Comparing Predictive Accuracy

Francis X. Diebold
Department of Economics, University of Pennsylvania, Philadelphia, PA 19104-6297, and
National Bureau of Economic Research, Cambridge, MA   02138

Roberto S. Mariano
Department of Economics, University of Pennsylvania, Philadelphia, PA   19104-6297

We propose and evaluate explicit tests of the null hypothesis of no difference in the accuracy of two competing forecasts. In contrast to previously developed tests, a wide variety of accuracy measures can be used (in particular, the loss function need not be quadratic and need not even be symmetric), and forecast errors can be non-Gaussian, nonzero mean, serially correlated, and contemporaneously correlated. Asymptotic and exact finite-sample tests are proposed, evaluated, and illustrated.

KEY WORDS: Economic loss function; Exchange rates; Forecast evaluation; Forecasting; Nonparametric tests; Sign test.

Prediction is of fundamental importance in all of the sciences, including economics. Forecast accuracy is of obvious importance to users of forecasts because forecasts are used to guide decisions. Forecast accuracy is also of obvious importance to producers of forecasts, whose reputations (and fortunes) rise and fall with forecast accuracy. Comparisons of forecast accuracy are also of importance to economists more generally who are interested in discriminating among competing economic hypotheses (models). Predictive performance and model adequacy are inextricably linked—predictive failure implies model inadequacy.

Given the obvious desirability of a formal statistical procedure for forecast-accuracy comparisons, one is struck by the casual manner in which such comparisons are typically carried out. The literature contains literally thousands of forecast-accuracy comparisons; almost without exception, point estimates of forecast accuracy are examined, with no attempt to assess their sampling uncertainty. On reflection, the reason for the casual approach is clear: Correlation of forecast errors across space and time, as well as several additional complications, makes formal comparison of forecast accuracy difficult. Dhrymes et al. (1972) and Howrey, Klein, and McCarthy (1974), for example, offered pessimistic assessments of the possibilities for formal testing.

In this article we propose widely applicable tests of the null hypothesis of no difference in the accuracy of two competing forecasts. Our approach is similar in spirit to that of Vuong (1989) in the sense that we propose methods for measuring and assessing the significance of divergences between models and data. Our approach, however, is based directly on predictive performance, and we entertain a wide class of accuracy measures that users can tailor to particular decision-making situations. This is important because, as is well known, realistic economic loss functions frequently do not conform to stylized textbook favorites like mean squared prediction error (MSPE). [For example, Leitch and Tanner (1991)

and Chinn and Meese (1991) stressed direction of change, Cumby and Modest (1987) stressed market and country timing, McCulloch and Rossi (1990), and West, Edison, and Cho (1993) stressed utility-based criteria, and Clements and Hendry (1993) proposed a new accuracy measure, the generalized forecast-error second moment.] Moreover, we allow for forecast errors that are potentially non-Gaussian, nonzero mean, serially correlated, and contemporaneously correlated.

We proceed by detailing our test procedures in Section 1. Then, in Section 2, we review the small extant literature to provide necessary background for the finite-sample evaluation of our tests in Section 3. In Section 4 we provide an illustrative application, and in Section 5 we offer conclusions and directions for future research.

1. TESTING EQUALITY OF FORECAST ACCURACY

Consider two forecasts, $\{\hat{y}_{it}\}_{t=1}^{T}$ and $\{\hat{y}_{jt}\}_{t=1}^{T}$, of the time series $\{y_t\}_{t=1}^{T}$. Let the associated forecast errors be $\{e_{it}\}_{t=1}^{T}$ and $\{e_{jt}\}_{t=1}^{T}$. We wish to assess the expected loss associated with each of the forecasts (or its negative, accuracy). Of great importance, and almost always ignored, is the fact that the *economic* loss associated with a forecast may be poorly assessed by the usual *statistical* metrics. That is, forecasts are used to guide decisions, and the loss associated with a forecast error of a particular sign and size is induced directly by the nature of the decision problem at hand. When one considers the variety of decisions undertaken by economic agents guided by forecasts (e.g., risk-hedging decisions, inventory-stocking decisions, policy decisions, advertising-expenditure decisions, public-utility rate-setting decisions, etc.), it is clear that the loss associated with a particular forecast error is in general an asymmetric function of the error and, even if symmetric, certainly need not conform to stylized textbook examples like MSPE.

Thus, we allow the time-t loss associated with a forecast (say i) to be an arbitrary function of the realization and prediction, $g(y_t, \hat{y}_{it})$. In many applications, the loss function will be a direct function of the forecast error; that is, $g(y_t, \hat{y}_{it}) = g(e_{it})$. To economize on notation, we write $g(e_{it})$ from this point on, recognizing that certain loss functions (like direction-of-change) do not collapse to $g(e_{it})$ form, in which case the full $g(y_t, \hat{y}_{it})$ form would be used. The null hypothesis of equal forecast accuracy for two forecasts is $E[g(e_{it})] = E[g(e_{jt})]$, or $E[d_t] = 0$, where $d_t \equiv [g(e_{it}) - g(e_{jt})]$ is the loss differential. Thus, the "equal accuracy" null hypothesis is equivalent to the null hypothesis that the population mean of the loss-differential series is 0.

1.1 An Asymptotic Test

Consider a sample path $\{d_t\}_{t=1}^T$ of a loss-differential series. If the loss-differential series is covariance stationary and short memory, then standard results may be used to deduce the asymptotic distribution of the sample mean loss differential. We have

$$\sqrt{T}(\bar{d} - \mu) \xrightarrow{d} N(0, 2\pi f_d(0)),$$

where

$$\bar{d} = \frac{1}{T} \sum_{t=1}^{T} [g(e_{it}) - g(e_{jt})]$$

is the sample mean loss differential,

$$f_d(0) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma_d(\tau)$$

is the spectral density of the loss differential at frequency 0, $\gamma_d(\tau) = E[(d_t - \mu)(d_{t-\tau} - \mu)]$ is the autocovariance of the loss differential at displacement τ, and μ is the population mean loss differential. The formula for $f_d(0)$ shows that the correction for serial correlation can be substantial, even if the loss differential is only weakly serially correlated, due to cumulation of the autocovariance terms.

Because in large samples the sample mean loss differential $\bar{d}$ is approximately normally distributed with mean μ and variance $2\pi f_d(0)/T$, the obvious large-sample $N(0, 1)$ statistic for testing the null hypothesis of equal forecast accuracy is

$$S_1 = \frac{\bar{d}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{T}}},$$

where $\hat{f}_d(0)$ is a consistent estimate of $f_d(0)$.

