Regression and Causation

by

C. Glymour, R. Scheines, P. Spirtes, C. Meek

October 1994

Report CMU-PHIL-60

Carnegie Mellon

Philosophy
Methodology
Logic

Pittsburgh, Pennsylvania 15213-3890

Regression and Causation

Clark Glymour, Richard Scheines, Peter Spirtes and Christopher Meek
Carnegie Mellon University

Abstract

In both linear and nonlinear multiple regression, when regressors are correlated the existence of an unmeasured common cause of regressor X_i and outcome variable Y may bias estimates of the influence of *other* regressors, X_k; variables having no influence on Y whatsoever may thereby be given significant regression coefficients. The bias may be quite large. Simulation studies show that standard regression model specification procedures make the same error. The strategy of regressing on a larger set of variables and checking stability may compound rather than remedy the problem. A similar difficulty in the estimation of the influence of other regressors arises if some X_i is an effect rather than a cause of Y. The problem appears endemic in uses of multiple regression on uncontrolled variables, and unless somehow corrected appears to invalidate many scientific uses of regression methods. We describe an implementation in the TETRAD II program of a model specification algorithm that avoids these and certain other errors in large samples. We recommend that such an algorithm be applied before regression is used to estimate influence.[1]

Correspondence: C. Glymour, Department of Philosophy, Carnegie Mellon University
Pittsburgh, Pa. 15213
E-mail: cg09@andrew.cmu.edu

1. Introduction

Regression models are commonly used to estimate the "influence" that regressors X have on an outcome variable, Y.[2] If the relations among the variables are linear then for each Xj the expected change in Y that would be produced by a unit change in Xi if all other X variables were forced to be constant can be represented by a coefficient, say aj. It is obvious and widely noted (see, for example, Fox, 1984) that the regression estimate of aj will be incorrect if Xi and Y have one or more unmeasured common causes, or in conventional statistical terminology, the estimate will be biased and inconsistent if the error variable for Y is correlated with Xi. To avoid such errors, it is sometimes recommended (Pratt and Schlaifer, 1988) that investigators enlarge the set of potential regressors and determine if the regression coefficients for the original regressors remain stable, in the hope that confounding common causes, if any, will thereby be measured and revealed. Regression estimates are known often to be unstable when the number of regressors is enlarged, because, for example, additional regressors may be common causes of previous regressors and the outcome variable (Mosteller and Tukey, 1977). The stability of a regression coefficients for X when other regressors are added is taken to be evidence that X and the outcome variable have no common cause. Mosteller and Tukey did not regard this technique or any other as adequate to deal with the problem of unmeasured common causes in regression, and chiefly on the basis of such considerations they warned that:

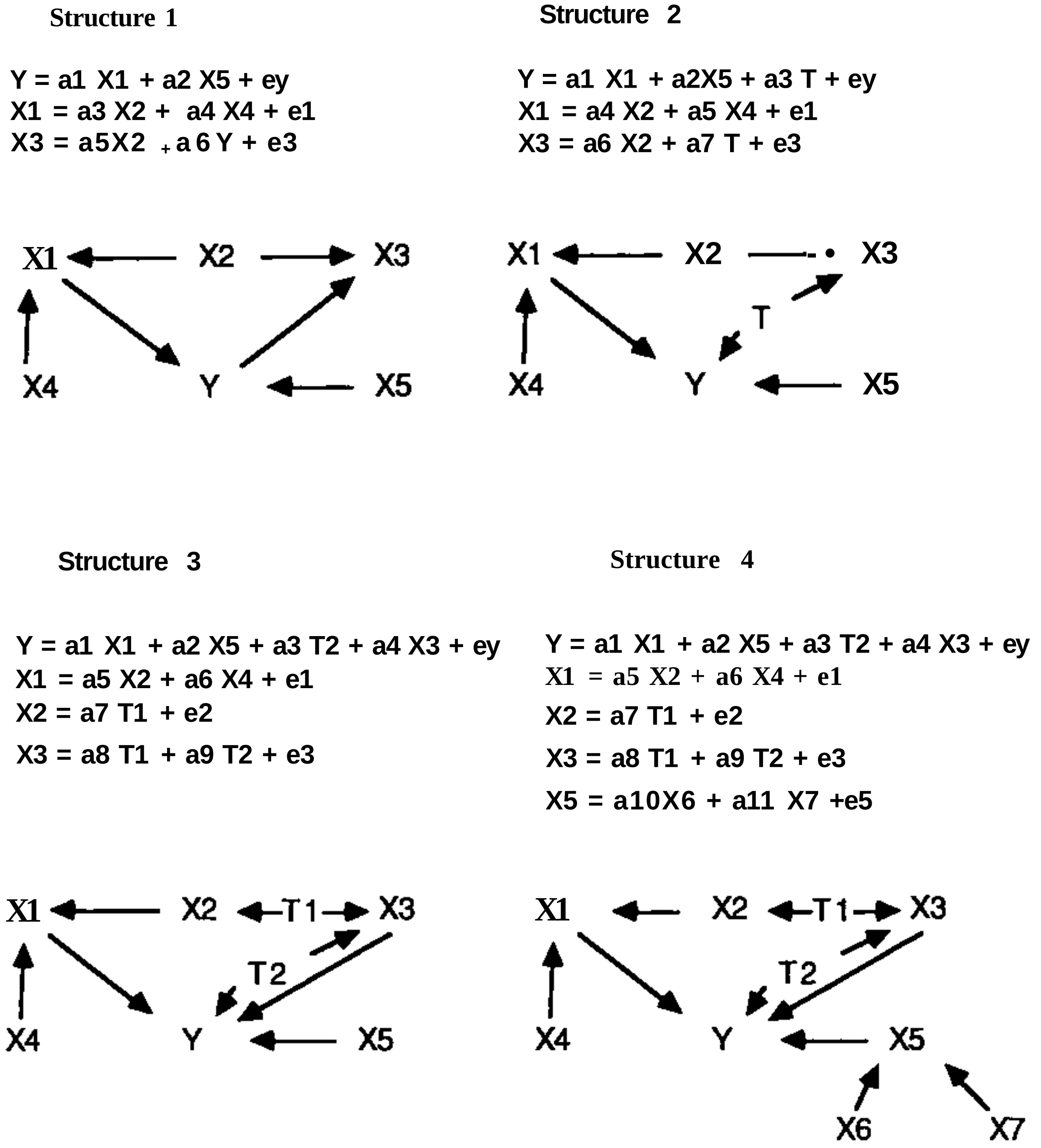
> George Box has [almost] said "The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively." These words of caution about "natural experiments" are uncomfortably strong. Yet in today's world we see no alternative to accepting them as, if anything, too weak.(p. 320)

