariance matrix (20) for the variables *X, Y, Z* is represented either by the diagram in panel (a) or the one in panel (b), where the flow of "causality" is reversed.

(20)

	X	*Y*	*Z*
X	1	.46	.50
Y	.46	1	.42
Z	.50	.42	1

Figure 8. Graphs (a) and (b) have the same covariance matrix. Both are complete: there is an arrow from every variable to every other variable. The numbers on the arrows are path coefficients, that is, standardized regression coefficients.



For a second example of indeterminacy when the graph is complete, consider four variables *X, Y, Z, W* with covariance matrix Σ given by

$$(21) \qquad \Sigma = \begin{pmatrix} 1 & 3/4 & 3/4 & 3/4 \\ 3/4 & 1 & 3/4 & 3/4 \\ 3/4 & 3/4 & 1 & 3/4 \\ 3/4 & 3/4 & 3/4 & 1 \end{pmatrix}$$

Figure 9. Two complete path diagrams, and a factor analysis model, all having the same covariance matrix.

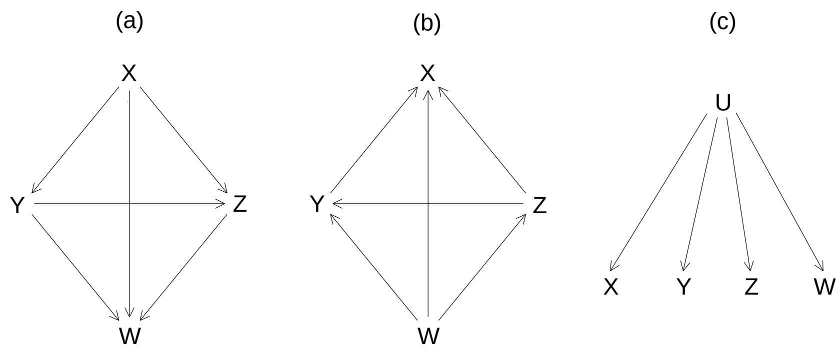


Figure 9 shows two complete path diagrams, both of which are compatible with the given covariance matrix. In the left hand panel, X is exogenous, and "causes" Y; then X and Y "cause" Z; finally X, Y, Z "cause" W. In panel (b), the flow of "causality" is reversed. The equations corresponding to the left hand panel are given as (22); panel (b) is described in (23).

$$
\begin{aligned}
Y &= \frac{3}{4} X + \delta_1 \\
Z &= \frac{3}{7} X + \frac{3}{7} Y + \delta_2 \\
W &= \frac{3}{10} X + \frac{3}{10} Y + \frac{3}{10} Z + \delta_3
\end{aligned}
\tag{22}
$$

$$
\begin{aligned}
Z &= \frac{3}{4} W + \varepsilon_1 \\
Y &= \frac{3}{7} Z + \frac{3}{7} W + \varepsilon_2 \\
X &= \frac{3}{10} Y + \frac{3}{10} Z + \frac{3}{10} W + \varepsilon_3
\end{aligned}
\tag{23}
$$

The covariance matrix Σ is also compatible with the factor analysis model (24), where the unobservable exogenous variable U causes all four observables (right hand panel of Figure 9).

$$(24) \quad X = U + \zeta_1, \quad Y = U + \zeta_2, \quad Z = U + \zeta_3, \quad W = U + \zeta_4$$

In each system of equations (22–23–24), the error terms are assumed to be independent and normally distributed with mean 0; error terms are independent of the exogenous variable. As a technical matter, the covariance matrix (20) is faithfully represented by both graphs in Figure 8. Likewise, the covariance matrix (21) is faithful to Figure 9(a) and to 9(b). Proofs may be based on (48) below.

To sum up, if a covariance matrix is faithful to a complete graph (with all pairs of vertices joined by arrows), it is faithful to many such graphs. Then correlational methods cannot tell the causes from the effects. SGS techniques work best when the graph is sparse, that is, relatively few pairs of vertices are joined by arrows (section 6).

11.4 Identifiability and consistency

The focus continues to be on linear models. In statistical terminology, models are "identifiable" when they make different predictions about observables. For example, suppose you have two models for your data. If, for all data sets,

$$P(\text{data} \mid \text{model 1}) = P(\text{data} \mid \text{model 2}),$$

there is an obvious problem-- the data cannot distinguish between the models. If a path model is complete, or the faithfulness assumption is not imposed, then the graph underlying a covariance matrix is not identifiable; that is the message of sections 11.1-3. By way of illustration, the models in Figure 7 are identifiable only if faithfulness holds.

However, even if we assume that a covariance matrix is faithful to a graph that is not complete, there may be several such graphs (SGS p. 89). For example, the following three graphs can generate the same covariance matrix:

$$X \to Y \to Z \qquad X \leftarrow Y \to Z \qquad X \leftarrow Y \leftarrow Z$$

Thus, SGS do not seem to have succeeded in defining a class of graphs and covariance matrices for which identifiability holds (SGS p. 194).

In statistical terminology, estimators are "consistent" provided that, as the sample gets larger and larger, these estimators come closer and closer to the population parameters. If the parameters are not identifiable, however, consistency is problematic.

SGS seem to claim that their algorithms will find all the path diagrams compatible with a given covariance matrix. However, the theorems suggest that the algorithms will at best find one such graph. SGS also seem to claim that their algorithms are consistent. However, without an identifiability theory for linear models, they cannot really be talking about consistency.

Statisticians do have the weaker notion of "Fisher consistency," named after R. A. Fisher: when applied to data for the whole population, an estimator should reproduce the population parameters exactly. Theorems like 5.1 in SGS (p. 405) seem to demonstrate the analog of Fisher consistency, rather than anything stronger. Such theorems show that, given the population covariance matrix, the algorithms will produce one graph consistent with that matrix.

