WI

SSRI
Workshop Series
8120

UNIVERSITY OF WISCONSIN-MADISON

SSRI Workshop Series

THE GENERAL EQUIVALENCE OF GRANGER AND SIMS CAUSALITY

Gary Chamberlain

8120

SOCIAL SYSTEMS RESEARCH INSTITUTE

Social Systems Research Institute

University of Wisconsin

THE GENERAL EQUIVALENCE OF GRANGER AND SIMS CAUSALITY

Gary Chamberlain

8120

June 1981

April 1980
Revised: July 1980,
June 1981

THE GENERAL EQUIVALENCE OF GRANGER AND SIMS CAUSALITY

by

Gary Chamberlain[1]
University of Wisconsin-Madison
and
National Bureau of Economic Research

ABSTRACT

Linear predictor definitions of causality are not adequate for discrete data. The paper extends the Granger and Sims definitions by using conditional independence instead of linear predictors. The extended definition of "y does not cause x" is that x is independent of past y conditional on past x. This is stronger than the strict exogeneity condition that y be independent of future x conditional on current and past x. Under a weak regularity condition, however, if y is independent of future x conditional on current and past x <u>and</u> past y, then y does not cause x.

THE GENERAL EQUIVALENCE OF GRANGER AND SIMS CAUSALITY

By

Gary Chamberlain[1]

1. INTRODUCTION

Let $[(x_t, y_t), t = \ldots, -1, 0, 1, \ldots]$ be a collection of random variables on a common probability space -- a stochastic process. Granger [6] defined "y does not cause x" as follows: the (minimum mean square error) linear predictor of x_{t+1} based on $x_t, x_{t-1}, \ldots, y_t, y_{t-1}, \ldots$ is identical to the linear predictor based on $x_t, x_{t-1}, \ldots$ alone.[2] Sims [10] defined x to be strictly exogenous relative to y if the linear predictor of y_t based on $\ldots, x_{t-1}, x_t, x_{t+1}, \ldots$ is identical to the linear predictor based on $x_t, x_{t-1}, \ldots$. Sims [10] showed that these two definitions are equivalent.[3] This is a beautiful result; we would like to know whether it still holds if linear predictors are replaced by a more general form of dependence.

In applying these ideas to longitudinal data on individuals, the prevalence of qualitative variables argues for considering models based on the entire conditional distribution instead of looking only at linear predictors. Suppose that y_{it} is zero or one, indicating, for example, whether or not individual i was employed in period t. We observe $(x_{i1}, y_{i1}, \ldots, x_{iT}, y_{iT})$ for i=1, ..., N individuals, and we regard these vectors as independent and identically distributed (i.i.d.) observations from the joint distribution of $(x_1, y_1, \ldots, x_T, y_T)$. Let $t = 1$ be the first period of the individual's (economic) life. Consider the following specification for the conditional probability that y_{it} equals one:

$$P(y_{it} = 1 | x_{i1}, \ldots, x_{iT}, c_i) = P(y_{it} = 1 | x_{i1}, \ldots, x_{it}, c_i),$$

where c is a latent variable that represents unmeasured characteristics of the individual; c is assumed to be constant over the sample period.

If c is independent of the x's, then, dropping the i subscripts, we have

$$P(y_t = 1 | x_1, \ldots, x_T) = \int P(y_t = 1 | x_1, \ldots, x_t, u) \, dP(c \leq u)$$

$$= P(y_t = 1 | x_1, \ldots, x_t),$$

so that x is strictly exogenous. However, if $P(c \leq u | x_1, \ldots, x_T) \neq P(c \leq u)$, then in general $P(c \leq u | x_1, \ldots, x_T) \neq P(c \leq u | x_1, \ldots, x_t)$; a latent variable that is constant over time is generally related to all of the x_t's if it is related to any of them. In that case

$$P(y_t = 1 | x_1, \ldots, x_T) = \int P(y_t = 1 | x_1, \ldots, x_t, u) \, dP(c \leq u | x_1, \ldots, x_T)$$

$$\neq P(y_t = 1 | x_1, \ldots, x_t).$$

Hence the failure of strict exogeneity indicates that the latent variable is not independent of the measured x's. Is there an extension of Granger's definition of "y does not cause x" that will imply $P(y_t = 1 | x_1, \ldots, x_T) = P(y_t = 1 | x_1, \ldots, x_t)$?

In the Granger definition, instead of requiring that $y_t, y_{t-1}, \ldots$ not contribute to the linear predictor of x_{t+1} given $x_t, x_{t-1}, \ldots$, we shall require that x_{t+1} be conditionally independent of $y_t, y_{t-1}, \ldots$.

DEFINITION 1: (G) -- x_{t+1} is independent of $y_t, y_{t-1}, \ldots$ conditional on $x_t, x_{t-1}, \ldots$ for all t.

In the Sims definition, instead of requiring that $x_{t+1}, x_{t+2}, \ldots$ not contribute to the linear predictor of y_t given $x_t, x_{t-1}, \ldots$, we shall require that y_t be conditionally independent of $x_{t+1}, x_{t+2}, \ldots$.

DEFINITION 2: (S) -- y_t is independent of $x_{t+1}, x_{t+2}, \ldots$ conditional on $x_t, x_{t-1}, \ldots$ for all t.