We wish to show that the problems of regression methods for assessing causal influence are even more fundamentally flawed. And we offer an alternative procedure for such inferences that under quite general conditions is demonstrably reliable when given population correlations, and that, in cases where the truth is known independently, has proved reliable and informative on simulated and empirical data sets.

2. When Regression Fails to Measure Influence

It does not seem to be recognized that when regressors are statistically dependent the existence of an unmeasured common cause of regressor X_i and outcome variable Y may bias estimates of the influence of *other* regressors, X_k; variables having no influence on Y whatsoever, nor even a common cause with Y, may thereby be given significant regression coefficients. The error may be quite large even in the large sample limit. The strategy of regressing on a larger set of variables and checking stability may compound rather than remedy this problem. A similar difficulty may arise if one of the measured candidate regressors is an *effect*, rather than a cause, of Y, a circumstance that we think may sometime occur in uncontrolled studies.

To illustrate the problem, consider the following linear structures, where for concreteness we specify that that exogenous and error variables are all uncorrelated and jointly normally distributed, the error variables have zero means, and linear coefficients are not zero. Only the X variables are assumed to be measured. Each set of linear equations is accompanied by a directed graph illustrating the assumed causal and functional dependencies among the non-error variables:



Structure 1

$Y = a1\ X1 + a2\ X5 + ey$
$X1 = a3\ X2 + a4\ X4 + e1$
$X3 = a5X2 + a6\ Y + e3$

Structure 2

$Y = a1\ X1 + a2X5 + a3\ T + ey$
$X1 = a4\ X2 + a5\ X4 + e1$
$X3 = a6\ X2 + a7\ T + e3$

Structure 3

$Y = a1\ X1 + a2\ X5 + a3\ T2 + a4\ X3 + ey$
$X1 = a5\ X2 + a6\ X4 + e1$
$X2 = a7\ T1 + e2$
$X3 = a8\ T1 + a9\ T2 + e3$

Structure 4

$Y = a1\ X1 + a2\ X5 + a3\ T2 + a4\ X3 + ey$
$X1 = a5\ X2 + a6\ X4 + e1$
$X2 = a7\ T1 + e2$
$X3 = a8\ T1 + a9\ T2 + e3$
$X5 = a10X6 + a11\ X7 + e5$

Figure 1

In large samples, for data from each of these structures linear multiple regression will give all variables in the set {X1, X2, X3, X5} non-zero regression coefficients, even though X2 has no direct influence on Y in any of these structures, and X3 has no direct or indirect influence direct or indirect on Y in structures 1, and 2, and the effect of X3 in structures 3 and 4 is confounded by an unmeasured common cause. The regression estimates of the influences of X2 and X3 will in all four cases be incorrect. If a specification search for regressors had selected X1 or X5 alone, or both variables, a regression on these variables singly or together would give consistent, unbiased estimates of their influence on Y. But the textbook procedures in commercial statistical packages will in all of these cases fail to identify {X1} or {X5} or {X1, X5} as the appropriate subset of regressors.

It is easy to produce examples of the difficulty by simulation. Using structure 1, twenty sets of values for the linear coefficients were generated, half positive and half negative, each with absolute value greater than .5. For each system of coefficient values a random sample of 5,000 units was generated by substituting those values for the coefficients of structure 1 and using unit exogenous and error variances and zero exogenous means[3]. Each sample was given to MINITAB, and in all cases the program found that {X1, X2, X3, X5} is the set of regressors with significant regression coefficients. In addition, in MINITAB the STEPWISE procedure, selection by Mallow's Cp, and selection by adjusted R^2 all always selected either the set {X1, X2, X3, X5} or the set {X1, X2, X3, X4, X5} (although in some cases they disagreed on which of these sets was selected.)

The difficulty can be remedied if one measures all common causes of the outcome variable and the candidate regressors, and if none of the candidate regressors are effects, rather than causes, of the outcome variable, but unfortunately nothing in regression methods informs one as to when these conditions obtain. The addition of extra candidate regressors may create the problem rather than remedy it; in the four structures illustrated, if X3 were not measured the regression estimate of X2 would be consistent and unbiased.

The problem we have illustrated is quite general; it will bias the estimate of the influence of any regressor Xk that causes or has a common unmeasured cause with any regressor Xi such that Xi and Y have an unmeasured common unmeasured cause (or X_i is an effect of Y). Depending on the true structure and coefficient values the error may be quite large. It is easy to construct cases in which a variable with no influence on the outcome variable has a standardized regression coefficient larger than any other single regressor. Completely parallel problems arise for categorical data.

