Variance-Ratio Tests: Small-Sample Properties with an Application to International Output Data

Stephen G. Cecchetti; Pok-sang Lam


Stable URL:  
http://links.jstor.org/sici?sici=0735-0015%28199404%2912%3A2%3C177%3AVTSPWA%3E2.0.CO%3B2-9

Variance-Ratio Tests: Small-Sample Properties With an Application to International Output Data

Stephen G. Cecchetti and Pok-sang Lam
Department of Economics, Ohio State University, Columbus, OH 43210

Two aspects of statistical inference using variance-ratio statistics are studied, (1) the accuracy of asymptotic approximations in small samples and (2) the size distortion arising from searching over many horizons in deciding whether to reject a model. A joint test combining variance-ratio statistics at various horizons is proposed, and a Monte Carlo procedure for conducting exact inference is provided. The real output data of nine countries are used to discuss these issues.

KEY WORDS: Asymptotic testing; Breaking trends; Joint testing; Markov switching; Monte Carlo; Size distortion.

The empirical finance and macroeconomics literature is filled with attempts to measure the persistence in economic time series. Many studies use the variance-ratio (VR) statistic for this purpose. In finance, Poterba and Summers (1988), Lo and MacKinlay (1988), and Richardson and Stock (1990) all used it to detect deviations of stock prices from a random-walk process. In the business-cycle literature, Cochrane (1988), Cogley (1990), Campbell and Mankiw (1987a,b; 1989), Perron (in press), and Kormendi and Meguire (1990) used it to measure the long-run effects of shocks on the path of real output.

The VR statistic, which is the sample variance of the \( k \)-period difference divided by \( k \) times the sample variance of the first difference, is useful in empirical finance when the null hypothesis is that shock prices follow a random walk. As Poterba and Summers (1988), Richardson and Stock (1990), and others emphasized, the VR statistic provides an intuitively appealing way to search for mean reversion, an interesting deviation from the random-walk hypothesis. In this context, use of the VR statistic has several advantages. First, Faust (1992) showed that it can be viewed as a likelihood ratio statistic for testing a Gaussian random-walk null against an alternative that the series is a Gaussian autoregressive (AR) process in first differences. Faust's conclusion implies that, when the VR statistic is used to test for a random walk, the test has optimal power asymptotically against a particular alternative that is often of substantial interest. Lo and MacKinlay (1989), using Monte Carlo experiments, demonstrated that the VR statistic is more powerful than either the Dickey–Fuller test or the Box–Pierce \( Q \) test for several interesting alternatives.

The VR statistic is also useful for computing the persistence of real output. First, as Cochrane (1988) emphasized, it is a measure of the spectral density at frequency 0, so it yields an estimate of the size of the random-walk component. Except for the choice of horizon, the VR statistic is nonparametric, so it has advantages over other methods. For example, Campbell and Mankiw (1987a) suggested measuring persistence by estimating the autoregressive moving average (ARMA) representation of output and then examining the implied moving average representation of the estimated process. But, as Cochrane (1988) pointed out, this methodology relies on short-horizon autocorrelations to infer long-run phenomenon. Furthermore, Christiano and Eichenbaum (1990) showed that slight changes in parametric assumptions about the ARMA process yield dramatically different conclusions about persistence. The VR statistic is immune to these problems.

Finally, Perron (in press) argued that the VR statistic is particularly useful when the time series contains a break in the level of its deterministic trend. For this case, he demonstrates that the VR statistic continues to be a consistent measure of the spectral density at frequency 0 but that standard unit-root tests are biased toward nonrejection of the unit-root hypothesis. Although it is possible to provide valid inference for such standard tests, doing so requires knowledge of the trend-break date, as well as tabulation of the critical values of the test statistics (Perron 1989). Neither of these is necessary if one uses the VR statistic.

Much of the applied literature using VR statistics employs asymptotic approximation to conduct inference. For testing the random-walk null, Lo and MacKinlay (1988, 1989) and Richardson and Stock (1990) both derived and applied asymptotic distributions for the statistic under different sets of assumptions. In the business-cycle literature, Campbell and Mankiw (1987a) and Perron (in press) used the asymptotic results of Priestly (1982) to derive standard errors for null models more general than a random walk. Both Lo and MacKinlay (1989) and Richardson and Stock (1990) provided extensive evidence on the properties of their asymptotic approximations in small samples.

Our first goal is to study the accuracy of the Priestley

approximation, about which very little is known, in small samples. We provide evidence for null models that include higher-order AR components as well as regime switches, both of which are plausible when considering output data. By expanding the set of null models examined, we extend the work of both Lo and MacKinlay (1989) and Richardson and Stock (1990). Specifically, we study four models, all of which we estimate using annual U.S. per capita real gross national product (GNP) data from 1871 to 1985—(1) a simple AR model with a deterministic trend (the trend model), (2) an autoregressive model with a unit root [the autoregressive integrated moving average (ARIMA) model], (3) Lam’s (1990) generalization of the trend model in which the trend switches according to a Markov process, and (4) Hamilton’s (1989) extension of the ARIMA model in which the drift switches according to a Markov process.

Our second goal is to address the inadequacies in the common practice of testing models by using a sequence of VR statistics at various horizons. The first problem with this method is the potential size distortion arising from the search process. In addition, sequentially testing individual horizons fails to take account of the covariance between VR statistics at different horizons. To address the first deficiency, we evaluate the size distortion associated with the common practice by using Monte Carlo experiments based on models of U.S. annual real output data. To tackle the second problem, we devise a joint test that combines the information contained in statistics at several horizons.

To summarize our results, first we find that VR tests based on asymptotic approximations are often misleading. There are size distortions, and the magnitude and direction are sensitive to both the horizon over which the VR statistic is computed and the method used for calculating the standard errors. In addition, we show that there is substantial size distortion in the common practice of searching over many horizons to decide whether to reject a model. These results highlight the virtue of using Monte Carlo methods to conduct inference.

We illustrate our findings by examining the per capita real GNP data for nine countries, including the United States. Here we show that the rejection of the deterministic trend model is not nearly as dramatic as indicated by the asymptotic tests. Furthermore, using our joint test, we find that, with the exception of the trend model for Norway, none of the four-model—nine-country combinations can be rejected at the 5% level of significance. In contrast to the claims of Perron (in press) our failure to reject the Hamilton or Lam models suggests that in practice VR statistics might not be particularly helpful in measuring persistence in the presence of breaks.

The remainder of this article is composed of six sections. Section 1 describes the models we study, and presents estimates for each country. Section 2 uses the estimates of the models based on U.S. data to study the size distortion of asymptotic tests based on a sequence of univariate statistics at various horizons. Section 3 presents our test using VR statistics at different horizons simultaneously. Section 4 considers the small-sample $p$ values and estimates of the joint-test statistic for each model and each country. Section 5 discusses the relationship between our results and those in the existing literature. Section 6 contains concluding remarks.

1. ALTERNATIVE MODELS: DESCRIPTION AND ESTIMATES

We begin with a description of the models we are interested in examining, followed by a presentation of estimates of the models using output data from nine countries. Our discussion uses the framework of Hamilton and Lam, which assumes that output is the sum of a general autoregressive component and a trend that switches between a normal and an unusual state based on a simple Markov process.

