1. INTRODUCTION

Predictive ability of an estimated model is critical not only to the quality of model forecasts but also to the adequacy of the model for policy analysis.

Despite the obvious desirability of formal testing procedures, earlier efforts at assessing the forecast accuracy of an estimated model revolved around the calculation of summary forecast error statistics - e.g., see Klein (1991); Clements and Hendry (1993); Wallis (1995); Newbold, Harvey, and Leybourne (1999); and Stock and Watson (1999). Resort to such an informal approach may be ascribed mainly to the complexities involved in dealing with sampling uncertainties and correlations that are present in forecast errors.

Efforts towards the construction of appropriate statistical tests of forecast accuracy started with loss functions that are quadratic in forecast errors. Furthermore forecast errors are assumed to be Gaussian and serially uncorrelated. More recent activity has led to tests under more relaxed conditions – where loss functions may be non-quadratic and asymmetric and forecast errors need not be Gaussian. These tests generally are based on large-sample asymptotic analysis – with some limited experimental studies of their size and power properties in small-samples.

This paper reviews such significance tests. The review is necessarily outpaced by ongoing research; it is necessarily limited as well – focusing on statistical tests of the accuracy of forecasts and tests comparing accuracy of competing forecasts.

2. SIGNIFICANCE TESTS OF FORECAST ACCURACY

Tests of forecast accuracy can be either model-based or model-free. The first type assumes that the econometric model is parametric, estimated from a given sample data and both the data and model are available for testing forecast accuracy. Tests of this genre have been developed for large macroeconometric models based on deterministic and stochastic simulations of the estimated model - see Mariano and Brown (1983); Chong and Hendry (1986); Pagan (1989); Mariano and Brown (1991); and Wallis (1995).

For model-free tests, on the other hand, we assume that the information set at the disposal of the investigator consists only of a set of forecasts and actual values of the predictand. We also consider the case where we have two or more sets of forecasts from competing
models, together with actual values, and wish to determine the relative accuracies of the forecasts.

Most of the tests we discuss here are model-free. These tests are intimately related to encompassing tests that have been developed in the econometric literature (e.g., Clements and Hendry, 1998) and are covered by Harvey and Newbold in this volume. Two very important topics – asymptotic inference for predictive ability and reality checks for data snooping – are not discussed as extensively to avoid duplication with other chapters in this volume.

Our order of presentation of model-free tests of forecast accuracy follows the historical development of the literature – starting with tests developed under squared-error loss and zero-mean, serially uncorrelated forecast errors and proceeding to more recent tests that are asymptotically valid under more general conditions allowing loss functions other than the quadratic and covering situations when forecast errors are non-Gaussian, non-zero-mean, serially correlated, and contemporaneously correlated.

### 2.1 Morgan-Granger-Newbold Test

For our discussion, assume that the available information consists of the following:

- actual values \( \{ y_t : t = 1, 2, 3, \ldots, T \} \)
- two forecasts: \( \{ \hat{y}_{t1} : t = 1, 2, 3, \ldots, T \} \) and \( \{ \hat{y}_{t2} : t = 1, 2, 3, \ldots, T \} \)

Define the forecast errors as

\[
e_{it} = \hat{y}_{it} - y_t, \quad \text{for } i = 1, 2.
\]

The loss associated with forecast \( i \) is assumed to be a function of the actual and forecast values only through the forecast error, \( e_{it} \), and is denoted by

\[
g(y_t, \hat{y}_{it}) = g(\hat{y}_{it} - y_t) = g(e_{it}).
\]

Typically, \( g(e_{it}) \) is the square (squared-error loss) or the absolute value (absolute error loss) of \( e_{it} \). Other measures that introduce alternative functional dependence or asymmetry in the loss function or express loss in terms of economic costs are discussed in, for example, Clements and Hendry (1993), Pesaran and Timmermann (1994), Christoffersen and Diebold (1996), Granger and Pesaran (1999), and Stock and Watson (1999). See also the chapter by Pesaran in this volume.