A further problem both for regression and for regression model selection procedures arises with small samples and large numbers of variables. A test of the hypothesis that a regression coefficient is zero, for example, is essentially a test of the hypothesis that the partial correlation of the regressor and the outcome variable vanishes when all of the remaining regressors are controlled for. The power of the test for a fixed sample size decreases as the number of control variables increases. A procedure that could use tests of lower order vanishing partial correlations to locate regressors that actually influence the outcome variable would in some cases be more reliable.

In the next section we will describe a rigorous solution to these problems. The solution also addresses the usual concern that the dependency between a regressor and an outcome variable may be confounded by an unmeasured common cause of both.

3. **Graphical Causal Models and Algorithms**

Ever since Sewell Wright's (1934) work, causal dependencies among variables in a population of units have occasionally been represented by directed graphs in which vertices represent random variables and a directed edge from one variable X to another Y indicates that even if all other variables considered were forced to be constant, some variation in X would produce variation in Y. Kiiveri and Speed (1982) made explicit the formal connection between directed acyclic graphs representing causal structure and distributional properties of a population of units so structured. One equivalent of their axiom (where we indicate sets of variables by boldface) is:

> The **Markov Condition.** Directed acyclic graph G and a probability distribution P on the vertices V of G satisfy the Markov condition if and only if for every W in V, W is independent of the set of its non-descendants given its immediate parents.

The Markov Condition entails that the joint density for P can be "factorized," that is, written as a product of marginal and conditional densities obtained by multiplying the exogenous densities by the conditional densities of their immediate children, and so on. Because of the factorization, for directed graphical models of discrete variables if the sampling distribution is multinomial a maximum likelihood estimate of the distribution can be obtained directly without the need for iterative procedures required to estimate many log-linear models.

In addition to the Markov Condition we assume:

The Faithfulness Condition: Graph G and distribution P on the vertices of G are faithful provided every conditional independence relation in P follows from the Markov Condition for G.

We believe these conditions are tacitly assumed in many statistical models of causal influence, and in regression models in particular. It is easy enough to produce populations in which they are violated, for example: (i) If some measured variables are deterministic functions of others, the Faithfulness Condition may be violated; (ii) if the population is a mixture of units with different causal structures the Markov Condition will be violated; (iii) if the population is a mixture of linear systems with the same causal structure but different values of linear coefficients, the Markov Condition may be violated; and, (iv) in a population of linear systems with the same structure, exogenous variances and linear coefficients, if the linear coefficients satisfy special constraints then the Faithfulness Condition may be violated. Case (iv) arises in linear models, for example, when linear coefficient values for different dependencies have the exact values required to cancel one another, as in figure 2 when a = - be.

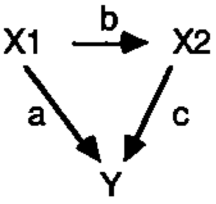


Figure 2

It can be shown that (Spirtes, Scheines and Glymour, 1992)

Theorem 1: If P is faithful to some directed acyclic graph, then P, G satisfy the Markov and Faithfulness Conditions if and only if

(i) for all vertices, X, Y of G, X and Y are adjacent if and only if X and Y are dependent conditional on *every* set of vertices of G that does not include X or Y; and

(ii) for all vertices X, Y, Z such that X is adjacent to Y and Y is adjacent to Z and X and Z are not adjacent, X -> Y <- Z is a subgraph of G if and only if X, Z are dependent conditional on *every* set containing Y but not X or Z.

Consideration of part (i) of this theorem explains why in structure 1 of figure 1 regression procedures incorrectly select X2 as a variable directly influencing Y: The structure and distribution satisfy the Markov and Faithfulness conditions, but linear regression takes a variable X_i to influence Y provided the partial correlation of X_i and Y controlling for all other **X** variables does not vanish. Part (i) of Theorem 1 shows that the regression criterion is insufficient. It follows immediately from Theorem 1 that, assuming the Markov and Faithfulness Conditions, regression of Y on a set **X** of variables will only yield an unbiased estimate of the influences of the **X** variables provided in the true structure no **X** variable is the effect of Y or has a common unmeasured cause with Y.

Theorem 1 immediately suggests an algorithm for inferring the set of directed acyclic graphs that a given distribution is faithful to when the two conditions are met (Spirtes, Glymour and Scheines, 1990):

SGS algorithm:

```
for each pair of variables check whether they are dependent conditional on every
subset of the remaining variables; if so put an edge between them;

for each triple X adjacent to Y adjacent to Z with X, Z not adjacent, check if X, Z
are dependent conditional on every set of the remaining variables that contains Y;
if so orient the adjacencies into Y otherwise not;

output all graphs in which the orientations of all remaining edges do not produce
cycles or new collisions A -> B <- C unless A and C are adjacent.
```

Given the population correlations, or correct decisions about statistical dependencies in the population, the algorithm will return correct information about the causal structure when there are no unmeasured common causes. SGS is not, however, feasible save for small numbers of variables, since all subsets of regressors must be checked, and the computational and statistical requirements of the algorithm therefore increase exponentially with the number of variables, regardless of the true structure. Another procedure, the PC algorithm (Spirtes, Glymour and Schemes, 1991), while less intuitive in statement, is asymptotically input/output equivalent to SGS (assuming the Markov and Faithfulness conditions) and for fixed low degree of the graph faithful to the distribution, PC requires exponentially fewer conditional independence tests than SGS. With adequate sample sizes, PC will run on a hundred or more variables for samples from sparse graphs.[4]

The algorithms described assume that no unmeasured common causes contribute to statistical dependencies among the measured variables. Building on work of Verma (1990), Spirtes (1992) found a generalization of the PC algorithm that is feasible for sparse graphs and in the large sample limit returns correct information about structure whether or not unmeasured common causes act, provided the entire system of variables, including any unknown common causes, satisfies the Markov and Faithfulness Conditions-of course, the marginal distribution over the measured variables need not satisfy either condition. Even the statement of the "Fast Causal Inference" (FCI) algorithm requires an intricate set of graph-theoretic concepts. A statement of the algorithm and proofs of the correctness are given in Spirtes (1992) and in Spirtes, Scheines and Glymour (1992). For an illustration of the power of the procedure, consider the following quite imaginary structure, in which the variables in boxes are unmeasured:

