Disproving Causal Relationships Using Observational Data

Henry L. Bryant[a], David A. Bessler[b], and Michael S. Haigh[c]

[a] Research Assistant Professor and [b] Professor
Department of Agricultural Economics
Texas A&M University
2124 TAMUS
College Station, TX 77843-2124
Tel: 979-845-5913
Fax: 979-845-3140

[c] Associate Chief Economist
Office of the Chief Economist
U.S. Commodity Futures Trading Commission
1155 21st Street, N.W., Washington, DC
Tel: 202-418-5063
Fax: 202-418-5660

Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Long Beach, California, July 23-26, 2006

Disclaimer: The views expressed in this paper are those of the authors and do not, in any way, reflect the views or opinions of the U.S. Commodity Futures Trading Commission.

Disproving Causal Relationships Using Observational Data

Abstract: Economic theory is replete with causal hypotheses that are scarcely tested because economists are generally constrained to work with observational data. This article describes the use of causal inference methods for testing a hypothesis that one random variable causes another. Contingent on a sufficiently strong correspondence between the hypothesized cause and effect, an appropriately related third variable can be employed for such a test. The procedure is intuitive, and is easy to implement. The basic logic of the procedure naturally suggests strong and weak grounds for rejecting the hypothesized causal relationship. Monte Carlo results suggest that weakly-grounded rejections are unreliable for small samples, but reasonably reliable for large samples. Strongly-grounded rejections are highly reliable, even for small samples.

Introduction and Background

Questions of causality have been central to economics from its beginnings, as the title of Adam Smith's *An Inquiry into the Nature and Causes of the Wealth of Nations* clearly indicates. Economic theory is, at its most fundamental level, a body of hypotheses regarding causal relationships among economic variables – endowments, production, exchange, and consumption of goods, rates of exchange between goods and stores of value, aggregations of such quantities, and the evolution of these quantities over time. Does an increase in income cause a person to consume more of a good? Does an increase in the money supply cause higher aggregate output?

Causal relationships in economics are not contemplated merely to satisfy idle intellectual curiosity. An understanding of such relationships is necessary if one wishes to address counterfactual questions of economic policy, and successfully impact the level of one quantity through the direct manipulation of another (Reiss and Cartwright, 2003). Hoover (2001) goes so far as to state that the uncovering of causal relationships among economic variables for the purpose of policy making is the "ultimate justification" for the

study of macroeconomics. Would an increase in the Federal Funds rate cause a decrease in the rate of general price inflation? Will greater availability of public transport cause reduced aggregate fuel consumption, helping us meet a policy objective of reducing releases of green house gases?

Even though questions of causality are an integral part of economic theory, the practice of economic measurement has had an uneasy relationship with the matter. Haavelmo (1944) and other Cowles Commission econometricians devised structural equation models that explicitly represented hypothesized causal relationships. Explicit causal interpretations of these types of models have largely fallen out of favor, however, and they are today often interpreted simply as compact representations of joint probability distributions (Pearl, 2000). A perceived empirical failure of such structural modeling efforts motivated the extensive adoption of multivariate time series methods that have no clear causal interpretation (Heckman, 2000).

Econometricians' reluctance to draw strong, explicit conclusions regarding causality no doubt stems in large part from the observational data with which we are forced to work. Mill (1884) regarded causal inference using observational data as impossible, a sentiment that has been shared by many economists since. Even when confronted with empirical results that seem inconsistent with the causal content of economic theory, econometricians will generally assign blame to auxiliary hypotheses rather than questioning the theory itself (Blaug, 1992). Thus economic theories are generally "confirmed" of "verified"; rigorous testing of hypothesized causal relations in economics is sorely limited.

Despite Mill's beliefs on the matter, many scholars have begun to infer causal relationships from observational data. Reichenbach (1956) proposed that causal relationships among random variables have specific implications for associated statistical independence relations. Hausman (1983) provides an early acknowledgment in economics that such causal inference should be feasible. More recently, several algorithms for conducting such inference have been proposed (Spirtes, Glymour and Scheines, 2000; Glymour and Cooper, 1999; and Pearl, 2000). This literature has focused on a definition of cause that stresses "manipulation" rather than statistical regularity or prediction. So that variable A is said to cause variable B if and only if one can manipulate B by manipulating A (Woodward 2003, chapter 2 and Pearl 2000, page 85). Manipulation-based definitions of causality are an improvement on prediction-based definitions, as they admit the possibility of latent variables. For example, a drop in the reading on a barometer (a change in position of a needle on a barometric pressure scale) predicts stormy weather in the future; however, one doesn't believe that by manipulating the barometer, by physically moving the needle on the instrument, one can affect the status of future rain conditions. One recognizes the existence of a latent variable (atmospheric pressure) that affects both the barometer reading and weather.