We denote the loss differential between the two forecasts by
and say that the two forecasts have equal accuracy if and only if the loss differential has zero expectation for all t.

So, we would like to test the null hypothesis

\[ H_0 : E(d_t) = 0 \] for all t

versus the alternative hypothesis

\[ H_1 : E(d_t) = \mu \neq 0. \]

For our first test, consider the following assumptions:

A1. Loss is quadratic.

A2. The forecast errors are
  a. zero mean,
  b. Gaussian,
  c. serially uncorrelated,
  d. contemporaneously uncorrelated.

Maintaining assumptions A1 and A2.a – A2.c, Granger and Newbold (1986) develop a test for equal forecast accuracy based on the following orthogonalization (see Morgan, 1939-1940)

\[ x_t = e_{1t} + e_{2t}, \]

\[ z_t = e_{1t} - e_{2t}. \]

In this case, the null hypothesis of zero mean loss differential is equivalent to equality of the two forecast error variances or, equivalently, zero covariance between \( x_t \) and \( z_t \), since it follows directly from the definition of \( x_t \) and \( z_t \) that

\[ \text{cov}(x_t, z_t) = E(e_{1t}^2 - e_{2t}^2). \]

Hence the test statistic is

\[ \text{MGN} = r / [(1 - r^2) / (T - 1)]^{1/2} \]

where

\[ d_t = g(e_{1t}) - g(e_{2t}) \]
\[ r = \frac{x'z}{[(x'x)(z'z)]^{1/2}} \]

and x and z are the Tx1 vectors with t-th elements \( x_t \) and \( z_t \), respectively. Under the null hypothesis of no zero covariance between \( x_t \) and \( z_t \), MGN has a t-distribution with \( T-1 \) degrees of freedom.

Note that this test is based on the maintained assumption that forecast errors are white noise – hence, the test is applicable only to one-step predictions. Also, the test is valid as a test of equality of forecast accuracy only under squared error loss.

2.2 Variations of the MGN Test

Harvey, Leybourne and Newbold (HLN, 1997) set up the MGN test in a regression framework:

\[ x_t = \beta z_t + \epsilon_t \]

and note that the MGN test statistic is exactly the same as that for testing the null hypothesis that \( \beta = 0 \) in the above regression. Specifically,

\[ \text{MGN} = \frac{b}{(s^2 / z'z)^{1/2}} \]

where

\[ b = \frac{x'z}{z'z} \]

\[ s^2 = (x - bz)'(x - bz)/(T - 1). \]

Thus, the test would work – and in fact, be uniformly most powerful unbiased – in the ideal situation where assumptions A1 and A2.a – A2.c all hold.

When the forecast errors come from a heavy-tailed distribution, however, HLN argue that the estimate of the variance of \( \beta \), which appears in the denominator of MGN, is biased and recommends the following modification of MGN:

\[ \text{MGN}^* = \frac{b}{[(\Sigma z_t^2 \hat{\epsilon}_t^2) / (\Sigma z_t^2)^2]^{1/2}}. \]

where \( \hat{\epsilon}_t \) is the calculated OLS residual at time t. In this modification, a White-correction for heteroskedasticity is utilized to estimate the variance of \( \beta \). Thus, in fact, the modification is a correction for heteroskedasticity, not fat tails, in the distribution of the forecast errors.
HLN further suggest comparing $\text{MGN}^*$ with critical values of the t-distribution with T-1 degrees of freedom.

The simulation study in HLN verifies that the original MGN test has empirical sizes that are equal to nominal sizes when forecast errors are drawn from a Gaussian distribution. However, when the forecast error generating process is $t_6$ (t-distribution with six degrees of freedom), the MGN test becomes seriously oversized – with the deficiency getting worse as sample size increases.