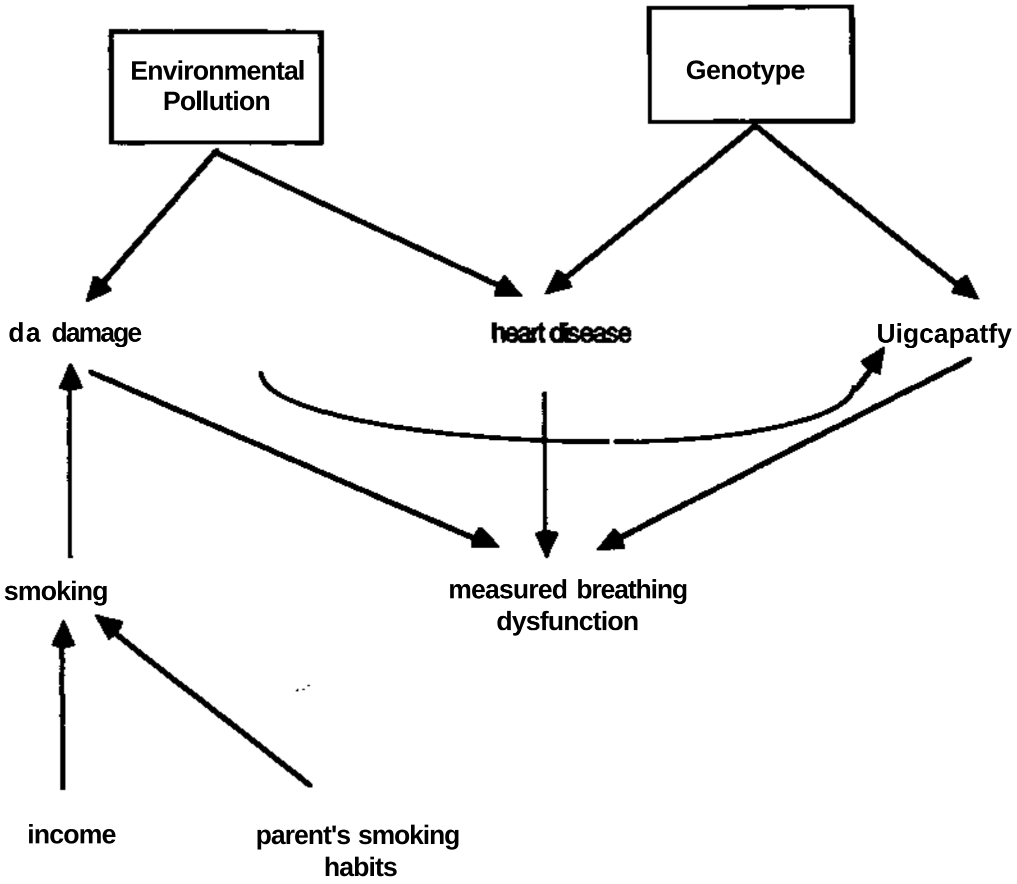


Figure 3

Given large sample correlations from a linear structure represented by this graph, normally distributed exogenous and error variables, and uncorrelated errors, the FCI algorithm recovers the structure almost uniquely. The output is:

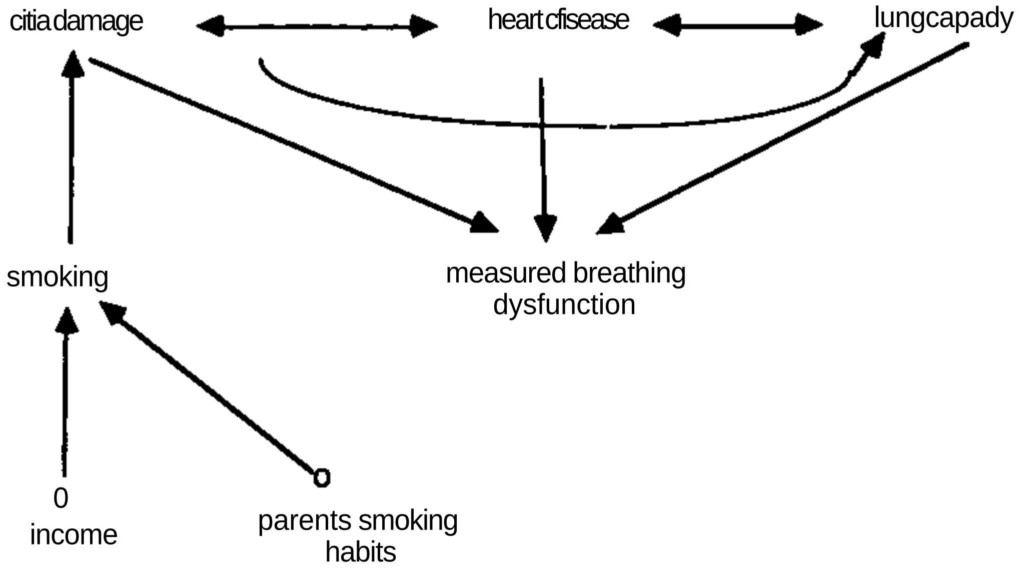


Figure 4

The double headed arrows indicate the presence of unmeasured common causes; the "o" at the ends of two edges indicate that the procedure cannot determine whether or not those ends should have an arrowhead.