Let $y_t$ be the log of real GNP, $z_t$ a possibly nonstationary AR process, and $n_t$ a Markov trend. Then we can write

$$y_t = z_t + n_t, \quad (1)$$

$$\phi(L)z_t = \epsilon_t, \quad (2)$$

and

$$n_t = n_{t-1} + \alpha_0 + \alpha_1 S_t, \quad (3)$$

where $\phi(L)$ is a polynomial in the lag operator $L$, $\epsilon_t$ is an iid normal random variable with variance $\sigma^2$, and $S_t$ is a Markov random variable that takes on values of 0 or 1 with transition probabilities

$$Pr(S_t = 1|S_{t-1} = 1) = p$$

$$Pr(S_t = 0|S_{t-1} = 1) = 1 - p$$

$$Pr(S_t = 1|S_{t-1} = 0) = 1 - q$$

$$Pr(S_t = 0|S_{t-1} = 0) = q. \quad (4)$$

Our interpretation of the Markov process is that the economy is in a normal state with growth rate $\alpha_0$ when $S_t = 0$, but it is in an unusual state with growth rate $(\alpha_0 + \alpha_1)$ when $S_t = 1$.

The four models we consider are all special cases of (1)-(4). The trend model is the case in which $\phi(L)$ contains no unit root and $\alpha_1 = 0$. The ARIMA model is obtained by restricting $\phi(L)$ to contain a single unit root and assuming that $\alpha_1 = 0$. When $\phi(L)$ has a unit root and $\alpha_1 \neq 0$, we obtain the Hamilton model. Finally, the Lam model is the case in which $\phi(L)$ has no unit root and $\alpha_1 \neq 0$. For the purposes of estimation, we assume that $\phi(L)$ is a fourth-order polynomial, so $z_t$ is an AR(4).

We study annual data from 1871 to 1985 for nine countries—Australia, Canada, Denmark, France, Italy, Norway, Sweden, the United Kingdom, and the United States. The data series are constructed by splicing current data from the Organization for Economic Cooperation and Development’s Main Economic Indicators database to the series of Maddison (1982). We estimate each model for each country by maximum likelihood using the filtering algorithms developed by Lam (1990) and based on the work of Hamilton (1989).

Table 1 presents estimates of the dates in which trend breaks are likely to have occurred, together with estimates of the transition probabilities and the two growth-rate parameters—$p$, $q$, $\alpha_0$, and $\alpha_1$. There are several interesting
Table 1. Estimates of the Markov Switching Models

<table>
<thead>
<tr>
<th>Model</th>
<th>q</th>
<th>p</th>
<th>Normal state average growth</th>
<th>Unusual state average growth</th>
<th>Years of unusual average growth</th>
<th>Percent of var(Δyt) explained by Markov switching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>( (S_t = 0, \alpha_0) )</td>
<td>( (S_t = 1, \alpha_0 + \alpha_1) )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hamilton</td>
<td>.98</td>
<td>.24</td>
<td>1.38</td>
<td>- 9.82</td>
<td>1882, 1892–1993, 1895</td>
<td>18.96</td>
</tr>
<tr>
<td>Lam</td>
<td>.94</td>
<td>.39</td>
<td>2.77</td>
<td>- 4.99</td>
<td>1882, 1892–1993, 1895, 1914, 1931, 1944–46</td>
<td>27.01</td>
</tr>
<tr>
<td>Canada</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hamilton</td>
<td>.96</td>
<td>.42</td>
<td>3.04</td>
<td>-10.78</td>
<td>1886, 1897, 1914, 1921, 1931–1933</td>
<td>33.54</td>
</tr>
<tr>
<td>Lam</td>
<td>.94</td>
<td>.24</td>
<td>2.92</td>
<td>- 9.93</td>
<td>1886, 1897, 1899, 1914, 1920–1921, 1931–1933</td>
<td>30.76</td>
</tr>
<tr>
<td>Denmark</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hamilton</td>
<td>.95</td>
<td>.28</td>
<td>3.16</td>
<td>-10.05</td>
<td>1915, 1917–1918, 1921, 1940–1941, 1945</td>
<td>54.96</td>
</tr>
<tr>
<td>Lam</td>
<td>.95</td>
<td>.28</td>
<td>3.16</td>
<td>- 9.99</td>
<td>1915, 1917–1918, 1921, 1940–1941, 1945</td>
<td>56.62</td>
</tr>
<tr>
<td>France</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hamilton</td>
<td>.99</td>
<td>.00</td>
<td>1.53</td>
<td>38.71</td>
<td>1946</td>
<td>32.55</td>
</tr>
<tr>
<td>Lam</td>
<td>.99</td>
<td>.00</td>
<td>1.40</td>
<td>38.60</td>
<td>1946</td>
<td>32.24</td>
</tr>
<tr>
<td>Italy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hamilton</td>
<td>.98</td>
<td>.50</td>
<td>2.58</td>
<td>-21.80</td>
<td>1919, 1944–1945</td>
<td>49.10</td>
</tr>
<tr>
<td>Lam</td>
<td>.98</td>
<td>.33</td>
<td>3.67</td>
<td>-21.61</td>
<td>1919, 1944–1945</td>
<td>48.14</td>
</tr>
<tr>
<td>Norway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hamilton</td>
<td>.95</td>
<td>.37</td>
<td>3.15</td>
<td>-9.81</td>
<td>1917–1918, 1921, 1931, 1940, 1944</td>
<td>57.94</td>
</tr>
<tr>
<td>Lam</td>
<td>.95</td>
<td>.37</td>
<td>4.03</td>
<td>-8.69</td>
<td>1917–1918, 1921, 1931, 1940, 1942, 1944</td>
<td>63.76</td>
</tr>
<tr>
<td>Sweden</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hamilton</td>
<td>.98</td>
<td>.00</td>
<td>2.40</td>
<td>-9.49</td>
<td>1917, 1931</td>
<td>30.47</td>
</tr>
<tr>
<td>Lam</td>
<td>.98</td>
<td>.00</td>
<td>2.77</td>
<td>-9.15</td>
<td>1917, 1931</td>
<td>31.46</td>
</tr>
<tr>
<td>United Kingdom</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hamilton</td>
<td>.97</td>
<td>.00</td>
<td>1.60</td>
<td>-8.27</td>
<td>1920, 1922</td>
<td>28.89</td>
</tr>
<tr>
<td>Lam</td>
<td>.96</td>
<td>.00</td>
<td>1.81</td>
<td>-7.55</td>
<td>1920, 1922, 1927</td>
<td>31.04</td>
</tr>
<tr>
<td>United States</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hamilton</td>
<td>.98</td>
<td>.49</td>
<td>2.36</td>
<td>-15.44</td>
<td>1930–1932, 1946</td>
<td>37.36</td>
</tr>
<tr>
<td>Lam</td>
<td>.98</td>
<td>.50</td>
<td>2.67</td>
<td>-14.71</td>
<td>1930–1932, 1946</td>
<td>35.97</td>
</tr>
</tbody>
</table>

NOTE: The years labeled "unusual average growth" are those in which the filtered estimates \( \Pr(S_t = 1|y_0, \ldots, y_{t-1}) > .5 \).

results. First, we find that \( q \), the conditional probability of remaining in a normal state, is at least .94 for every case considered, but the estimated values for \( p \), the conditional probability of remaining in an unusual state, are generally quite small. Second, even though the trend breaks are infrequent, they are large on average. The difference between the mean growth rate in the two states ranges from 7.76% for Australia to 37.20% for France, both using the Lam model. The size of these differences implies that the Markov component of the estimated stochastic process explains a significant portion of the fluctuations in annual GNP growth rate—ranging from 16.3% for the United States to 63.8% for Norway, again both using the Lam model.