Following standard practice, we obtain a consistent estimate of $2\pi f_d(0)$ by taking a weighted sum of the available sample autocovariances,

$$2\pi \hat{f}_d(0) = \sum_{\tau=-(T-1)}^{(T-1)} 1\left(\frac{\tau}{S(T)}\right) \hat{\gamma}_d(\tau),$$

where

$$\hat{\gamma}_d(\tau) = \frac{1}{T} \sum_{t=|\tau|+1}^{T} (d_t - \bar{d})(d_{t-|\tau|} - \bar{d}),$$

$1(\tau/S(T))$ is the lag window, and $S(T)$ is the truncation lag.

To motivate a choice of lag window and truncation lag that we have often found useful in practice, recall the familiar result that optimal k-step-ahead forecast errors are at most $(k - 1)$-dependent. In practical applications, of course, $(k - 1)$-dependence may be violated for a variety of reasons. Nevertheless, it seems reasonable to take $(k - 1)$-dependence as a reasonable benchmark for a k-step-ahead forecast error (and the assumption may be readily assessed empirically). This suggests the attractiveness of the uniform, or rectangular, lag window, defined by

$$1\left(\frac{\tau}{S(T)}\right) = 1 \quad \text{for} \quad \left|\frac{\tau}{S(T)}\right| \leq 1$$
$$= 0 \quad \text{otherwise.}$$

$(k - 1)$-dependence implies that only $(k - 1)$ sample autocovariances need be used in the estimation of $f_d(0)$ because all the others are 0, so $S(T) = (k - 1)$. This is legitimate (i.e., the estimator is consistent) under $(k - 1)$-dependence so long as a uniform window is used because the uniform window assigns unit weight to all included autocovariances.

Because the Dirichlet spectral window associated with the rectangular lag window dips below 0 at certain locations, the resulting estimator of the spectral density function is not guaranteed to be positive semidefinite. The large positive weight near the origin associated with the Dirichlet kernel, however, makes it unlikely to obtain a negative estimate of $f_d(0)$. In applications, in the rare event that a negative estimate arises, we treat it as 0 and automatically reject the null hypothesis of equal forecast accuracy. If it is viewed as particularly important to impose nonnegativity of the estimated spectral density, it may be enforced by using a Bartlett lag window, with corresponding nonnegative Fejer spectral window, as in the work of Newey and West (1987), at the cost of having to increase the truncation lag "appropriately" with sample size. Other lag windows and truncation lag selection procedures are of course possible as well. Andrews (1991), for example, suggested using a quadratic spectral lag window, together with a "plug-in" automatic bandwidth selection procedure.

1.2 Exact Finite-Sample Tests

Sometimes only a few forecast-error observations are available in practice. One approach in such situations is to bootstrap our asymptotic test statistic, as done by Mark (1995). Ashley's (1994) work is also very much in that spirit. Little is known about the first-order asymptotic validity of the bootstrap in this situation, however, let alone higher-order asymptotics or actual finite-sample performance. Therefore, it is useful to have available exact finite-sample tests of predictive accuracy, to complement the asymptotic test presented previously. Two powerful such tests are based on the observed loss differentials (the sign test) or their ranks (Wilcoxon's signed-rank test). [These tests are standard, so our discussion is terse. See, for example, Lehmann (1975) for details.]

1.2.1 The Sign Test. The null hypothesis is a zero-median loss differential: $\text{med}(g(e_{it}) - g(e_{jt})) = 0$. Note that the null of a zero-median loss differential is not the same

as the null of zero difference between median losses; that is, $\text{med}(g(e_{it}) - g(e_{jt})) \neq \text{med}(g(e_{it})) - \text{med}(g(e_{jt}))$. For that reason, the null differs slightly in spirit from that associated with our earlier discussed asymptotic test statistic S_1, but it nevertheless has an intuitive and meaningful interpretation—namely, that $P(g(e_{it}) > g(e_{jt})) = P(g(e_{it}) < g(e_{jt}))$.

If, however, the loss differential is symmetrically distributed, then the null hypothesis of a zero-median loss differential corresponds precisely to the earlier null because in that case the median and mean are equal. Symmetry of the loss differential will obtain, for example, if the distributions of $g(e_{it})$ and $g(e_{jt})$ are the same up to a location shift. Symmetry is ultimately an empirical matter and may be assessed using standard procedures. We have found roughly symmetric loss-differential series to be quite common in practice.

The construction and intuition of a test statistic are straightforward. Assuming that the loss-differential series is iid (and we shall relax that assumption shortly), the number of positive loss-differential observations in a sample of size T has the binomial distribution with parameters T and $\frac{1}{2}$ under the null hypothesis. The test statistic is therefore simply

$$S_2 = \sum_{t=1}^{T} I_+(d_t),$$

where

$$I_+(d_t) = 1 \quad \text{if } d_t > 0$$
$$= 0 \quad \text{otherwise.}$$

Significance may be assessed using a table of the cumulative binomial distribution. In large samples, the studentized version of the sign-test statistic is standard normal:

$$S_{2a} = \frac{S_2 - .5T}{\sqrt{.25T}} \stackrel{a}{\sim} N(0, 1).$$

1.2.2 Wilcoxon's Signed-Rank Test. A related distribution-free procedure that *requires* symmetry of the loss differential (but can be more powerful than the sign test in that case) is Wilcoxon's signed-rank test. We again assume for the moment that the loss-differential series is iid. The test statistic is

$$S_3 = \sum_{t=1}^{T} I_+(d_t) \, \text{rank}(|d_t|),$$

the sum of the ranks of the absolute values of the positive observations. The exact finite-sample critical values of the test statistic are invariant to the distribution of the loss differential—it need be only zero-mean and symmetric—and have been tabulated. Moreover, its studentized version is asymptotically standard normal,

$$S_{3a} = \frac{S_3 - \frac{T(T+1)}{4}}{\sqrt{\frac{T(T+1)(2T+1)}{24}}} \stackrel{a}{\sim} N(0, 1).$$

1.3 Discussion

Here we highlight some of the virtues and limitations of our tests. First, as we have stressed repeatedly, our tests are valid for a very wide class of loss functions. In particular,

the loss function need not be quadratic and need not even be symmetric or continuous.

Second, a variety of realistic features of forecast errors are readily accommodated. The forecast errors can be nonzero-mean, non-Gaussian, and contemporaneously correlated. Allowance for contemporaneous correlation, in particular, is important because the forecasts being compared are forecasts of the *same* economic time series and because the information sets of forecasters are largely overlapping so that forecast errors tend to be strongly contemporaneously correlated.