For the modified MGN test, the simulation results are mixed. As theory suggests, the modified test registered the correct size when the sample size was large. However, oversizing shows up when samples are small, both when the forecast error distribution is Gaussian or $t_6$. In fact, in the latter case (with small samples), the modified test performed worse than the original MGN test.

These simulation results led HLN to consider yet another variation of the MGN test – a nonparametric approach using Spearman’s rank test for correlation between $x$ and $z$. Strictly speaking, just like the MGN and MGN* tests, this variation would be valid for testing one-period forecasts in situations where forecast errors are white noise.

The real drawback of these variations of the MGN test is the limitation of their applicability to one-step predictions and to squared error loss. Also, there are inherent difficulties in extending the procedure to multi-period forecast horizons.

### 2.3 Meese-Rogoff Test

Meese and Rogoff (1988) developed a test of equal forecast accuracy when the forecast errors are serially and contemporaneously correlated (relaxing assumptions A2.c and A2.d, but maintaining assumptions A1, A2.a and A2.b). They assume squared error loss and base their test directly on the sample covariance between $x$ and $z$.

Given the maintained assumptions, the following result holds under the hypothesis of equal forecast accuracy:

$$T^{1/2} \hat{\gamma}_{xz}(0) \rightarrow N(0, \Omega) \text{ in distribution}$$

where

$$\hat{\gamma}_{xz}(0) = \frac{x'z}{T}$$

$$\Omega = \sum_{k=-\infty}^{\infty} [\gamma_{xx}(k) \gamma_{zz}(k) + \gamma_{xz}(k) \gamma_{zx}(k)]$$
\[
\gamma_{xz}(k) = \text{cov}(x_t, z_{t-k}), \quad \gamma_{zx}(k) = \text{cov}(z_t, x_{t-k}), \\
\gamma_{xx}(k) = \text{cov}(x_t, x_{t-k}), \quad \text{and} \quad \gamma_{zz}(k) = \text{cov}(z_t, z_{t-k}).
\]

The own- and cross-autocovariances can be estimated consistently by their sample counterparts and a consistent estimate of \( \Omega \) is

\[
\hat{\Omega} = \sum_{k=-m(T)}^{m(T)} (1 - |k|/T) [\hat{\gamma}_{xx}(k)\hat{\gamma}_{zz}(k) + \hat{\gamma}_{xz}(k)\hat{\gamma}_{zx}(k)].
\]

The truncation lag \( m(T) \) increases with sample size \( T \) but at a slower rate.

The statistic for Meese-Rogoff's test is then

\[
\text{MR} = \hat{\rho}_{xz}(0)/\left(\hat{\Omega}/T\right)^{1/2}.
\]

### 2.4 Diebold-Mariano Test

Diebold and Mariano (DM, 1995) consider model-free tests of forecast accuracy that are directly applicable to non-quadratic loss functions, multi-period forecasts, and forecast errors that are non-Gaussian, non-zero-mean, serially correlated, and contemporaneously correlated.

The basis of the test is the sample mean of the observed loss differential series \( \{d_t : t = 1, 2, 3, \ldots, T\} \), when assumptions A1 and A2, a-d need not hold.

Assuming covariance stationarity and other regularity conditions on the process \( \{d_t\} \), we use the standard result that

\[
T^{1/2}(\overline{d} - \mu) \rightarrow N(0, 2\pi f_d(0)), \text{ in distribution}
\]

where \( f_d(\lambda) \) is the spectral density of \( \{d_t\} \) and \( \overline{d} \) is the sample mean differential:

\[
f_d(\lambda) = (1/2\pi) \sum_{k=-\infty}^{\infty} \gamma_d(k) \exp(-ik\lambda), \quad -\pi \leq \lambda \leq \pi
\]

\[
\overline{d} = \frac{1}{T} \sum_{t=1}^{T} [g(e_{1t}) - g(e_{2t})]/T
\]
and \( \gamma_d(k) \) is the autocovariance of \( d_t \) at displacement \( k \):

\[
\gamma_d(k) = E[(d_t - \mu)(d_{t-k} - \mu)].
\]