The sixth column of Table 1 reports the years in which we estimate the economies to have been in the unusual regime—where \( S_t \) is likely to have been equal to 1—which we take to be dates when our filtered estimate of \( \Pr(S_t = 1) \) exceeds .5. We find that, although most countries experienced the unusual regime both sometime during the early 1930s (the Great Depression) and following the end of World War II (1944, 1945, or 1946), the switches generally occurred at different times in different countries, suggesting that different mechanisms are at work. We also note that, in contrast to Perron's (1989) focus on a trend break during the Great Depression, our procedure locates a second break for the United States in 1946.

2. VARIANCE-RATIO STATISTICS

Our next task is to examine the statistics commonly used in attempts to discriminate among the models discussed in Section 1. We begin with the definition of the VR statistic, followed by an examination of its small-sample properties.

The VR is the variance of the \( k \)th difference of a time series, divided by \( k \) times the variance of its first difference. See, for example, Cochrane (1988), Campbell and Mankiw (1989), Cogley (1990), Kormendi and Meguire (1990), and Perron (in press). Following this work, we can write the VR as

\[
VR(k) = \frac{V_k}{V_1},
\]
where \( V_k = \frac{1}{2} \text{Var}(y_t - y_{t-k}). \) It is straightforward to show that \( VR(k) \) can be written as the weighted sum of correlations between \( \Delta y_t \) and \( \Delta y_{t-j} \) and that the limit of \( VR(k) \) as \( k \) goes to infinity is the scaled spectral density of \( \Delta y_t \) at the zero frequency. Cochrane (1988) showed that, if the series \( y_t \) is first-difference stationary, then \( VR(\infty) \) is the portion of the variance in \( \Delta y_t \) that can be attributed to the random-walk component. On the other hand, Quah (1992) demonstrated that any time series can be decomposed into an arbitrarily smooth unit-root process and a stationary process, regardless of the value of the spectral density at frequency 0. His result implies that the interpretation of the spectral density at frequency 0 as the size of a permanent component is valid only if the latter is restricted to be a random walk.

These \( VR \)'s are population measures, and in practice they are estimated from the data. Following the literature, we examine the sample analogs, \( \hat{VR}(k) = \hat{V}_k / \hat{V}_1 \), which we compute using Cochrane's bias adjustment (see Cochrane 1988).

Conducting statistical inference based on the estimate \( \hat{VR}(k) \) requires that we compute its sampling distribution. Lo and MacKinlay (1988) derived an asymptotic distribution for the case of a random-walk null and normal iid innovations. They noted that, for any given \( k \), \( \hat{V}_k \) can be expressed approximately as a linear combination of the first \( k \)-sample autocovariances of the innovations. A standard central limit theorem provides the asymptotic distribution of the sample autocovariances of the normal iid innovations. From these Lo and MacKinlay derived the asymptotic distribution of both \( \hat{V}_k \) and \( \hat{V}_1 \). Because the innovations are normal iid, \( \hat{V}_1 \) is the efficient estimator of the innovation variance. This implies that the covariance between \( \hat{V}_k \) and \( \hat{V}_1 \) is 0. These results, together with a linear approximation of \( \hat{V}_k / \hat{V}_1 \), are used to derive the asymptotic distribution of the \( VR \) statistic. This distribution is normal, mean 1, and variance \( 2(2k - 1)(k - 1) / 3kT \).

The asymptotic distribution of \( Lo \) and MacKinlay is derived under the assumption that \( k \) is fixed and \( T \) goes to infinity. Richardson and Stock (1990) contended that, when \( k \) is large and \( T \) is small, \( Lo \) and MacKinlay's asymptotic distribution is not an accurate approximation. They went on to derive an asymptotic distribution, also under the random-walk null but with both \( k \) and \( T \) going to infinity and \( k/T \) going to a constant. They showed that under this set of assumptions, the asymptotic distribution is not normal but can be represented as a functional of Brownian motion, which can be evaluated numerically by a sequence of Monte Carlo experiments. They showed that their approximation performs better than that of \( Lo \) and MacKinlay.

The business-cycle literature generally examines a more general null model than the random walk. Researchers in this area have relied on the only known result for asymptotic inference, the one due to Priestley (1982). He showed that, as \( k \) and \( T \) both go to infinity and \( k/T \) goes to 0, \( \hat{VR}(k) \) is normally distributed with mean \( VR(k) \) and variance \( (4k/3T)VR(k) \). Note that when the null model is a random walk so that \( VR(k) \) is 1 and \( k \) is large this variance is similar to the one derived by \( Lo \) and MacKinlay.

When \( VR(k) \) is known under the null, Priestley's result gives the asymptotic variance of the \( VR \) statistic and can be used for inference. This is the method employed by Campbell and Mankiw (1987b). When the null hypothesis fails to specify a unique population value for the \( VR \) statistic, this technique cannot be used. In this case, Perron (in press) suggested the simple alternative of replacing \( VR(k) \) with its sample estimate and estimating the asymptotic variance of \( VR(k) \) by \( (4k/3T)VR(k) \).

We now proceed to use Monte Carlo techniques to compare these asymptotic approximations to the empirical distributions of the estimated \( VR \) statistics. We do this for each of the four models described in Section 1, estimated using U.S. data, and for \( k \) from 2 to 40. The Monte Carlo experiments used to compute the empirical distributions are based on the maximum likelihood estimates of the model, 115 time series observation, and 10,000 trials. [Technical descriptions and tables are available from the authors.]

We find that the \( VR \) statistics are biased in small samples. For the ARIMA and Hamilton models, the medians of the empirical distribution are less than the population values, so there is a downward bias. The bias is small for small \( k \), but it increases with \( k \). When \( k \) is 40, it is 20% and 28% of the population values for the ARIMA model and the Hamilton model, respectively. Campbell and Mankiw (1989) found similar bias for the random-walk model. Confirming the results of Christiano and Eichenbaum (1990), the medians of the empirical distributions are above the population values for the trend model, so there is an upward bias. Again the bias is small for small \( k \), but it increases with \( k \). When \( k \) is 40, the bias is as much as 57% of the population value.

We also find that the empirical distributions have a large degree of positive skewness, confirming the results of both \( Lo \) and MacKinlay (1989) and Richardson and Stock (1990). The skewness is more pronounced for large \( k \). For example, when \( k \) is 40, the distance between the 95th percentile and the median is about three times the distance between the median and the 5th percentile for the ARIMA, Hamilton, and Lam models.