We shall show that (G) implies (S). So in our example, we would need to check whether x_{t+1} is independent of $y_t, \ldots, y_1$ conditional on $x_t, \ldots, x_1$. The inference problem is simplest when x is also a binary variable; x_t could indicate whether or not the individual was in a training program in period t; or in a sample of married women, x_t could indicate whether or not there was a birth in period t. Then the joint distribution of $(x_1, y_1, \ldots, x_T, y_T)$ is given by a set of multinomial probabilities, with each individual falling in one of 2^{2T} cells. The hypotheses (G) and (S) specify that the cell probabilities are specified functions of fewer than $2^{2T} - 1$ parameters. Given a random sample of size N from such a distribution, the asymptotic inference problem as $N \to \infty$ for fixed T is straightforward.[4]

If T = 2, then (S) requires that

$$P(y_1 = 1|x_1 = 0, x_2 = 0) = P(y_1 = 1|x_1 = 0, x_2 = 1),$$

$$P(y_1 = 1|x_1 = 1, x_2 = 0) = P(y_1 = 1|x_1 = 1, x_2 = 1).$$

It is simple to check that (G) imposes precisely the same restrictions; but when $T > 2$, (G) imposes more restrictions than (S). We shall present a counterexample to show that, in contrast to the linear predictor case, (S) need not imply (G).[5] The counterexample works for the following reason: if a random variable is uncorrelated with each of two other random variables, then it is uncorrelated with every linear combination of them; but if it is independent of each of the other random variables, it need not be independent of every function of them.

There is a modification of the Sims definition which, given a regularity condition, is equivalent to (G).

DEFINITION 3: (S') -- y_t is independent of $x_{t+1}, x_{t+2}, \ldots$ conditional on $x_t, x_{t-1}, \ldots, y_{t-1}, y_{t-2}, \ldots$.

In order to state the regularity condition, let F be the set of random variables of the form $z = 1$ if $(x_s, y_s, \ldots, x_{s-j}, y_{s-j}) \in B$, $z = 0$ otherwise, where s and j are arbitrary integers and B is a Borel set.

CONDITION (R): $$\lim_{k\to\infty} E(z|x_t, x_{t-1}, \ldots, y_{-k}, y_{-k-1}, \ldots)$$

$$= E(z|x_t, x_{t-1}, \ldots)$$

for all $z \in F$ and all t.

Then (G) is equivalent to (S') if (R) holds. (S') is a tractable modification of (S); they are equivalent in the linear predictor case if a condition corresponding to (R) holds.[6] Condition (R) requires that the current effect of y's from the distant past vanishes; similar assumptions are routine in the analysis of aggregate time-series data. In the longitudinal example, (R) holds automatically since y_t is degenerate prior to the "birth" of the individual. The important point here is that we are not making any stationarity assumptions, and so we are free to assign x_t and y_t degenerate distributions prior to some starting point for the process.

In our example, (G) and (S') imply precisely the same restrictions on the multinomial cell probabilities. It makes no difference which version we choose to test. Now suppose that $t = 1$ is not the starting point for the process, so that we are missing some observations. Then it is possible that we shall reject (G) or (S') simply because we have not included enough lags. Furthermore, tests of (G) and (S') are no longer equivalent, since the bias from truncating the lag distribution may be different in the two cases.

Suppose that

$$P(y_t = 1 | \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, c) = P(y_t = 1 | x_t, \ldots, x_{t-M}, c) \quad (t > M);$$

for example, it may be that children do not affect the woman's employment status once they are in school, so that M corresponds to approximately six years. If c is independent of the x's, then (S) holds and

$$P(y_t = 1 | x_1, \ldots, x_T) = P(y_t = 1 | x_t, \ldots, x_{t-M})$$

implies testable restrictions if $T > M + 1$. Rejection of (S) implies rejection of (G), whereas we may be unable to construct a valid direct test of (G) (or (S')) due to bias from truncating the lag distributions. If the process started at $t = -J \leq 0$, then

$$P(x_{t+1} = 1|x_t, \ldots, x_1, y_t, \ldots, y_1) = P(x_{t+1} = 1|x_t, \ldots, x_1)$$

does not, in general, hold for any t, even if (G) holds and y depends on only M lagged values of x.

An additional problem is to choose correct functional forms for the conditional distributions under the null hypothesis. Suppose that x is binary and let $t = 1$ be the starting point for the process. Then the multinomial distribution provides a completely general specification for $P(x_{t+1} = 1|x_t, \ldots, x_1)$; but if x is continuous, then specifying the conditional distribution will require a restrictive functional form. It may be easier to justify functional form restrictions in either the Granger or the Sims version of the test. The functional form issue is important; if the regression function is not linear, then the linear predictor form of (G) (or (S)) will generally fail to hold even though (G) holds. The past y's (or future x's) will help to correct for the error in approximating the regression function.

If we impose no regularity conditions, then we require a stronger version of the Sims definition: y_t must be independent of $x_{t+1}, x_{t+2}, \ldots$ conditional on $x_t, x_{t-1}, \ldots$ and conditional on any subset of $y_{t-1}, y_{t-2}, \ldots$.

