Tests for Predictive Relationships Between Time Series Variables: A Monte Carlo Investigation

CHARLES R. NELSON and G. WILLIAM SCHWERT*

Bivariate time series models have been used extensively to analyze the relationship between pairs of economic variables. Various tests have been proposed that can be used to examine the adequacy of specific models. The empirical literature is noteworthy for the frequency with which different authors using different tests reach different conclusions, and for the apparent lack of evidence for certain relationships strongly suggested by economic theory. The objective of this study is to use Monte Carlo methods to examine the size and power of alternative tests, and to relate these findings to the analytical structure of the tests.

KEY WORDS: Monte Carlo; ARIMA models; Causality tests.

1. INTRODUCTION

Numerous empirical studies have appeared in recent years that purport to test for the existence and direction of "causal" relationships among monetary, macroeconomic, and financial variables. Notable in this literature is the frequency with which different investigators examining the same basic data report contradictory results. Feige and Pearce (1979) find no evidence of causal relationships between the supply of money and aggregate nominal income, in strong contradiction to the conclusions reached earlier by Sims (1972). Pierce (1977a) studies relationships among pairs of various monetary, financial, and macroeconomic variables and finds little evidence of causation, even in cases such as growth of demand deposits and the yield on Treasury bills, which theory suggests are importantly related. Similarly puzzling is that Feige and Pearce report no significant relationship between growth of the money supply and the inflation rate.

Since different investigators use different testing procedures, the explanations for these inconsistencies presumably lie in the relative size and power of the alternative test statistics. This question has been actively debated in recent years. Sargent (1976, p. 233), Pierce and Haugh (1977), Pierce (1977a,b), Sims (1977), Hsiao (1979, 1981), Wallis (1977), Schwert (1979), and Jacobs, Leamer, and Ward (1979) discuss the relative merits of different tests, but the differences among these tests have not been quantified. Since this article was first written, we have seen articles by Geweke, Meese, and Dent (1979), Guilkey and Salemi (1979), and Geweke (1981a,b), who are examining related issues.

The objective of this paper is to investigate the sampling distributions of alternative tests in the context of a bivariate time series model that includes independence, one-way causation, and feedback as special cases. We are aware that the concept of causation developed by Granger (1969) is a purely predictive one and may not in some circumstances coincide with the concept of causation discussed by philosophers of science. Zellner (1979) and Nelson (1979) discuss this distinction in detail. Nevertheless, the Granger concept of causality does correspond to restrictions on bivariate time series models that are often of interest to economists.

2. ANALYTICAL RELATIONSHIPS AMONG ALTERNATIVE TESTS

A time series $\{X_t\}$ is said to "cause" another time series $\{Y_t\}$ in the sense defined by Granger (1969) if past values of X are useful in predicting Y_t when the past values of Y have been taken into account. The definition is symmetric for Y causing X, and feedback is said to exist if causality is present in both directions.

The bivariate time series model

$$\begin{bmatrix} 1 & -\gamma_{12} \\ -\gamma_{21} & 1 \end{bmatrix} \begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} .5 & 0 \\ 0 & .5 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}, \quad (2.1)$$

is used to generate data in the simulation experiments, where u_{1t} and u_{2t} are serially uncorrelated pseudorandom

* Charles R. Nelson is Professor, Department of Economics, University of Washington, Seattle, WA 98195 and G. William Schwert is Associate Professor, Graduate School of Management, University of Rochester, NY 14627. The National Science Foundation provided support for this research. Comments from Ray Fair, Martin Geisel, Clive Granger, Randall Olson, Franz Palm, David Pierce, Charles Plosser, Stephen Ross, Donald Rubin, Christopher Sims, Arnold Zellner and the participants at the N.B.E.R. Conference on Time Series Analysis in Madison, Wisconsin, in May 1979, and at the Warwick Summer Econometric Workshop in July 1979 are gratefully acknowledged. Previous versions of this article had the title "Tests for Granger/Wiener Causality: A Monte Carlo Investigation."

© Journal of the American Statistical Association
March 1982, Volume 77, Number 377
Applications Section

normal deviates with $\text{cov}(u_{1t}, u_{2t}) = 0$.[1] The parameters γ_{12} and γ_{21} are varied across our experiments to create different bivariate relationships between Y and X. By choosing γ_{12} nonzero, we simulate a situation in which X helps to predict Y; a nonzero γ_{21} implies that Y helps to predict X; and by setting γ_{12} and γ_{21} both equal to zero, Y and X are independent. This is easy to see in terms of the reduced form for (2.1)

$$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} v_{1t} \\ v_{2t} \end{bmatrix}, \qquad (2.2)$$

which is a first-order vector autoregression. The reduced form coefficients ϕ_{12} and ϕ_{21} indicate whether lagged values of Y and X are useful for improving the predictions of X and Y, respectively.

The tests discussed in the remainder of this section use restrictions implied by the various bivariate structures for (a) the reduced form (2.2); (b) the univariate autoregressive-moving average (ARMA) representations of Y and X; (c) the cross-correlations between the residuals from the univariate ARMA representations for Y and X; (d) the two-sided distributed lag regressions between univariate ARMA residuals; and (e) the two-sided distributed lag regressions between Y and X. All of these tests have been used to determine relationships between pairs of economic time series variables.