Furthermore, the asymptotic standard errors of Priestley, computed using population \( VR \)'s, are inaccurate estimates of the standard deviations of the empirical distributions (the true standard errors). When \( k \) is small, the asymptotic standard errors greatly exaggerate the true standard errors. When \( k \) is 2, for example, the asymptotic standard errors are at least two times the true standard errors for all models. The asymptotic standard errors become more accurate as \( k \) increases—for \( k = 40 \), for example, they overstate the true standard errors by about 19% for the trend model, understate the true standard errors by about 6%, 6%, and 16%, respectively, for the ARIMA model, the Hamilton model, and the Lam model.

All of these properties of the empirical distributions suggest that inference based on the asymptotic result of Priestley will be misleading. We use a Monte Carlo procedure to evaluate exactly how inaccurate it will be. This procedure, except for the difference in the data-generating process and the method of constructing standard errors, is the same as the one used by \( Lo \) and MacKinlay (1989). Specifically, for each
draw of 115 observations from each of the four models, and for each value of \( k \), we compute a \( t \) ratio as the sample VR minus its population value divided by the Priestley standard error, calculated using the population VR \((k)\). We repeat this experiment 10,000 times. The frequency of finding the absolute value of the \( t \) ratio larger than the two-tailed critical value of 1.96 gives the empirical size of a 5% test. Analogously, setting the critical value to 1.645 gives the empirical size of a 10% test.

Three results are worth noting. First, when \( k \) is small, the empirical size of both a 5% test and a 10% test is effectively 0 for all models. The reason is that when \( k \) is small, the asymptotic standard errors greatly overstate the true standard errors. In the work of Lo and MacKinlay (1989), by contrast, the asymptotic approximation is more accurate for smaller \( k \). This difference is not surprising because the asymptotic theory of Lo and MacKinlay assumes a fixed \( k \), but Priestley assumes that \( k \) goes to infinity.

Second, in all models the empirical size of the asymptotic test comes almost exclusively from rejection in the upper tail—the asymptotic test seldom rejects in the lower tail. This is mainly due to the large positive skewness of the empirical distribution. The inability to reject in the lower tail was emphasized by Lo and MacKinlay (1989) in the context of the random-walk null. They show that when \( k/T \) exceeds \( \frac{1}{2} \) the \( t \) statistic based on their asymptotic distribution is always larger than −1.74 so that the random-walk null will never be rejected by a lower-tail test of 10%. Our result shows that their finding carries over to higher-order AR and Markov-switching models. In addition, we note that for the trend model this phenomenon is partly due to the upward bias in the empirical distribution.

Third, we observe that, when \( k \) is large the asymptotic test rejects far too often for the trend model, despite the fact that the empirical size comes exclusively from the upper tail. The major reason, again, is that there is an upward bias in the small-sample distribution and the bias grows as \( k \) increases. When \( k \) is 40, for example, the empirical size is 12.4% for a 5% test and 19.5% for a 10% test.

We repeat the Monte Carlo procedure, computing the asymptotic error using the Perron method. Again we find substantial size distortion but for different reasons. When a small VR is realized, the estimated asymptotic standard error will be small, making it easy to reject a model. On the other hand, when a large variance ratio is realized, the estimated standard error is large, making it difficult to reject the model. As a result, unlike the test that uses population values to estimate standard errors, the empirical size of the test comes almost entirely from the lower tail. For the ARIMA, Hamilton, and Lam models, the downward bias in the small-sample distribution makes the lower tail a likely event, and the result is overwhelming overrejections. For a 5% test and \( k \) equal to 40, the empirical size is about 25% for the Hamilton model and about 20% for both the ARIMA model and the Lam model. On the other hand, for the trend model the lower tail is an unlikely event due to the upward bias of the empirical distribution. As a result, the empirical size is virtually 0 for the trend model.

We draw the following general conclusion about the use of tests based on the asymptotic standard errors of Priestley. The size distortion of the test is often quite large. For small \( k \), the test almost never rejects for any models. For large \( k \), the direction of the distortion depends on whether the population VR's or sample VR's are used to estimate the standard errors. When population VR's are used, the test rejects too often for the trend model and not often enough for the other three models. When the sample VR's are used, it rejects not often enough for the trend model but too often for the other three models. In other words, for large \( k \) the test based on population VR's incorrectly favors models with nonstationary components—the ARIMA, Hamilton, and Lam models—over the trend model, but the reverse is true for tests based on sample VR's.

3. JOINTLY TESTING VR STATISTICS OF MANY HORIZONS

We now proceed to examine procedures that use VR statistics at more than one horizon simultaneously. We begin by studying a commonly used procedure that examines sequentially whether any individual statistic rejects a particular null model and then proceed to propose a joint test that combines all of the information from several VR statistics into a scalar.

First, we look at tests that employ the statistics in a way that is broadly consistent with its use in the existing literature. In past work, such as that of Summers and Poterba (1988) and Cecchetti, Lam, and Mark (1990), a model is considered rejected when at least some of the VR statistics provide evidence against it. Strictly speaking, this method of testing penalizes the null hypothesis because it is subject to a sequence of tests, and the critical value of each test does not take into account the nature of the procedure.

We use Monte Carlo experiments to examine the empirical size of a sequential test of single horizons. For each time series draw of length 115, we compute \( V_R(k) \), for \( k \) from 2 to 40. From 10,000 trials we tabulate the 95th percentile and 5th percentile \( V_R(k) \) for each \( k \) and examine the percentage of the trials that fall outside these bounds for at least one horizon. This gives the empirical size of the test that rejects the model if at least one of the VR statistics rejects the model at a 10% size. We repeat the experiments both for a test of 5% nominal size, and for a test that considers only five horizons, \( k = \{2, 5, 10, 20, 40\} \).

The results are reported in Table 2. The table shows that this test procedure exhibits enormous size distortion regardless of the nominal size and model examined. A test procedure that has a size of 5% individually has an effective size of at least 16% when all 40 horizons are used. When only 5 horizons are used, the size distortion decreases somewhat, but a test that has a size of 5% individually still has an effective size of at least 13%.

One solution to the size distortion problem inherent in the sequential testing procedure is to construct a scalar measure that combines the VR statistics of various horizons and then to use Monte Carlo methods to compute the \( p \) value of the
resulting statistic. The measure we propose takes account of the correlations between VR statistics at various horizons and weights them according to their variances. Specifically, we consider

\[ S(\tau) = \{VR(\tau) - E[VR(\tau)]\}^2 \Sigma^{-1}(\tau) \{VR(\tau) - E[VR(\tau)]\}, \]

where \( E \) is the expectation operator, \( VR \) is a column vector of a sequence of VR statistics \( VR(\tau) = \{VR(2), VR(3), \ldots VR(\tau)\}^T \), \( E[VR(\tau)] \) is the expected value of \( VR(\tau) \), and \( \Sigma(\tau) \) is a measure of the covariance matrix of \( VR(\tau) \).

When the vector of VR statistics \( VR(\tau) \) is distributed as a multivariate normal—as it is asymptotically—then \( S(\tau) \) is distributed as chi-squared with \( (\tau - 1) \) df. But the results in Section 2 suggest that this asymptotic approximation is likely to be very misleading in samples of the size we have available. Consequently, we study the empirical distribution of \( S(\tau) \) using Monte Carlo techniques.