Prior substantive knowledge about causal order can be integrated with each of these algorithms. If it is known, for example, that C occurs later than A and B, then no sets including C are used in testing whether A and B should be adjacent; if it is known that A precedes B, then an edge between A and B must be oriented into B.

4. The Solution and Its Application

Assuming appropriate variables have been measured and the population sampled satisfies the Markov and Faithfulness conditions, the PC and FCI algorithms offer a straightforward method for addressing the particular difficulties with regression noted in the first two sections.

We begin by noting that for the twenty samples from structure 1, in every case our implementation of the PC algorithm-which of course assumes there are no latent variables- selects {X1, X5} as the variables that directly influence Y. Our implementation of the FCI algorithm, which makes no such assumption, in every case says that X1 directly influences Y, that X5 may, and that the other variables do not.

In each of the other three structures in figure 1 with sufficiently large samples multiple regression methods will make comparable errors, always including X2 and X3 among the "significant" or "best" or "important" variables. In contrast the FCI algorithm will give the following results:

Structure	Direct Influence	No Direct Influence	Undetermined
2	X1	X2, X3, X4	X5
3	X1	X4	X2, X3, X5
4	X1, X5	X4, X6, X7	X2, X3

In all of these cases the procedure either correctly determines that X2 and X3 have no influence on Y, or determines that the issue cannot be decided.

4.1 Example 1: Components of the Armed Forces Qualification Test

The AFQT is a test battery used by the United States armed forces. It has a number of component tests, including those listed below:

Arithmetical Reasoning (AR)
Numerical Operations (NO)
Word Knowledge (WK)

In addition a number of other tests, including those listed below, are not part of the AFQT but are correlated with it and with its components:

Mathematical Knowledge (MK)
Electronics Information (EI)
General Science (GS)
Mechanical Comprehension (MC)

Given scores for these 8 measures on 6224 armed forces personnel, 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. The covariance matrix is given below

n = 6224

af	no	wk	ar	mk	ei	me	gs
253.985							
29.649	51.7649						
60.3604	6.29317	41.967					
57.6566	14.5143	16.0226	40.9329				
29.3763	18.2701	13.2055	20.6052	40.7386			
36.2318	2.10733	22.6958	16.3664	12.1773	63.1039		
35.8244	4.45539	17.4155	20.3952	16.459	35.1981	62.9647	
38.251	5.61516	27.1492	14.7402	14.8442	29.9095	26.6842	48.93

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 adjacent to AFQT, and hence the only variables that can be components of AFQT. (Spirtes, Glymour, Scheines and Sorensen, 1990)5 in this case regression methods probably fail because of structural problems like those discussed in the first section of this paper.

4.2 Example 2: Causes of Spartina Biomass

A recent textbook on regression (Rawlings 1988) skillfully illustrates regression principles and techniques for a biological study in which it is reasonable to think there is a causal process at work relating the variables. According to Rawlings, Linthurst (1979) obtained five samples of Spartina grass and soil from each of nine sites on the Cape Fear Estuary of North Carolina. Besides the mass of the grass (bio), fourteen variables were measured for each sample:

1. Free Sulfide (h2s)
2. Salinity (sal)
3. Redox potentials at ph 7 (eh7)
4. Soil pH in water (ph)

5. Buffer acidity at pH 6.6 (buf)
6. Phosphorus concentration (p)
7. Potassium concentration (k)
8. Calcium concentration (ca)
9. Magnesium concentration (mg)
10. Sodium concentration (na)
11. Manganese concentration (mn)
12. Zinc concentration (zn)
13. Copper concentration (cu)
14. Ammonium concentration (nh4)

The correlation matrix is as follows:

bio	h2s	sal	eh7	ph	buf	p	k	ca	mg	na	mn	zn	cu	nh4
1.0														
.33	1.0													
-.10	.10	1.0												
.05	.40	.31	1.0											
.77	.27	-.05	.09	1.0										
-.73	-.37	-.01	-.15	-.95	1.0									
-.35	-.12	-.19	-.31	-.40	.38	1.0								
-.20	.07	-.02	.42	.02	-.07	-.23	1.0							
.64	.09	.09	-.04	.88	-.79	-.31	-.26	1.0						
-.38	-.11	-.01	.30	-.18	.13	-.06	.86	-.42	1.0					
-.27	0.00	.16	.34	-.04	-.06	-.16	.79	-.25	-.90	1.0				
-.35	.14	-.25	-.11	-.48	.42	.50	-.35	-.31	-.22	-.31	1.0			
-.62	-.27	-.42	-.23	-.72	.71	.56	.07	-.70	.35	.12	.60	1.0		
.09	.01	-.27	.09	.18	-.14	-.05	.69	-.11	.71	.56	-.23	.21	1.0	
-.63	-.43	-.16	-.24	-.75	.85	.49	-.12	-.58	.11	-.11	.53	.72	.93	1.0

The aim of the data analysis was to determine for a later experimental study which of these variables most influenced the biomass of Spartina in the wild. Greenhouse experiments would then try to estimate causal dependencies out of the wild. In the best case one might hope that the statistical analyses of the observational study would correctly select variables that influence the growth of Spartina in the greenhouse. In the worst case, one supposes, the observational study would find the wrong causal structure, or would find variables that influence growth in the wild (e.g., by inhibiting or promoting growth of a competing species) but have no influence in the greenhouse.

Using the SAS statistical package, Rawlings first analyzed only six variables: the outcome variable, bio, together with sal, pH, k, na and zn. He found that

(i) in a multiple regression of bio on sal, pH, k, na and zn only pH has a significant regression coefficient;

ii) backward elimination of one variable at a time yields a best model with pH and k as the only regressors;

iii) all subsets, i.e., all possible regressions, yields a best model with pH and na as the only regressors;

Rawlings subsequently considered all fifteen variables, analyzing the variable set first with a multiple regression and then with two stepwise regression procedures from the SAS package. A search through all possible subsets of regressors was not carried out, presumably because the candidate set of regressors is too large. The results are as follows:

(iv) a multiple regression of bio on all other variables gives only k and cu significant regression coefficients;

(v) the two stepwise regression procedures[6] both yield a model with pH, mg, ca and cu as the only regressors, and regression on just these variables gives them all significant coefficients;

(vi) simple regressions one variable at a time identify pH, buf, ca, zn and nh4.