Moreover, the asymptotic test statistic S_1 can of course handle a *serially* correlated loss differential. This is potentially important because, as discussed earlier, even optimal forecast errors are serially correlated in general. Serial correlation presents more of a problem for the exact finite-sample test statistics S_2 and S_3 and their asymptotic counterparts S_{2a} and S_{3a} because the elements of the set of all possible rearrangements of the sample loss differential series are *not* equally likely when the data are serially correlated, which violates the assumptions on which such randomization tests are based. Nevertheless, serial correlation may be handled via Bonferroni bounds, as suggested in a different context by Campbell and Ghysels (1995). Under the assumption that the forecast errors and hence the loss differential are $(k - 1)$-dependent, each of the following k sets of loss differentials will be free of serial correlation: $\{d_{ij,1}, d_{ij,1+k}, d_{ij,1+2k}, \ldots\}$, $\{d_{ij,2}, d_{ij,2+k}, d_{ij,2+2k}, \ldots\}, \ldots, \{d_{ij,k}, d_{ij,2k}, d_{ij,3k}, \ldots\}$. Thus, a test with size bounded by α can be obtained by performing k tests, each of size α/k, on each of the k loss-differential sequences and rejecting the null hypothesis if the null is rejected for *any* of the k samples. Finally, it is interesting to note that, in multistep forecast comparisons, forecast-error serial correlation may be a "common feature," in the terminology of Engle and Kozicki (1993), because it is induced largely by the fact that the forecast horizon is longer than the interval at which the data are sampled and may therefore not be present in loss *differentials* even if present in the forecast errors themselves. This possibility can of course be checked empirically.

2. EXTANT TESTS

In this section we provide a brief description of three existing tests of forecast accuracy that have appeared in the literature and will be used in our subsequent Monte Carlo comparison.

2.1 The Simple *F* Test: A Naive Benchmark

If (1) loss is quadratic and (2) the forecast errors are (a) zero mean, (b) Gaussian, (c) serially uncorrelated, or (d) contemporaneously uncorrelated, then the null hypothesis of equal forecast accuracy corresponds to equal forecast error variances [by (1) and (2a)], and by (2b)–(2d), the ratio of sample variances has the usual *F* distribution under the null hypothesis. More precisely, the test statistic

$$F = \frac{\frac{e_i' e_i}{T}}{\frac{e_j' e_j}{T}} = \frac{e_i' e_i}{e_j' e_j}$$

is distributed as $F(T, T)$, where the forecast error series have been stacked into the $(T \times 1)$ vectors e_i and e_j.

Test statistic F is of little use in practice, however, because the conditions required to obtain its distribution are too restrictive. Assumption (2d) is particularly unpalatable for reasons discussed earlier. Its violation produces correlation between the numerator and denominator of F, which will not then have the F distribution.

2.2 The Morgan–Granger–Newbold Test

The contemporaneous correlation problem led Granger and Newbold (1977) to apply an orthogonalizing transformation due to Morgan (1939–1940) that enables relaxation of Assumption (2d). Let $x_t = (e_{it} + e_{jt})$ and $z_t = (e_{it} - e_{jt})$, and let $x = (e_i + e_j)$ and $z = (e_i - e_j)$. Then, under the maintained Assumptions (1) and (2a)–(2c), the null hypothesis of equal forecast accuracy is equivalent to zero correlation between x and z (i.e., $\rho_{xz} = 0$) and the test statistic

$$ MGN = \frac{\hat{\rho}_{xz}}{\sqrt{\frac{1-\hat{\rho}_{xz}^2}{T-1}}} $$

is distributed as Student's t with $T - 1$ df, where

$$ \hat{\rho}_{xz} = \frac{x'z}{\sqrt{(x'x)(z'z)}} $$

(e.g., see Hogg and Craig 1978, pp. 300–303).

Let us now consider relaxing the Assumptions (1) and (2a)–(2c) underlying the Morgan–Granger–Newbold (MGN) test. It is clear that the entire framework depends crucially on the assumption of quadratic loss (1), which cannot be relaxed. The remaining assumptions, however, can be weakened in varying degrees; we shall consider them in turn.

First, it is not difficult to relax the unbiasedness Assumption (2a), while maintaining Assumptions (1), (2b), and (2c). Second, the normality Assumption (2b) may be relaxed, while maintaining (1), (2a), and (2c), at the cost of substantial tedium involved with accounting for the higher-order moments that then enter the distribution of the sample correlation coefficient (e.g., see Kendall and Stuart 1979, chap. 26). Finally, the no-serial-correlation Assumption (2c) may be relaxed in addition to the no-contemporaneous-correlation Assumption (2d) while maintaining (1), (2a), and (2b), as discussed in Subsection 2.3.

2.3 The Meese–Rogoff Test

Under Assumptions (1), (2a), and (2b), Meese and Rogoff (1988) showed that

$$ \sqrt{T}\hat{\gamma}_{xz} \xrightarrow{d} N(0, \Sigma), $$

where $\hat{\gamma}_{xz} = x'z/T$, $\Sigma = \sum_{\tau=-\infty}^{\infty} [\gamma_{xx}(\tau)\gamma_{zz}(\tau) + \gamma_{xz}(\tau)\gamma_{zx}(\tau)]$, $\gamma_{xz}(\tau) = \text{cov}(x_t, z_{t-\tau})$, $\gamma_{zx}(\tau) = \text{cov}(z_t, x_{t-\tau})$, $\gamma_{xx}(\tau) = \text{cov}(x_t, x_{t-\tau})$, and $\gamma_{zz}(\tau) = \text{cov}(z_t, z_{t-\tau})$. This is a well-known result (e.g., Priestley 1981, pp. 692–693) for the distribution of the sample cross-covariance function, $\text{cov}(\hat{\gamma}_{xz}(s), \hat{\gamma}_{xz}(u))$, specialized to a displacement of 0.

A consistent estimator of Σ is

$$ \hat{\Sigma} = \sum_{\tau=-S(T)}^{S(T)} \left[1 - \frac{|\tau|}{T} \right] [\hat{\gamma}_{xx}(\tau)\hat{\gamma}_{zz}(\tau) + \hat{\gamma}_{xz}(\tau)\hat{\gamma}_{zx}(\tau)], $$

where

$$ \hat{\gamma}_{xz}(\tau) = \frac{1}{T}\sum_{t=\tau+1}^{T} x_t z_{t-\tau}, \qquad \tau \geq 0 $$
$$ = \hat{\gamma}_{zx}(-\tau) \quad \text{otherwise,} $$
$$ \hat{\gamma}_{zx}(\tau) = \frac{1}{T}\sum_{t=\tau+1}^{T} z_t x_{t-\tau}, \qquad \tau \geq 0 $$
$$ = \hat{\gamma}_{xz}(-\tau) \quad \text{otherwise,} $$
$$ \hat{\gamma}_{xx}(\tau) = \frac{1}{T}\sum_{t=\tau+1}^{T} x_t x_{t-\tau}, $$
$$ \hat{\gamma}_{zz}(\tau) = \frac{1}{T}\sum_{t=\tau+1}^{T} z_t z_{t-\tau}, $$

and the truncation lag $S(T)$ grows with the sample size but at a slower rate. Alternatively, following Diebold and Rudebusch (1991), one may use the closely related covariance matrix estimator,

$$ \hat{\Sigma}^* = \sum_{\tau=-S(T)}^{S(T)} [\hat{\gamma}_{xx}(\tau)\hat{\gamma}_{zz}(\tau) + \hat{\gamma}_{xz}(\tau)\hat{\gamma}_{zx}(\tau)]. $$

Either way, the test statistic is

$$ MR = \frac{\hat{\gamma}_{xz}}{\sqrt{\hat{\Sigma}/T}}. $$

Under the null hypothesis and the maintained Assumptions (1), (2a), and (2b), MR (Meese–Rogoff) is asymptotically distributed as standard normal.