DEFINITION 4: (S") -- y_t is independent of $x_{t+1}, x_{t+2}, \ldots$ conditional on $x_t, x_{t-1}, \ldots, Y_t$ for every Y_t and all t, where Y_t is a subset of $y_{t-1}, y_{t-2}, \ldots$.

We shall show that (G) is equivalent to (S"). In the definition of (S"), it is sufficient that the conditional independence holds for Y_t equal to the null set and all sets of the form $Y_t = [y_{t-1}, \ldots, y_{t-k}]$, $k = 1,2, \ldots$;

but we shall present counterexamples to show that, in the absence of (R), (S) + (S') does not imply (G).

Our proofs are simple applications of the following fundamental property of conditional expectation:

$$E(y|x) = E[E(y|x, z)|x].$$

All of our results and proofs continue to hold as stated when x_t and y_t are vectors of finite dimension.

The conditional independence property can be weakened by considering regression functions:

DEFINITION 5: (G_R) -- $E(x_{t+1}|x_t, x_{t-1}, \ldots, y_t, y_{t-1}, \ldots)$

$$= E(x_{t+1}|x_t, x_{t-1}, \ldots)$$

for all t.

DEFINITION 6: (S_R) -- $E(y_t|\ldots, x_{t-1}, x_t, x_{t+1}, \ldots)$

$$= E(y_t|x_t, x_{t-1}, \ldots)$$

for all t.

The Granger version states that x_{t+1} is "mean independent" of current and past y conditional on current and past x. The Sims version states that y_t is mean independent of future x conditional on current and past x. We shall use counterexamples to show that there are no equivalencies in the regression case. The counterexamples work because mean independence is not a symmetric relationship. If x is uncorrelated with y, then y is uncorrelated with x; if x is independent of y, then y is independent of x; but if $E(x|y) = E(x)$, it need not be true that $E(y|x) = E(y)$.

2. THE MAIN RESULTS

THEOREM 1: (G) implies (S).

PROOF: This is a special case of Theorem 4, which is proved below.

THEOREM 2: (S) does not imply (G).

PROOF: Consider the following counterexample: let y_1, y_2 be independent Bernoulli random variables with $P(y_t = 1) = P(y_t = -1) = 1/2$, $t = 1,2$. Let $x_3 = y_1 y_2$. Then y_1 is independent of x_3 and y_2 is independent of x_3. Let all of the other random variables be degenerate (equal to zero, say). Then (S) holds but x_3 is clearly not independent of y_1, y_2 conditional on $x_2, x_1, \ldots$.

Q.E.D.

For a stationary counterexample, we can let y_t be i.i.d. for $t = \ldots, -1,0,1, \ldots$ with $P(y_t=1) = P(y_t=-1) = 1/2$. Then set $x_t = y_{t-1}y_{t-2}$. One can check that y_t is independent of $\ldots, x_{t-1}, x_t, x_{t+1}, \ldots$, so that (S) holds. One can also check that the x_t are independent of each other; hence (G) requires that $P(x_{t+1} = 1 | x_t, x_{t-1}, \ldots, y_t, y_{t-1}) = 1/2$, which clearly does not hold. For a stationary, nondeterministic counterexample, we can set $x_t = y_{t-1}y_{t-2} + u_t$, where the u_t are independent of each other and of all of the y_t.

The proof of Theorem 3 will require two auxiliary definitions and a lemma. Let k be some positive integer.

DEFINITION 7: (G_k') - - x_{t+1} is independent of $y_t, y_{t-1}, \ldots, y_{t-k+1}$ conditional on $x_t, x_{t-1}, \ldots, y_{t-k}, y_{t-k-1}, \ldots$ for all t.[7]

DEFINITION 8: (S_k') -- y_t is independent of $x_{t+1}, x_{t+2}, \ldots, x_{t+k}$ conditional on $x_t, x_{t-1}, \ldots, y_{t-1}, y_{t-2}, \ldots$ for all t.

LEMMA: (G_k') is equivalent to (S_k').[8]

PROOF: All of the equalities in our proofs hold with probability one. The proofs are based on induction and

$$E(z|\mathcal{G}_1) = E[E(z|\mathcal{G}_2)|\mathcal{G}_1],$$

where z is an integrable random variable and $\mathcal{G}_1$, $\mathcal{G}_2$ are σ-fields (information sets) with $\mathcal{G}_1 \subset \mathcal{G}_2$.

1. We shall show first that (G_k') implies (S_k'). Let B and B_s be Borel sets and let D = 1 if $y_t \in B$, D = 0 otherwise; $D_s = 1$ if $x_{t+s} \in B_s$, $D_s = 0$ otherwise. Let $\mathcal{N}_t = \sigma(x_t, x_{t-1}, \ldots, y_{t-1}, y_{t-2}, \ldots)$ be the σ-field generated by these variables. Clearly (G_1') implies (S_1'); also (G_k') implies (G_j') if $j \leq k$. (This is based on the following result: let $\mathcal{G}_j$, j=1, ..., 4 be σ-fields; then $\mathcal{G}_1$ is independent of $\mathcal{G}_2$ conditional on $\sigma(\mathcal{G}_3 \cup \mathcal{G}_4)$ if and only if

$$P(A_1|\sigma(\mathcal{G}_2 \cup \mathcal{G}_3 \cup \mathcal{G}_4)) = P(A_1|\sigma(\mathcal{G}_3 \cup \mathcal{G}_4)) \tag{2.1}$$

for all $A_1 \in \mathcal{G}_1$ (Chow and Teicher [2], Theorem 1.i, p. 217); (2.1) holds if G_1 is independent of $\sigma(G_2 \cup G_3)$ conditional on G_4, for then both sides of the equality in (2.1) are equal to $P(A_1|G_4)$.) So (G_k') implies (S_1').