Seven different methods and six different results; what is one to think? This analysis is supplemented by a ridge regression, which increases the stability of the estimates of coefficients, but the results for the point at issue-identifying the important variables--are much the same as with least squares. Rawlings also provides a principal components factor analysis and various geometrical plots of the components. The result is only a clustering of regression variables. He reports that "None of the results was satisfying to the biologist; the inconsistencies of the results were confusing and variables expected to be biologically important were not showing signfiicant effects." (p. 361).

If we apply the PC algorithm to the Linthurst data then judged by this sample there is one extremely robust conclusion: the only variable that may *directly* influence biomass in this sample[7] is pH; pH is distinguished from all other variables by the fact that the correlation of every other variable with bio vanishes when pH is controlled for.[8] The relation is not symmetric; the correlation of pH and bio, for example, does not vanish when buf is controlled. The algorithm finds pH to be the only variable adjacent to bio no matter whether we use a significance level of.05 to test for vanishing partial correlations, or a level of 0.1, or a level of 0.2. We find the same result at all of these significance levels if we include only the five variables in Rawling's first set of candidate regressors or if we include all fourteen of the regressors. In all of these cases, the PC algorithm or the FCI algorithm yield the result that pH and only pH can be directly connected with bio. Of course, over a larger range of values of the variables there is little reason to think that biomass depends linearly on the regressors, or that factors that have no influence in producing variation within this sample would continue to have no influence. Nor can our analysis determine whether the relationship between pH and biomass is confounded by one or more unmeasured common causes, but the principles of the theory suggest that in this case that is unlikely. If pH and biomass have a common unmeasured cause T, say, and any other variable, Z, among the 13 others either causes pH or has a common unmeasured cause with pH, then Z and biomass should be correlated conditional on pH, which appears not to be the case.

The program and theory lead us to expect that if pH is forced to have some constant value like those in the sample, then manipulations of other variables within the ranges evidenced in the sample will have no effect on the growth of Spartina. Lindhurst's thesis confirms the prediction. In an experiment Lindhurst describes, samples of Spartina were collected from a salt marsh creekbank and transplanted to a greenhouse . Using a 3 X 4 X 2 (ph X salinity X aeration) randomized complete block design with four blocks, the plants were given a common nutrient solution with varying values pH and salinity and aeration. The aeration variable turned out not to matter in this experiment. Acidity values were pH 4, 6 and 8. Salinity of the nutrient solutions was adjusted to 15, 25 35 and 45 %o. In contrast, in the observational data Rawlings reports (p. 358), almost all salinity measurements are around 30--the extremes are 24 and 38. Compared to the experimental study rather retricted variation was observed in the wild sample. The observed values of pH in the wild, however, are clustered at the two extremes; only four observations with within half a pH unit of 6, and no observations at all occurred at pH values between 5.6 and 7.1.

Lindhurst found experimentally that growth varied with salinity at ph 6 but not at the other pH values, 4 and 8, while growth varied with pH at all values of salinity.(po. 104). Each variable was correlated with plant mineral levels. Lindhurst considered a variety of mechnisms by which extreme pH values might control plant growth:

> At pH 4 and 8, salinity had little effect on the performance of the species. The pH appeared to be more dominant in determining the growth response. (p.108)
>
> The overall effect of pH at the two extremes is suggestive of damage to the root directly, thereby modifying its membrane permeability and subsequently its capacity for selective uptake. (p. 109).

A comparison of the observational and experimental data suggests that the PC prediction was essentially correct and can be extrapolated through the variation in the populations sampled in the wild, but cannot be extrapolated through pH values that approach neutrality.

In this case multiple regression of biomaas on all of the other measured variables probably failed because the sample size was small compared to the number of variables, so that the tests for vanishing partial regression coefficients had little power. In contrast, the TETRAD procedure never needed to test for more than a second order vanihsing partial correlation.

4.3 Example 3: The Effects on Foreign Investment on Third World Political Forms

Timberlake and Williams (1984) used regression to claim foreign investment in third-world countries promotes dictatorship. They measured political exclusion (po) (i.e., dictatorship), foreign investment penetration in 1973 (fi), energy development in 1975 (en), and civil liberties (cv). Civil liberties was measured on an ordered scale from 1 to 7, with lower values indicating *greater* civil liberties. Their correlations for 72 "non-core" countries are

po	1.0	-.175	-.480	.868
fi		1.0	.330	-.391
en			1.0	-.430
cv				1.0

Their inference is unwarranted. Their model and the TETRAD II model (obtained with the PC algorithm using a .12 significance level to test for vanishing partial correlations) are shown in figure 5.[9]