2.1 Reduced-Form Tests

To test the null hypothesis of independence, we compare the values of the log likelihood function for the unconstrained reduced form (2.2) with that of the constrained reduced form implied by independence, where $\phi_{12} = \phi_{21} = \text{cov}(v_{1t}, v_{2t}) = 0$. The coefficients are estimated by least squares and the likelihood is evaluated using estimated variances and covariances of the reduced form residuals in the unconstrained case. Two times the difference in these log likelihoods should be approximately chi square with three degrees of freedom (df) if the null hypothesis of independence is true. Using preliminary simulation results, we multiply the likelihood ratio test by $(T - K)/T$, where T is the sample size and K is the number of parameters estimated in the unconstrained model.

To test the null hypothesis that X does not help predict Y, we regress Y_t on Y_{t-1} and X_{t-1} by least squares and examine the t ratio for $\hat{\phi}_{12}$ (which will not have a t distribution in small samples). Similarly, we examine the t ratio for $\hat{\phi}_{21}$ to test the null hypothesis that Y does not help predict X.

In applications, there has been a tendency to include a large number of parameters in bivariate autoregressive tests of "causality" to assure that there are not important omitted lags of either variable. To represent this tendency toward profligately parameterized models, we also conduct tests based on the assumption of a sixth-order vector autoregressive process. In this case, the likelihood ratio test statistic will be approximately chi square with 13 df.

To test the hypothesis that X does not help predict Y, we use the F test for the significance of the coefficients of lagged X when Y_t is the regressand; similarly, we use the F test for the coefficients of lagged Y when X_t is the regressand to test the hypothesis that Y does not help predict X. The resulting tests will be less powerful than those based on the exact reduced form.

2.2 Bivariate Structure and ARMA Representations

Restrictions on the orders and parameters of the ARMA representations of $\{Y_t\}$ and $\{X_t\}$ are implied by different parameter values for (2.1). It is straightforward to show that when there is feedback, both Y and X are ARMA (2, 1) processes with identical autoregressive coefficients (see, for example, Wallis 1977)

$$\begin{aligned} [1 - (\phi_{11} + \phi_{22})L - (\phi_{12}\phi_{21} - \phi_{11}\phi_{22})L^2]\, Y_t &= [1 - \theta_y L]a_t, \\ [1 - (\phi_{11} + \phi_{22})L - (\phi_{12}\phi_{21} - \phi_{11}\phi_{22})L^2]\, X_t &= [1 - \theta_x L]b_t, \end{aligned} \qquad (2.3)$$

where a_t and b_t are the univariate "innovations" for Y_t and X_t, respectively. If ϕ_{12} is zero, Y follows an AR(1) process; if ϕ_{21} is zero, X follows an AR(1) process. Therefore, if X helps predict Y, an ARMA(2,1) model should fit the data for Y better than does an AR(1) model, and if Y helps predict X, an ARMA(2,1) model is implied for X.

Zellner and Palm (1974) suggest that a comparison of the AR(1) model versus the ARMA(2,1) model can be used to determine whether there is a predictive relationship between Y and X. Wallis (1977) suggests that a test for feedback can be based on the bivariate ARMA model in (2.3); the likelihood for the restricted model can be compared with the likelihood for the unrestricted bivariate ARMA(2,1) model to test the hypothesis that there is feedback between Y and X. Unfortunately, in both of these tests the unrestricted ARMA(2,1) model does not have identified parameters under the null hypothesis that Y and X are unrelated, so the information matrix will be singular. Thus, there is reason to doubt that tests based on univariate or multivariate ARMA models can be used to test for predictive relationships.

2.3 Cross-Correlation of Univariate ARMA Residuals

It is straightforward to show that the univariate ARMA innovations in (2.3), a_t and b_t, will be cross-correlated

[1] The pseudorandom normal deviates are generated by Marsaglia's rectangular-wedge-tail method, incorporated in the program RANORM, obtained from the University of Chicago Computation Center. Kinderman and Ramage (1976) discuss some attractive properties of generators such as this. As an expedient device, it is assumed that $Y_0 = X_0 = 0$ in (2.1); then $T + 20$ observations for $\{Y_t\}$ and $\{X_t\}$ are computed so that the last T observations can be used in the tests. This procedure mitigates the arbitrary assumption about the initial conditions, Y_0 and X_0.

when there is a predictive relationship between Y and X. As Pierce and Haugh (1977) show, when cov $(a_t, b_{t-k}) \neq 0$ for some $k > 0$, then X helps predict Y, since a_t is the part of Y_t not predicted by past Y's. Similarly, when cov $(a_{t-k}, b_t) \neq 0$ for some $k > 0$ then Y helps predict X. In the case of nonzero γ_{12} and γ_{21} all of these covariances are nonzero and feedback is present. Finally, in the case where $\gamma_{12} = \gamma_{21} = 0$, all of these covariances are zero, which confirms the independence of Y and X.

In our experiments, we use values of γ_{12} equal to .5 and 1.0 when $\gamma_{21} = 0$, implying values of θ_y in (2.3) equal to .382 and .234, respectively. The correlation between a_t and b_{t-k} is $(.437) \cdot (.382)^k$ for $k \geq 0$ when γ_{12} is .5, and $(.685) \cdot (.234)^k$ when γ_{12} is 1.0. Note that the cross-correlations decay geometrically at a rate of θ_y in this case, so that evidence of the predictive relation is concentrated at the first lag.