The Diebold-Mariano test statistic is

\[
DM = \frac{\bar{d}}{2\pi \hat{f}_d(0)/T}^{1/2}
\]

where \( \hat{f}_d(0) \) is a consistent estimate of \( f_d(0) \). The null hypothesis that \( E(d_t) = 0 \) for all \( t \) is rejected in favor of the alternative hypothesis that \( E(d_t) \neq 0 \) when \( DM \), in absolute value, exceeds the critical value of a standard unit Gaussian distribution.

Consistent estimators of \( f_d(0) \) can be of the form

\[
\hat{f}_d(0) = \frac{1}{2\pi} \sum_{k=-m(T)}^{m(T)} w(k/m(T)) \hat{\gamma}_d(k)
\]

where

\[
\hat{\gamma}_d(k) = \frac{1}{T} \sum_{t=|k|+1}^{T} (d_t - \bar{d})(d_{t-|k|} - \bar{d}).
\]

\( m(T) \), the bandwidth or lag truncation, increases with \( T \) but at a slower rate, and \( w(\bullet) \) is the weighting scheme or kernel (see, for example, Andrews (1991) for econometric applications.) One weighting scheme, called the truncated rectangular kernel and used in Diebold and Mariano (1995), is the indicator function that takes the value of unity when the argument has an absolute value less than one:

\[
w(x) = I(|x| < 1).
\]

The Monte Carlo analysis in DM, provides some insight into the finite-sample behavior of the DM test statistic, relative to MGN and MR. The experimental design covers a variety of contemporaneous and serial correlation in the forecast errors. In addition to Gaussian distributions for forecast errors, the t-distribution with 6 degrees of freedom was also used to represent fat tails in the distribution of forecast errors. For comparison purposes, only quadratic loss is considered in the analysis.

In the case of Gaussian forecast errors, the main results are:

1. MGN is robust to contemporaneous correlation and it remains correctly sized as long as there is no serial correlation in the forecast errors. With serial correlation present, the empirical size of MGN exceed nominal size.
2. MR is unaffected by contemporaneous and serial correlation in large samples, as expected. In the presence of serial correlation, it tends to be oversized in small samples. Asymptotic behavior sets in rather quickly; the test shows approximately the correct size for sample sizes exceeding 64.

3. The behavior of DM is similar to that of MR: robust to contemporaneous and serial correlation in large samples and oversized in small samples. Empirical size converges to nominal size a bit more slowly than for MR.

When forecast errors are non-Gaussian, MGN and MR are drastically missized in large as well as small samples. DM, on the other hand, maintains approximately correct size for all but very small sample sizes.

### 2.5 Small-Sample Modifications to the Diebold-Mariano Test

Harvey, Leybourne, and Newbold (HLN, 1997) propose a small-sample modification of the Diebold-Mariano test. The modification revolves around an approximately unbiased estimate of the variance of the mean loss differential when forecast accuracy is measured in terms of mean-squared prediction error and h-step ahead forecast errors are assumed to have zero autocorrelations at order h and beyond.

Since optimal h-step ahead predictions are likely to have forecast errors that are MA(h-1) – a moving average process of order h-1, HLN assume that for h-step ahead forecasts, the loss differential \( d_t \) has autocovariance

\[
\gamma(k) = 0, \text{ for } k \geq h.
\]

In this case, for \( 0 \leq k \leq h \)

\[
\hat{\gamma}_k(k) = (1/T) \sum_{t=k+1}^T (d_t - \bar{d})(d_{t-k} - \bar{d}).
\]

The exact variance of the mean loss differential is

\[
V(\bar{d}) = (1/T)[\gamma_0 + (2/T) \sum_{k=1}^{h-1} (T-k)\gamma_k].
\]