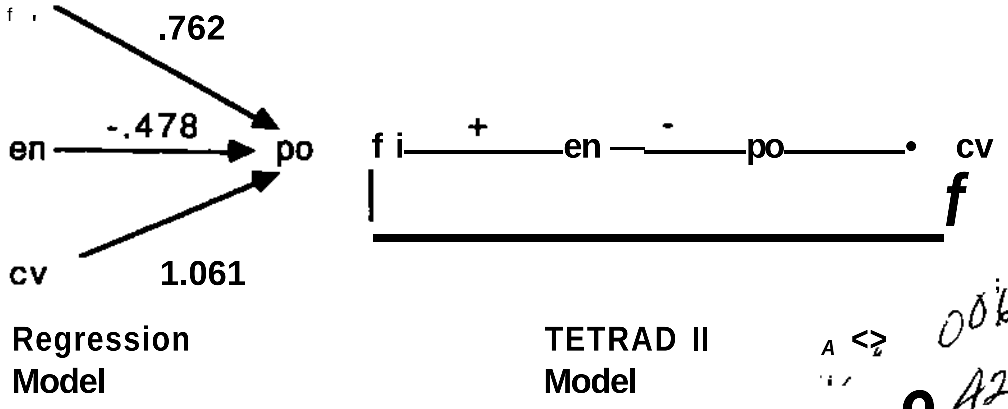


Figure 5

Neither PC nor FCI will orient the fi-en and en-po, edges, or determine whether they are due to at least one unmeasured common cause. Maximum likelihood estimates of the TETRAD II model require that the influence of fi on po (if any) be negative, and the model easily passes a likelihood ratio test with the EQS program. If the TETRAD II model is correct, Timberlake and William's regression model appears to be a case in which an effect of the outcome variable is taken as a regressor, as in structure 1 of figure 1.

4.4 Example 4: College Plans

Sewell and Shah (1968) studied five variables from a sample of 10,318 Wisconsin high school seniors. The variables and their values are:

sex (sex), measured by 0 for male and 1 for female;
Intelligence Quotient (iq), measured from least to highest by numerical values;
college plans (cp), measured by 0 for yes and 1 for no;
parental encouragement (pe), measured by 0 for low and 1 for high;
socioeconomic status (ses), measured from least to highest by numerical values 0,1, 2, 3.

They offer the following causal hypothesis:

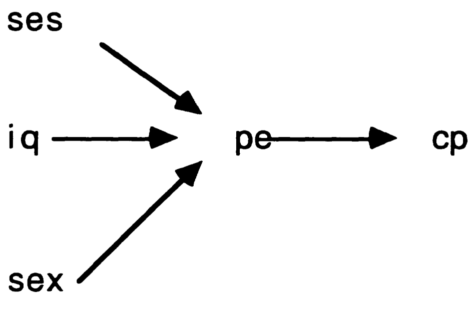


Figure 6

The data were reanalyzed by Fienberg (1977), who attempted to give a causal interpretation using log-linear models, but found a model that could not be given a graphical interpretation.

Given prior information that orders the variables by time as follows

1 sex
2 iq pe ses
3 cp

so that later variables cannot be specified to be causes of earlier variables, the TETRAD II output with the PC algorithm is the structure:

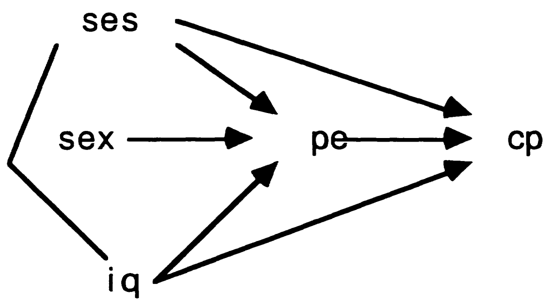


Figure 7

The program cannot orient the edge between iq and ses. It seems very unlikely that the child's intelligence causes the family socioeconomic status, and the only sensible interpretation is that ses causes iq, or they have a common unmeasured cause. Choosing the former, we have a directed graph whose joint distribution can be estimated directly from the sample. We find, for example, that the maximum likelihood estimate of the probability that males have college plans is .72, while the probability for females is .68. Judged by this sample the probability a child with low IQ, no parental encouragement, and low socioeconomic status plans to go to college is .011; more distressing, the probability that a child otherwise in the same conditions but with a high IQ plans to go to college is only .124.

4.5 Example 5. More Simulated Data

The short-run reliabiities of search procedures that make unpredictable sequential tests of hypotheses can, so far as we know, only be established by simulation studies. To illustrate what can (and should) be done in this regard we generated data from the graph of figure 8:

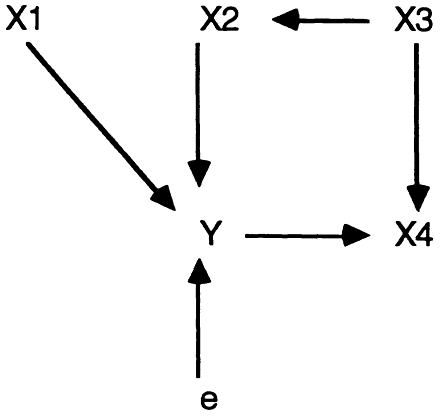


Figure 8

For both the linear and the discrete cases, one hundred trials were run at each of sample sizes 2,000 and 10,000 using the SGS algorithm. Results were scored separately for errors concerning the existence and the directions of edges, and for correct choice of regressors. Recall that an edge existence error of commission (Co) occurs when any pair of variables are adjacent in the output but not in the pattern of the graph in figure 5. An edge direction error of commission occurs when any arrowhead not in the pattern of (b) occurs in the output in an edge occurring in the pattern of (b). Errors of omission (Om) are defined analogously in each case. The results are tabulated as the average over the trial distributions of the ratio of the number of actual errors to the number of possible errors of each kind. The proportion of trials in which both (Both) actual causes of Y were correctly identified (with no incorrect causes), and in which one (One) but not both causes of Y were correctly identified (again with no incorrect causes) were recorded for each sample size:

Variable	#trials	n	%Edge Existence		%Edge Direction		%Both	%One
Type			Errors		Errors		Correct	
			Co	Om	Co	Cm	Cause(s)	
Linear	100	2000	1.4	3.6	3.0	5.4	85.7	3.6
Linear	100	10,000	1.6	1.0	2.7	2.2	90	7
Discrete	100	2000	0.6	16.6	29.5	21.8	38	34
Discrete	100	10,000	1.2	7.4	30.0	9.1	60	25

For purposes of prediction and policy, the numbers in the last two columns suggest that the procedure quite reliably finds real causes of the outcome variable when the statistical assumptions of the simulations are met, the sample is large and a causal structure like that in figure 14 obtains.