If the innovations $\{a_t\}$ and $\{b_t\}$ could be observed or calculated from $\{Y_t\}$ and $\{X_t\}$, then tests of association could be based on sample cross-correlations between the innovations. In particular, sample cross-correlations between independent random series are asymptotically normal with mean zero and standard deviation $(T - |k|)^{-1/2}$ for lag k, and they are independent across lags (see Bartlett 1955 or Hannan 1970). Further, the same result holds for random series with some nonzero cross-correlations, as long as the true cross-correlations are zero in the range of lags being considered. Therefore, to test the null hypothesis that Y and X are independent, one would calculate the statistic

$$\sum_{k=-K_1}^{K_2} (T - |k|) r_{ab}(k)^2, \tag{2.4}$$

where $r_{ab}(k)$ denotes the sample correlation between a_t and b_{t+k}, which would be approximately distributed as $\chi^2(K_1 + K_2 + 1)$ under the null hypothesis. Similarly, to test the null hypothesis that X does not help predict Y one would calculate the statistic in (2.4) for one side of the cross-correlation function.

Operationally, the innovations are not available since the parameters of the univariate ARMA representations are unknown; instead, we use residuals from fitted ARMA models. Haugh (1972, 1976) has shown that the result in (2.4) continues to hold when residuals are used instead of innovations if Y and X are independent; therefore, the test for independence remains valid. Haugh suggests a formula that appears to be different from (2.4). The difference arises from the definition of the estimator of the cross-correlation coefficient at lag k. Thus, the statistic in (2.4) is equivalent to Haugh's test. However, Durbin's (1970) analysis of the distribution of test statistics based on residuals implies that when Y and X are not independent, the variance of the statistic in (2.4) is smaller than the variance of $\chi^2(K_1 + K_2 + 1)$. Thus, if X helps predict Y, the one-sided test as to whether Y helps predict X (i.e., $K_1 = -1$, $K_2 > 0$) will have an overstated significance level. Similarly, the power of such tests is reduced because of the Durbin problem. Further discussion of this fact is found in Pierce (1977a,b), Sims (1977), and Pierce and Haugh (1977), but as far as we know the magnitude of this phenomenon has not been examined previously.

In examining the results of our experiments, it will be interesting to see the empirical significance levels and power of the residual cross-correlation tests in situations where Y and X are not independent. We will also be interested in comparing tests based on true innovations as opposed to fitted residuals.

The test statistics we calculate use six leads or six lags in (2.4), since a generous number of lags are typically included when these tests are used in practical situations. The innovations are calculated recursively assuming that $a_0 = b_0 = 0$. The residuals are calculated two different ways: (a) the appropriate ARMA model (either ARMA(2,1) or AR(1)) is estimated for both Y and X; and (b) an AR(6) model is estimated for both Y and X, since high-order autoregressive models are often used to approximate general ARMA models.

2.4 Regressions Between Univariate ARMA Residuals

Since both $\{a_t\}$ and $\{b_t\}$ are nonautocorrelated, though in general they are cross-correlated, the regression relations between them are of the form

$$\begin{aligned} a_t &= \sum_{k=-\infty}^{\infty} \psi_{1k} b_{t-k} + e_{at}, \\ b_t &= \sum_{k=-\infty}^{\infty} \psi_{2k} a_{t-k} + e_{bt}, \end{aligned} \tag{2.5}$$

where $\psi_{1k} = \rho_{ab}(-k) \cdot \sigma_a/\sigma_b$, $\psi_{2k} = \rho_{ab}(k) \cdot \sigma_b/\sigma_a$, $\rho_{ab}(k)$ is the correlation between a_t and b_{t+k}, and where e_a and e_b will be autocorrelated in general. Thus, the analysis of the cross-correlation functions in the previous section implies analogous results for the dynamic regressions in (2.5).

Under the hypothesis of independence, e_a and e_b will be nonautocorrelated so that a test of independence can be based on the F statistic for significance of either of the regressions in (2.5). Under the null hypothesis of one-way predictive ability, say from X to Y, the F statistic for the one-sided regressions of a_t on future b's or b_t on past a's can be used. Therefore, tests analogous to the cross-correlation tests can be based on simple F statistics from the analogous regressions, and these tests are exact, not large-sample approximations.

When ARMA residuals are used in place of true innovations, the tests are no longer exact and the implications of Durbin's analysis are again relevant. Further discussion of these regression tests can be found in Granger (1973) and Schwert (1979). We implement these tests using six lags or leads, and we examine tests using the true innovations, AR(1) or ARMA(2,1) residuals, and AR(6) residuals.