Hoover (2001) considers the use of causal inference methods in economics. Swanson and Granger (1997), Demiralp and Hoover (2003), and Hoover (2005) describe the use of causal inference algorithms for inferring contemporaneous causal relations among variables in vector autoregressions. Demiralp, Hoover, and Perez (2006) advance a bootstrap method for assessing the confidence that can be placed on such results.

Applications in economics include Akleman, et al. (1999), Haigh and Bessler (2003), and Bessler (2003).

The designers of causal inference algorithms seem to intend them to be used in a manner that might be described as "data mining" or "machine learning". In such use, observations of a large number of potentially related variables are assembled, and a causal structure among those variables is inferred. Most proposed algorithms conduct this overall inference by sequentially conducting several individual tests of conditional independence among the variables. This multiple testing leads to criticism that the overall probability of an error is unknown, and possibly unreasonably high, particularly for a large system. Casual experimentation with the algorithms using data sets with a moderate to high number of variables suggests that results are indeed fragile, and reversals of the direction of causal flow are not uncommon as one changes the algorithms' parameters. Demiralp and Hoover (2003) investigate such issues using Monte Carlo methods, and find that the probabilities of such errors are sensitive to the peculiarities of the data sets and are difficult to quantify.

In this paper, we explore the use of causal inference methods for testing specific hypothesized causal relations – H_0: *A* causes *B*. Employing these methods in this way entails advantages over the more typical method of application. First, relatively small numbers of causally related variables are needed, and the researcher needs not observe all potentially causally relevant variables. Observing some variable *C* that is causally related in a certain way to *A* and *B* allows rejection H_0, regardless of what other causally-related variables may or may not exist. Second, this narrow focus implies that there are only a limited number of ways in which latent, concomitant variables might influence the

observed variables involved in the test. This allows us to numerically estimate the size of the test. Third, testing such a hypothesis with respect to a particular *C* involves only three individual tests of unconditional independence. Due to this simplicity, the test can be easily conducted without using specialized computer software, and the researcher is fully aware of the basis for a particular conclusion.

After describing the application of causal inference methods to test specific causal hypotheses, we present the results of a Monte Carlo evaluation of the size of such tests. We then illustrate the method, rejecting a counter-intuitive causal hypothesis relating to Peltzman's (1975) traffic safety study.

The Casual Inference Algorithm Test of H_0: *A* Causes *B*

The method that we now describe is essentially a subset of the method described in Spirtes, Meek, and Richardson (1999). They present the Causal Inference (CI) algorithm, which is appropriate for inferring causal relationships among random variables when zero or more of those variables are unobserved.[1] This is a subset in the sense that we concern ourselves only with inference over sets of three observed variables (*A*, *B*, and *C*), and only with particular combinations of independence tests whose results may allow us to reject H_0: *A* causes *B*, based on evidence provided by *C*. This greatly reduces the complexity of the inference procedure and the potential computational burden. We attempt to provide as intuitive an explanation of the concepts as possible; the rigorous development of these concepts is presented in the original source and references cited therein.

[1] They present a "fast" version of their algorithm (the "FCI" algorithm), which is implemented in the Tetrad II, III, and IV computer programs. We have no need here for the extra steps that they take to reduce the computational burden when considering large numbers of variables.

Causal relationships between two random variables in an underlying linear causal structure are graphically represented by an edge, or line, connecting those variables. An arrowhead on one end of an edge indicates the direction of causation. We assume that there are no cycles present in the system. An example graph is presented in Figure 1, in which a latent variable *L* causes observed variables *A* and *B*, and *B* is additionally influenced by *C*.

We do not assume that all causally relevant variables are observed. This implies that the independence relations over the observed variables will be consistent with multiple underlying full causal structures. In graphs over only observable variables, such observationally equivalent causal structures will be represented using edges with circles on one end. An edge $A \text{ o}\!\!\rightarrow B$ indicates that either *A* causes *B*, or they share a latent common cause, or both.