Now we shall assume that (G_k') implies (S_j') for some $j \epsilon [1, \ldots, k-1]$, and we shall show that (G_k') and (S_j') imply (S_{j+1}'). Let $N_{t,j} = \sigma(N_t \cup \sigma(x_{t+1}, \ldots, x_{t+j}))$.

$$(2.2) \qquad E(DD_{j+1}|N_{t,j}) = E(D|N_t)\ E(D_{j+1}|N_{t,j})$$

(by (G_{j+1}') and (S_j'));

$$E(DD_1 \ldots D_{j+1}|N_t) = E[(D_1 \ldots D_j)E(DD_{j+1}|N_{t,j})|N_t]$$

$$= E(D|N_t)\ E(D_1 \ldots D_{j+1}|N_t)$$

(by (2.2) and the fact that $E(D|N_t)$ is measurable N_t). Hence (G_k') and (S_j') imply (S_{j+1}'), and so our result follows by induction.

2. We shall complete the proof by showing that (S_k') implies (G_k'). Let $D = 1$ if $x_{t+1} \epsilon B$, $D = 0$ otherwise; $D_s = 1$ if $y_{t-s} \epsilon B_s$, $D_s = 0$ otherwise. Let $M_{t,k} = \sigma(x_t, x_{t-1}, \ldots, y_{t-k}, y_{t-k-1}, \ldots)$. (S_k') implies that x_{t+1} is independent of y_{t-k+1} conditional on $M_{t,k}$. We shall base our induction on the assumption that x_{t+1} is independent of $y_{t-j}, \ldots, y_{t-k+1}$ conditional on $M_{t,k}$ for some $j \epsilon [1, \ldots, k-1]$.

$$(2.3) \qquad E(DD_{j-1}|M_{t,j}) = E[DE[D_{j-1}|\sigma(M_{t,j} \cup \sigma(x_{t+1}))]|\ M_{t,j}]$$

$$= E(D_{j-1}|M_{t,j})\ E(D|M_{t,k})$$

(by (S_j') and the induction assumption);

$$E(DD_{j-1} \cdots D_{k-1}|\mathcal{M}_{t,k}) = E[(D_j \cdots D_{k-1})E(DD_{j-1}|\mathcal{M}_{t,j})|\mathcal{M}_{t,k}]$$

$$= E(D|\mathcal{M}_{t,k})\ E(D_{j-1} \cdots D_{k-1}|\mathcal{M}_{t,k})$$

(by (2.3)). So x_{t+1} is independent of $y_{t-j+1}, y_{t-j}, \ldots, y_{t-k+1}$ conditional on $\mathcal{M}_{t,k}$, and our result follows by induction.

Q.E.D.

THEOREM 3: (G) is equivalent to (S') if (R) holds.

PROOF: (S') is equivalent to (S_k') holding for all $k \epsilon [1,2, \ldots]$, since the sets $[x_{t+1} \epsilon B_1, \ldots, x_{t+k} \epsilon B_k]$ form a π-system generating $\sigma(x_{t+1}, x_{t+2}, \ldots)$ (Chow and Teicher [2], Theorem 1.iii, p. 217). By the Lemma, (S_k') holding for all k is equivalent to (G_k') holding for all k. (G) implies that (G_k') holds for all k. So we only need to show that (G_k') holding for all k implies (G).

Let D = 1 if $x_{t+1} \epsilon B$, D = 0 otherwise; $D_s = 1$ if $y_{t-s} \epsilon B_s$, $D_s = 0$ otherwise. Let $\mathcal{J}_t = \sigma(x_t, x_{t-1}, \ldots)$. We need to show that for any $j \epsilon [0,1, \ldots]$,

(2.4) $$E(DD_0 \cdots D_j|\mathcal{J}_t) = E(D|\mathcal{J}_t)\ E(D_0 \cdots D_j|\mathcal{J}_t)$$

for all t. If (G_k') holds for all k, then

(2.5) $$E(DD_0 \cdots D_j|\mathcal{M}_{t,k}) = E(D|\mathcal{M}_{t,k})\ E(D_0 \cdots D_j|\mathcal{M}_{t,k})$$

for all $k > j$. Taking the limit as $k \to \infty$ in (2.5) and applying (R) gives (2.4).

Q.E.D.

Recall that $M_{t,k} = \sigma(x_t, x_{t-1}, \ldots, y_{t-k}, y_{t-k-1}, \ldots)$ and set $M_{t,\infty} = \bigcap_{k=0}^{\infty} M_{t,k}$.