11.5 Methodological contributions

There is a connection between the theory of "Directed Acyclic Graphs" (DAGs) and conditional independence of random variables. (See Darroch et al. 1980, Kiiveri and Speed 1982, Speed and Kiiveri 1986; Pearl 1986, Pearl 1988, Verma and Pearl 1990, Geiger 1990, Pearl and Verma 1991.) Much of this work is reviewed in SGS and CG. However, the mathematics of nonlinear causal diagrams seems to be irrelevant to the big question: how do we infer causation from association?

Most the applications in SGS are linear, i.e., based on path models. The "nonlinear causal diagrams" turn out to be multinomial models for categorical data; examples are on pp. 147–51. The issues about causation are quite similar to those for linear models, although the technical details are different. The real applications in CG all seem to involve linear models.

This section will focus on path models. To describe the novelty in the SGS approach to estimation, suppose you have data from a path model, and wish to estimate the model. Consider two cases:

Case I. You know the classification of variables as to level: that is, you know which variables are at level 0, which are at level 1, and so forth.

Case II. You do not know the classification of variables as to level.

In Case I, SGS have little to tell us about estimation (as to confounding, see section 12.1). Some of their algorithms seem to be equivalent to regression, others may be less efficient. In Case II, SGS try to estimate the classification of variables as well as the path coefficients. That is the methodological contribution. To estimate the classification, SGS must impose the faithfulness assumption (section 11.2). It is disappointing that SGS do not pin down the sense in which their algorithms are successful (section 11.4).

12. More examples, and some theory

Section 12.1 explains how the faithfulness assumption and conditional independence are supposed to eliminate confounding. Section 12.2 discusses omitted variables. Sections 12.3–5 revisit two examples from a more mathematical perspective; the idea is to show the limits of correlational methods.

12.1 Faithfulness, conditional independence, and confounding

The problems created by unobservable variables are well known. As indicated above, SGS handle such problems by imposing the faithfulness assumption. More specifically, the assumption is used to rule out confounding. If confounding can be eliminated, the goal is in sight -- association may soon be converted into causation. This section, which is based on work by Jamie Robins (personal communication), examines the logic in more detail. Also see (Pearl and Verma 1991).

With some models, exact conditional independence forces a choice:

- either there is no confounding by unmeasured common causes,
- or the faithfulness assumption is violated.

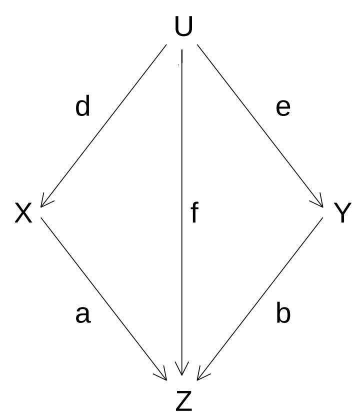
Near-independence is not good enough; associations may then be entirely spurious. Thus, causal inferences made by the SGS technique need exact conditional independence as well as the faithfulness assumption.

This use of the faithfulness assumption has some theoretical interest. However, in order to base empirical work on such mathematical ideas, it would seem necessary to resolve the following questions, which SGS have not addressed:

- Can the basic models be validated?
- Can exact conditional independence be demonstrated?
- Given exact independence, why is exact cancellation of confounded effects overwhelmingly less likely than the total absence of such effects?

As a practical matter, exact independence seems quite unusual. However, the theory is worth understanding, and an example will make the position clearer. Figure 10 shows a relatively simple diagram where faithfulness and conditional independence would eliminate confounding. The arrows denote causation, not mere association. Variables *X, Y, Z* are observable; *U* is unobservable. Such unobservables are also called "confounders" or "unmeasured common causes." The joint distribution is normal, and variables are standardized to have mean 0 and variance 1.

Figure 10. The faithfulness assumption, conditional independence, and confounding. Variables *X, Y, Z* are observable; *U* is unobservable. Arrows represent causation, not just association. The lower-case letters on the arrows denote path coefficients. If a path coefficient vanishes, the corresponding arrow must be deleted.



	U	X	Y	Z	
	U	1			
(25)	X	d	1		
	Y	e	de	1	
	Z	$f+ad+be$	$a+bde+fd$	$b+ade+fe$	1

The covariance matrix for all 4 variables is shown in (25).[28] Of course, only the covariance matrix (26) of the observables (X,Y,Z) can be estimated from the data. In particular, de is determined from the observables, as $\mathrm{cov}(X,Y)$.

(26)

	X	Y	Z
X	1		
Y	de	1	
Z	$a+bde+fd$	$b+ade+fe$	1

It may help to review the idea of faithfulness, in the context of our example. Faithfulness is an assumption about unobservables; more specifically, it is a constraint on the relationship between the full covariance matrix (25) and the graph in Figure 10. The assumption amounts to this: independence relationships (conditional and unconditional) are determined by the presence or absence of arrows in the diagram, not specific parameter values.

In particular, if the covariance matrix (25) is faithful to the diagram in Figure 10, you cannot set any of the path coefficients to 0, except by deleting the corresponding arrow. An arrow from X to Z, say, entails that X has some causal effect on Z, no matter how small that effect may turn out to be.

I return to more conventional issues. In our example, the parameter of interest is b, the causal effect of Y on Z. Due to the unmeasured confounder U, a regression of Z on X and Y produces a biased estimate of b. By a slightly tedious calculation, the coefficient of Y in the regression equation is

$$b + fe(1-d^2)/(1-d^2e^2). \tag{27}$$

(For details on multiple regression, see the appendix.) The bias in the regression estimate is the second term in (27). From a slightly different perspective, $\mathrm{cov}(Y,Z)$ in (26) measures the total association between Y and Z. Part of this association is real: b measures the causal effect of Y on Z. Alas, part of the association is spurious: $ade + fe$ represents the effects of the confounder U.

[28] Covariance matrices are symmetric; only the lower triangular part is shown. Entries are assumed to be positive but less than 1. The matrix is assumed to be positive definite.