It is easy to show that, if the null hypothesis and Assumptions (1), (2a), (2b), and (2c) are satisfied, then all terms in Σ are 0 except $\gamma_{xx}(0)$ and $\gamma_{zz}(0)$ so that MR coincides asymptotically with MGN. It is interesting to note also that reformulation of the test in terms of correlation rather than covariance would have enabled Meese and Rogoff to dispense with the normality assumption because the sample autocorrelations are asymptotically normal even for non-Gaussian time series (e.g., Brockwell and Davis 1992, pp. 221–222).

2.4 Additional Extensions

In Subsection 2.3, we considered relaxation of Assumptions (2a)–(2c), one at a time, while consistently maintaining Assumption (1) and consistently relaxing Assumption (2d). Simultaneous relaxation of multiple assumptions is possible within the MGN orthogonalizing transformation framework but much more tedious. The distribution theory required for joint relaxation of (2b) and (2c), for example, is complicated by the presence of fourth-order cumulants in the distribution of the the sample autocovariances, as shown, for example, by Hannan (1970, p. 209) and Mizrach (1991). More importantly, however, *any* procedure based on the MGN orthogonalizing transformation is inextricably wed to the assumption of quadratic loss.

the observed loss differentials (the sign test) or their ranks

3. MONTE CARLO ANALYSIS

3.1 Experimental Design

We evaluate the finite-sample size of test statistics F, MGN, MR, S_1, S_2, S_{2a}, S_3, and S_{3a} under the null hypothesis and various of the maintained assumptions. The design includes a variety of specifications of forecast-error contemporaneous correlation, forecast-error serial correlation, and forecast-error distributions. To maintain applicability of all test statistics for comparison purposes, we use quadratic loss; that is, the null hypothesis is an equality of MSPE's. We emphasize again, however, that an important advantage of test statistics S_1, S_2, S_{2a}, S_3, and S_{3a} in substantive economic applications—and one not shared by the others—is their direct applicability to analyses with nonquadratic loss functions.

Consider first the case of Gaussian forecast errors. We draw realizations of the bivariate forecast-error process, $\{e_{it}, e_{jt}\}_{t=1}^T$, with varying degrees of contemporaneous and serial correlation in the generated forecast errors. This is achieved in two steps. First, we build in the desired degree of contemporaneous correlation by drawing a (2×1) forecast error innovation vector u_t from a bivariate standard normal distribution, $u_t \sim N(0_2, I_2)$, and then premultiplying by the Choleski factor of the desired contemporaneous innovation correlation matrix. Let the desired correlation matrix be

$$\mathbf{R} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \qquad \rho \in [0, 1).$$

Then the Choleski factor is

$$\mathbf{P} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1 - \rho^2} \end{bmatrix}.$$

Thus, the transformed (2×1) vector $\nu_t = Pu_t \sim N(0_2, \mathbf{R})$. This operation is repeated T times, yielding $\{\nu_{it}, \nu_{jt}\}_{t=1}^T$.

Second, (moving average) MA(1) serial correlation (with parameter θ) is introduced by taking

$$\begin{bmatrix} e_{it} \\ e_{jt} \end{bmatrix} = \begin{bmatrix} \frac{1+\theta L}{\sqrt{1+\theta^2}} & 0 \\ 0 & \frac{1+\theta L}{\sqrt{1+\theta^2}} \end{bmatrix} \begin{bmatrix} \nu_{it} \\ \nu_{jt} \end{bmatrix}, \qquad t = 1, \ldots, T.$$

We use $\nu_0 = 0$. Multiplication by $(1 + \theta^2)^{-1/2}$ is done to keep the unconditional variance normalized to 1.

We consider sample sizes of $T = 8, 16, 32, 64, 128, 256$, and 512, contemporaneous correlation parameters of $\rho = 0$, .5, and .9, and MA parameters of $\theta = 0, .5, .9$. Simple calculations reveal that ρ is not only the correlation between ν_i and ν_j, but also the correlation between the forecast errors e_i and e_j so that varying the correlation of ν_i and ν_j through $[0, .9]$ effectively varies the correlation of the observed forecast errors through the same range.

We also consider non-Gaussian forecast errors. The design is the same as for the Gaussian case described previously but driven by fat-tailed variates $(u_{it}^*, u_{jt}^*)'$ [rather than $(u_{it}, u_{jt})'$], which are independent standardized t random variables with 6 df. The variance of a $t(6)$ random variable is 3/2. Thus,

standardization amounts to dividing the $t(6)$ random variable by $\sqrt{3/2}$.

Throughout, we perform tests at the $\alpha = .1$ level. When using the exact sign and signed-rank tests, restriction of nominal size to precisely 10% is impossible (without introducing randomization), so we use the obtainable exact size closest to 10%, as specified in the tables. We perform at least 5,000 Monte Carlo replications. The truncation lag is set at 1, reflecting the fact that the experiment is designed to mimic the comparison of two-step-ahead forecast errors, with associated MA(1) structure.

3.2 Results

Results appear in Tables 1–6, which show the empirical size of the various test statistics in cases of Gaussian and non-Gaussian forecast errors as the degree of contemporaneous correlation, the degree of serial correlation, and sample size are varied.

Let us first discuss the case of Gaussian forecast errors. The results may be summarized as follows:

1. F is correctly sized in the absence of both contemporaneous and serial correlation but is missized in the presence of either contemporaneous or serial correlation. Serial correlation pushes empirical size above nominal size, but contemporaneous correlation pushes empirical size *drastically* below nominal size. In combination, and particularly for large ρ and θ, contemporaneous correlation dominates and F is undersized.

2. MGN is designed to remain unaffected by contemporaneous correlation and therefore remains correctly sized so long as $\theta = 0$. Serial correlation, however, pushes empirical size above nominal size.

3. As expected, MR is robust to contemporaneous and serial correlation in large samples, but it is oversized in small samples in the presence of serial correlation. The asymptotic distribution obtains rather quickly, however, resulting in approximately correct size for $T > 64$.