We assume (**A1**) that Reichenbach's (1956) principle of the common cause holds: two variables are statistically dependent only if one variable causes the other, or they share one or more common causes.[2] The causation between two variables may be mediated by other variables – if *A* causes *C* and *C* causes *B*, we assume that *A* and *B* will be statistically dependent. We also assume (**A2**) that two variables that share a common cause will not be rendered statistically independent by peculiar, precisely offsetting structural parameters. That is to say, independence relations reflect the underlying causal structure.[3]

[2] This is generalized in the more recent causal inference literature as the causal Markov assumption, which extends the basic principal to accommodate conditional statistical independence between two indirectly causally related variables, where the conditioning is over a common cause or a mediating variable.
[3] This is referred to as the faithfulness condition in Spirtes, Glymour and Scheines (2000) and Glymour and Cooper (1999). Pearl (2000) calls this the stability condition.

Given our assumptions, we can infer something regarding the causal structures that underlie three observed variables based on their unconditional independence relations. Suppose that A and B are statistically independent (denoted $A \perp B$). It follows from **A1** and **A2** that we must reject all causal structures represented by the graph A o→ Z o→ B for any (possibly empty) set of variables Z, and we must therefore reject H_0: A causes B.

Suppose that A and B are not statistically independent (denoted $A \not\perp B$). There then exist four possible combinations of independence relations between the pairs $\{A, C\}$ and $\{B, C\}$.

Case 1: $A \perp C$, and $B \not\perp C$. A and B are casually related by **A1** and **A2**, as are B and C. By **A1**, **A2** and the independence of A and C we must conclude that A cannot cause C, either directly or indirectly, and that they cannot share a common cause. The underlying causal structure must be a member of the class of structures represented by the graph A o→ B ←o C. H_0 therefore cannot be rejected.

Case 2: $A \not\perp C$, and $B \perp C$. By the same logic presented in Case 1, we conclude that C o→ A ←o B. There is no possibility that A is a cause of B. H_0 is therefore rejected.

Case 3: $A \perp C$, and $B \perp C$. In this case C is not causally related to A and B by **A2**, and provides no information regarding the causal connection between them as it did in the first two cases. We therefore have no basis for rejecting H_0.

Case 4: A !⊥ C, and B !⊥ C. Again the logic of Cases 1 and 2 cannot be applied. There is no basis for rejecting H_0.[4]

There are thus two combinations of independence relations among *A*, *B*, and *C* that are sufficient to reject H_0: *A* causes *B*: either 1) $A \perp B$ or 2) A !⊥ B, A !⊥ C, and $B \perp C$. The interesting grounds for rejecting H_0 is, of course, finding A !⊥ B, A !⊥ C, and $B \perp C$. We refer to *C* as a *test instrument*, and to a *C* such that A !⊥ C, and $B \perp C$ as an *evidential test instrument*.

This can be easily understood on an intuitive level by examining graphs representing hypothetical causal structures in which *A* does cause *B*, as illustrated in Figure 2. If we observe some variable *C* that is causally related to *A*, then as *C* varies there should be some extent of corresponding variation in *B*. This is because either C indirectly causes *B* (as in panel i), or because *A* and *B* share a common cause (as in panels ii and iii). Stated differently, if no correspondence between *B* and is *C* is observed, then either *A* and *C* are not causally related, or *A* does not cause *B*.

Since the data are observational, it not possible to manipulate *A* and monitor *B* for possible changes. Essentially, what is required is a test instrument that is either naturally manipulating (in some sense) *A*, or providing evidence of some manipulation of *A*.[5] The selection of the instrument will be informed by the researcher's knowledge of the

[4] In the full Causal Inference algorithm (or its fast counterpart), tests for conditional independence of pairs of the observed variables would be conducted in this case, where the conditioning would be on the third variable. The causal Markov condition mentioned in note 2 would be invoked, rather than our **A1**. In such cases, however, there is no possibility of rejecting H_0. For example, if *A* and *C* are independent conditional on *B* would be consistent with the classes of causal structures A o→ B o→ C, A ←o B ←o C, and A ←o B o→ C. Any variable may be a cause of any other (either directly or indirectly).