The goal is to separate the real part of the association from the spurious part. The familiar obstacle: we have only (26), not (25). And (26) does not suffice to separate $b + ade + fe$ into its components. But, SGS might say, suppose that X and Z are conditionally independent given Y:

(28) $\quad \mathrm{cov}(X,Z \mid Y) = 0.$

By (48) below, this means

(29) $\quad \mathrm{cov}(X,Z) = \mathrm{cov}(X,Y) \times \mathrm{cov}(Y,Z).$

A bit of algebra based on (25) shows that (29) is equivalent to

(30) $\quad a(1 - d^2e^2) + df = de^2f.$

Although de is known and $0 < de < 1$, there are many possible ways to solve equation (30). At this point, SGS would invoke the faithfulness assumption, concluding that

(31) $\quad a = 0$ and $f = 0.$

The implication: we have to remove the arrow from X to Z, as well as the arrow from U to Z.

Confounding has now been eliminated. On this basis, $\mathrm{cov}(Y,Z) = b$; the whole of the association is real, and regression produces an unbiased estimate for the causal effect of Y on Z. At last, association has been converted into causation. Of course, quite a lot of causality was built into Figure 10 from the beginning-- by assumption.

Those were the implications of exact conditional independence. On the other hand, suppose we have approximate conditional independence: $\mathrm{cov}(X,Z \mid Y) = .00001$. Now the faithfulness assumption has no force. Given the covariances in (26), we can match them by suitable choice of the other parameters, even if $a = b = 0$.[29]

[29] This matching assumes, for instance, that any two of the variables have positive covariance given the third. To avoid violating the faithfulness assumption, if you set a and b to 0, erase the corresponding arrows; if that is distasteful, set a and b to small but positive values. The SGS logic would apply to a wide variety of diagrams; however, an arrow from Y to X, no matter how small the coefficient, spoils the show.

With approximate conditional independence, observed associations can be entirely spurious. Thus, even in the realm of mathematics, faithfulness and conditional independence preclude confounding only when the independence is exact. To make the contrast sharper, let us assume faithfulness.

- If $\mathrm{cov}(X,Z|Y) = 0$, then the association between Y and Z is purely causal; the effects of the unmeasured common cause U do not confound the relationship between Y and Z.

- If $\mathrm{cov}(X,Z|Y) = .00001$, then confounding by unmeasured common causes may account for all of the observed association between Y and Z.

Apparently, converting association into causation is still quite a dicey affair.

A similar problem must be considered when estimating path models from data (section 11). Exact conditional independence, together with the faithfulness assumption, often permits us to identify the path diagram from the covariance matrix. However, approximate conditional independence is not enough: then, the covariance matrix will be faithful to a variety of complete graphs.

A final example is the Timberlake-Williams model (section 10 above). This models explains political exclusion (PO) in terms of foreign investment (FI), energy development (EN) and civil liberties (CV); the sample correlation matrix was shown in Table 6. Consider three scenarios for the "true" correlation matrix ρ.

(i) Suppose ρ happens to equal the sample correlation matrix. Then, faithfulness obtains.

(ii) Suppose the true correlation ρ(PO,FI) between foreign investment and political exclusion happens to vanish exactly. Then, the Timberlake-Williams model violates the faithfulness condition; presumably, that is SGS's real complaint.

(iii) If ρ(PO,FI) = .00001, faithfulness is restored. According to the SGS criteria, Timberlake and Williams are back in business.

Within the framework of path models, scenario (ii) cannot be rejected at conventional significance levels; neither can (iii); and (i) represents our best estimate, subject to large uncertainties. SGS seize on hypothesis (ii), the only one that legitimates their critique. They are balking at shadows.

It may be possible to strengthen the faithfulness assumption, to rule out near-cancellation of confounded effects. However, developing such a theory will run into technical difficulties. In particular, stronger assumptions are likely to reduce the stock of path models that can be identified, because fewer covariance matrices would then be faithful to their graphs.

12.2 Omitted variables

The problem of omitted variables was raised by Cliff Clogg at the Notre Dame conference, and this section paraphrases one of his points. There is a response variable *Y*, with explanatory variables *X* and *Z*; these may be construed as vectors. Suppose the data are generated according to the "true" model (32T).

(32*T*) $Y = X\beta + Z\gamma + \varepsilon$ (32*R*) $Y = X\beta_R + \delta$

The parameter vectors β and γ are unknown, and to be estimated from data by regression; it is β that is of primary interest. Subjects are assumed to be independent and identically distributed; (X, Z) and the error term ε are independent and jointly normal; all variables have expected value 0. Consider too the "restricted" model (32R), where β_R is defined so that $E\{Y|X\} = X\beta_R$. The constituents of (32R) may be computed from the true model.[30]

In principle, the variables *X, Y* and *Z* are all observable; *X* and *Z* may be correlated. However, investigators who do not know that *Z* is relevant may fit the restricted model R rather than the true model T. If so, the estimate of β can be quite biased. In the vernacular, β_R includes the effect of *X* on *Y* through *Z*. The covariance matrix of (*X*, *Y*) cannot distinguish between the two

[30] Indeed, $\beta_R = \beta + \alpha$ where α is obtained by the regression of $Z\gamma$ on *X*. In other terms, $Z\gamma = X\alpha + \eta$, where η is normal with mean 0, independent of *X*. Then $\delta = \varepsilon + \eta$. It may be seen that α depends linearly on γ.

models, because the matrix can be generated by either model. Therefore, no statistical procedure based on that matrix can tell you whether the restricted model is right or wrong.[31]

12.3 On the direction of causality

This section uses "cause" in its ordinary (perhaps undefinable) meaning, not as shorthand for certain kinds of covariation. I return to Judea Pearl's example, shown in Figure 2(a). Given the covariance matrix for X, Y and Z, the SGS algorithm will produce the graph shown in panel (a). If you tell the algorithm that omitted variables are a possibility, it will tell you that Y cannot cause X or Z.