4. The behavior of S_1 is similar to that of MR. S_1 is robust to contemporaneous and serial correlation in large samples, but it is oversized in small samples, with nominal and empirical size converging a bit more slowly than for MR.

5. The Bonferroni bounds associated with S_2 and S_3 work well, with nominal and empirical size in close agreement throughout. Moreover, the asymptotics on which S_{2a} and S_{3a} depend obtain quickly.

Now consider the case of non-Gaussian forecast errors. The striking and readily apparent result is that F, MGN, and MR are drastically missized in large as well as small samples. S_1, S_{2a}, and S_{3a}, on the other hand, maintain approximately correct size for all but the very small sample sizes. In those cases, S_2 and S_3 continue to perform well. The results are well summarized by Figure 1, p. 261, which charts the dependence of F, MGN, MR, and S_1 on T for the non-Gaussian case with $\rho = \theta = .5$.

Table 1.   Empirical Size Under Quadratic Loss, Test Statistic F

		Gaussian			Fat-tailed		
T	ρ	θ = .0	θ = .5	θ = .9	θ = .0	θ = .5	θ = .9
8	.0	9.85	12.14	14.10	14.28	15.76	17.21
8	.5	7.02	9.49	11.42	9.61	11.64	13.02
8	.9	.58	1.26	1.86	.57	1.13	1.79
16	.0	9.83	12.97	14.85	16.47	18.59	19.78
16	.5	7.30	10.11	11.89	11.14	13.55	14.94
16	.9	.47	.99	1.55	.34	.70	1.13
32	.0	9.88	12.68	14.34	18.06	19.55	20.35
32	.5	6.98	9.50	11.22	21.30	21.00	21.37
32	.9	.23	.55	1.00	.01	.07	.23
64	.0	9.71	13.05	14.62	29.84	29.72	29.96
64	.5	6.48	9.25	10.62	23.48	23.93	24.15
64	.9	.16	.47	.79	.02	.12	.29
128	.0	10.30	13.41	14.99	30.34	30.95	31.26
128	.5	7.01	10.13	11.64	24.89	25.01	25.16
128	.9	.16	.50	.74	.11	.44	.73
256	.0	10.01	13.05	14.65	31.07	31.12	31.24
256	.5	7.37	10.31	11.78	25.48	25.45	25.70
256	.9	.19	.51	.80	.51	1.13	1.44
512	.0	10.22	13.51	15.25	31.45	32.38	32.60
512	.5	7.53	10.16	11.49	26.35	26.92	16.95
512	.9	.18	.50	.85	.81	1.58	2.06

NOTE: T is sample size, ρ is the contemporaneous correlation between the innovations underlying the forecast errors, and θ is the coefficient of the MA(1) forecast error. All tests are at the 10% level. 10,000 Monte Carlo replications are performed.

Table 2.   Empirical Size Under Quadratic Loss, Test Statistic MGN

		Gaussian			Fat-tailed		
T	ρ	θ = .0	θ = .5	θ = .9	θ = .0	θ = .5	θ = .9
8	.0	10.19	14.14	17.94	18.10	21.89	25.65
8	.5	9.96	14.66	18.61	16.00	20.51	24.19
8	.9	9.75	14.53	18.67	11.76	16.31	20.00
16	.0	10.07	14.34	17.54	20.33	24.54	27.08
16	.5	9.56	14.37	17.95	37.15	36.18	25.66
16	.9	10.02	14.70	18.20	12.01	16.76	19.81
32	.0	9.89	15.04	18.00	22.94	26.32	28.72
32	.5	10.08	15.11	17.95	20.23	23.76	26.20
32	.9	9.59	15.32	18.25	12.75	17.78	20.54
64	.0	10.09	15.37	17.99	24.56	28.15	30.00
64	.5	9.95	15.18	18.15	21.10	25.18	27.28
64	.9	10.26	15.67	18.49	12.98	18.09	20.53
128	.0	9.96	15.09	17.59	26.47	29.50	30.94
128	.5	10.23	15.07	17.48	23.62	26.82	28.51
128	.9	10.11	15.05	18.05	14.34	18.89	21.56
256	.0	10.28	15.62	18.37	27.39	30.74	32.46
256	.5	10.60	16.02	18.44	23.81	28.38	30.31
256	.9	10.11	15.48	17.91	14.15	19.43	22.03
512	.0	10.12	15.34	17.68	27.64	30.55	32.14
512	.5	10.05	14.96	17.66	24.10	27.40	29.28
512	.9	9.90	15.09	17.53	14.78	19.16	21.49

NOTE: T is sample size, ρ is the contemporaneous correlation between the innovations underlying the forecast errors, and θ is the coefficient of the MA(1) forecast error. All tests are at the 10% level. 10,000 Monte Carlo replications are performed.

Table 3. Empirical Size Under Quadratic Loss, Test Statistic MR

T	ρ	Gaussian			Fat-tailed		
		θ = .0	θ = .5	θ = .9	θ = .0	θ = .5	θ = .9
8	.0	9.67	19.33	22.45	16.16	25.26	27.62
8	.5	9.50	19.00	22.07	14.81	24.50	26.99
8	.9	9.66	19.51	22.85	11.23	21.28	24.14
16	.0	9.62	13.92	14.72	19.94	22.56	23.06
16	.5	10.02	13.88	14.96	17.70	21.04	21.26
16	.9	10.04	13.82	14.94	11.76	15.68	16.70
32	.0	9.96	10.98	11.12	22.78	22.86	21.72
32	.5	9.68	11.46	11.66	19.78	20.32	20.14
32	.9	9.86	11.62	11.96	12.42	13.54	13.46
64	.0	10.32	11.02	11.04	24.50	22.60	21.58
64	.5	9.84	10.56	10.64	21.44	19.48	18.84
64	.9	9.58	10.58	10.34	13.38	13.38	13.20
128	.0	9.78	10.54	10.44	25.86	22.90	21.54
128	.5	10.02	11.04	11.18	22.76	20.26	19.44
128	.9	10.76	11.28	11.38	13.44	13.52	12.92
256	.0	10.04	9.90	9.58	27.16	23.74	22.70
256	.5	10.32	9.92	9.82	24.00	20.50	19.18
256	.9	9.92	10.16	10.34	13.38	12.70	12.24
512	.0	9.94	10.48	10.56	26.92	23.40	21.78
512	.5	9.52	10.56	10.48	23.56	20.52	19.36
512	.9	9.80	9.82	9.88	13.96	12.98	12.74

NOTE: T is sample size, ρ is the contemporaneous correlation between the innovations underlying the forecast errors, and θ is the coefficient of the MA(1) forecast error. All tests are at the 10% level. At least 5,000 Monte Carlo replications are performed.