2.5 Two-Sided Regression Tests

Sims (1972) proves that the regression of Y_t on past, current, and future values of X

$$Y_t = \sum_{k=-\infty}^{\infty} \beta_k X_{t-k} + \eta_t, \tag{2.6}$$

will not include the future values of X if and only if Y does not help predict X. In our experiments with two-sided regression tests we use six leads and lags

$$Y_t = \alpha_1 + \sum_{k=-6}^{6} \beta_{1k} X_{t-k} + \eta_{1t}, \tag{2.7a}$$

$$X_t = \alpha_2 + \sum_{k=-6}^{6} \beta_{2k} Y_{t-k} + \eta_{2t}. \tag{2.7b}$$

To test the hypothesis of independence, we calculate the F statistic for the significance of the regression in (2.7a), which is distributed as F (13, T − 26) under the null hypothesis when the regression disturbances are serially independent. To test the hypothesis that Y does not help predict X, we compare the F statistic for the significance of the six lead coefficients in (2.7a), which is distributed as F (6, T − 26) under the null hypothesis when the regression disturbances are serially independent. Similarly, the test that X does not help predict Y is based on the six lead coefficients in (2.7b).

It is important to note that the disturbances in (2.7a) or (2.7b) will generally be serially correlated. Under the null hypothesis that Y does not help predict X, the error term in (2.7a) will be first-order autoregressive with a coefficient of .5, as can be seen from the structural equation (2.1)

$$Y_t = \frac{\gamma_{12}}{(1 - .5L)} X_t + \frac{1}{(1 - .5L)} u_{1t} \,. \tag{2.8}$$

Thus, the disturbance η_{1t} in (2.7a) follows an AR(1) process when Y and X are independent, or when there is a one-way predictive relation from X to Y. When feedback is present, the autocorrelation structure of η_{1t} is much more complicated.

To correct for serial correlation in the disturbances of (2.7a) and (2.7b), we use a one-step second-order Cochrane-Orcutt (1949) procedure. An AR(2) model is estimated for the residuals; then the data for Y and X are transformed using the AR(2) filter; finally, the regression model is reestimated using the transformed data. The tests are based on the final regressions. This procedure yields asymptotically valid tests in our experiments, although it will not correct for the serial correlation of the disturbances when there is feedback between Y and X. Because of this serial correlation correction, the df in the denominator of the F statistics are T − 28 instead of T − 26.

3. SMALL-SAMPLE PROPERTIES OF ALTERNATIVE TESTS

Our sampling experiments are based on data generated by the five hypothetical structures depicted in Table 1 that correspond to various values of γ_{12} and γ_{21} selected to provide examples of independence, one-way predictive relations, and feedback.

Table 1. Hypothetical Structures Used in Sampling Experiments

	Structural Coefficients Equation (2.1)		Reduced-Form Coefficients Equation (2.2)			
Structure	γ_{12}	γ_{21}	ϕ_{11}	ϕ_{12}	ϕ_{22}	ϕ_{21}
I	0	0	.5	0	.5	0
II	.5	0	.5	.25	.5	0
III	1.0	0	.5	.5	.5	0
IV	.37	.37	.58	.21	.58	.21
V	.85	.085	.54	.46	.54	.046

Structure I corresponds to the situation in which X and Y are independent. Structures II and III represent cases where X helps predict Y with different strengths. Structures IV and V are feedback situations with the strength being symmetric in IV, while only weak feedback from Y to X is present in V. Each of these structures is used to generate Y and X series of lengths 50, 100, and 200 observations.

3.1 Tests of Independence Versus Feedback

When the null hypothesis is that Y and X are independent, the investigator will presumably employ test statistics that are two-sided, so that a predictive relation in either direction will lead to rejection of the null hypothesis. Table 2 presents the frequencies of rejection in these tests when conducted at a presumed 5 percent significance level over 500 replications. If the true probability of rejection is 5 percent, then these empirical frequencies have a standard error of about 1 percent.[2] To increase the comparability of the experiments across different structures and different sample sizes, the same seed is used to start the normal random number generator for all of the experiments. However, this also means that the results for different experiments are not independent.

Structure I allows us to estimate the size of the various tests. The reduced-form tests, based on the large-sample distribution of the likelihood ratio statistic, reject about 5 percent of the time when the df correction is used. However, there seems to be a tendency for the rejection frequencies to rise as the sample size increases. When the standard likelihood ratio test statistic is used without the df adjustment, the reduced-form tests reject far too frequently. This is an especially serious problem for the over-parameterized bivariate AR(6) model with a sample size of 50, which rejects 25.2 percent of the time when Y and X are independent.

[2] Every trial is an independent realization from a binomial distribution with a probability of rejection equal to p. The proportion of rejections, $\hat{p}$ is an unbiased estimator of p, and the variance of the sample proportion is $p(1 - p)/500$. Thus, for $p = .05$, the standard error of $\hat{p}$ is 0.00975.