The original DM test would estimate this variance by

\[
\hat{V}(\bar{d}) = (1/T)[\hat{\gamma}_0 + (2/T) \sum_{k=1}^{h-1} (T-k)\hat{\gamma}_k].
\]
\[
\hat{\gamma}^*(k) = T \hat{\gamma}(k)/(T - k).
\]

With \( d \) based on squared prediction error, HLN obtain the following approximation to the expected value of \( \hat{V}(\hat{d}) \):

\[
E(\hat{V}(\hat{d})) \sim V(\hat{d})[T + 1 - 2h + h(h - 1)/T]/T,
\]

and therefore suggest modifying the DM test statistic to

\[
DM^* = DM /[(T + 1 - 2h + h(h - 1)/T)/T]^{1/2}.
\]

Further, HLN suggest comparing \( DM^* \) with critical values from the t-distribution with \((T-1)\) degrees of freedom, instead of the standard unit normal distribution.

HLN also perform a Monte Carlo experiment with these tests. Their simulation study allows for forecast horizons up to 10 periods and uses expected squared prediction error as the criterion of forecast quality. Their main findings are:

1. The tendency of the original DM test to be oversized becomes more severe as the forecast horizon grows, especially in smaller samples. Table 1 in HLN illustrates this problem for samples as small as 16 observations when dealing with one-step ahead forecasts and for samples with up to 128 observations for four-step ahead forecasts.

2. In all cases, the modified test performs considerably better, although it also tends to be oversized, but not as much. Thus, the finite-sample modification suggested by HLN for the DM test provides an important (although not complete) size correction.

3. These findings also apply when forecast errors are contemporaneously correlated, serially correlated, or generated by heavy-tailed distribution (in the study, the t-distribution with six degrees of freedom was used).

4. HLN also provide simulation results on power comparisons – restricting the scope to one-step predictions and quadratic loss and two forecast error distributions – Gaussian and \( t_6 \). Under Gaussian error distributions, the original MGN test is considerably more powerful than the modified DM test when the sample is small (up to \( T = 32 \)). For larger sample sizes, the original MGN and modified DM tests show just about equal power while the modified MGN rank test is somewhat inferior. Under a \( t_6 \) error distribution, the rank test is more powerful than the modified DM test for all sample sizes and all levels of contemporaneous correlation between forecast errors.
2.6 Nonparametric Tests

We summarize three nonparametric tests that have been proposed in the literature.

The first is the standard sign test which is based on the assumption that the loss differential series is independent and identically distributed, and tests the null hypothesis that the median of the loss-differential distribution is equal to zero. Under this null hypothesis, the number, \( N \), of positive loss-differentials in a sample of size \( T \) has a binomial distribution with the number of trials equal to \( T \) and success probability equal to \( \frac{1}{2} \). The test statistic in this case is

\[
\text{SIGN} = \frac{(N - 0.5T)}{0.5^{1/2}}
\]

which is asymptotically \( N(0, 1) \) as sample size increases.

Wilcoxon’s signed rank test also has been used. This test considers the sum of the ranks of the absolute values of positive forecast loss differentials (\( d_t \)):

\[
\text{SR} = \sum I(d_t > 0) \text{rank } |d_t|
\]

where \( I(d_t > 0) \) is the indicator function taking the value one when \( d_t > 0 \).

For loss differentials that are independent, identically distributed with a symmetric distribution around zero, the exact critical values of SR are tabulated in standard texts on nonparametric methods. Under the null hypothesis, it is also known that, as \( T \) goes to infinity,

\[
[S_R - T(T+1)/4]/[T(T+1)2T+1]/24]^{1/2} \rightarrow N(0,1), \text{ in distribution.}
\]

Pesaran and Timmermann (1992) focus on the sign of the predictand \( y_t \), and a nonparametric test is developed based on the number of correct predicted signs in the forecast series of size \( T \). The maintained assumptions are that the distributions of the predictand and predictor are continuous, independent and invariant over time.