DEFINITION 9: (G') -- x_{t+1} is independent of $y_t, y_{t-1}, \ldots$ conditional on $M_{t,\infty}$.

COROLLARY 1: (G') is equivalent to (S').

PROOF: Repeat the proof of Theorem 3 up to (2.5) with $M_{t,\infty}$ replacing J_t. Since $M_{t,0} \supset M_{t,1} \supset \ldots$, we have

$$\lim_{k\to\infty} E(z|M_{t,k}) = E(z|M_{t,\infty})$$

for any integrable random variable z. (Reversed martingale; see Doob [3], Theorem 4.3, p. 331.) Taking the limit as $k \to \infty$ in (2.5) gives the result.

Q.E.D.

We shall say that the $[x_t, y_t]$ process is mixing from the left if

$$\lim_{k\to\infty} E(z|x_{-k}, y_{-k}, x_{-k-1}, y_{-k-1}, \ldots) = E(z)$$

for $z \in F$.[9] So (R) requires that the process be mixing from the left conditional on $x_t, x_{t-1}, \ldots$ for all t.

Let $T = \bigcap_{k=0}^{\infty} \sigma(x_{-k}, y_{-k}, x_{-k-1}, y_{-k-1}, \ldots)$ be the left tail σ-field of the [x,y] process. T is degenerate if it contains only sets with probability measure zero or one; an equivalent characterization is that a random variable z is measurable with respect to T only if z equals a constant with probability one. A standard regularity condition requires that T be degenerate -- see

Rozanov [9], p. 178. Since

$$\lim_{k\to\infty} E(z|x_{-k}, y_{-k}, x_{-k-1}, y_{-k-1}, \ldots) = E(z|\mathcal{T}),$$

the process is mixing from the left when $\mathcal{T}$ is degenerate. It may seem plausible that (R) also holds when $\mathcal{T}$ is degenerate, but this is false. For a counterexample, let x_1 be independent of x_0 with $P(x_1=1) = P(x_1=-1) = 1/2$, and let x_0 have a uniform distribution on $(0,1]$; let $x_t = 0$ for $t \neq 0$ or 1. We can express x_0 in terms of its nonterminating dyadic expansion: $x_0 = \Sigma_{k=1}^{\infty} d_k(x_0)/2^k$, where each $d_k(x_0)$ is 0 or 1. Let $y_{-k} = x_1[2d_k(x_0)-1]$ for $k \geq 1$ and $y_{-k} = 0$ for $k < 1$. Then $y_{-1}, y_{-2}, \ldots$ is a sequence of i.i.d. random variables, and Kolmogorov's zero-one law implies that $\mathcal{T}$ is degenerate. But $x_1 = y_{-k}/[2d_k(x_0) - 1]$ for any $k \geq 1$, and so

$$\lim_{k\to\infty} E(x_1|x_0, y_{-k}) = x_1.$$

Since $E(x_1|x_0) = 0$, we see that (R) does not hold.

2.1 *An Equivalence Without Regularity Conditions*

The following equivalence does not require (R):

THEOREM 4: (G) is equivalent to (S").

PROOF: The proof follows that of the Lemma quite closely.

1. We shall show first that (G) implies (S"). Let $\mathcal{J}'_t = \sigma(x_t, x_{t-1}, \ldots, y_{t_1}, y_{t_2}, \ldots)$, where $[t_1, t_2, \ldots]$ is some, possibly finite, subset of $[t-1, t-2, \ldots]$. Let (S''_k) be the following property:

$$(S''_k):\quad P(y_t \in B, x_{t+1} \in B_1, \ldots, x_{t+k} \in B_k | \mathcal{J}'_t)$$

$$= P(y_t \in B | \mathcal{J}'_t) P(x_{t+1} \in B_1, \ldots, x_{t+k} \in B_k | \mathcal{J}'_t)$$

for all $\mathcal{J}'_t$ and all t. (S") is equivalent to (S''_k) holding for $k = 1,2, \ldots.$ Let $D = 1$ if $y_t \in B$, $D = 0$ otherwise; $D_s = 1$ if $x_{t+s} \in B_s$, $D_s = 0$ otherwise. (G) implies (S''_1). Assume that (S''_k) holds for some $k \in [1,2, \ldots]$. Let $\mathcal{J}'_{t,k} = \sigma(\mathcal{J}'_t \cup \sigma(x_{t+1}, \ldots, x_{t+k}))$.

$$(2.6)\qquad E(DD_{k+1}|\mathcal{J}'_{t,k}) = E[DE[D_{k+1}|\sigma(\mathcal{J}'_{t,k} \cup \sigma(y_t))]|\ \mathcal{J}'_{t,k}]$$

$$= E(D_{k+1}|\mathcal{J}'_{t,k})\ E(D|\mathcal{J}'_t)$$

(by (G) and (S''_k));

$$E(DD_1 \ldots D_{k+1}|\mathcal{J}'_t) = E[(D_1 \ldots D_k)\ E(DD_{k+1}|\mathcal{J}'_{t,k})|\mathcal{J}'_t]$$

$$= E(D|\mathcal{J}'_t)\ E(D_1 \ldots D_{k+1}|\mathcal{J}'_t)$$

(by (2.6)). Hence (G) and (S''_k) imply (S''_{k+1}), and our result follows by induction.