5. Conclusion

In the absence of very strong prior 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 should never be used at all in any contexts where causal inferences are at stake. We recommend that in such contexts the unconfounded causes of the outcome variable, if any, be identified by the PC or FCI algorithms or some equally reliable procedure, and multiple regression applied using the variables thus selected.

References

Fienberg, S. E. (1977). *The Analysis of Cross-classified Categorical Data,* M.I.T. Press: Cambridge, Mass.

Fox, J. (1984). *Linear Statistical Models and Related Methods,* John Wiley and Sons, New York.

Freedman, D. (1983). A note on screening regression equations. *American Statistician,* 37, 152-155.

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

Kiiveri, H., Speed, T., and Carlin, J. (1984). "Recursive Causal Models." *Journal of the Australian Mathematical Society,* Vol. 36, pp. 30-52.

Lauritzen, S., Speed, T. and Vijayan, K. *Decomposable Graphs and Hypergraphs,* preprint 1978, no. 9, Institute of Mathematical Statistics, University of Copenhagen.

Linthurst, R. (1979) Aeration, Nitrogen, PH and Salinity as Factors Affecting Spartina Alterniflora Growth and Dieback. Ph.D Thesis, North Carolina State University, Raleigh.

Mosteller, G. and Tukey, (1977) J. *Data Analysis and Regression,* Addison Wesley, New York.

Pearl. J. (1988). *Probabilistic Reasoning in Intelligent Systems,* Morgan and Kaufman: San Mateo.

Pratt, J.W., and Schlaifer, R. (1988). "On the Interpretation and Observation of Laws," *Journal of Econometrics,* Vol. 39, pp. 23-52.

Rawlings, J. *Applied Regression Analysis,* Wadsworth, Belmont, Ca., 1988.

Scheines, R., Spirtes, P., Glymour, G.f and Sorensen, S. (1990). "Causes of Success and Satisfaction Among Naval Recruiters," *Report to the Naval Personnel Research Development Center,* July.

Sewell, W. H. and Shah, V. P. (1968). Social class, parental encouragement, and educational aspirations. *American Journal of Sociology* 73. 559-572

Spirtes, P., Glymour, C, and Scheines, R (1991). "An Algorithm for Fast Recovery of Sparse Causal Graphs." *Social Science Computer Review,* 9 62-72.

Spirtes, P., Glymour, C, and Scheines, R. (1990). "Causality from Proability", in G. McGee, ed., *Evolving Knowledge,* Pitman.

Spirtes, P., Glymour, C, and Scheines, R. (1992) *Causality, Prediction and Search,* Springer, New York, forthcoming.

Spirtes, P. (1992) "Bulding Models with Latent Variables" in B. Skyrms, ed., *Proceedings to the International Congress on Logic, Methodology and Philosophy of Science,* Upsalla, forthcoming.

Verma, T. and Pearl, J. (1990). "Equivalence and Synthesis of Causal Models," *Proceedings of the Sixth Conference on Uncertainty in Aritificial Intelligence,* Cambridge, Mass.

Wright, S. (1934). "The Method of Path Coefficients", *Annals of Mathematical Statistics,* 5, pp. 161-215.

[1] Research for this paper was supported in part by contract number N00114-89-J-1964 from the Navy Personnel Research and Development Center and the Office of Naval Research, and by a grant from the program on Knowledge, Models and Cognitive Systems of the National Science Foundation. We thank Steve Sorensen for valuable help with data.

[2] In linear regression, we understand the "direct influence" of Xj on Y to mean (i) the change in value of a variable Y that would be produced in each member of a population by a unit change in Xj, with all other X variables forced to be unchanged. Other meanings might be given, for example: (ii) the population *average* change in Y for unit change in Xi, with all other X variables forced to be unchanged; (Hi) the change in Y in each

member of the population for unit change in Xi; (iv) the population average change in Y for unit change in Xi; etc. Under interpretations (iii) and (iv) the regression coefficient is an unreliable estimate whenever Xi also influences other regressors that influence Y. Interpretation (ii) is equivalent to (i) if the units are homogeneous and the stochastic properties are due to sampling; Otherwise, regression will be unreliable under interpretation (i) except in special cases, e.g., when the linear coefficients, as random variables, are independently distributed (in which case the analysis given in this paper still applies (Glymour, Spirtes and Scheines, 1991a)).

[3] We used the Monte procedure in TETRAD II (Spirtes, Scheines, Meek and Glymour, 1991).

[4] The PC and FCI algorithms are implemented in the TETRAD II program using Fisher's Z and a normal table to test for vanishing partial correlations in the linear case, and G^2 (Fienberg, 1977) to test for conditional independence among discrete variables. The default significance level is .05. At present the program is available to anyone with a Unix machine with PASCAL compiler and network connection by writing Richard Scheines at rs2l@andrew.cmu.edu

[5] In fact, we were inadvertently misinformed that all seven tests are components of AFQT and we first discovered otherwise with the SGS algorithm.

[6] The "maximum R-square" and "stepwise" options in PROC REG in the SAS program.

[7] Although the definition of the population in this case is unclear, and must in any case be drawn quite narrowly.

[8] More exactly, at .05, with the exception of mg the partial correlation of every regressor with bio vanishes when some set containing pH is controlled for; the correlation of mg with bio vanishes when ca is controlled for.

[9] Searches at lower significance levels remove the adjacency between fi and en.