Table 2. Two-Sided Tests (percent rejections at the 5% level)

	Structures						
	I (X,Y independent) *Sample Size*			*II ($X \rightarrow Y$)* *Sample Size*	*III ($X \rightarrow Y$)* *Sample Size*	*IV ($X \leftrightarrow Y$)* *Sample Size*	*V ($X \leftrightarrow Y$)* *Sample Size*
Tests Based on	*50*	*100*	*200*	*50*	*50*	*50*	*50*
Reduced Form							
Bivariate AR(1)	5.4	6.2	6.8	85.6	100.0	100.0	100.0
Bivariate AR(6)	3.4	3.6	4.2	21.2	89.6	73.4	89.8
Two-Sided Regressions							
AR(2) Cochrane-Orcutt	17.0	10.4	6.2	**	**	**	**
ARMA Residual Cross Correlations							
True Innovations	5.4	5.0	4.4	47.0	96.8	81.4	95.8
ARMA(2,1) Residuals	4.6	5.0	3.8	38.6	95.2	79.8	93.6
AR(6) Residuals	3.8	4.2	3.6	22.8	89.8	68.4	89.8
ARMA Residual Regressions							
True Innovations	6.0	6.0	5.0	33.0	87.0	63.8	85.2
ARMA(2,1) Residuals	5.4	5.8	5.2	28.2	84.6	63.6	82.2
AR(6) Residuals	5.2	4.8	4.6	24.6	80.4	55.8	78.4

NOTE: These are tests of whether Y and X are independent at all leads and lags. See Table I for a description of the structures that are the basis for the experimental design. Briefly, Structure I represents independence, Structures II and III represent X helping to predict Y at different strengths, and Structures IV and V represent feedback between X and Y. Thus, Structure I provides estimates of size and Structures II–V provide estimates of power. The different tests are described in detail in Section 2 of the article. The rejection frequencies for Structures II–V with sample sizes 100 and 200 are all close to 100, so they are not shown.

** These entries are omitted because the size of this test seems to be incorrect.

The two-sided regression tests using the second-order Cochrane-Orcutt procedure to correct for residual autocorrelation rejects far too frequently for samples of size 50 or 100 observations, but the size of the test is approximately correct for a sample size of 200.

The estimated sizes of the residual cross-correlation tests are within 1.5 standard errors of 5 percent for all three sample sizes, although there is some tendency for the frequency of rejection to decrease as the sample size increases. The estimated sizes of the residual regression tests are even closer to 5 percent. This is not surprising since the regression test based on the true innovations is the only test among the set we are analyzing that is exact in finite samples.

The results for Structures II through V give estimates of power for tests that have the correct size under various alternative hypotheses. Since the rejection frequencies are all close to 100 percent for sample sizes of 100 and 200, these results are not shown in Table 2. The largest contrasts in Table 2 are seen in the experiment with Structure II and 50 observations. Recall that in Structure II X helps predict Y, but the relationship is not as strong as in Structure III. The reduced form based on the bivariate AR(1) specification is the exact reduced-form vector ARMA representation for the data we use in our experiments; therefore, it is not surprising that this test offers the highest power. The test based on the overspecified bivariate AR(6) representation does not reject as often as the test based on the correct bivariate AR(1) model.

The ARMA residual cross-correlation tests seem to offer greater power than do the comparable ARMA residual regression tests. As expected, the tests based on the true innovations have more power than the tests based on the AR(1) or ARMA(2,1) residuals, reflecting the fact that the ARMA parameters must be estimated in order to perform the latter test. Similarly, the tests based on the AR(1) or ARMA(2,1) residuals are more powerful than those based on the AR(6) residuals. The ARMA residual tests seem to be more powerful than the bivariate AR(6) reduced-form test when the structure of the ARMA model is known, even if the ARMA parameters must be estimated.

The rejection frequencies for the two-sided regression tests are not reported for Structures II through V, because it is apparent that the nominal size of this test is incorrect when smaller sample sizes are available.

Thus, among the tests which have the correct 5 percent significance level, the exact reduced-form test is most powerful, followed by the ARMA residual cross-correlation and regression tests and the overparameterized reduced-form test, with the ordering of the latter tests depending on the true parameters.

3.2 Tests of One-Way Predictive Relations

Part A of Table 3 presents rejection frequencies for tests of the null hypothesis that X does not help predict Y against the alternative that X does help predict Y. The reduced-form tests are based on the t ratio for lagged X in the regression of Y_t on one lag of Y and one lag of X, or the F ratio for all six lags of X in regressions that include six lags of Y. Two-sided regression tests are based on the joint significance of all six *lead* coefficients in a regression of X_t on current, six lags, and six leading values of Y. The ARMA residual tests are based on cross-correlations and regressions of Y innovations or residuals on six past innovations or residuals for X. Part B of Table 3 contains rejection frequencies for comparable tests of the hypothesis that Y does not help predict X.

In Part A of Table 3, the results for Structure I provide estimates of the size of the tests. In Part B, Structures

Table 3. Tests of One-Way Predictive Relations (percent rejections at the 5% level)