Table 4. Empirical Size Under Quadratic Loss, Test Statistic S_1

T	ρ	Gaussian			Fat-tailed		
		θ = .0	θ = .5	θ = .9	θ = .0	θ = .5	θ = .9
8	.0	31.39	31.10	31.03	31.62	29.51	29.07
8	.5	31.37	30.39	29.93	31.21	29.71	29.36
8	.9	31.08	30.19	30.18	31.18	30.12	29.75
16	.0	20.39	19.11	18.94	19.26	18.50	18.32
16	.5	20.43	19.52	18.86	19.57	17.67	17.63
16	.9	20.90	19.55	19.59	20.15	18.38	18.16
32	.0	12.42	12.28	12.18	11.30	11.64	11.56
32	.5	13.32	13.22	12.94	11.54	10.66	10.84
32	.9	12.60	13.38	13.22	11.16	11.22	11.50
64	.0	12.47	12.11	11.94	12.44	11.62	11.36
64	.5	12.76	12.49	12.35	12.10	12.26	12.10
64	.9	12.21	12.23	12.03	13.00	12.36	12.16
128	.0	11.72	11.94	12.04	11.48	10.72	10.28
128	.5	11.44	11.72	11.60	10.84	10.96	10.96
128	.9	11.76	11.26	11.34	11.50	10.66	10.86
256	.0	11.11	10.65	10.66	12.06	11.67	11.79
256	.5	10.90	10.39	10.48	12.16	11.46	11.60
256	.9	10.69	10.79	10.75	11.51	11.59	11.16
512	.0	11.15	10.67	10.63	10.06	9.46	9.62
512	.5	10.90	10.39	10.49	9.94	9.66	9.76
512	.9	10.31	10.09	10.05	10.12	10.12	10.06

NOTE: T is sample size, ρ is the contemporaneous correlation between the innovations underlying the forecast errors, and θ is the coefficient of the MA(1) forecast error. All tests are at the 10% level. At least 5,000 Monte Carlo replications are performed.

Table 5.   Empirical Size Under Quadratic Loss, Test Statistics S_2 and S_{2a}

T	ρ	Gaussian			Fat-tailed		
		$\theta = .0$	$\theta = .5$	$\theta = .9$	$\theta = .0$	$\theta = .5$	$\theta = .9$
		S_2, nominal size = 25%					
8	.0	22.24	22.48	22.38	23.94	23.46	23.34
8	.5	22.14	23.46	22.16	23.08	24.80	23.06
8	.9	22.24	23.02	22.66	22.92	23.26	22.86
		S_2, nominal size = 14.08%					
16	.0	13.46	13.26	13.14	13.62	13.06	13.76
16	.5	14.22	13.46	12.92	13.70	13.24	13.62
16	.9	13.08	13.84	13.28	12.86	13.06	13.20
		S_2, nominal size = 15.36%					
32	.0	14.36	14.52	14.28	14.54	14.32	14.30
32	.5	14.36	14.06	13.94	15.08	14.36	15.02
32	.9	14.68	14.62	13.46	14.94	14.76	14.52
		S_{2a}, nominal size = 10%					
64	.0	9.72	9.92	9.42	9.68	10.36	10.44
64	.5	9.66	10.34	9.68	9.52	10.06	10.00
64	.9	10.84	9.46	10.34	9.40	8.98	10.02
		S_{2a}, nominal size = 10%					
128	.0	11.62	11.62	11.84	12.22	12.20	11.42
128	.5	11.66	11.62	11.90	12.06	11.94	11.44
128	.9	11.22	11.72	11.28	12.06	10.76	11.40

NOTE: T is sample size, ρ is the contemporaneous correlation between the innovations underlying the forecast errors, and θ is the coefficient of the MA(1) forecast error. At least 5,000 Monte Carlo replications are performed.

Table 6.   Empirical Size Under Quadratic Loss, Test Statistics S_3 and S_{3a}

T	ρ	Gaussian			Fat-tailed		
		$\theta = .0$	$\theta = .5$	$\theta = .9$	$\theta = .0$	$\theta = .5$	$\theta = .9$
		S_3, nominal size = 25%					
8	.0	22.50	22.92	22.90	23.26	23.34	21.96
8	.5	22.98	22.26	23.06	23.42	23.86	22.88
8	.9	23.16	22.36	24.24	24.26	23.32	23.34
		S_3, nominal size = 10.92%					
16	.0	10.62	10.06	10.40	10.16	10.42	9.84
16	.5	10.38	10.92	10.32	10.54	10.94	10.34
16	.9	10.64	10.18	9.62	10.58	10.96	10.64
		S_3, nominal size = 10.12%					
32	.0	10.72	10.28	9.30	9.90	10.00	9.98
32	.5	10.56	10.00	10.02	10.40	10.64	10.30
32	.9	10.92	10.44	10.30	10.46	9.96	10.70
		S_{3a}, nominal size = 10%					
64	.0	9.38	9.54	9.16	9.64	9.24	8.84
64	.5	9.80	10.02	9.66	9.58	8.82	8.78
64	.9	9.90	9.24	9.68	9.92	9.78	10.00
		S_{3a}, nominal size = 10%					
128	.0	9.94	9.70	9.12	9.82	9.04	8.46
128	.5	9.52	10.00	9.32	10.08	9.24	9.20
128	.9	9.46	9.64	9.42	9.28	9.22	9.26

NOTE: T is sample size, ρ is the contemporaneous correlation between the innovations underlying the forecast errors, and θ is the coefficient of the MA(1) forecast error. At least 5,000 Monte Carlo replications are performed.
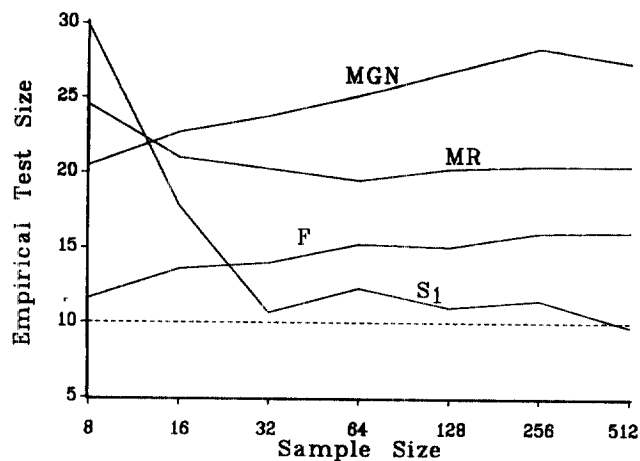
Figure 1.   Empirical Size, Four Test Statistics: Fat-Tailed Case; Theta = Rho = .5.

Figure 3.   Loss Differential (forward—random walk).