Let

\[
p = \text{sample proportion of times that the sign of } y_t \text{ is correctly predicted,}
\]

\[
\pi_1 = \Pr(y_t > 0)
\]

\[
\pi_2 = \Pr(\hat{y}_t > 0)
\]

\[
p_1 = \text{sample proportion of times that actual } y \text{ is positive}
\]
\[ p_2 = \text{sample proportion of times that forecast } y \text{ is positive.} \]

Under the null hypothesis that \( \hat{y}_i \) and \( y_i \) are independently distributed of each other (so that the forecast values have no ability to predict the sign of \( y_i \)), then the number of correct sign predictions in the sample has a binomial distribution with \( T \) trials and success probability equal to

\[ \pi_* = \pi_1 \pi_2 + (1 - \pi_1)(1 - \pi_2). \]

If \( \pi_1 \) and \( \pi_2 \) are known (for example, if the distributions are symmetric around zero), the test statistic is simply

\[ \text{PTK} = \frac{p - \pi_*}{[\pi_*(1 - \pi_*)/T]^{1/2}}. \]

When \( \pi_1 \) and \( \pi_2 \) are not known, they can be estimated by the sample proportions \( p_1 \) and \( p_2 \), so that \( \pi_* \) can be estimated by

\[ p_* = p_1 p_2 + (1 - p_1)(1 - p_2). \]

The test statistic in this case, which Pesaran and Timmermann show to converge in distribution to \( N(0,1) \) under the null hypothesis, is

\[ \text{PTNK} = (p - p_*)([\text{vâr}(p) - \text{vâr}(p_*)]^{1/2} \]

where

\[ \text{vâr}(p) = p_*(1 - p_*)/T \]

\[ \text{vâr}(p_*) = (2p_1 - 1)^2 p_2(1 - p_2)/T + (2p_2 - 1)^2 p_1(1 - p_1)/T + 4 p_1 p_2 (1 - p_1)(1 - p_2)/T^2. \]

Pesaran and Timmermann generalize this test to situations where there are two or more meaningful categories for the actual and forecast values of the predictand. They also remark that, in the case of two categories, the square of PTNK is asymptotically equal to the chi-squared statistic in the standard goodness-of-fit test using the 2x2 contingency table categorizing actual and forecast values by sign. The authors apply their methodology to analyze reported price changes in the Industrial Trends Survey of the Confederation of British Industries (CBI) and demand data from business surveys of the French manufacturing industry conducted by the Institut de la Statistique et des Etudes Economiques (INSEE).

2.7 West’s Asymptotic Inference on Predictive Ability
While the Diebold-Mariano test applies to forecasts from models whose parameters have been estimated, forecast uncertainty due to parameter estimation is not taken into account explicitly. West (1996) provides the formal asymptotic theory for inference about moments of smooth functions of out-of-sample predictions and prediction errors – thus establishing a framework for testing predictive ability of models. The machinery allows for non-nested and nonlinear models and dependence of predictions on estimated regression parameters. The analysis is carried out in environments where there is a long time series of predictions that have been made from a sequence of base periods. Furthermore, if predictions are based on regression estimates, it is also assumed that these regression estimates come from a long time series. In West and McCracken (1998), the framework is utilized to develop regression-based tests of hypotheses about out-of-sample prediction errors, when predictions depend on estimated parameters. The hypotheses tested include (1) zero-mean prediction error and (2) zero correlation between a prediction error and a vector of predictors.

2.8 White’s Reality Check P-Value

When using a data set repeatedly for purposes of inference – say, for evaluating the predictive ability of a large number of alternative models – there is a possibility of getting apparently satisfactory results simply because of chance. Such data-snooping bias has been recognized, especially in empirical finance. White (2000) builds on Diebold and Mariano (1995) and West (1996) and develops what he calls his Reality Check Bootstrap methodology for purposes of testing “whether a given model has predictive superiority over a benchmark model after accounting for the effects of data-snooping.”