[5] Our "test instrument" is very similar to the "switch variable" or "experimental handle" defined in Reiss (2003). We do not, however, require that *C* cause *A* as in his **EH1**. Such a *C* would be sufficient, but a *C* that shares a common cause with *A* will also serve. Our test instrument is also similar to the instruments used in instrumental variables estimation, as discussed by Reiss.

underlying problem. The researcher will likely have in mind some alternative hypothesized causal structure(s) wherein *B* causes *A* or they share a common cause (i.e., H_0 is false). A good choice of a *C* would, under some plausible alternative(s), be believed likely causally unrelated to *B* except via *A*. The existence of *any* evidential test instrument requires rejection of H_0, even though there will doubtlessly exist numerous non-evidential test instruments.[6]

A test of H_0: *A* causes *B* is readily operational in the linear, normal case using Fisher's z-test of correlations. Suppose that the underlying causal structure is linear in that it can be represented by a recursive structural equation model

(1) $$X = \Gamma_0 + \Gamma_1 X + \varepsilon$$

where *X* is a vector of random variables (both observed and unobserved), Γ_0 is a conformable coefficient vector, Γ_1 is a conformable triangular (for some ordering of the variables in *X*) coefficient matrix with non-zero terms corresponding to the directed edges in the corresponding graph, and ε is a conformable vector of independent normal errors. The variables in the non-zero terms on the right-hand side of this equation cause the variables on the left hand side, but the converse in not true. The test then consists simply of computing all three correlation coefficients, conducting Fisher's z-test on each, and determining if either of the conditions sufficient for rejecting H_0 are true: either 1) $\rho_{AB} = 0$ or 2) $\rho_{AB} \neq 0$, $\rho_{AC} \neq 0$, and $\rho_{BC} = 0$.

The first condition, $\rho_{AB} = 0$, is an inherently weak basis for concluding that *A* causes *B* is false. This is because the burden of proof in the z-test is opposite of that needed for the causal hypothesis – we would reject H_0: *A* causes *B* based on failing to

[6] This is, of course, Popper's swan argument. Popper would likely not approve of the inductive nature of the procedure described here, however.

reject H_0: $\rho_{AB} = 0$. We henceforth refer to this as a *weak-basis rejection* of H_0: *A* causes *B*. If *A* does, in fact, cause *B*, but the correspondence is weak, we are likely to often fail to reject $\rho_{AB} = 0$, especially for small samples. Due to these weak-form rejections, the overall size of the test of CI Algorithm test is approximately bound from below by one minus the power of the z-test.

A rejection of H_0: *A* causes *B* due to finding $\rho_{AB} \neq 0$, $\rho_{AC} \neq 0$, and $\rho_{BC} = 0$ (Case 2 above) is a *strong-basis rejection*. In this case, the burden proof in the z-tests is such that we are confident that the pairs {*A*, *B*} and {*A*, *C*} are indeed causally related. Moreover, any correspondence between *B* and *C* should be evident if *A* causes *B*, despite the fact that the burden of proof is opposite of that which is desired in the z-test of H_0: $\rho_{BC} = 0$. This is because the process that leads to a strong-basis rejection reflects a self-correcting mechanism that reduces the probability of a type II error in this latter z-test. The risk of such an error is greatest when $|\rho_{BC}|$ and the available sample are both small. As the sample gets smaller, however, the sample correlation coefficients r_{AB} and r_{AC} must be larger before we are convinced that the corresponding population correlation coefficients are not zero. On average, this corresponds to higher population correlation coefficients, including ρ_{BC} if *A* causes *B* is true. Thus the test of H_0: $\rho_{BC} = 0$ is not conducted in circumstances where it is highly susceptible to type II errors. Strong-basis rejections are therefore likely to reflect the underlying causal structure, and are unlikely to result from the inherent difficulty in discerning weak causal relationships using observational data.

Monte Carlo Simulations

To test the empirical size of the CI algorithm test, we generate a large number of random systems in which the null hypothesis is true, and observe the frequency with which it is rejected. The random systems all feature three observed variables, over which there are 25 possible acyclic causal structures.[7] In eight of these structures, A causes B directly (i.e., there is an edge $A \rightarrow B$), and in one structure A causes B indirectly ($A \rightarrow C \rightarrow B$).

The observed variables are assumed to be a subset of a larger causal structure that includes zero or more unobserved variables. There is an infinite number of possible full causal structures that might be considered, however there is a finite number of possible sets of independence relations among the observed variables. A latent variable that is a cause of only one of the observed variables does not impact these independence relations. Latent variables that do not cause any of the observed variables (but may be caused by them) would similarly not impact causal inference over the observed variables. Also, the independence relations among the observed variables will be identical whether two observed variables share a single latent common cause or share more than one such common cause.