2. We shall complete the proof by showing that (S") implies (G). Let $D = 1$ if $x_{t+1} \in B$, $D = 0$ otherwise; $D_s = 1$ if $y_{t-s} \in B_s$, $D_s = 0$ otherwise. Let $\mathcal{J}_t = \sigma(x_t, x_{t-1}, \ldots)$. For any $k \in [0,1, \ldots]$, (S") implies that x_{t+1} is independent of y_{t-k} conditional on $\mathcal{J}_t$; in fact, only (S) is needed for this result. We shall base our induction on the assumption that x_{t+1} is independent of $y_{t-j}, \ldots, y_{t-k}$ conditional on $\mathcal{J}_t$ for some $j \in [1, \ldots, k]$. Let $\mathcal{J}^*_t = \sigma(x_t, x_{t-1}, \ldots, y_{t-j}, \ldots, y_{t-k})$.

$$(2.7) \qquad E(DD_{j-1}|\mathcal{J}^*_t) = E[DE[D_{j-1}|\sigma(\mathcal{J}^*_t \cup \sigma(x_{t+1}))]|\mathcal{J}^*_t] = E(D_{j-1}|\mathcal{J}^*_t)\; E(D|\mathcal{J}_t)$$

(by (S") and the induction assumption);

$$E(DD_{j-1} \cdots D_k|\mathcal{J}_t) = E[(D_j \cdots D_k)\, E(DD_{j-1}|\mathcal{J}^*_t)|\mathcal{J}_t]$$

$$= E(D|\mathcal{J}_t)\; E(D_{j-1} \cdots D_k|\mathcal{J}_t)$$

(by (2.7)). So x_{t+1} is independent of $y_{t-j+1}, \ldots, y_{t-k}$ conditional on $\mathcal{J}_t$, and induction shows that x_{t+1} is independent of $y_t, \ldots, y_{t-k}$ conditional on $\mathcal{J}_t$. Then our result follows since this holds for any $k \in [0,1, \ldots\}$.

Q.E.D.

It is clear from our proof that an apparently weaker version of (S") will imply (G): it is sufficient to assume that y_t is independent of x_{t+1}, x_{t+2}, ... conditional on x_t, x_{t-1}, ..., Y^*_t for every Y^*_t and all t, where $Y^*_t = [y_{t-1}, \ldots, y_{t-k}]$ for some $k \in [1,2, \ldots]$. However, in the absence of (R), we cannot further weaken (S"). Consider the following counterexample:

$y_t = y_0$ for all t,

$x_1 = y_0$, $x_t = 0$ for $t \neq 1$.

Then (S') holds since the distribution of y_t conditional on any non-null subset of previous y's is degenerate; but (G) clearly does not hold. Note that (S) does not hold.

There remains the possibility that (S) + (S') implies (G); but consider the following counterexample:

$y_0, \ldots, y_{2k-1}$ are independent for some $k \geq 3$ with

$$P(y_s = 1) = P(y_s = -1) = 1/2 \quad (s = 0, \ldots, 2k - 1);$$

$$y_{2kt+s} = y_s \text{ for } s = 0, \ldots, 2k-1 \text{ and for all } t;$$

$$x_{2k} = y_k y_{2k-1}, \quad x_t = 0 \text{ for } t \neq 2k.$$

Then (S') holds since the distribution of y_t conditional on y_{t-2k} is degenerate. Also (S) holds since y_k is independent of x_{2k} and y_{2k-1} is independent of x_{2k}. Clearly (G) does not hold. Furthermore, y_t is independent of $x_{t+1}, x_{t+2}, \ldots$ conditional on $x_t, x_{t-1}, \ldots, y_{t-1}, \ldots, y_{t-k+2}$ for all t. So it appears that no further simplification of (S") is possible.

2.2 *The Linear Predictor Case*

We shall conclude this section by relating our results to the linear predictor case.

COROLLARY 2: If the joint distribution of $[x_s, y_s, \ldots, x_{s-k}, y_{s-k}]$ is multivariate normal for all integers s and k, then (G), (S), and (S") are equivalent.

PROOF: In the proof that (S") implies (G), only (S) is needed to show that x_{t+1} is independent of y_{t-k} conditional on J_t for $k \in [0,1,\ldots]$. By joint normality, the pair-wise independence implies that x_{t+1} is independent of $\sigma(y_t, y_{t-1}, \ldots)$ conditional on J_t. Hence (S) implies (G). By Theorem 4, (G) implies (S"). Clearly (S") implies (S).

Q.E.D.

Assume that x_t and y_t have finite variance for all t. Consider the Hilbert space of random variables generated by the linear manifold spanned by the variables $[(x_t, y_t), t = \ldots, -1,0,1, \ldots]$, closed with respect to convergence in mean square. We include also a constant (1) in the space. The inner product is $\langle z_1,z_2\rangle = E(z_1z_2)$. Then the linear predictor of a random variable based on a set of random variables is the projection of the random variable on the closed linear subspace generated by the set of random variables and 1.