We propose the following procedure to study the sampling distribution of \( S(\tau) \). First, we calculate \( E[VR(\tau)] \) and \( \Sigma(\tau) \). Because these cannot be calculated analytically, we approximate them using Monte Carlo techniques. We begin by generating 20,000 random series of 115 time series observations that conform to the model. From the distribution of VR statistics we compute the mean vector and covariance matrix, \( \overline{VR}(\tau) \) and \( \overline{\Sigma}(\tau) \), which we take as the true population values.

Next, we once again generate 10,000 random series of 115 time series observations. For each, we compute a value for the statistic \( S(\tau) \), using \( \overline{VR}(\tau) \) and \( \overline{\Sigma}(\tau) \), and tabulate the distribution. The value of \( S(\tau) \) implied by the data is computed by setting \( VR(\tau) \) equal to the estimates from the data, again using the Monte Carlo estimates of \( \overline{VR}(\tau) \) and \( \overline{\Sigma}(\tau) \).

We note a simple and intuitive interpretation of \( S(\tau) \). In the Appendix we show that the calculated \( S(\tau) \) is numerically identical to that of the quadratic sum of deviations of the first \( (\tau - 1) \) autocorrelations from their population values, weighted by the inverse of the covariance matrix of the sample autocorrelations, so our procedure can be interpreted as testing whether these autocorrelations taken together conform to a particular model.

Our test can be viewed as a generalization of the one proposed by Diebold (1989). He studied the small-sample properties of a statistic that jointly tests whether a sequence of VR’s all equal 1. Using Monte Carlo experiments, he showed that the test is more powerful against fractionally integrated alternatives than a two-tailed VR test at any individual horizon. Because his null hypothesis is that all VR’s equal 1, his test applies only to the random-walk null. By contrast, our technique is applicable to any arbitrarily chosen null.

### 4. EMPIRICAL RESULTS FOR OUTPUT DATA FROM NINE COUNTRIES

Our final task is to apply the analysis of Sections 2 and 3 to an example of interest to macroeconomists. The issue we examine is whether measures such as these can be used to distinguish the models we propose in Section 1, which clearly have very different degrees of persistence.

We begin by presenting the estimates of \( VR(k) \) from the data, \( \overline{VR}(k) \), together with the population values calculated from the four models. Table 3 reports these values for \( k = (5, 10, 40) \), as well as the population value for \( VR(\infty) \). Although the four models all have similar medium-run properties, the long-horizon measures of persistence are very different. To see this, note that \( VR(5) \) and \( VR(10) \) are similar across models for a given country but that \( VR(40) \) ranges substantially.

The large difference in long-horizon VR’s across models means that we are attempting to discriminate among processes that are quite different. This is similar to attempts to differentiate a random walk from an AR process with a coefficient of .95. Although it may be difficult to do this in practice, it is not theoretically impossible. Blough (1992) and Cochrane (1991) demonstrated that unit-root testing is theoretically impossible because the generic null of the unit-root hypothesis contains within it the case of a nearly stationary process that is indistinguishable from a stationary process in any finite sample. Their results imply that any test for a unit root must be based on additional restrictions. We select the processes to be discriminated among by applying maximum likelihood estimation to Equations (1)–(4). If our approach is to be viewed as a unit-root test, then the procedure of selecting the four specific processes represents a set of implicit restrictions to be tested. But testing a generic unit-root null is not our purpose. Instead, we are attempting to discriminate among four specific processes that are convenient and useful for empirical work.

Section 2 concludes that asymptotic tests have substantial size distortion in small samples. In particular, the asymptotic test with standard errors computed using the population VR’s incorrectly favors unit-root models over the trend model. We proceed to illustrate this point by comparing the inferences based on the asymptotic and small-sample distributions for each of the 36 model/country combinations we estimated in Section 1. To this end, we compute the VR’s from the data for \( k = 2 \) to 40. Then, for each country and model, we take the maximum likelihood estimates in Section 1 as the data-generating process and compute the empirical distribution for each \( \overline{VR}(k) \), using the Monte Carlo experiments described in Section 2. Comparing this distribution to \( \overline{VR}(k) \) from the data allows us to compute the empirical \( p \)-value—that is,
Table 3. Variance-Ratio Statistics: Estimates From the Data and Values Implied by the Models