For these reasons, we specify systems in which there are three possible latent variables, L_{AB}, L_{AC}, and L_{BC}, each of which is a latent common cause of the two indicated observed variables. Each of these may or may not be present in a system. There are thus 2^3 possible arrangements of latent common causes that may accompany the 9 causal structures among observed variables in which A causes B. We therefore consider $9 \times 2^3 = 72$ causal structures, which fully represent all possible patterns of independence relations

[7] Each of the three possible edges has one of three possible states: absent, pointing in one direction, or pointing in the opposite direction. There are therefore $3^3 = 27$ possible causal structures, two of which involve cycles: $A \rightarrow B \rightarrow C \rightarrow A$ and $A \leftarrow B \leftarrow C \leftarrow A$.

among observed variables. In each trial, we randomly select one of these 72 causal structures with equal probability.

For each of one hundred thousand trials, equation (1) is parameterized to reflect the selected causal structure. Without loss of generality, Γ_0 is a zero vector for all trials. Following Demiralp and Hoover (2003), the parameters of the Γ_1 matrix are selected to reflect three signal strengths. Individual elements are drawn from a U(0,*d*) distribution, where *d* is calibrated so that the mean parameter value will result in one of three desired population correlation coefficients between two variables[8]. These correlation levels are set at 0.25, 0.50, and 0.75 to reflect low, medium, and high signal strengths, respectively. For each observation in each trial, ε in (1) is drawn, with individual elements independently distributed as N(0,1). Equation (1) is then solved for $X' = [A, B, C, L_{AB}, L_{AC}, L_{BC}]$ for that observation.

Finally, for each trial we apply the CI algorithm test of H_0: *A* causes *B,* described in the previous section, to the observed variables *A*, *B*, and *C*. The numbers of failures to reject, weak-basis rejections, and strong-basis rejections are tabulated. All rejections constitute type I errors, as H_0 is true by design in all systems. We consider sample sizes of 50, 100, 250, 500, and 1,000. For each sample size, one hundred thousand trials are conducted. All z-tests are initially conducted using an alpha value of 0.10.

The proportions of trials that result in weak-basis and strong-basis rejections are reported in Table 1. For the medium and high signal strengths, the proportions of weak-basis rejections are reasonably low for all sample sizes, generally falling below the nominal size of 0.10 employed in the underlying z-tests. For large samples (by social

[8] When one is the sole cause of the other, in the sense of equation (1).

science standards), the proportions of weak-basis rejections are low, even when the signal strength is low. However, as expected, the limited power of the z-test at low signal strengths results in large proportions of weak-basis rejections for smaller samples. These results confirm the difficulty of detecting a weak causal relationship between two variables using observational data. This suggests the importance of disclosing the nature of a rejection (weak-basis or strong-basis) when applying the CI algorithm test, especially when working with small samples.

By contrast, the proportion of strong-basis rejections is reasonably low for all sample sizes and all signal strengths. In all cases the proportion of strong-basis rejections is below the proportion of weak-basis rejections. Performance again improves moderately as sample size increases, and improves dramatically as signal strength increases. The proportion of strong-basis rejections is almost always below the nominal size of 0.10 used in the underlying z-tests, and the researcher can thus have a much greater degree of confidence in such a result.

The above results suggest that it may be desirable to employ a higher nominal size in the underlying z-tests, in order to reduce the probability of making a type II error in the initial test of H_0: $\rho_{AB} = 0$. This might especially be advisable when the sample correlation coefficient between *A* and *B* is low, giving the researcher the opportunity to strongly reject H_0: *A* causes *B*. This would certainly reduce the incidence of weak-basis rejections, but what effect would this have on incidence of strong-basis rejections? A strong-basis rejection will not be possible if *either* a type I *or* type II error is made in the underlying z-tests. Adjusting the alpha value reduces the probability of one type of error,

but increases the probability of the other. Strong-basis rejections are thus naturally robust to the confidence level employed in the underlying independence tests.

We conduct a second set of simulations that use an alpha value of 0.20 in the underlying z-tests, with results presented in Table 2. The incidence of weak-basis rejections is indeed substantially improved for small sample sizes and low signal strengths, and the proportions of strong-basis rejections are very similar to the previous results in all cases. The proportions of strong-basis rejections are only worse for low signal strengths and the smallest two sample sizes, and then only marginally so. The higher alpha value for the z-tests thus affords a greater degree of confidence in weak-form rejections, while generally preserving the degree of confidence in strong-form rejections.