4. AN EMPIRICAL EXAMPLE

We shall illustrate the practical use of the tests with an application to exchange-rate forecasting. The series to be forecast, measured monthly, is the three-month change in the nominal dollar/Dutch guilder end-of-month spot exchange rate (in U.S. cents, noon, New York interbank), from 1977.01 to 1991.12. We assess two forecasts, the "no change" (0) forecast associated with a random-walk model and the forecast implicit in the three-month forward rate (the difference between the three-month forward rate and the spot rate).

The actual and predicted changes are shown in Figure 2. The random-walk forecast, of course, is just constant at 0, whereas the forward market forecast moves over time. The movements in both forecasts, however, are dwarfed by the realized movements in exchange rates.

We shall assess the forecasts' accuracy under absolute error loss. In terms of point estimates, the random-walk forecast is more accurate. The mean absolute error of the random-walk forecast is 1.42, as opposed to 1.53 for the forward market forecast; as one hears so often, "The random walk wins." The

loss-differential series is shown in Figure 3, in which no obvious nonstationarities are visually apparently. Approximate stationarity is also supported by the sample autocorrelation function of the loss differential, shown in Figure 4, which decays quickly.

Because the forecasts are three-step-ahead, our earlier arguments suggest the need to allow for at least two-dependent forecast errors, which may translate into a two-dependent loss differential. This intuition is confirmed by the sample autocorrelation function of the loss differential, in which sizable and significant sample autocorrelations appear at lags 1 and 2 and nowhere else. The Box–Pierce χ^2 test of jointly zero autocorrelations at lags 1 through 15 is 51.12, which is highly significant relative to its asymptotic null distribution of χ^2_{15}. Conversely, the Box–Pierce χ^2 test of jointly zero autocorrelations at lags 3 through 15 is 12.79, which is insignificant relative to its null distribution of χ^2_{13}.

We now proceed to test the null of equal expected loss. F, MGN, and MR are inapplicable because one or more of their

Figure 2.   Actual and Predicted Exchange-Rate Changes. The solid line is the actual exchange-rate change. The short dashed line is the predicted change from the random-walk model, and the long dashed line is the predicted change implied by the forward rate.
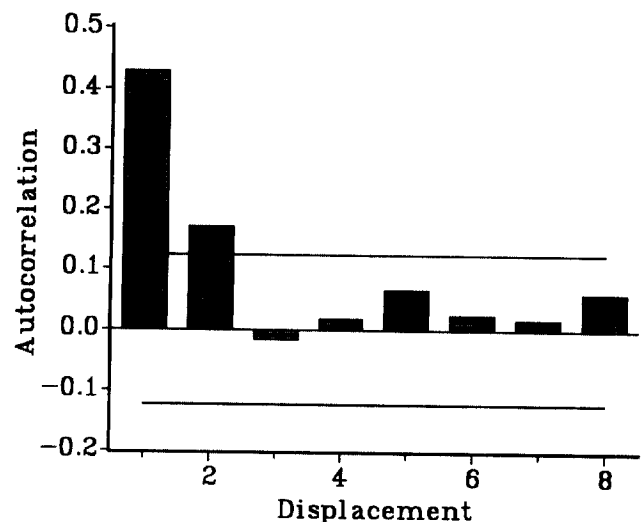

Figure 4.   Loss Differential Autocorrelations. The first eight sample autocorrelations are graphed, together with Bartlett's approximate 95% confidence interval.

maintained assumptions are explicitly violated. We therefore focus on our test statistic S_1, setting the truncation lag at two in light of the preceding discussion. We obtain $S_1 = -1.3$, implying a p value of .19. Thus, for the sample at hand, we do not reject at conventional levels the hypothesis of equal expected absolute error—the forward rate is not a statistically significantly worse predictor of the future spot rate than is the current spot rate.

5. CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

We have proposed several tests of the null hypothesis of equal forecast accuracy. We allow the forecast errors to be non-Gaussian, nonzero mean, serially correlated, and contemporaneously correlated. Perhaps most importantly, our tests are applicable under a very wide variety of loss structures.

We hasten to add that comparison of forecast accuracy is but one of many diagnostics that should be examined when comparing models. Moreover, the superiority of a particular model in terms of forecast accuracy does not necessarily imply that forecasts from other models contain no additional information. That, of course, is the well-known message of the forecast combination and encompassing literatures; see, for example, Clemen (1989), Chong and Hendry (1986), and Fair and Shiller (1990).

Several extensions of the results presented here appear to be promising directions for future research. Some are obvious, such as generalization to comparison of more than two forecasts or, perhaps most generally, multiple forecasts for each of multiple variables. Others are less obvious and more interesting. We shall list just a few:

1. Our framework may be broadened to examine not only whether forecast loss differentials have nonzero mean but also whether other variables may explain loss differentials. For example, one could regress the loss differential not only on a constant but also on a "stage of the business cycle" indicator to assess the extent to which relative predictive performance differs over the cycle.
2. The ability to formally compare predictive accuracy afforded by our tests may prove useful as a model-specification diagnostic, as well as a means to test both nested and nonnested hypotheses under nonstandard conditions, in the tradition of Ashley, Granger, and Schmalensee (1980) and Mariano and Brown (1983).
3. Explicit account may be taken of the effects of uncertainty associated with estimated model parameters on the behavior of the test statistics, as shown by West (1994).

Let us provide some examples of the ideas sketched in 2. First, consider the development of a test of exclusion restrictions in time series regression that is valid *regardless* of whether the data are stationary or cointegrated. The desirability of such a test is apparent from works like those of Stock and Watson (1989), Christiano and Eichenbaum (1990), Rudebusch (1993), and Toda and Phillips (1993), in which it is simultaneously apparent that (a) it is difficult to

determine reliably the integration status of macroeconomic time series and (b) the conclusions of macroeconometric studies are often critically dependent on the integration status of the relevant time series. One may proceed by noting that tests of exclusion restrictions amount to comparisons of restricted and unrestricted sums of squares. This suggests estimating the restricted and unrestricted models using part of the available data and then using our test of equality of the mean squared errors of the respective one-step-ahead forecasts.

As a second example, it would appear that our test is applicable in nonstandard testing situations, such as when a nuisance parameter is not identified under the null. This occurs, for example, when testing for the appropriate number of states in Hamilton's (1989) Markov-switching model. In spite of the fact that standard tests are inapplicable, certainly the null and alternative models may be estimated and their out-of-sample forecasting performance compared rigorously, as shown by Engel (1994).

In closing, we note that this article is part of a larger research program aimed at doing model selection, estimation, prediction, and evaluation using the relevant loss function, whatever that loss function may be. This article has addressed evaluation. Granger (1969) and Christoffersen and Diebold (1994) addressed prediction. These results, together with those of Weiss and Andersen (1984) and Weiss (1991, 1994) on estimation under the relevant loss function will make feasible recursive, real-time, prediction-based model selection under the relevant loss function.