We can always construct a multivariate normal process $[x_t^*, y_t^*]$ with the same means and covariances as the $[x_t, y_t]$ process. If the linear predictor of y_t based on all x is identical to the linear predictor based on $x_t, x_{t-1}, \ldots$, then y_t^* is independent of $x_{t+1}^*, x_{t+2}^*, \ldots$ conditional on $x_t^*, x_{t-1}^*, \ldots$. Then Corollary 2 implies that x_{t+1}^* is independent of $y_t^*, y_{t-1}^*, \ldots$ conditional on $x_t^*, x_{t-1}^*, \ldots$. Hence the linear predictor of x_{t+1} based on $x_t, x_{t-1}, \ldots, y_t, y_{t-1}, \ldots$ is identical to the linear predictor based on $x_t, x_{t-1}, \ldots$. In a similar fashion we can establish the converse: if the linear predictor of x_{t+1} based on current and lagged x and y is identical to the linear predictor based on current and lagged x, then Corollary 2 implies that the linear predictor of y_t based on all x is identical to the linear predictor based on current and lagged x.

So Corollary 2 implies the original Sims result on the equivalence of the linear predictor versions of (G) and (S). We have the additional result that the linear predictor version of (G) or (S) implies the linear predictor version of (S").

Now consider the linear predictor version of (S'). Let $H(1, x_s, x_{s-1}, \ldots, y_j, y_{j-1}, \ldots)$ denote the closed linear subspace spanned by those variables; let $E^*(z|1, x_s, x_{s-1}, \ldots, y_j, y_{j-1}, \ldots)$ denote the projection of z on the closed linear subspace. Let F_L be the set of random variables that are contained in $H(1, x_s, y_s, \ldots, x_{s-j}, y_{s-j})$ for some integers s and j.

CONDITION (R_L): $\lim_{k\to\infty} E^*(z|x_t, x_{t-1}, \ldots, y_{-k}, y_{-k-1}, \ldots) = E^*(z|x_t, x_{t-1}, \ldots)$

for all $z \in F_L$ and all t.

By considering a multivariate normal process with the same means and covariances as the $[x_t, y_t]$ process, we can show that the linear predictor version of (G_k') is equivalent to the linear predictor version of (S_k'). Then an argument similar to that used in the proof of Theorem 3 shows that the linear predictor version of (G) is equivalent to the linear predictor version of (S') if (R_L) holds.

Suppose that we transform the random variables so that $E(x_t) = E(y_t) = 0$ for all t. Let $T_L = \bigcap_{k=0}^{\infty} H(x_{-k}, y_{-k}, x_{-k-1}, y_{-k-1}, \ldots)$. In the terminology of Rozanov [9], the $[x_t, y_t]$ process is linearly regular if $T_L = 0$.[10] This condition is similar to (R_L), but in fact does not imply (R_L). For a counterexample, let x_t be i.i.d. with mean zero and variance one for $t \leq 0$; let $x_t = 0$ for $t > 0$. Let $y_{-t} = \Sigma_{j=0}^{t-1} x_{-j}$ for $t \geq 1$; $y_{-t} = 0$ for $t < 1$. Then the projection of x_0 on $H(x_{-1}, x_{-2}, \ldots, y_{-k}, y_{-k-1}, \ldots)$ equals x_0 for any $k > 1$, since $x_0 = y_{-k} - (x_{-1} + \ldots + x_{-k+1})$. The projection of x_0 on $H(x_{-1}, x_{-2}, \ldots)$ is 0; hence (R_L) does not hold. Since $y_{-k-1} - y_{-k} = x_{-k}$, we have $H(x_{-k}, y_{-k}, x_{-k-1}, y_{-k-1}, \ldots) = H(y_{-k}, x_{-k}, x_{-k-1}, \ldots)$. If we use $[x_0, x_1, \ldots]$ as an orthonormal basis, then the coordinates of a point in $H(y_{-k}, x_{-k}, x_{-k-1}, \ldots)$ form a square-summable sequence whose first k elements are equal. Hence $T_L = \bigcap_{k=0}^{\infty} H(y_{-k}, x_{-k}, x_{-k-1}, \ldots) = 0$, and the counterexample is complete.

3. REGRESSION

Our causality definitions have used conditional independence instead of zero partial correlation. The conditional independence property can be weakened by using the regression function versions (G_R) and (S_R). The following example will show that neither of these definitions implies the other.

Let u_t be i.i.d. for all t with $P(u_t > 0) = 1$. Let v_t be i.i.d., independent of the u's, with $P(v_t = 1) = P(v_t = -1) = 1/2$. Set $x_t = u_{t-1}v_{t-1}$, $y_t = u_t$. Then with probability one

$$E(x_{t+1} | x_t, x_{t-1}, \ldots, y_t, y_{t-1}, \ldots)$$

$$= E(u_t v_t | u_t) = u_t E(v_t) = 0.$$

Hence (G_R) holds; but

$$E(y_t | \ldots, x_{t-1}, x_t, x_{t+1}, \ldots)$$
$$= E(u_t | u_t v_t) = |u_t v_t| = y_t,$$

$$E(y_t | x_t, x_{t-1}, \ldots) = E(y_t).$$

So (G_R) does not imply (S_R).