In the example, X, Y and Z are the only observables, and their covariance matrix is faithful to the graph in Figure 2(a). I claim that such information cannot by itself determine the direction of the causal flow. To substantiate this claim, I now construct two theories. In both, the observables X, Y and Z will have the same covariance matrix, faithful to the graph in Figure 2(a). However, the direction of the causal flow will be different in the two theories.

Theory #1

I first generate X, Z, U as independent N(0,1) variables; U is an unobservable error term. (If you want to intervene and change X or Z, now is your moment.) Then

(33) $$Y = X + Z + U.$$

According to theory #1, X and Z cause Y, as suggested by Figure 2(a).

---

[31] See (Clogg and Haritou 1994), who make the following very interesting point. Adding variables that are correlated with ε can also bias the estimate of β; this "included variable" bias can be just as troublesome as the more familiar "omitted variable" bias: the latter problem cannot be solved by throwing variables into the model. The SGS treatment of omitted variables was discussed in section 12.1 above.

Theory #2

I first generate Y as N(0,3). (If you want to intervene and change Y, now is your moment.) After a suitable pause, so that time's arrow will delineate the flow of causality, I generate the errors V_1, V_2 and V_3 as independent N(0,1/3) variables, and then produce X, Z and U according to

$$
(34)\qquad
\begin{aligned}
X &= \frac{1}{3}Y + V_1 - V_2\\
Z &= \frac{1}{3}Y + V_2 - V_3\\
U &= \frac{1}{3}Y + V_3 - V_1
\end{aligned}
$$

In this theory, Y causes X and Z.

As far as the observables are concerned-- namely, the joint distribution of X, Y and Z-- theories #1 and #2 agree. Furthermore, the joint distribution is faithful to the graph in Figure 2(a). But the direction of causality is determined neither by the data nor by the mathematics. With correlational methods, causality follows from the assumptions about the unobservables.

12.3 The AFQT problem

SGS seem to claim that, as a demonstrable mathematical fact, their procedures will find the right answers:

> Assuming the right variables have been measured, there is a straightforward solution to these problems: apply the PC, FCI, or other reliable algorithm, and appropriate theorems from the preceding chapters, to determine which X variables influence the outcome Y, which do not, and for which the question cannot be answered.... then estimate the dependencies by whatever methods seem appropriate and apply the results of the previous chapter to obtain predictions of the effect of manipulating the X variables. No extra theory is required. We will give a number of illustrations.... (SGS p. 242)

The first example given by SGS to illustrate this claim is AFQT (section 9 above). To demonstrate that SGS are exaggerating more than a little, I pose a sharp mathematical question with the essential features of the AFQT problem. Then, I show the question to be undecidable by correlational methods. (Of course, when applied to the real example, both SGS and ordinary least squares made the right guess.)

To set up the question, assume that X and Y are random variables; X is a vector, Y is scalar.

(35) Y is a linear combination of X's, with fixed weights.

(36) The observables are Y and $V_1, \ldots, V_7$.

Some V's are X's, some V's are ringers. (A "ringer" is a variable that does not enter into the linear combination for Y.) There are also unobservables, including the X's that are not V's. Assume too that

(37) The full joint distribution is multivariate normal, with mean 0.

You are given the covariance matrix for the observables, but not the full covariance matrix. The problem is to say which of the V's are X's and which are ringers. I claim this problem is not solvable, because I can produce two different theories leading to different classifications of the V's, but having the same joint distribution for the observables.

Theory #1

I use the covariance matrix for the 7 observable subtests V_1 = NO,... V_7 = GS together with the 3 unobservable subtests, CS, AS and PC. (The subtests are listed in Table 8, section 12.5 below). The full distribution is defined to be jointly normal, and all variables have mean 0. Let $Y = .5 \times \text{NO} + \text{AR} + \text{WK} + \text{PC}$, where NO, AR and WK are observable but PC is unobservable. In this theory, V_1, V_2, V_3 are X's, the remaining V's are ringers. This theory happens to have been more or less correct, prior to 1989: see equation (42) in section 12.5.

Theory #2

Again, I use the covariance matrix for the 7 observable subtests V_1 = NO,... V_7 = GS together with the other 3 unobservable subtests CS, AS, PC. I create an auxiliary variable U, which is independent of the 10 subtests and has small variance. The distribution of these 11 variables is defined to be jointly normal, and all variables have mean 0. There are three additional unobservables, defined as follows:

(38) $T_1 = .25(\mathrm{AR} + \mathrm{NO}) + .5\mathrm{PC} + U,$

(39) $T_2 = .25(\mathrm{WK} + \mathrm{NO}) + .5\mathrm{PC} + U,$

(40) $T_3 = .75(\mathrm{AR} + \mathrm{WK}) - 2U.$

Let

(41) $Y = T_1 + T_2 + T_3.$

In theory #2, T_1, T_2, T_3 are the unobservables; all the V's are ringers. The auxiliary variables U, CS, AS, PC serve only to define the joint distribution.

Theory #1 and theory #2 provide the same joint distribution for the observables. Therefore, no statistical procedure based on the joint distribution-- like the SGS algorithms or any other correlational methods-- can adjudicate between the two theories.

This section and the previous one demonstrate the obvious: you cannot infer cause and effect relationships by doing arithmetic on a correlation matrix, because association is not causation. The mathematical development in SGS avoids such problems only by imposing more or less arbitrary conditions (like faithfulness) on unobservable variables, as discussed in sections 11.2 and 12.1.

In the present section, neither theory #1 nor theory #2 fits into the SGS framework: Y is a deterministic function of the explanatory variables, with no stochastic error term; see (35). Furthermore, if U and PC are treated as variables rather than error terms in (38-39-40), the joint distribution in theory #2 is, presumably, unfaithful to its causal graph. Similar comments apply to the previous section.

12.5 Institutional background on the AFQT

The "Armed Services Vocational Aptitude Battery" (ASVAB) has ten subtests, including the seven listed in Table 4, section 9 above. All ten are shown in Table 8.

Table 8. The ten subtests in ASVAB.
The first seven were analyzed by SGS.