Of course, sampling variation will prevent the researcher from knowing a system's true underlying signal strength. The low signal strength columns in the tables thus reveal conservative levels of confidence that can be reported for weak-basis and strong-basis rejections. The confidence level for a strong-basis rejection is simply the proportion of such rejections reported in the tables. The confidence level for a weak-basis rejection should be the sum of the proportions of weak- and strong-basis rejections, as the researcher would presumably have accepted the stronger evidence.

Application

We intentionally apply the method to a hypothesis that most would consider intuitively false. Peltzman (1975) finds that alcohol consumption is significantly correlated with the total motor vehicle death rate. This is likely due to the fact that

increased alcohol consumption is associated with increased incidence of impaired driving. An alternative hypothesis however, might be that increased traffic deaths lead to increased grief and stress, which in turn lead to increased alcohol consumption. We test this latter hypothesis. We employ a test instrument that we expect to be causally related to the death rate, but not necessarily to alcohol consumption: average motor vehicle speed.

We employ annual observations from 1947 through 1993 of the total (both occupant and pedestrian) number of traffic fatalities divided by the total vehicle mileage (*DEATH*), average urban vehicle speed (*SPEED*), and average annual per capita alcohol consumption (*ALCOHOL*).[9] Augmented Dickey-Fuller tests indicate that *ALCOHOL* is mean stationary, that *DEATH* is trend stationary, and that *SPEED* is nonstationary in levels, but stationary in first differences. Furthermore, all series exhibit autoregressive characteristics. To accommodate these features of the data, we follow Swanson and Granger (1997), Demiralp and Hoover (2003), Haigh and Bessler (2003), and Hoover (2005) by conducting causal inference over filtered data. We estimate a vector-autoregression (VAR) in first differences, with constants, using a Schwarz (1978) information criterion-minimizing lag length of one. The innovations from this VAR are all mean stationary, and approximately normally distributed. We cannot reject normality for *DEATH* and *ALCOHOL* innovations using Jarque-Bera tests (p-values of 0.82 and 0.76, respectively). Normality for *SPEED* innovations is not rejected at a p-value of 0.91 if a single outlier reflecting the lowering of the national speed limit in 1974 to 55 miles

[9] Total traffic fatalities are taken from the National Safety Council's *Accident Facts* (through 1974), and the US Dept. of Transportation (DOT), National Highway Traffic Safety Admin., Fatal Accident Reporting System (after 1974). Total vehicle mileage and speed data are from US DOT, Federal Highway Admin., *Highway Statistics*. Alcohol consumption data are from US Dept. of Health and Human Services, Public Health Service, National Institute on Alcohol Abuse and Alcoholism, *Surveillance Report*, Dec. 1995.

per hour is omitted. The sample correlation coefficient between the putative cause and putative effect is r(*DEATH, ALCOHOL*) = 0.341. For 45 observations, we reject H0: $\rho = 0$ at the 20% level using Fisher's z-test for sample correlation coefficients greater than 0.198. Thus *DEATH* and *ALCOHOL* are causally-related, given Reichenbach's common cause principal, and there is not a weak correspondence that requires a weak-basis rejection of H_0: *DEATH* causes *ALCOHOL*. The sample correlation coefficients between the test instrument and putative cause is r(*SPEED, DEATH*) = 0.386. We are therefore confident that *SPEED* is causally-related to *DEATH*. Finally, the sample correlation coefficient between the test instrument and the putative effect is r(*SPEED*, *ALCOHOL*) = 0.136; they are not causally-related. *SPEED* is thus an evidential test instrument that informs a strong-basis rejection of H_0: *DEATH* causes *ALCOHOL*. Using Table 2, and conservatively assuming low signal strength in the underlying causal relations, we reject H_0 at approximately the 11% level of significance.

On an intuitive level, we can make an assumption about the direction of causal flow between *DEATH* and the test instrument, *SPEED* – most people would believe that increased vehicle speeds lead to greater traffic fatalities on average. Given this, if the hypothesis that increased traffic fatalities caused increased alcohol consumption was true, then there should be a significant correspondence between average vehicle speeds and alcohol consumption, as the former would indirectly cause the latter. No such correspondence is observed, and must reject the hypothesis. This example illustrates that causal hypotheses can be rejected at conventional levels of confidence, even when experimental manipulation is not possible, and only small numbers of observations are available.