	Structures														
	I (X,Y independent) *Sample Size*			*II (X → Y)* *Sample Size*			*III (X → Y)* *Sample Size*			*IV (X ↔ Y)* *Sample Size*			*V (X ↔ Y)* *Sample Size*		
Tests Based on	*50*	*100*	*200*	*50*	*100*	*200*	*50*	*100*	*200*	*50*	*100*	*200*	*50*	*100*	*200*
Part A: Tests of X Predicting Y															
Reduced Form															
One Lag	7.0	6.2	6.0	32.2	57.6	86.8	46.8	75.6	95.0	21.6	39.8	70.8	45.4	73.4	94.2
Six Lags	7.0	5.6	5.0	14.0	26.0	53.4	15.8	35.0	73.8	12.2	19.6	36.2	15.8	35.2	73.6
Two-Sided Regressions															
AR(2) Cochrane-Orcutt	9.8	5.6	4.6	**	26.8	51.8	**	34.0	71.6	**	18.2	33.8	**	33.8	71.2
ARMA Residual Cross Correlations															
True Innovations	3.2	4.6	4.8	8.0	18.0	35.8	6.0	14.2	29.4	4.6	8.0	15.2	5.6	13.2	27.2
ARMA(2,1) Residuals	3.2	4.6	4.4	4.8	11.6	32.8	2.0	5.0	18.4	2.8	4.0	9.0	2.0	4.2	16.8
AR(6) Residuals	2.6	3.6	4.0	3.8	11.6	28.4	2.4	3.2	13.4	1.4	2.8	4.0	2.2	3.2	11.8
ARMA Residual Regressions															
True Innovations	6.2	4.8	5.6	8.6	18.0	37.6	7.0	13.6	31.0	6.0	7.6	15.4	6.6	13.0	27.8
ARMA(2,1) Residuals	5.8	4.6	5.4	6.0	14.4	34.4	4.2	7.0	22.0	3.4	5.2	9.4	3.8	7.4	18.4
AR(6) Residuals	5.8	4.8	4.2	5.8	13.4	28.8	3.0	3.6	14.4	2.6	2.6	4.6	3.0	3.4	12.8
Part B: Tests of Y Predicting X															
Reduced Form															
One Lag	6.4	5.8	5.0	6.8	6.0	4.6	8.0	6.6	6.2	22.2	41.2	71.6	6.4	8.0	12.8
Six Lags	8.2	5.8	5.6	5.8	7.8	5.4	5.8	6.4	6.6	8.8	17.8	38.6	6.6	6.2	7.6
Two-Sided Regressions															
AR(2) Cochrane-Orcutt	9.6	7.6	4.6	9.8	8.4	6.0	8.6	6.8	6.8	**	19.0	38.8	**	7.8	7.6
ARMA Residual Cross Correlations															
True Innovations	4.2	4.2	3.8	4.4	6.2	3.2	4.0	6.8	4.4	5.6	10.2	15.0	3.8	7.0	4.6
ARMA(2,1) Residuals	4.6	4.2	4.0	3.6	3.8	3.0	4.0	3.4	2.2	2.4	5.2	10.8	2.4	1.6	2.0
AR(6) Residuals	3.2	4.4	3.6	1.4	2.0	0.8	0.2	0.4	0.2	**	**	**	**	**	**
ARMA Residual Regressions															
True Innovations	7.4	5.8	4.8	5.2	7.0	4.0	6.0	7.6	5.4	6.0	9.8	16.0	5.8	7.0	5.2
ARMA(2,1) Residuals	7.6	5.4	4.6	5.8	5.4	5.2	7.4	3.8	3.6	4.0	7.0	12.2	3.6	1.6	3.0
AR(6) Residuals	6.0	5.2	5.6	3.2	2.6	1.8	0.6	0.4	0.2	**	**	**	**	**	**

NOTE: Part A contains tests of whether lagged *X* improves the prediction of *Y*, given lagged values of *Y*. Part B contains tests of whether lagged *Y* improves the prediction of *X*, given lagged values of *X*. See Table I for a description of the structures which are the basis for the experimental design. Briefly, Structure I represents independence, Structures II and III represent *X* predicting *Y* at different strengths, and Structures IV and V represent feedback between *X* and *Y*. Thus, for Part A, Structure I provides estimates of size and Structures II–V provide estimates of power; for Part B, Structures I–III provide estimates of size and Structures IV and V provide estimates of power. The different tests are described in detail in Section 2 of the paper.

** These entries are omitted because the size of the test appears to be incorrect.

I, II, and III provide estimates of size, since these are all cases where Y does not help predict X. Structures II and III provide an opportunity to examine the Durbin problem with ARMA residual tests, because X helps to predict Y in these cases. In Structure I, where Y and X are independent, most of the tests have rejection frequencies within two standard deviations of 5 percent, except for the two-sided regression test for a sample size of 50. There are a few cases where a particular test rejects too often or too infrequently in one part of Table 3, but the comparable test in the other part of the Table is close to the nominal 5 percent significance level. The rejection frequencies seem to be slightly low for the residual cross-correlation tests, but none of the deviations is more than two standard deviations below the nominal 5 percent significance level.

In Part B of Table 3, the results for Structures II and III appear to be similar to the results for Structure I for the reduced-form tests, the two-sided regression tests, and for the ARMA tests based on the true innovations. These are all tests that should not be affected by the fact that X helps predict Y. On the other hand, most of the ARMA residual tests appear to be affected by the Durbin problem, particularly when the AR(6) model is used to estimate the residuals. In Structure III, where the predictive relation between Y and X is strongest, the rejection frequencies for the AR(6) residual cross-correlation and regression tests are less than 1 percent for all sample sizes. Interestingly, the Durbin problem does not seem to be as serious for the residual regression tests when the correct ARMA(2,1) model is used to estimate the residuals; in Structure II all of the rejection frequencies are slightly above 5 percent and in Structure III the rejection frequencies are not more than 1.5 standard deviations lower than the nominal 5 percent significance level. Thus, it seems that the Durbin problem does not seriously affect the size of the residual regression tests when the correct form of the ARMA model is known.