1.	Numerical Operations	NO
2.	Word Knowledge	WK
3.	Arithmetical Reasoning	AR
4.	Mathematical Knowledge	MK
5.	Electronics Information	EI
6.	Mechanical Comprehension	MC
7.	General Science	GS
8.	Coding Speed	CS
9.	Auto & Shop Information	AS
10.	Paragraph Comprehension	PC

Notes: ASVAB Form 17, July 1990.

Until January, 1989 the AFQT was computed as follows:

(42) $\quad AFQT = .5 \times NO + AR + WK + PC.$

After that date, NO was replaced by MK; a "verbal" score VE was defined as VE = WK + PC; and terms were standardized to have mean 0 and variance 1 on some calibration data-- the "NORC 1985 sample." AFQT was redefined as

(43) $\quad AFQT = MK_Z + AR_Z + 2 \times VE_Z,$

where the subscript Z denotes standardization. Throughout the period, raw scores were by Congressional requirement converted to percentiles based on the NORC sample. One hopes the data used by SGS come from 1988 or before, since they pick up formula (42) rather than (43); section 9 above.[32]

[32] SGS appear to be considering raw scores, and I follow suit. The material in this section was reported by Larry Hanser, personal communication; he refers to (Welsh et al. 1990, esp. Table 3 on p. 5), and (Eitelberg 1988, esp. p. 73).

13. Responses

Formal statistical inference is, by its nature, conditional. If assumptions A, B, C, ... hold, then H can be tested against the data. However, if A, B, C, ... remain in doubt, so must inferences about H. Indeed, the statistical calculations may prove to be quite misleading.

Many assumptions are made but only a few are tested. Those made without testing are called "maintained hypotheses." They are usually statistical and often rather technical -- linearity, independence, exogeneity, etc. Careful scrutiny of such assumptions would therefore seem to be a critical part of empirical work.

In the social sciences, however, statistical assumptions are rarely made explicit, let alone validated. Questions provoke reactions that cover the gamut from indignation to obscurantism. *We know all that. Nothing is perfect. Linearity has to be a good first approximation. The assumptions are reasonable. The assumptions don't matter. The assumptions are conservative. You can't prove the assumptions are wrong. The biases will cancel. We can model the biases. We're only doing what everybody else does. Now we use more sophisticated techniques. What would you do? The decision-maker has to be better off with us than without us. We all have mental models, not using a model is still a model.*

With the SGS approach, responses are more subtle but no more empirical. Proponents often seem to take a Bayesian stance: faithfulness is justified on the grounds that the exceptional cases have measure 0, and must therefore be viewed as negligible *a priori*.[33] However, the SGS approach is frequentist not Bayesian; the simulations, being done on finite-state computers, must concentrate in a set of measure 0; and the SGS class of models has measure 0 within larger classes of models. Indeed, from my perspective, the whole class of path models seems rather unlikely-- given the intensity of the research effort and the paucity of convincing examples. The assumptions that diagrams are sparse and faithful stretch credibility even further.

---

[33] The "measure" here is the uniform distribution in Euclidean space, e.g., length, area, volume.... Mathematicians call the uniform distribution "Lebesgue measure," in honor of Henri Lebesgue (1875-1941) who developed its mathematical foundations. The SGS argument (p. 95) seems to be a variation on Laplace's "principle of insufficient reason" (Stigler 1986, 127).

Attempts have also be made to justify the faithfulness assumption by appeals to continuity. If a covariance matrix is unfaithful, small changes to parameter values make it faithful. However, the same argument can be turned against correlational methods. For example, if a covariance matrix is faithful to an incomplete graph, small changes to hidden parameters make the graph complete and vitiate the SGS search procedures. Section 12.1 points to another kind of instability in the SGS framework. The continuity defense (like the Bayesian argument) reflects an aesthetic judgment about modeling styles. Taste is no substitute for empirical verification.

The SGS criteria for causality may also be defended as follows-- it is unlikely that anything could produce the patterns of intercorrelation identified by SGS, other than causation; thus, correlational methods shift the burden of argument. Figures 5 and 6 should dispose of this idea. In real examples, the patterns identified by the SGS search algorithms can hardly represent cause-and-effect relationships. The burden would seem to be on the modelers: how can they recommend an algorithm that gives such results?

Proponents of modeling can also be heard to argue that all of us make assumptions about unobservables. However, what is unobservable with one design may become observable with another. And some investigators still deal with unobservables the hard way-- by doing the right studies. For example, take Fisher's "constitutional hypothesis:" there may be a genetic factor that predisposes you to smoke and to get lung cancer, heart disease, etc.[34] This putative genetic factor is the unobservable common cause for smoking and illness.

The epidemiologists did not deal with the constitutional hypothesis by introducing special assumptions. Instead, they studied the matter empirically, using data from twin studies. For a recent report on the Swedish twin registry, see (Floderus et al. 1988). On the Finnish twin registry, see (Kaprio and Koskenvuo 1989). Data on the Danish twin registry are fragmentary. There are forthcoming data on the U.S. twin registry, which are quite strong (Dorit Carmelli, personal communication). The numbers on lung cancer are suggestive, but still small-- this is a rare disease, even among smokers. The data on heart disease and total mortality, however, make the constitutional hypothesis untenable.

[34] See SGS pp. 298-9, CG p. 32.

13.1 A comment from Judea Pearl

Judea Pearl writes that

> Correlation-based model-searching schemes produce causal inferences with only limited guarantees. Yet such schemes have potential, if conducted under conditions that screen out accidental independencies while maintaining structural independencies-- for example, longitudinal studies under slightly varying conditions. This assumes, of course, that under such varying conditions the parameters of the model will be perturbed, while its structure remains stable. Maintaining such delicate balance under changing conditions may be hard in real-life studies. However, considering the alternative of resorting to controlled, randomized experiments, such longitudinal studies are still an exciting opportunity.
>
> Additionally, any investigator who is searching for a causal model knowing that the parameters might be tied together by some hidden equation, like (17) [section 11.2], is wasting time (and public funds). Such a model, even if correct, is bound to be useless, because without the assumption of autonomy (i.e., that each parameter can be perturbed without altering the others), the model cannot predict the effect of interventions or other changes.... [personal communication]

Also see (Pearl 1993; Pearl and Wermuth 1993).