<table>
<thead>
<tr>
<th>Measure</th>
<th>Estimate from data</th>
<th>Model</th>
<th>Trend</th>
<th>ARIMA</th>
<th>Lam</th>
<th>Hamilton</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VR(5)</td>
<td>1.090</td>
<td></td>
<td>1.019</td>
<td>1.091</td>
<td>1.057</td>
<td>1.007</td>
</tr>
<tr>
<td>VR(10)</td>
<td>1.170</td>
<td></td>
<td>.956</td>
<td>1.100</td>
<td>1.071</td>
<td>.967</td>
</tr>
<tr>
<td>VR(40)</td>
<td>1.912</td>
<td></td>
<td>.662</td>
<td>1.112</td>
<td>.965</td>
<td>.949</td>
</tr>
<tr>
<td>VR(∞)</td>
<td>—</td>
<td></td>
<td>.000</td>
<td>1.115</td>
<td>.539</td>
<td>.943</td>
</tr>
<tr>
<td>Canada</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VR(5)</td>
<td>1.051</td>
<td></td>
<td>1.097</td>
<td>1.112</td>
<td>1.142</td>
<td>1.253</td>
</tr>
<tr>
<td>VR(10)</td>
<td>.865</td>
<td></td>
<td>.931</td>
<td>1.133</td>
<td>.878</td>
<td>1.309</td>
</tr>
<tr>
<td>VR(40)</td>
<td>.881</td>
<td></td>
<td>.337</td>
<td>1.149</td>
<td>.545</td>
<td>1.351</td>
</tr>
<tr>
<td>VR(∞)</td>
<td>—</td>
<td></td>
<td>.000</td>
<td>1.154</td>
<td>.448</td>
<td>1.366</td>
</tr>
<tr>
<td>Denmark</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VR(5)</td>
<td>.911</td>
<td></td>
<td>.777</td>
<td>.842</td>
<td>1.395</td>
<td>1.532</td>
</tr>
<tr>
<td>VR(10)</td>
<td>.946</td>
<td></td>
<td>.691</td>
<td>.820</td>
<td>1.428</td>
<td>1.734</td>
</tr>
<tr>
<td>VR(40)</td>
<td>1.104</td>
<td></td>
<td>.428</td>
<td>.808</td>
<td>1.142</td>
<td>1.922</td>
</tr>
<tr>
<td>VR(∞)</td>
<td>—</td>
<td></td>
<td>.000</td>
<td>.804</td>
<td>.918</td>
<td>1.986</td>
</tr>
<tr>
<td>France</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VR(5)</td>
<td>1.817</td>
<td></td>
<td>1.751</td>
<td>1.789</td>
<td>1.725</td>
<td>1.722</td>
</tr>
<tr>
<td>VR(10)</td>
<td>1.680</td>
<td></td>
<td>1.784</td>
<td>2.024</td>
<td>1.867</td>
<td>2.023</td>
</tr>
<tr>
<td>VR(40)</td>
<td>2.198</td>
<td></td>
<td>.867</td>
<td>2.209</td>
<td>1.080</td>
<td>2.283</td>
</tr>
<tr>
<td>VR(∞)</td>
<td>—</td>
<td></td>
<td>.000</td>
<td>2.271</td>
<td>.317</td>
<td>2.371</td>
</tr>
<tr>
<td>Italy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VR(5)</td>
<td>1.405</td>
<td></td>
<td>1.196</td>
<td>1.316</td>
<td>1.465</td>
<td>1.909</td>
</tr>
<tr>
<td>VR(10)</td>
<td>1.421</td>
<td></td>
<td>1.048</td>
<td>1.281</td>
<td>1.503</td>
<td>2.312</td>
</tr>
<tr>
<td>VR(40)</td>
<td>1.822</td>
<td></td>
<td>.629</td>
<td>1.259</td>
<td>1.306</td>
<td>2.701</td>
</tr>
<tr>
<td>VR(∞)</td>
<td>—</td>
<td></td>
<td>.000</td>
<td>1.252</td>
<td>.868</td>
<td>2.834</td>
</tr>
<tr>
<td>Norway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VR(5)</td>
<td>1.165</td>
<td></td>
<td>.887</td>
<td>1.110</td>
<td>1.788</td>
<td>2.098</td>
</tr>
<tr>
<td>VR(10)</td>
<td>1.159</td>
<td></td>
<td>.825</td>
<td>1.187</td>
<td>2.027</td>
<td>2.711</td>
</tr>
<tr>
<td>VR(40)</td>
<td>1.909</td>
<td></td>
<td>.531</td>
<td>1.255</td>
<td>1.788</td>
<td>3.584</td>
</tr>
<tr>
<td>VR(∞)</td>
<td>—</td>
<td></td>
<td>.000</td>
<td>1.277</td>
<td>1.248</td>
<td>3.931</td>
</tr>
<tr>
<td>Sweden</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VR(5)</td>
<td>1.213</td>
<td></td>
<td>1.017</td>
<td>1.139</td>
<td>.981</td>
<td>1.105</td>
</tr>
<tr>
<td>VR(10)</td>
<td>1.082</td>
<td></td>
<td>.842</td>
<td>1.171</td>
<td>.878</td>
<td>1.145</td>
</tr>
<tr>
<td>VR(40)</td>
<td>1.039</td>
<td></td>
<td>.302</td>
<td>1.197</td>
<td>.526</td>
<td>1.179</td>
</tr>
<tr>
<td>VR(∞)</td>
<td>—</td>
<td></td>
<td>.000</td>
<td>1.205</td>
<td>.304</td>
<td>1.191</td>
</tr>
<tr>
<td>United Kingdom</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VR(5)</td>
<td>1.283</td>
<td></td>
<td>1.244</td>
<td>1.302</td>
<td>1.224</td>
<td>1.313</td>
</tr>
<tr>
<td>VR(10)</td>
<td>.982</td>
<td></td>
<td>1.002</td>
<td>1.198</td>
<td>.946</td>
<td>1.234</td>
</tr>
<tr>
<td>VR(40)</td>
<td>1.224</td>
<td></td>
<td>.465</td>
<td>1.145</td>
<td>.496</td>
<td>1.192</td>
</tr>
<tr>
<td>VR(∞)</td>
<td>—</td>
<td></td>
<td>.000</td>
<td>1.127</td>
<td>.289</td>
<td>1.178</td>
</tr>
<tr>
<td>United States</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VR(5)</td>
<td>1.529</td>
<td></td>
<td>1.501</td>
<td>1.548</td>
<td>1.744</td>
<td>1.765</td>
</tr>
<tr>
<td>VR(10)</td>
<td>1.150</td>
<td></td>
<td>.867</td>
<td>1.470</td>
<td>1.751</td>
<td>1.875</td>
</tr>
<tr>
<td>VR(40)</td>
<td>.469</td>
<td></td>
<td>.204</td>
<td>1.429</td>
<td>1.410</td>
<td>1.967</td>
</tr>
<tr>
<td>VR(∞)</td>
<td>—</td>
<td></td>
<td>.000</td>
<td>1.416</td>
<td>1.017</td>
<td>1.997</td>
</tr>
</tbody>
</table>

NOTE: Calculations are based on the discussion in Section 2.

the percentage of the empirical distribution that lies below the value of the $\text{VR}(k)$ from the data. For each model and country, we then compute the asymptotic $p$ value, which is the probability that a standard normal random variable is less than the sample $\text{VR}(k)$ minus the population $\hat{\text{VR}}(k)$, divided by Priestley's standard errors computed using population VR's.

A striking result is that of the 1,404 individual tests (39 horizons for each of the four models and nine countries), only 15 result in rejection at the 5% level using the empirical distributions. Overall, for most countries it is virtually impossible to distinguish these models at commonly used levels of significance using VR statistics.

By comparison, using the asymptotic $p$ values, 113 of the 1,404 cases are rejected at 5% level nearly 10 times as many. Of these 113 cases, 103 are from the trend model, and these are mainly at long horizons. Furthermore, tests based on empirical $p$ values reveal no rejections of the trend model at the 5% level. These comparisons confirm the findings of Section 2 that asymmetric tests incorrectly favor unit-root models over the trend model when population VR's are used to estimate the standard errors.

Our misgivings about the sequential nature of these tests lead us to examine results using the joint-test statistic proposed in Section 3. Using the techniques described there, we construct estimates of $S(\tau)$ in Equation (6) for each of the nine countries and the four models. Then, we use the Monte Carlo procedures to calculate the empirical $p$ value of each of the estimated statistics. The results for $\tau$ equal to 10 and 40 are reported in Table 4.

As we expect, the models fare much better under this joint test than under the sequential tests. The trend model can be rejected at the 10% level for several horizons for Norway, Sweden, and Italy under the sequential test. Using $S(40)$, the trend model can still be rejected for Norway at the 10% level but not for Italy or Sweden.

Although the Hamilton model and the Lam model can be rejected at the 10% level for Denmark and Norway using the sequential test procedure at short horizons, the results in Table 4 show that these two models look much better under $S(10)$—they are not rejected at the 5% level. For Australia, Canada, Denmark, France, Sweden, and the United States, the joint test does not reject any of the models at the 10% level. At the 5% level, and with $\tau$ set equal to 40, the only model that can be rejected is the trend model for Norway. Overall, these results point to the fact that improperly combining information from VR statistics at different horizons can greatly exaggerate our ability to discriminate among models with very different long-run properties. The overall picture from Table 4 is that we can only reject one model for one country.