We proceed with a broad outline of the method and refer the reader to White (2000) and Sullivan, Timmermann, and White (1999) for technical details.

If $f_j$ measures performance (not loss) of the jth forecasting model relative to the benchmark, the methodology tests the null hypothesis that the maximum of $E(f_j)$, over all models under study, is less than or equal to zero. Correction for data-snooping is achieved by extending the maximization of $E(f_j)$ over all relevant forecasting models. Rejection of the null hypothesis supports the proposition that the best model has superior performance relative to the benchmark.

Suppose there are L models under study and T sample observations on the performance measures $f_{jt}$ ($j = 1, 2, 3, \ldots, L; t = 1, 2, 3, \ldots, T$). White’s test statistic is

$$\max_j T^{1/2} \tilde{f}_j,$$

where

$$\tilde{f}' = \sum f_j' / T = (\tilde{f}_1', \tilde{f}_2', \tilde{f}_3', \ldots, \tilde{f}_L'),$$

and $f_i' = (f_{1i}, f_{2i}, f_{3i}, \ldots, f_{Li})$. 

12
Under appropriate assumptions (e.g., in West (1996) and White (2000)), for some positive definite $\Omega$,

$$T^{1/2}(\hat{f} - E(f_i)) \to N(0, \Omega).$$

White (2000) shows further that his test statistic converges in distribution to the maximum element of a random vector with distribution $N(0, \Omega)$. The distribution of this maximum is not known for the general case. Thus, White uses resampling techniques to determine asymptotic p-values of his test.

White points out that a Monte Carlo p-value can be obtained. One first computes a consistent estimate $\hat{\Omega}$ through block resampling or block subsampling. The next step requires repeated draws from $N(0, \hat{\Omega})$, followed by calculation of the p-value from the pseudo-sampling distribution of the maximum component of the $N(0, \hat{\Omega})$ draws.

For his analysis, White suggests using a bootstrap reality check p-value. The approach resamples from the empirical distribution of $f_i$ using the “stationary bootstrap” procedure of Politis and Romano (PR, 1994). Earlier development in the bootstrap literature has led to the moving blocks method (Kuensch (1989) and Liu and Singh (1992)) for resampling from a stationary dependent process. This technique constructs a resample from blocks of observations with fixed length and random starting indices. PR modify this technique by introducing randomness in the length of the moving blocks. The lengths are assumed to follow a geometric distribution with mean value that increases with sample size. PR call their method “stationary bootstrap” because of the fact that their resampled time-series is stationary (conditional on the original data). PR also show that their method produces valid bootstrap approximations for means of $\alpha$-mixing processes. The procedure also avoids certain biases associated with fixed lengths in the moving blocks bootstrap.


3. APPLICATIONS

There is a myriad of empirical papers dealing with forecast evaluation. Here, we briefly review four recent ones that apply the testing methodologies that we have discussed.

In studying the predictability of nominal exchange rates, Mark (1995) estimates regressions of multi-period changes in the log exchange rate on the deviation of the log exchange rate from its “fundamental value.” For operational purposes, the fundamental
value is defined as a linear combination of log relative money stocks and log relative income. Four currencies are analyzed: the Canadian dollar, the German deutsche mark, the Swiss franc, and the Japanese yen. The Diebold-Mariano test is used to compare the predictive ability of the estimated regressions with a driftless random walk and leads to the general conclusion that out-of-sample point forecasts of the regressions based on fundamentals outperform forecasts from the random walk model at longer horizons. “While short-horizon changes (in log exchange rate) tend to be dominated by noise, this noise is apparently averaged out over time, thus revealing systematic exchange-rate movements (over longer horizons) that are determined by economic fundamentals. These findings are noteworthy because it has long been thought that log exchange rates are unpredictable.”