14. Other literature

There is an extensive literature on the evaluation of models, going back at least to the Keynes-Tinbergen exchange (Keynes 1939 and 1940, Tinbergen 1940). Also see (Liu 1960) and (Lucas 1976). For more recent discussions, with other citations to the literature, see (Freedman 1987 and 1991). Many authors have tried to explain the basis for inferring causation by using regression. See, for example, (Pratt and Schlaifer 1984 and 1988), or (Holland 1986 and 1988). Of enthusiastic views on social-science modeling, there is no shortage; see, for instance, (Smelser and Gerstein 1986) or (Bartels and Brady 1993). For a recent discussion of causal modeling, see (Cox and Wermuth 1993).

15. Conclusions

SGS have not succeeded in clarifying the circumstances under which causal inferences can be drawn from observed associations, nor have they invented a reliable engine for performing this feat. Their algorithms have some technical interest, but will make causal inferences only when causation is assumed in the first place. To be more explicit: If we assume that the arrows in a path diagram represent causation rather than association, and we also assume that the path diagram can be estimated from data, then indeed SGS can infer causation from association. The faithfulness assumption and exact conditional independence will together eliminate certain kinds of confounding. Even so, causality is assumed into the picture at the beginning, not proved in at the end. As Nancy Cartwright says, "No causes in, no causes out."[35]

The larger problem remains. Can quantitative social scientists infer causality by applying statistical technology to correlation matrices? That is not a mathematical question, because the answer turns on the way the world is put together. As I read the record, correlational methods have not delivered the goods. We need to work on measurement, design, theory. Fancier statistics are not likely to help much.

---

[35] (Cartwright 1989, Chapters 2 and 3). Also see (Pearl and Verma 1991).

Appendix: Regression and conditioning

For ease of reference, this appendix presents the usual formulas for computing regressions, and conditional covariances. I begin with regression. Suppose ξ and η are random variables; ξ may be a row vector. We seek the column vector β of regression coefficients for η on ξ. Let $C = \mathrm{E}\{\xi'\xi\}$ and $D = \mathrm{E}\{\xi'\eta\}$; the prime denotes matrix transposition. Assume C is positive definite. Then

(44) $$\beta = C^{-1}D.$$

Now $\eta = \xi\beta + u$, where u is automatically orthogonal to ξ. The mean square of u may be computed as follows:

(45) $$\mathrm{E}(u^2) = \mathrm{E}(\eta^2) - \beta' C\beta.$$

If ξ and η have mean 0, then $C = \mathrm{cov}(\xi)$ and $D = \mathrm{cov}(\xi,\eta)$; also, $\mathrm{E}(u) = 0$. Likewise, if some component of ξ is a non-zero constant, $\mathrm{E}(u) = 0$. If now the variables are jointly normal, u is independent of ξ.

I turn to estimation. Recall equation (2), repeated here for ease of reference.

(2) $$Y = X\beta + \varepsilon.$$

In this equation, X is the "design matrix," representing the explanatory variables. There is one row for each unit in the study, and one column for each variable; the entry in the ith row and jth column represents the jth variable, as observed on the ith unit in the study; X may include a column of 1's, if there is to be an intercept in the equation. Y is a column vector representing the dependent variable, whose ith component represents the value of Y for the ith unit in the study. ε is also a column vector, with one component for each unit in the study, representing the impact on Y of chance factors unrelated to X. Typically, there will be many fewer parameters than data points, so β has relatively few components.

The ordinary least squares estimator for β is denoted by a hat, and may be computed as

(46) $$\hat{\beta} = (X'X)^{-1}X'Y.$$

The covariance matrix for $\hat{\beta}$, conditional on the design matrix, is computed as

$$(47)\quad \operatorname{cov}(\hat{\beta}|X) = (X'X)^{-1}\operatorname{var}(\varepsilon_i|X).$$

Of course, (46) is related to (44); this is seen by defining (ξ,η) as a row chosen at random from (X,Y).

The "predicted values" and "residuals" are defined as follows: $\hat{Y} = X\hat{\beta}$ and $e = Y - \hat{Y}$. The residuals are automatically orthogonal to X. The residual sum of squares, minimized by the choice of β, is RSS $= \|e\|^2 = \sum_i e_i^2$. Then $\operatorname{var}(\varepsilon_i|X)$ in (47) may be estimated as RSS/$(n-p)$, where n is the number of data points and p is the number of explanatory variables. Variances will be found along the diagonal of the covariance matrix, and the standard error is computed as the square root of the variance. In deriving these formulas, it is assumed that given X, the components of ε are conditionally independent and identically distributed, with mean 0.

Suppose the model has an intercept. Then R^2 may be defined as $R^2 = \operatorname{var}\{\hat{Y}\}/\operatorname{var}\{Y\}$, where, e.g.,

$$var\{Y\} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \quad\text{and}\quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

If all variables have mean 0, then R^2 may be computed as $\hat{\beta}'X'X\hat{\beta}/(n\times\operatorname{var}\{Y\})$.

The usual formula for computing conditional covariances may be presented as follows. Let $n > 2$. Suppose $X_1, X_2, \ldots, X_n$ are jointly normal. We seek the conditional covariance of X_1 and X_2 given $X_3, X_4, \ldots, X_n$. Let Σ be the covariance matrix of $X_3, X_4, \ldots, X_n$. Let κ_1 be the covariance of X_1 with $X_3, X_4, \ldots, X_n$; let κ_2 be the covariance of X_2 with $X_3, X_4, \ldots, X_n$. We view κ_1 and κ_2 as $n-2 \times 1$ column vectors. The conditional covariance is given by

$$(48)\quad \operatorname{cov}(X_1,X_2|X_3,\ldots,X_n) = \operatorname{cov}(X_1,X_2) - \kappa_1'\Sigma^{-1}\kappa_2.$$

The prime denotes matrix transposition. Details on the material in this appendix may be found in standard texts, for instance, (Rao 1973).