5. INTERPRETATIONS AND COMPARISONS WITH PREVIOUS RESULTS

Our results have implications for three types of past research—studies of trend breaks, work on model selection, and research into model discrimination. Perron (1989, in press) concluded that in small samples a stationary time series with trend breaks looks very much like a unit-root process, so it is difficult to distinguish between the two using conventional methods. Perron (in press) considered two types of trend breaks, one in the level and one in the slope of the series. He then examined the VR statistics because, with proper
Table 4. P Values for Joint Tests, S(τ)

<table>
<thead>
<tr>
<th>Maximum horizon (τ)</th>
<th>Model</th>
<th>Trend</th>
<th>ARIMA</th>
<th>LAM</th>
<th>Hamilton</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Australia</td>
<td>.0008</td>
<td>.0005</td>
<td>.0008</td>
<td>.0064</td>
</tr>
<tr>
<td>10</td>
<td>Canada</td>
<td>.3088</td>
<td>.4550</td>
<td>.2759</td>
<td>.4258</td>
</tr>
<tr>
<td>40</td>
<td>Denmark</td>
<td>.0048</td>
<td>.0923</td>
<td>.3920</td>
<td>.4482</td>
</tr>
<tr>
<td>10</td>
<td>France</td>
<td>.2336</td>
<td>.2872</td>
<td>.5945</td>
<td>.5814</td>
</tr>
<tr>
<td>40</td>
<td>Italy</td>
<td>.0695</td>
<td>.0290</td>
<td>.3896</td>
<td>.3807</td>
</tr>
<tr>
<td>10</td>
<td>Norway</td>
<td>1.0000</td>
<td>.1444</td>
<td>.8460</td>
<td>.9219</td>
</tr>
<tr>
<td>40</td>
<td>Sweden</td>
<td>.0274</td>
<td>.3981</td>
<td>.4773</td>
<td>.4628</td>
</tr>
<tr>
<td>10</td>
<td>United Kingdom</td>
<td>.9148</td>
<td>.9350</td>
<td>.9785</td>
<td>.9891</td>
</tr>
<tr>
<td>40</td>
<td>United States</td>
<td>.7921</td>
<td>.7935</td>
<td>.7758</td>
<td>.7935</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.0372</td>
<td>.0544</td>
<td>.0379</td>
<td>.0465</td>
</tr>
</tbody>
</table>

Detrending, the asymptotic distribution of the statistics is unaffected by these breaks. Perron proceeded to point out that the estimated VR's for U.S. GNP, both before and after his detrending procedure, are small. From this he concluded that output is stationary in levels (around a breaking trend).

Our examination of the two regime-switching models can be viewed as an extension of Perron's work to international data but with several advantages. First, Perron's method of detrending requires that he determine the break dates, which necessarily involves an arbitrary judgment. Furthermore, as the results in Table 1 indicate, Perron's two types of trend breaks appear inconsistent with our filtered estimates, which are based on maximum likelihood estimation. Finally, his statistical inference is based on Priestley's standard errors computed using sample VR's. But, as we have shown, this procedure incorrectly favors models with less persistence.

Although our results are clearly consistent with Perron's contention that the data may be stationary around a breaking trend, we conclude that this statement is too narrow. The Lam model, which is generally consistent with the international data, also implies that output is stationary in levels around a breaking trend. We have also shown that the Hamilton model, the unit-root analog of the Lam model, fits the data nearly as well. Furthermore, the models without trend breaks—the trend model and the ARIMA model—are not rejected by the data. Our conclusion is that we cannot rule out any of these models and Perron's claims for his view are too strong.

We can also compare our results to those of Kormendi and Meguire (1990), who, using data from a wide variety of countries, concluded that there is overwhelming evidence in favor of the existence of a unit root in log output. They found that an AR(3) model in first differences, our ARIMA model, is "almost always more 'probable'" (p. 91) than an AR(3) model with a trend, a model similar to our trend model. Their conclusion is based on comparing the p values of VR(k), for k = 20 and k = 50, computed from an empirical distribution. The more probable model is the one with a p value closest to .5 for the given horizon.

A similar comparison of the models we consider can be carried out by using the experiments reported in Section 4. We would reach a conclusion analogous to theirs: The "best" models in this sense are the ARIMA and Lam models. But such a statement could be misleading because each model implies an entire autocorrelation structure, not just values of the VR statistic at several isolated horizons. Our joint test takes into account the entire sequence of VR statistics, considering all of the autocorrelations together. The results in Table 4 show that, although the ARIMA model is still the "best" model in the sense of Kormendi and Meguire, the evidence is far from overwhelming.

Finally, our results can be compared with those of Christiano and Eichenbaum (1990) on discriminating unit-root models from trend-stationary models. Using a method similar to the one in Section 4, they concluded that we cannot tell whether there is a unit root in U.S. quarterly data. Our conclusion, that we cannot accurately discriminate among models with very different long-run properties, extends theirs to the nonlinear models of Hamilton and Lam, as well as to annual data on nine countries. In addition, we note that all of the models we consider have similar implications for output persistence over horizons of 5 to 10 years. If, as is often the case, one is interested in changes in forecast of output levels at business-cycle frequencies, then our results suggest that it makes little difference which model one chooses.

6. CONCLUSION

This article examines two aspects of statistical inferences based on VR statistics, (1) the accuracy of asymptotic approximations when applied to small samples and (2) the size distortion arising from searching over many horizons in testing a model. Our Monte Carlo experiments show that asymptotic tests are often misleading and that there is substantial size distortion in searching over many horizons to decide whether to reject a model. We propose a joint test that incorporates the correlations between VR statistics of various
horizons and shows how to draw valid inferences using the Monte Carlo method.

Our empirical application to the real GNP data of nine countries confirms the results of our Monte Carlo experiments, which indicate substantial size distortions. We find that the evidence against the trend model is far less dramatic than is suggested from asymptotic inference based on standard errors computed using population VR’s. Our joint test provides even less evidence against this model. Our main empirical conclusion is that correct inference based on VR statistics does not allow us to discriminate among unit-root and trend-stationary models. This is true, contrary to the claim of Perron (in press), even if we account for trend breaks.

Our conclusions on size distortion serve as a warning to researchers using VR statistics to conduct inference. They imply that VR tests should be conducted using Monte Carlo methods. Our findings also imply that when VR’s of more than one horizon are tested researchers need to be specific about the rule used to reject a model and evaluate the correct critical values accordingly. In these situations, we recommend using our joint test. It is intuitive, and its critical values are straightforward to compute.

ACKNOWLEDGMENTS

We thank Timothy Cogley, Mario Crucini, Francis Diebold, Andrew Lo, Charles Nelson, Alan Viard, two anonymous referees, an associate editor, and the participants of a seminar at Ohio State for comments on an earlier draft. Cecchetti acknowledges financial support from the National Science Foundation.

APPENDIX: THE EQUIVALENCE OF THE JOINT VR TEST AND THE JOINT TEST OF AUTOCORRELATIONS

In this appendix, we show that the empirical distribution of the statistic \( S(\tau) \) is numerically identical to the quadratic sum of the deviations of the first \((\tau - 1)\) autocorrelations from their population values, weighted by their covariance matrix. To see this, first define \( \mathbf{P}(j) \) as a \( j \) vector of the first \( j \) autocorrelations of the \( i \)th draw of a time series process and \( \mathbf{P} \) as a \( k \) vector of ones. We can write the vector of VR statistics \( \mathbf{VR}(\tau) \) computed from the \( i \)th draw as

\[
\mathbf{VR}(\tau) = \mathbf{WP}(\tau - 1) + \mathbf{I}^{-1},
\]

where \( W \) is a triangular, full-rank matrix implied by the sample analog of the VR statistic expressed as the weighted sum of autocovariances between \( \Delta y_t \) and \( \Delta y_{t-j} \). It follows that the sample mean of the above expression is just

\[
\frac{1}{N}\left\{ \mathbf{WP}(\tau - 1) + \mathbf{I}^{-1} \right\}
\]

\[
+ \mathbf{I}^{-1} \ldots \mathbf{WP}(\tau - 1) + \mathbf{I}^{-1} \right\}
\]

\[
= \mathbf{WP}(\tau - 1) + \mathbf{I}^{-1},
\]