REFERENCES

Andrews, D. W. K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Ashley, R. (1994), "Postsample Model Validation and Inference Made Feasible," unpublished manuscript, Virginia Polytechnic Institute, Dept. of Economics.

Ashley, R., Granger, C. W. J., and Schmalensee, R. (1980), "Advertising and Aggregate Consumption: An Analysis of Causality," *Econometrica*, 48, 1149–1167.

Brockwell, P. J., and Davis, R. A. (1992), *Time Series: Theory and Methods* (2nd ed.), New York: Springer-Verlag.

Campbell, B., and Ghysels, E. (1995), "Is the Outcome of the Federal Budget Process Unbiased and Efficient? A Nonparametric Assessment," *Review of Economics and Statistics*, 77, 17–31.

Chinn, M., and Meese, R. A. (1991), "Banking on Currency Forecasts: Is Change in Money Predictable?" unpublished manuscript, University of California, Berkeley, Graduate School of Business.

Chong, Y. Y., and Hendry, D. F. (1986), "Econometric Evaluation of Linear Macroeconomic Models," *Review of Economic Studies*, 53, 671–690.

Christiano, L., and Eichenbaum, M. (1990), "Unit Roots in Real GNP: Do We Know, and Do We Care?" *Carnegie-Rochester Conference Series on Public Policy*, 32, 7–61.

Christoffersen, P., and Diebold, F. X. (1994), "Optimal Prediction Under Asymmetric Loss," Technical Working Paper 167, National Bureau of Economic Research, Cambridge, MA.

Clemen, R. T. (1989), "Combining Forecasts: A Review and Annotated Bibliography" (with discussion), *International Journal of Forecasting*, 5, 559–583.

Clements, M. P., and Hendry, D. T. (1993), "On the Limitations of Comparing Mean Square Forecast Errors" (with discussion), *Journal of Forecasting*, 12, 617–676.

Cumby, R. E., and Modest, D. M. (1987), "Testing for Market Timing Ability: A Framework for Forecast Evaluation," *Journal of Financial Economics*, 19, 169–189.

Diebold, F. X., and Rudebusch, G. D. (1991), "Forecasting Output with the Composite Leading Index: An Ex Ante Analysis," *Journal of the American Statistical Association*, 86, 603–610.

Dhrymes, P. J., Howrey, E. P., Hymans, S. H., Kmenta, J., Learner, E. E., Quandt, R. E., Ramsey, J. B., Shapiro, H. T., and Zarnowitz, V. (1972), "Criteria for Evaluation of Econometric Models," *Annals of Economic and Social Measurement*, 1, 291–324.

Engel, C. (1994), "Can the Markov Switching Model Forecast Exchange Rates" *Journal of International Economics*, 36, 151–165.

Engle, R. F., and Kozicki, S. (1993), "Testing for Common Features," *Journal of Business & Economic Statistics*, 11, 369–395.

Fair, R. C., and Shiller, R. J. (1990), "Comparing Information in Forecasts From Econometric Models," *American Economic Review*, 80, 375–389.

Granger, C. W. J. (1969), "Prediction With a Generalized Cost of Error Function," *Operational Research Quarterly*, 20, 199–207.

Granger, C. W. J., and Newbold, P. (1977), *Forecasting Economic Time Series*, Orlando, FL: Academic Press.

Hamilton, J. D. (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57, 357–384.

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

Hogg, R. V., and Craig, A. T. (1978), *Introduction to Mathematical Statistics* (4th ed.), New York: MacMillan.

Howrey, E. P., Klein, L. R., and McCarthy, M. D. (1974), "Notes on Testing the Predictive Performance of Econometric Models," *International Economic Review*, 15, 366–383.

Kendall, M., and Stuart, A. (1979), *The Advanced Theory of Statistics* (Vol. 2, 4th ed.), New York: Oxford University Press.

Lehmann, E. L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden-Day.

Leitch, G., and Tanner, J. E. (1991), "Econometric Forecast Evaluation: Profits Versus the Conventional Error Measures," *American Economic Review*, 81, 580–590.

Mariano, R. S., and Brown, B. W. (1983), "Prediction-Based Test for Misspecification in Nonlinear Simultaneous Systems," in *Studies in Econometrics, Time Series and Multivariate Statistics, Essays in Honor of T. W. Anderson*, eds. T. Amemiya, S. Karlin, and L. Goodman, New York: Academic Press, pp. 131–151.

Mark, N. (1995), "Exchange Rates and Fundamentals: Evidence on Long-Horizon Predictability," *American Economic Review*, 85, 201–218.

McCulloch, R., and Rossi, P. E. (1990), "Posterior, Predictive, and Utility-Based Approaches to Testing the Arbitrage Pricing Theory," *Journal of Financial Economics*, 28, 7–38.

Meese, R. A., and Rogoff, K. (1988), "Was it Real? The Exchange Rate-Interest Differential Relation Over the Modern Floating-Rate Period," *Journal of Finance*, 43, 933–948.

Mizrach, B. (1991), "Forecast Comparison in L_2," unpublished manuscript, Wharton School, University of Pennsylvania, Dept. of Finance.

Morgan, W. A. (1939–1940), "A Test for Significance of the Difference Between the Two Variances in a Sample From a Normal Bivariate Population," *Biometrika*, 31, 13–19.

Newey, W., and West, K. (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Priestley, M. B. (1981), *Spectral Analysis and Time Series*, New York: Academic Press.

Rudebusch, G. D. (1993), "The Uncertain Trend in U.S. Real GNP," *American Economic Review*, 83, 264–272.

Stock, J. H., and Watson, M. W. (1989), "Interpreting the Evidence on Money-Income Causality," *Journal of Econometrics*, 40, 161–181.

Toda, H. Y., and Phillips, P. C. B. (1993), "Vector Autoregression and Causality," *Econometrica*, 61, 1367–1393.

Vuong, Q. H. (1989), "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses," *Econometrica*, 57, 307–334.

Weiss, A. A. (1991), "Multi-step Estimation and Forecasting in Dynamic Models," *Journal of Econometrics*, 48, 135–149.

———— (1994), "Estimating Time Series Models Using the Relevant Cost Function," unpublished manuscript, University of Southern California, Dept. of Economics.

Weiss, A. A., and Andersen, A. P. (1984), "Estimating Forecasting Models Using the Relevant Forecast Evaluation Criterion," *Journal of the Royal Statistical Society*, Ser. A, 137, 484–487.

West, K. D. (1994), "Asymptotic Inference About Predictive Ability," SSRI Working Paper 9417, University of Wisconsin-Madison, Dept. of Economics.

West, K. D., Edison, H. J., and Cho, D. (1993), "A Utility-Based Comparison of Some Models of Exchange Rate Volatility," *Journal of International Economics*, 35, 23–46.