References

Bartels, L. M. 1991. "Instrumental and 'quasi-instrumental' variables." *American Journal of Political Science 35* (1991): 777-800.

Bartels, L. M. and H. E. Brady. 1993. "The state of quantitative political methodology." In Ada W. Finifter, ed. *Political Science: The State of the Discipline II*. Washington, D.C.: American Political Science Association.

Blau, P. M. and O. D. Duncan. 1967. *The American Occupational Structure*. New York: Wiley.

CG, see Glymour (1993).

Cairns, J. 1978. *Cancer: Science and Society*. San Francisco: W. H. Freeman.

Cartwright, N. 1989. *Nature's Capacities and their Measurement.* Oxford: Clarendon Press.

Clogg, C. C. and A. Haritou. 1994. "The regression method of causal inference and a dilemma with this method." Technical report, Department of Sociology, Pennsylvania State University.

Cornfield, J., W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. 1959. "Smoking and lung cancer: recent evidence and a discussion of some questions." *Journal of the National Cancer Institute 22* (1959): 173-203.

Cox, D. R. and N. Wermuth. 1993. "Linear dependencies represented by chain graphs." *Statistical Science 8* (1993): 204-83 (with discussion).

Daggett, R. and D. Freedman. 1985. "Econometrics and the law: a case study in the proof of antitrust damages." In L. LeCam and R. Olshen, eds. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Vol. I, pp. 126-75. Belmont, Calif.: Wadsworth.

Darroch, J. N., S. L. Lauritzen and T. P. Speed. 1980. "Markov fields and log-linear interaction models for contingency tables." *Annals of Statistics 8* (1980): 522-39.

Desrosières, A. 1993. *La politique des grands nombres.* Paris: Éditions la Découverte.

Duncan, O. D. 1975. *Introduction to Structural Equation Models*. New York: Academic Press.

Ehrenberg, A. S. C. and J. A. Bound. 1993. "Predictability and Prediction." *Journal of the Royal Statistical Society, Series A, 156, Part 2* (1993): 167-206.

Eitelberg, M. J. 1988. *Manpower for military occupations*. Office of the Assistant Secretary of Defense (Force Management and Personnel).

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

Floderus, B., R. Cederlof, and L. Friberg. 1988. "Smoking and mortality: a 21-year follow-up based on the Swedish Twin Registry." *International Journal of Epidemiology 17* (1988): 332-40.

Freedman, D. 1983. "A note on screening regression equations." *The American Statistician 37* (1983): 152-55.

Freedman, D. 1987. "As others see us: A case study in path analysis." *Journal of Educational Statistics 12* (1987) number 2 (with discussion, whole issue).

Freedman, D. 1991. "Statistical models and shoe leather." In P. Marsden, ed. *Sociological Methodology 1991*. Washington, D.C.: American Sociological Association.

Freedman, D. and D. Lane. 1981. *Mathematical Methods in Statistics*. New York: W. W. Norton.

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

Geiger, D. 1990. *Graphoids: A Qualitative Framework for Probabilistic Inference.* Ph. D. Dissertation, UCLA, Department of Computer Science.

Glymour, C. 1993. "A review of recent work on the foundations of causal inference." Paper presented at the Notre Dame conference. Cited as CG.

Glymour, C., R. Scheines, P. Spirtes, and K. Kelly. 1987. *Discovering Causal Structure*. New York: Academic Press.

Hakama, M., M. Lehtinen, P. Knekt, A. Aromaa, P. Leinikki, A. Miettinen, J. Paavonen, R. Peto, and L. Teppo. 1993. "Serum antibodies and subsequent cervical neoplasms: a prospective study with 12 years of follow-up." *American Journal of Epidemiology 137* (1993): 166-70.

Hausman, J. 1978. "Specification Tests in Econometrics." *Econometrica 46* (1978): 1251-71.

Hofferth, S. L. and K. A. Moore. 1979. "Early childbearing and later economic well-being." *American Sociological Review 44* (1979): 784-815.

Holland, P. 1986. "Statistics and causal inference." *Journal of the American Statistical Association 81* (1986): 945-60.

Holland, P. 1988. "Causal Inference, Path Analysis, and Recursive Structural Equations Models." In C. Clogg, ed., *Sociological Methodology 1988*, pp. 449-484. Washington, D.C.: American Sociological Association.

International Agency for Research on Cancer. 1986. *Tobacco smoking*. Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans. Vol 38. IARC, Lyon, France.

Kaprio, J. and M. Koskenvuo. 1989. "Twins, smoking and mortality: a 12-year prospective study of smoking-discordant twin pairs." *Social Science and Medicine 29* (1989): 1083-89.

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

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

Kiiveri, H. and T. Speed. 1982. "Structural analysis of multivariate data: a review." In S. Leinhardt, ed., *Sociological Methodology 1982*. San Francisco: Jossey Bass.

Leamer, E. E. 1985. "Vector autoregressions for causal inference." In K. Brunner and A. Meltzer, eds. *Understanding Monetary Regimes*. Supplement to the *Journal of Monetary Economics.* Amsterdam: North-Holland.

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

Liu, T. C. 1960. "Under-identification, structural estimation, and forecasting." *Econometrica 28* (1960): 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.* Amsterdam: North-Holland. pp. 19-64, with discussion.

Maddala, G. S. 1992. *Introduction to Econometrics*. 2nd ed. McGraw-Hill.

Manski, C. F. 1993. "Identification Problems in the Social Sciences." In P. V. Marsden, ed. *Sociological Methodology 1993*. Oxford: Basil Blackwell. Pp. 1-56.