Next, write the sample covariance matrix, computed over \( N \) draws as

\[
\frac{1}{N} \left\{ \mathbf{WP}(\tau - 1) - \mathbf{WP}(\tau - 1) - \mathbf{WP} \right\}
\]

\[
\ldots \mathbf{WP}(\tau - 1) - \mathbf{WP} \right\}
\]

\[
\left\{ \mathbf{WP}(\tau - 1) - \mathbf{WP} \right\}
\]

\[
\ldots \mathbf{WP}(\tau - 1) - \mathbf{WP} \right\}
\]

\[
= \mathbf{WP}(\tau - 1) + \mathbf{I}^{-1},
\]

where \( \mathbf{P}(\tau - 1) \) is a vector of sample means of \( \mathbf{P}(\tau - 1) \) over \( N \) draws.

\[
\frac{1}{N} \left\{ \mathbf{WP}(\tau - 1) - \mathbf{WP}(\tau - 1) - \mathbf{WP} \right\}
\]

\[
\ldots \mathbf{WP}(\tau - 1) - \mathbf{WP} \right\}
\]

\[
\left\{ \mathbf{WP}(\tau - 1) - \mathbf{WP} \right\}
\]

\[
\ldots \mathbf{WP}(\tau - 1) - \mathbf{WP} \right\}
\]

\[
= \mathbf{WP}(\tau - 1) + \mathbf{I}^{-1},
\]

where \( \mathbf{P}(\tau - 1) \) is a vector of sample means of \( \mathbf{P}(\tau - 1) \) over \( N \) draws.

Next, write the sample covariance matrix, computed over \( N \) draws as

\[
\frac{1}{N} \left\{ \mathbf{WP}(\tau - 1) - \mathbf{WP}(\tau - 1) - \mathbf{WP} \right\}
\]

\[
\ldots \mathbf{WP}(\tau - 1) - \mathbf{WP} \right\}
\]

\[
\left\{ \mathbf{WP}(\tau - 1) - \mathbf{WP} \right\}
\]

\[
\ldots \mathbf{WP}(\tau - 1) - \mathbf{WP} \right\}
\]

\[
= \mathbf{WP}(\tau - 1) + \mathbf{I}^{-1},
\]

where \( \mathbf{P}(\tau - 1) \) is a vector of sample means of \( \mathbf{P}(\tau - 1) \) over \( N \) draws.

Using these two results, we can rewrite \( S(\tau) \) for the \( i \)th draw as

\[
S(\tau) = \left[ \mathbf{WP}(\tau - 1) - \mathbf{WP}(\tau - 1) \right]
\]

\[
\times \left[ \mathbf{WP}(\tau - 1) - \mathbf{WP}(\tau - 1) \right]
\]

\[
= \left[ \mathbf{WP}(\tau - 1) - \mathbf{WP}(\tau - 1) \right]
\]

\[
\times \mathbf{I}^{-1} \left[ \mathbf{WP}(\tau - 1) - \mathbf{WP}(\tau - 1) \right],
\]

which is the quadratic sum of the deviations of the first \((\tau - 1)\) sample autocorrelations from their population values, weighted by the inverse of their covariance matrix.

[Received October 1991. Revised September 1993.]

REFERENCES


Christiano, L. J., and Eichenbaum, M. (1990), "Unit Roots in GNP: Do We Know, and Do We Care?" Carnegie-Rochester Conference Series on Public Policy, 32, 7–62.


Lam, P. (1990), "The Hamilton Model With a General Autoregressive Component: Estimation and Comparison With Other Models of Economic
You have printed the following article:

**Variance-Ratio Tests: Small-Sample Properties with an Application to International Output Data**
Stephen G. Cecchetti; Pok-sang Lam
Stable URL:
http://links.jstor.org/sici?sici=0735-0015%28199404%2912%3A2%3C177%3AVTSPWA%3E2.0.CO%3B2-9

This article references the following linked citations. If you are trying to access articles from an off-campus location, you may be required to first logon via your library web site to access JSTOR. Please visit your library's website or contact a librarian to learn about options for remote access to JSTOR.

References

**The Relationship Between Power and Level for Generic Unit Root Tests in Finite Samples**
Stephen R. Blough
Stable URL:
http://links.jstor.org/sici?sici=0883-7252%28199207%2F09%297%3A3%3C295%3ATRBPAL%3E2.0.CO%3B2-9

**Permanent and Transitory Components in Macroeconomic Fluctuations**
John Y. Campbell; N. Gregory Mankiw
Stable URL:
http://links.jstor.org/sici?sici=0002-8282%28198705%2977%3A2%3C111%3APATCIM%3E2.0.CO%3B2-U

**Are Output Fluctuations Transitory?**
John Y. Campbell; N. Gregory Mankiw
Stable URL:
http://links.jstor.org/sici?sici=0033-5533%28198711%29102%3A4%3C857%3AAOFT%3E2.0.CO%3B2-F
Mean Reversion in Equilibrium Asset Prices
Stephen G. Cecchetti; Pok-Sang Lam; Nelson C. Mark
Stable URL:
http://links.jstor.org/sici? sici=0002-8282%28199006%2980%3A3%3C398%3AMRIEAP%3E2.0.CO%3B2-A

How Big Is the Random Walk in GNP?
John H. Cochrane
Stable URL:
http://links.jstor.org/sici? sici=0022-3808%28198810%2996%3A5%3C893%3AHBITRW%3E2.0.CO%3B2-P

International Evidence on the Size of the Random Walk in Output
Timothy Cogley
Stable URL:
http://links.jstor.org/sici? sici=0022-3808%28199006%2998%3A3%3C501%3AIEOTSO%3E2.0.CO%3B2-X

When are Variance Ratio Tests for Serial Dependence Optimal?
Jon Faust
Stable URL:
http://links.jstor.org/sici? sici=0012-9682%28199209%2960%3A5%3C1215%3AWAVRTF%3E2.0.CO%3B2-J

A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle
James D. Hamilton
Stable URL:
http://links.jstor.org/sici? sici=0012-9682%28198903%2957%3A2%3C357%3AANATTE%3E2.0.CO%3B2-2

A Multicountry Characterization of the Nonstationarity of Aggregate Output
Roger C. Kormendi; Philip Meguire
Stable URL:
http://links.jstor.org/sici? sici=0022-2879%28199002%2922%3A1%3C77%3AAMCOTN%3E2.0.CO%3B2-S
Stock Market Prices do not Follow Random Walks: Evidence from a Simple Specification Test
Andrew W. Lo; A. Craig MacKinlay
Stable URL:
http://links.jstor.org/sici?sici=0893-9454%28198821%291%3A1%3C41%3ASMPDNF%3E2.0.CO%3B2-U

The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis
Pierre Perron
Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28198911%2957%3A6%3C1361%3ATGCTOP%3E2.0.CO%3B2-W

The Relative Importance of Permanent and Transitory Components: Identification and Some Theoretical Bounds
Danny Quah
Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28199201%2960%3A1%3C107%3ATRIOPA%3E2.0.CO%3B2-5