Now let $x_t = u_{t-1}$, $y_t = u_t v_t$. With probability one

$$E(y_t | \ldots, x_{t-1}, x_t, x_{t+1}, \ldots)$$

$$= E(u_t v_t | u_t) = u_t \, E(v_t) = 0.$$

Hence (S_R) holds; but

$$E(x_{t+1}|x_t, x_{t-1}, \ldots, y_t, y_{t-1}, \ldots)$$

$$= E(u_t|u_t v_t) = |u_t v_t| = x_{t+1},$$

$$E(x_{t+1}|x_t, x_{t-1}, \ldots) = E(x_{t+1}).$$

So (S_R) does not imply (G_R). Furthermore, we have

$$E(y_t|\ldots, x_{t-1}, x_t, x_{t+1}, \ldots, y_{t_1}, y_{t_2}, \ldots) = 0,$$

where $[t_1, t_2, \ldots]$ is any subset of $[t-1, t-2, \ldots]$. So (S''_R) does not imply (G_R). It appears that there are no interesting equivalencies in the regression case.

4. CONCLUSION

If y_t is a binary variable, then the regression function for y_t conditional on $x_t, x_{t-1}, \ldots$ is generally not identical to the linear predictor of y_t based on $x_t, x_{t-1}, \ldots$. So even if y is independent of future x conditional on current and past x, it will generally not be true that x is strictly exogenous in the linear predictor sense; the future x's will help to correct for the error in approximating the nonlinear regression function. Suppose, for example, that t=1 is the starting point for the process and that

$$E(y_2|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 .$$

The linear predictor version of (G) or (S) requires zero partial correlation between x_1x_2 and x_3 given x_1 and x_2. This will be true if the joint distribution of (x_1,x_2,x_3) is multivariate normal or, more generally, if the regression function for x_3 conditional on x_1, x_2 is linear. I would not expect it to be true if x_t is a binary variable. For then the regression function is

$$E(x_3|x_1,x_2) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_1 x_2 ,$$

and γ_3 is generally not zero.

So we have considered extensions of the Granger and Sims definitions that use conditional independence instead of linear predictors. The extended Granger definition of "y does not cause x" is stronger than the condition that y be independent of future x conditional on current and past x; so noncausality is stronger than strict exogeneity. Under a weak regularity condition, however, if y is independent of future x conditional on current and past x <u>and</u> past y, then y does not cause x.

FOOTNOTES

[1] I am grateful to John Geweke, Kenneth Judd, James Kuelbs, Linda Rothschild, Charles Wilson, Arnold Zellner, and especially Christopher Sims for helpful discussions. Financial support was provided by the National Science Foundation (Grants No. SOC-7925959 and No. SES-8016383).

[2] Granger's initial definition was in terms of the regression function. He then specialized it to the linear predictor form. Granger also allowed for conditioning on information in addition to the current and past values of x and y. It is straightforward to modify our results to incorporate such additional conditioning.

[3] His proof assumes that the process is covariance stationary with no linearly deterministic component; also the process has an autoregressive representation. Hosoya [7] showed that these conditions are not necessary.

[4] See Rao [8], section 6b.

[5] A Co-Editor has informed me of a paper by Florens and Mouchart [4] which contains results similar to (G) implies (S) and (S) does not imply (G). Their paper relates the Granger definition to the concept of transitivity in sequential analysis.

[6] The linear predictor version of (S') has been used by Geweke, Meese, and Dent [5]. Under assumptions similar to those used by Sims [10], they show that the linear predictor version of (S') is equivalent to the linear predictor version of (G).

[7] This definition is related to Granger's [6] definition of causality lag.

[8] I am indebted to Christopher Sims for the observation that (G_k') and (S_k') are equivalent in the linear predictor case.

[9] Mixing is defined for stationary processes in Billingsley [1], p. 12.

[10] A stationary process is linearly regular if it does not have a deterministic component; see Rozanov [9], Chapter II, section 2. Sims [10] assumes that the process is linearly regular.

REFERENCES

[1] Billingsley, P.: Ergodic Theory and Information, Wiley, New York, 1965.

[2] Chow, Y.S., and H. Teicher: Probability Theory: Independence, Interchangeability, Martingales, Springer-Verlag, New York, 1978.

[3] Doob, J.L.: Stochastic Processes, Wiley, New York, 1953.

[4] Florens, J.P., and M. Mouchart: "A Note on Non-Causality," University of Aix-Marseille, unpublished manuscript, 1980.

[5] Geweke, J., R. Meese, and W.T. Dent: "Comparing Alternative Tests of Causality in Temporal Systems: Analytic Results and Experimental Evidence," University of Wisconsin-Madison, Social Systems Research Institute Workshop Series No. 7928, 1979.

[6] Granger, C.W.J.: "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," Econometrica, 37 (1969), 424-438.

[7] Hosoya, Y.: "On the Granger Condition for Non-Causality," Econometrica, 45 (1977), 1735-1736.

[8] Rao, C.R.: Linear Statistical Inference and Its Applications, Wiley, New York, 1973.

[9] Rozanov, Y.A.: Stationary Random Processes, Holden-Day, San Francisco, 1967.

[10] Sims, C.A.: "Money, Income, and Causality," The American Economic Review, 62 (1972), 540-552.