Meehl, P. 1954. *Clinical versus statistical prediction; a theoretical analysis and a review of the evidence*. University of Minnesota Press.

Moore, K. A. and S. L. Hofferth. 1980. "Factors affecting early family formation: a path model." *Population and Environment 3* (1980): 73-98.

Pearl, J. 1986. "Fusion, Propagation and Structuring in Belief Networks." *Artificial Intelligence 29* (1986): 241-288.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. San Mateo: Morgan Kaufmann Publishers, Inc.

Pearl, J. 1993. "Comment: Graphical models, causality and intervention." *Statistical Science 8* (1993): 266-73.

Pearl, J. 1994a. "On the statistical interpretation of structural equations." Technical Report, Computer Science Department, UCLA.

Pearl, J. 1994b. "On the identification of nonparametric structural equations." Technical Report, Computer Science Department, UCLA.

Pearl, J., D. Geiger and T. Verma (1989). "The Logic of Influence Diagrams." In R. M. Oliver and J. Q. Smith, eds. *Influence Diagrams, Belief Nets and Decision Analysis*, pp. 67-87. John Wiley & Sons, Ltd.

Pearl, J. and T. Verma. 1991. "A Theory of Inferred Causation." In J. A. Allen, R. Fikes, and E. Sandewall, eds., *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pp. 441-52. San Mateo, Calif.: Morgan Kaufmann.

Pearl, J. and N. Wermuth. 1993. "When Can Association Graphs Admit A Causal Explanation?" *Proceedings of the Fourth International Workshop on Artificial Intelligence and Statistics*, 141-150. To appear in P. Cheeseman and W. Oldford, eds. *Artificial Intelligence and Statistics*. Berlin: Springer-Verlag (1994).

Peto, R. and H. zur Hausen, eds. 1986. *Viral etiology of cervical cancer.* Cold Spring Harbor Laboratory. Banbury report No. 21.

Pratt, J. and R. Schlaifer. 1984. "On the nature and discovery of structure." *Journal of the American Statistical Association 79* (1984): 9-21.

Pratt, J. and R. Schlaifer. 1988. "On the interpretation and observation of laws." *Journal of Econometrics 39* (1988): 23-52.

Rao, C. R. 1973. *Linear Statistical Inference and Its Applications*. 2nd ed. Wiley.

Rindfuss, R. R., L. Bumpass, and C. St. John. 1980. "Education and Fertility: Implications for the roles women occupy." *American Sociological Review 45* (1980): 431-47.

Rindfuss, R. R., L. Bumpass, and C. St. John. 1984. "Education and the timing of motherhood: disentangling causation." *Journal of Marriage and the Family 46* (1984): 981-84.

SGS, see Spirtes, Glymour and Scheines (1993).

Seneta, E. 1987. Discussion. *Journal of Educational Statistics 12* (1987): 198-201.

Sherman, K. J., J. R. Daling, and J. Chu et al. 1991. "Genital warts, other sexually transmitted diseases, and vulvar cancer." *Epidemiology 2* (1991): 257-62

Simon, H. 1980. "The meaning of causal ordering." In R. K. Merton, J. S. Coleman, and P. H. Rossi, eds., *Qualitative and Quantitative Social Research*, pp. 65-81. New York: Free Press.

Smelser, N. J. and D. R. Gerstein. 1986. *Behavioral and social science: fifty years of discovery*. Washington, D.C.: National Academy Press.

Speed, T. P. and H. T. Kiiveri. 1986. "Gaussian Markov distributions over finite graphs." *Annals of Statistics 14* (1986): 138-50.

Spirtes, P., C. Glymour, and R. Scheines. 1993. *Causation, Prediction and Search*. Springer Lecture Notes in Statistics, No. 81. Cited as SGS.

Spirtes, P., R. Scheines, C. Glymour and C. Meek. 1993. *TETRAD II*. Documentation for Version 2.2. Technical Report, Department of Philosophy, Carnegie Mellon University, Pittsburgh, Penn.

Stigler, S. 1986. *The History of Statistics*. Harvard University Press.

Timberlake, M. and K. Williams. 1984. Dependence, political exclusion and government repression: some cross national evidence. *American Sociological Review 49* (1984): 141-46.

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

Verma, T. and J. Pearl. 1990. "Causal Networks: Semantics and Expressiveness." In R. Shachter, T.S. Levitt, and L.N. Kanal, eds., *Uncertainty in AI 4*, pp. 69-76. Elsevier Science Publishers.

Welsh, J. R., S. K. Kucinkas, and L. T. Curran. 1990. *Armed Services Vocational Battery (ASVAB): Integrative Review of Validity Studies*. Air Force Human Resources Laboratory Report AFHRL-TR-90-22.

White, H. 1980. "A heteroskedasticity-consistent estimator and a direct test for heteroskedasticity." *Econometrica 48* (1980): 817-838.

White, H. 1982. "Maximum likelihood estimation of misspecified models." *Econometrica 50* (1982): 1-25.

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* (1899): 249-95.

Acknowledgments

Many useful comments were made by Dick Berk, John Cairns, Cliff Clogg, Mark Hansen, Larry Hanser, Jerome Horowitz, Paul Humphreys, Ron Lee, Tony Lin, Bill Mason, Vaughn McKim, Judea Pearl, Diana Petitti, Jamie Robins, Tom Rothenberg, Terry Speed and Steve Turner. Amos Tversky's work on the paper amounted to collaboration.

Richard Scheines and Peter Spirtes deserve special mention. Through many exchanges on the topics of this paper, they were helpful, thoughtful, and surprisingly patient; which is not to imply that they will agree with anything I say.

Research partially supported by NSF Grant DMS 92-08677.

Muñoz, N., Bosch, F. X., Shah, K. V., Meheus, A., eds. 1992. *The epidemiology of human papillomavirus and cervical cancer*. Lyon: International Agency for Research on Cancer; New York: Distributed in the USA by Oxford University Press.

Ericsson, Neil R. 1995. *Testing Exogeneity*. Oxford University Press.