Kaufman and Stern (1999) analyze the effects of model specification, data aggregation, and data quality on the predictive ability of an equilibrium-correction time-series model for global temperature as explained by radiative forcing of greenhouse gases, anthropogenic sulphur emissions and solar irradiance. An important variable in the model is what the authors call the “global temperature reconstruction” which appears in both the cointegration relationship and the short-run dynamics of the model. Alternative estimated versions of the model (and hence, also temperature forecasts) result from alternative measures of this variable. Two measures were used in their paper – one based on a diffusion energy balance model (EBM) and the other on an atmosphere ocean general circulation model (GCM) of the climate. Utilizing absolute prediction error loss, the authors implement the Diebold-Mariano test and find that the predictive accuracy of the EBM construction is statistically superior to that of GCM.

Swanson and White (1997 a, b) study the usefulness of a variety of flexible specification, fixed specification, linear and nonlinear econometric models in predicting future values of nine macroeconomic variables. These variables are the civilian unemployment rate, Aaa corporate bond yield, industrial production index, nominal gross national product, real gross national product, corporate profits after taxes, personal consumption expenditures, change in business inventories, and net exports of goods and services. Using a variety of out-of-sample forecast-based model selection criteria and tests (including Granger-Newbold, Diebold-Mariano, the Wilcoxon signed rank test, and a contingency test based on forecast direction accuracy or confusion rate), the authors find in (1997a) that “multivariate adaptive linear vector autoregression models often outperform adaptive and nonadaptive univariate models, nonadaptive multivariate models, adaptive nonlinear models, and professionally available survey predictions.” In (1997 b), the authors point out that flexible artificial neural networks (ANN) appear to offer a viable alternative to fixed specification linear models (such as random walk, random walk with drift, and unrestricted vector autoregressive models), especially for forecast horizons greater than one period ahead. However, there is mixed evidence on the performance of the flexible ANN models used in the study – perhaps pointing to the need for extending the analysis to include more elaborate ANN models. The study also tends to support the notion that alternative cost functions (e.g., those based on market timing
and profitability instead of prediction error) can lead to conflicting ranking of the forecasting ability of the models.

Sullivan, Timmermann and White (STW, 1999) apply the Reality Check bootstrap methodology of White (2000) for significance testing of the predictive superiority of technical trading rules in the stockmarket after accounting for data-snooping over the universe from which the trading rules were drawn. Revisiting Brock, Lakonishok and LeBaron (BLL, 1992), STW expands the universe from the 26 trading rules considered by BLL to almost 8000 trading rules. These were drawn from previous academic studies and the literature on technical analysis, and were in use in a substantial part of the sample period. Two performance measures relative to a benchmark were used: market returns and a Sharpe ratio relative to a risk-free rate. STW reports the following major findings:

1. Over the sample period covered by BLL (1897 – 1986), the best technical trading rule generates superior performance even after accounting for data-snooping.

2. However, the best technical trading rule does not deliver superior performance when used to trade in the subsequent 10-year post-sample period.

3. When applied to Standard and Poor’s 500 futures contracts, the analysis taking into account data-snooping provides no evidence of superior performance of the best trading rule.

4. CONCLUSION

This chapter has reviewed the development of significance tests of forecast accuracy. Developments in the decade of the 1990s have yielded model-free testing procedures that are applicable in a wide variety of cases – where performance measures are non-quadratic and asymmetric, where forecast errors are serially correlated and contemporaneously correlated (when multivariate in nature), and where underlying distributions need not be Gaussian.

So now, there are no more excuses for settling for simply descriptive statistics when assessing the predictive ability of forecasting procedures. There are statistical tests of significance that can be used. And, as our selective examples illustrate, they have been used effectively in actual empirical investigation in numerous substantive applications in economics, business and other disciplines. But there remains a wide research frontier in this area, as more complications are taken into account in the testing environment for forecast accuracy – e.g., parameter estimation in West’s tests of predictive ability, nonstandard test statistics in White’s treatment of data snooping, and nonstationarity.


17