Conclusion

This article describes the use of causal inference methods for testing a hypotheses that one specific random variable causes another. This is in contrast to the more standard use of such methods, which entails searching for a full set of causal relationships among numerous variables. We describe how, contingent on a sufficiently strong correspondence between the hypothesized cause and effect, an appropriately related third variable can be employed in such a test. The procedure is easily understood. In the linear normal case, the procedure is easy to implement, involving only the evaluation of three sample correlation coefficients using Fisher's z-test.

The basic logic of the testing procedure naturally suggests strong and weak bases for rejecting hypothesized causal relationships. Monte Carlo results confirm that for small samples, rejections motivated by the two different bases warrant substantially different levels of confidence. When the strength of the underlying causal relations is low, particularly between the hypothesized cause and effect, a small number of observations can only provide a low degree of confidence in a rejection of the null hypothesis. By contrast, when the strength of the underlying causal relationships is high, particularly that between the hypothesized cause and effect, even a relatively small number of observations can be used to reject the null hypothesis with a high degree of confidence. The simulation results reveal that the size of the test with respect to these strong evidence rejections is almost always lower than the alpha level employed in the underlying z-tests. Future work should examine the robustness of the procedure to non-normality and non-linearity.

We illustrate the method using only 45 observations of U.S. traffic fatality rates, which are hypothesized to cause per capita alcohol consumption. Using average vehicle speeds as a test instrument, we are able to strongly-reject this counter-intuitive hypothesis.

Economic theory is replete with causal hypotheses that are scarcely tested because economists are generally constrained to work with observational data. The procedure described here should facilitate the testing of such hypothesis, affording applied economists the opportunity to more closely realize the ideals of scientific inquiry.

References

Akleman, Derya G.; Bessler, David A.; and Burton, Diane M. (1999): "Modeling Corn Exports and Exchange Rates with Directed Graphs and Statistical Loss Functions," in Clark Glymour and Gregory F. Cooper (eds.) *Computation, Causation, and Discovery*, Menlo Park: American Association for Artificial Intelligence and Cambridge: MIT Press.

Bessler, David A. (2003): "On World Poverty: Its Causes and Effects," Research Bulletin. Food and Agricultural Organization of the United Nations, Rome.

Blaug, Mark (1992): *The Methodology of Economics*. Cambridge: Cambridge University Press.

Danks, David (2005): "Scientific Coherence and the Fusion of Experimental Results." *British Journal for the Philosophy of Science*, forthcoming.

Dermilap, S. and Hoover, Kevin D. (2003): "Searching for the Causal Structure of a Vector Autoregression," *Oxford Bulletin of Economics and Statistics*, 65:745-767.

Demiralp, S., Hoover, K., and Perez, S. (2006): "A Bootstrap Method for Identifying and Evaluating a Structural Vector Autoregression," working paper.

Haigh, Michael S.; and Bessler, Davis A. (2003): "Causality and Price Discovery: An Application of Directed Acyclic Graphs," *Journal of Business*, forthcoming.

Hausman, Daniel M. (1983): "Are There Causal Relations Among Dependent Variables?" *Philosophy of Science*, 50:58-81.

Hayek, Friederich A. (1964): "The Theory of Complex Phenomena," in *Studies in Philosophy, Politics, and Economics*, Chicago: University of Chicago Press.

Hoover, Kevin D. (2001): *Causality in Macroeconomics*, Cambridge: Cambridge University Press.

Hoover, Kevin D. (2005): "Automatic Inference of the Contemporaneous Causal Order of a System of Equations," *Econometric Theory*, 21:69-77.

Mill, John Stewart (1884): *Collected Works: A System of Logic Ratiocinative and Inductive*,

Pearl, Judea (2000): *Causality*, Cambridge: Cambridge University Press.

Peltzman, S. (1975): "Effects of Automobile Safety Regulation," *Journal of Political Economy,* 83:677-725.

Reichenbach, H. (1956): *The Direction of Time*, Berkeley: University of California Press.

Reiss, Julian (2003): "Practice Ahead of Theory: Instrumental Variables, Natural Experiments, and Inductivism in Econometrics," working paper, London School of Economics.

Reiss, Julian; and Cartwright, Nancy (2003): "Uncertainty in Econometrics: Evaluating Policy Counterfactuals," working paper, London School of Economics.