The relative power of the tests is quite consistent across experiments with Structures II through V in Part A of

Table 3, which are all cases in which X helps predict Y. In particular, there is little to choose between the ARMA residual tests, neither of which is as powerful as the Cochrane-Orcutt version of the two-sided regression test or the reduced-form tests. The power of the two-sided regression Cochrane-Orcutt test is always comparable to that of the six lag reduced-form test, but both are dominated by the one lag reduced-form test that estimates the correct reduced form

There are several interesting aspects to the ARMA residual tests. First, the rejection frequencies are higher for Structure II than for III, even though past X are more frequently significant in predicting Y_t in Structure III. The explanation for this result is that the *lagged* residual cross correlations for Structure II are larger and more persistent than for III. Second, the Durbin problem is evident in the results for Structures IV and V, where there is feedback between Y and X. The rejection frequencies for the tests based on residuals are much lower than for the tests based on true innovations. In particular, when there is strong feedback in Structure IV, the tests based on AR(6) residuals never reject more than five percent of the time, which is the size of the test. Nevertheless, when the correct ARMA(2,1) model is used to estimate the residuals, the rejection frequencies are significantly greater than five percent for a sample size of 200 in both Structures IV and V.

In Part B of Table 3, Structures IV and V represent cases where the power of the tests can be compared under strong (IV) and weak (V) feedback. The most powerful test is the one-lag reduced-form test; in fact, this is the only test with a rejection frequency more than three standard errors greater than its apparent significance level for the weak feedback case (for example, 12.8 percent rejections with a sample size of 200). The power of the six-lag reduced-form test and the two-sided regression Cochrane-Orcutt test are comparable in both feedback cases, given that the significance level of the latter test appears to be too high for small sample sizes. However, neither of these tests appears to have power much greater than the level of the test in the weak feedback case.

Rejection frequencies for the ARMA residual tests, whether based on true innovations or on residuals from the fitted ARMA(2,1) model, rise above the 5 percent significance level when feedback is strong and the sample size is 200. However, the same tests based on the AR(6) residuals reject less frequently than the nominal significance level even in the case of strong feedback, again reflecting the seriousness of the Durbin problem.

4. CONCLUSIONS

As with any Monte Carlo study, our conclusions must be tempered by the limitations of the experimental design. For example, it is worth noting that tests based on the regression of one variable on past values of both variables are reduced-form tests only when, as in our case, the reduced-form vector process is purely autoregressive. In another situation, the correct reduced form might well include moving averages in past values of the reduced-form disturbance, perhaps the result of moving average disturbances in underlying structural equations. The correct reduced form would still be discoverable from the data alone, and tests for predictive relations would involve off-diagonal moving average as well as autoregressive parameters. If an investigator fitted a bivariate AR(1) model in such a case it would not yield reduced-form tests as it did in our situation; rather, it would simply be a misspecification. A high-order bivariate autoregressive representation could then be viewed as a computationally convenient, but presumably statistically inefficient, way of approximating the correct bivariate mixed ARMA process. Further experimentation to study the consequences of this kind of misspecification on the reduced-form tests might be worthwhile. Similarly, it would be interesting to examine the effect of varying number of lags used to compute all of the test statistics.

The structural system we use in our experiments implies a relatively strong contemporaneous relationship between the variables, which is why our two-sided tests in Table 2 appear to be much more powerful than the one-sided tests in Table 3. It would be interesting to see how the power of the tests would change if the structural model implied stronger lagged predictive relations.

Keeping in mind that our results are confined to variations in one particular system, we regard the evidence presented in this article as pointing strongly to the following generalizations. (a) The most powerful tests are those based on the correct reduced-form model for the variables; (b) power is lost when a test includes estimation of irrelevant parameters; (c) tests based on cross-correlations or regressions of univariate ARMA residuals are less powerful than parametric tests based on reduced-form models, and additional power is lost in having to rely on residuals from estimated models instead of the unobserved univariate innovations; (d) as implied by Durbin's analysis, tests based on residuals tend to have smaller size and little power in testing for predictive ability in one direction when predictive ability is present in the other direction; (e) the two-sided regression tests are sensitive to correct specification of autocorrelation of the regression errors, with incorrect specification resulting in the actual size exceeding the nominal significance level;[3] (f) when appropriate correction for residual autocorrelation is made in the two-sided regressions, so the size of the test is approximately correct, that test is about as powerful as the over-specified reduced-form test that includes a comparable number of parameters.

[3] Geweke, Meese, and Dent (1979) and Guilkey and Salemi (1979) have reached similar conclusions. Geweke et al. analyze simulations of the two-sided regression test using a Hannan-efficient correction for serially correlated disturbances and a sample size of 100. Thus, it appears that the problem of correcting for serial correlation is serious for samples as large as 100 observations when performing the two-sided regression test.

We argue that tests of predictive relations based on the parameters of univariate or multivariate ARMA models have unknown properties, since the information matrix of the unrestricted ARMA model is singular under the null hypothesis that Y and X are independent. The singularity of the information matrix occurs because there are redundant roots in the autoregressive and moving average polynomials of the unrestricted ARMA model when there is no relationship between the variables.