Schwarz, Gideon (1978): "Estimating the Dimensions in a Model," Annals of Statistics, 6, 461-464.

Spirtes, Peter; Glymour, Clark; and Scheines, Richard (2000): *Causation, Prediction, and Search*, Cambridge: The MIT Press.

Spirtes, Peter; Meek, Christopher; and Richardson, Thomas (1999): "An Algorithm for Causal Inference in the Presence of Latent Variables and Selection Bias," in Clark Glymour and Gregory F. Cooper (eds.) *Computation, Causation, and Discovery*, Menlo Park: American Association for Artificial Intelligence and Cambridge: MIT Press.

Swanson, Norman R.; and Granger, Clive W. J. (1997): "Impulse Response Functions Based on a Causal Approach to Residual Orthogonalization in Vector Autoregressions," *Journal of the American Statistical Association*, 92:357-367.

Woodward, James (2003): *Making Things Happen: A Theory of Causal Explanation*, New York: Oxford University Press.

Figure 1: An Example Graphical Representation of a Causal Structure

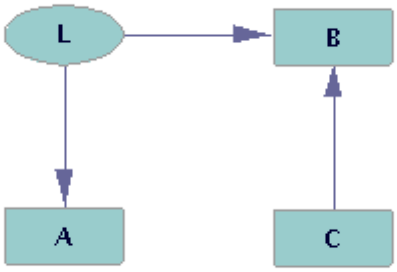
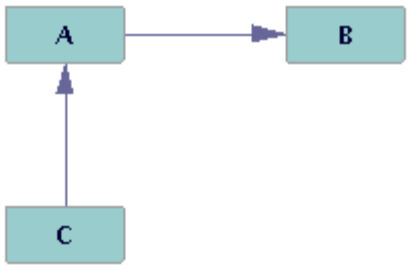
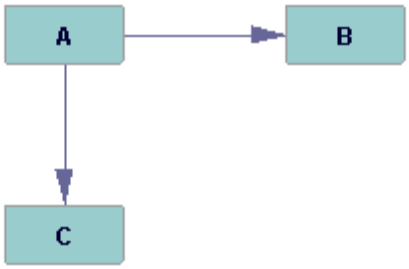


Figure 2: Graphs Representing Causal Structures in which A Causes B

(i)



(ii)



(iii)

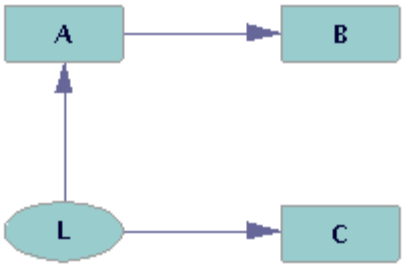


Table 1: Proportion of Rejections of a True H_0: A Causes B when Using an Alpha of 0.10 in Underlying z-tests.

	Low Signal Strength	*Medium Signal Strength*	*High Signal Strength*
N = 50			
Weak-basis Rejections	0.424	0.124	0.046
Strong-basis Rejections	0.094	0.043	0.013
N = 100			
Weak-basis Rejections	0.294	0.080	0.030
Strong-basis Rejections	0.106	0.035	0.009
N = 250			
Weak-basis Rejections	0.178	0.045	0.017
Strong-basis Rejections	0.100	0.025	0.006
N = 500			
Weak-basis Rejections	0.122	0.030	0.012
Strong-basis Rejections	0.085	0.020	0.005
N = 1,000			
Weak-basis Rejections	0.083	0.020	0.008
Strong-basis Rejections	0.071	0.016	0.004

Table 2: Proportion of Rejections of a True H_0: A Causes B when Using an Alpha of 0.20 in Underlying z-tests.

	Low Signal Strength	*Medium Signal Strength*	*High Signal Strength*
N = 50			
Weak-basis Rejections	0.331	0.094	0.035
Strong-basis Rejections	0.111	0.043	0.013
N = 100			
Weak-basis Rejections	0.228	0.061	0.023
Strong-basis Rejections	0.111	0.034	0.010
N = 250			
Weak-basis Rejections	0.136	0.034	0.013
Strong-basis Rejections	0.096	0.025	0.008
N = 500			
Weak-basis Rejections	0.093	0.023	0.009
Strong-basis Rejections	0.079	0.019	0.006
N = 1,000			
Weak-basis Rejections	0.064	0.016	0.006
Strong-basis Rejections	0.065	0.016	0.005