Our advice to empiricists is to base tests of predictive relations on reduced-form vector ARMA models, keeping in mind that the list of variables should include the entire set of variables that are part of the relevant information set. This strategy is simple to generalize to situations in which more than two variables are relevant, whereas the other procedures discussed in this paper would be difficult to implement with more than two variables. Further, regardless of the outcome of tests for possible predictive relations, the investigator is left with a model that is useful in forecasting and that exploits the predictive relations among the variables. The alternative testing procedures discussed in this article do not generally lead to operational forecasting models, particularly in cases where the hypothesis of one-way predictive ability is rejected in favor of feedback.

[*Received April 1979. Revised August 1981.*]

REFERENCES

BARTLETT, M.S. (1955), *Stochastic Processes*, London: Cambridge University Press.

COCHRANE, D., and ORCUTT, G.H. (1949), "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms," *Journal of the American Statistical Association*, 44, 32–61.

DURBIN, J. (1970), "Testing for Serial Correlation in Least-Squares Regression When Some of the Regressors are Lagged Dependent Variables," *Econometrica*, 38, 410–421.

FEIGE, EDGAR L., and PEARCE, DOUGLAS K. (1976), "Economically Rational Expectations: Are Innovations in the Rate of Inflation Independent of Innovations in Measures of Monetary and Fiscal Policy?," *Journal of Political Economy*, 84, 499–522.

——— (1979), "The Casual Causal Relationship Between Money and Income: Some Caveats for Time Series Analysis," *Review of Economics and Statistics*, 61, 521–533.

GEWEKE, JOHN (1981a), "A Comparison of Tests of the Independence of Two Covariance-Stationary Time Series," *Journal of the American Statistical Association*, 76, 363–373.

——— (1981b), "The Approximate Slopes of Econometric Tests," *Econometrica* (forthcoming).

GEWEKE, JOHN, MEESE, RICHARD, and DENT, WARREN (1979), "Comparing Alternative Tests of Causality in Temporal Systems: Analytic Results and Experimental Evidence," *Journal of Econometrics*, in press.

GRANGER, C.W.J. (1969), "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica*, 37, 424–438.

——— (1973), "Causality, Model Building and Control: Some Comments," unpublished manuscript, University of Nottingham.

GUILKEY, DAVID K., AND SALEMI, MICHAEL K. (1979), "Small Sample Properties of Three Tests for Granger-Causal Ordering in a Bivariate Stochastic System," unpublished manuscript, University of North Carolina.

HANNAN, E.J. (1970), *Multiple Time Series*, New York: John Wiley.

HAUGH, LARRY D. (1972), "The Identification of Time Series Interrelationships with Special Reference to Dynamic Regression," unpublished doctoral dissertation, University of Wisconsin, Dept. of Statistics.

——— (1976), "Checking the Independence of Two Covariance-Stationary Time Series: A Univariate Residual Cross Correlation Approach," *Journal of the American Statistical Association*, 71, 378–385.

HSIAO, CHENG (1979), "Autoregressive Modeling of Canadian Money and Income Data," *Journal of the American Statistical Association*, 74, 553–560.

——— (1981), "Autoregressive Modeling and Money-Income Causality Detection," *Journal of Monetary Economics*, 7, 85–106.

JACOBS, RODNEY L., LEAMER, EDWARD E., and WARD, MICHAEL P. (1979), "Difficulties With Testing for Causation," *Economic Inquiry*, 17, 401–413.

KINDERMAN, A.J., and RAMAGE, J.G. (1976), "Computer Generation of Normal Random Variables," *Journal of the American Statistical Association*, 71, 893–896.

NELSON, CHARLES R. (1979), "Discussion of the Zellner and Schwert Papers," *Carnegie-Rochester Conference Series on Public Policy* (supplement to *Journal of the Monetary Economics*), 10, 97–101.

PIERCE, DAVID A. (1977a), "Relationships—and the Lack Thereof—Between Economic Time Series, With Special Reference to Money and Interest Rates," *Journal of the American Statistical Association*, 72, 11–22.

——— (1977b), "Rejoinder," *Journal of the American Statistical Association*, 72, 24–26.

PIERCE, DAVID A., and HAUGH, LARRY D. (1977), "Causality in Temporal Systems: Characterizations and a Survey," *Journal of Econometrics*, 5, 265–294.

SARGENT, THOMAS J. (1976), "A Classical Econometric Model of the United States," *Journal of Political Economy*, 84, 207–237.

SCHWERT, G. WILLIAM (1979), "Tests of Causality: The Message in the Innovations," *Carnegie-Rochester Conference Series on Public Policy* (supplement to *Journal of Monetary Economics*), 10, 55–96.

SIMS, CHRISTOPHER A. (1972), "Money, Income, and Causality," *American Economic Review*, 62, 540–552.

——— (1977), "Comment," *Journal of the American Statistical Association*, 72, 23–24.

WALLIS, KENNETH F. (1977), "Multiple Time Series Analysis and the Final Form of Econometric Models," *Econometrica*, 45, 1481–1498.

ZELLNER, ARNOLD (1979), "Causality and Econometrics," *Carnegie-Rochester Conference Series on Public Policy* (supplement to *Journal of Monetary Economics*), 10, 9–54.

ZELLNER, ARNOLD, and PALM, FRANZ (1974), "Time Series Analysis and Simultaneous Equation Econometric Models," *Journal of Econometrics*, 2, 17–54.