Statistical Tests for Multiple Forecast Comparison

Roberto S. Mariano
(Singapore Management University & University of Pennsylvania)
Daniel Preve
(Uppsala University)

June 6-7, 2008
T.W. Anderson Conference, Stanford University

Why Test for Predictive Ability?

- Typically there are several plausible methods to forecast a time series.
- Since the future values of the time series are unknown, a portion of the observations from the estimation process is held back and the alternative methods are estimated over the shortened span of data. The estimates are then used to forecast the observations of the holdback period (typically a recursive scheme is used). The properties of the forecast errors of the alternative methods can now be compared.
- Early efforts at assessing forecast accuracy revolved around the calculation of error measures like MSE and MAD.
- With forecasts from several methods it is inevitable that the sample will show differences in forecast accuracy between the methods.
- How likely is this outcome due to pure chance, that is, is the observed difference statistically significant or not?

Tests for Equal Predictive Accuracy

If there are just *two* plausible models they can be put to a head-to-head test. Model-free (i.e. the models that generated the forecasts need not be available) tests that compare the forecasts of two alternative time series methods include:

- Morgan-Granger-Newbold (1977) & Meese-Rogoff (1988), tests for equal MSEs.
- Diebold-Mariano (1995).

For these tests H_0 says that the alternative methods are *equally* accurate on average.

The Diebold-Mariano (DM) Test

- Two alternative forecast methods.
- Two time series of forecast errors: $e_{i1}, ..., e_{iT}$ and $e_{j1}, ..., e_{jT}$.
- The quality of each forecast is evaluated by some loss function g of the forecast error.
- The null hypothesis of *equal predictive accuracy* is $E\, d_t = 0$ for all t, where $d_t = g(e_{it}) - g(e_{jt})$.
- Under fairly weak conditions $\frac{\bar{d}}{\sqrt{\hat{\omega}/T}}$ is asymptotically $N(0,1)$ under H_0.
- $\bar{d}$ is the sample mean of the *loss differential* series $d_1, ..., d_T$ and $\hat{\omega}$ is a consistent estimator of the asymptotic variance of $\sqrt{T}\bar{d}$.
- **Remark 1:** Harvey, Leybourne & Newbold (HLN) later addressed the finite-sample properties of the DM statistic, under the additional assumption that all autocovariances of $\{d_t\}$ beyond some lag length q are zero, and proposed a modified test based on an *approximately* unbiased estimator of Var $\bar{d}$.

Introductory Remarks

- Obvious desirability of formal testing procedures
- But earlier efforts at assessing forecast accuracy revolved around calculation of summary error statistics-mainly due to complexities in dealing with sampling uncertainties and correlations present in forecast errors

Introductory Remarks ... continue

- Formal testing approaches started with loss functions that are quadratic in forecast errors; & forecast errors are assumed to be Gaussian and serially uncorrelated

- More recent efforts – much more relaxed conditions
  - Loss functions may be nonquadratic and asymmetric
  - Forecast errors need not be Gaussian
  - Generally based on large-sample asymtotic analysis
  - With limited experimental studies on small-sample properties

Significance Tests of Forecast Accuracy

- Model-based tests
  - Assumes an econometric model, typically parametric
  - Model is estimated from a given data sample
  - Data and model are both available for testing forecast accuracy
  - Applied in large macroeconometric models, using deterministic and stochastic simulations of the estimated model

Significance Tests of Forecast Accuracy ... continue

- Model-free tests
  - Limited information set: set of forecasts and actual values of the predictand

Preliminaries (1)

- Available information: t=1,2,3, … T
  - Actual values y_t
  - Forecast i: $\hat{y}_{it}$, i=1,2
- Forecast errors: $e_{it} = \hat{y}_{it} - y_t$
- Loss depends on forecast and actual values only through the forecast error:

  $g(y_t, \hat{y}_{it}) = g(\hat{y}_{it} - y_t) = g(e_{it})$
- Loss differential between the two forecasts

  $d(t) = g(e_{1t}) – g(e_{2t})$

Preliminaries (2)

- Two forecasts have equal accuracy if and only if the loss differential has zero expectation for all t

- Hence, test

  H_0: $E(d_t) = 0$ for all t

  versus the alternative hypothesis

  H_1: $E(d_t) = \mu$, different from zero

Morgan-Granger-Newbold (MGN) Test (1977)

- Assume
  - A(1) Loss is quadratic
  - A(2) Forecast errors are (a) zero mean, (b) Gaussian, (c) serially uncorrelated
- Let

  $x_t = e_{1t} + e_{2t}$

  $z_t = e_{1t} - e_{2t}$

- Here, H_0 is equivalent to equality of the two forecast error variances, or, equivalently, zero correlation between x_t and z_t

Variations of MGN Test

- Harvey, Leybourne and Newbold (1997) regression set up

$$x_t = \beta z_t + \varepsilon_t$$

- The MGN test statistic is exactly the same as that for testing the null hypothesis that $\beta = 0$ in this regression.

Variations of MGN Test … continue

- When the forecast errors come from a heavy-tailed distribution , HLN argue that the estimate of the variance of b is biased and suggest utilizing a White-correction for heteroskedasticity to estimate the variance of b.

- Another HLN variation: Spearman's rank test for zero correlation between x and z

Variations of MGN Test … continue

- Real drawback of all these tests: limitation of applicability to one-step predictions and to squared error loss

Meese-Rogoff (MR) Test (1988)

- Now, forecast errors can be serially and contemporaneously correlated
- Still maintain assumptions A1, A2a, and A2b and assume squared error loss
- The MR test is based on the sample covariance between x_t and z_t

Diebold-Mariano (DM) Test (1995)

- Applicable to nonquadratic loss functions, multi-period forecasts, and forecast errors that are non-Gaussian, nonzero-mean, serially correlated, and contemporaneously correlated.

- Basis of the test: sample mean of the observed loss differential series
  - $\{d_t : t=1, 2, \ldots\}$

DM Test (2)

- Assuming covariance stationarity and other regularity conditions on the process $\{d_t\}$, then $T^{1/2}(\bar{d} - \mu)$ converges in distribution to $N(0, 2\pi f_d(0))$,
- $f_d(.)$ is the spectral density of $\{d_t\}$
- $\bar{d}$ is the sample mean loss differential

DM Test Statistic

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

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

of $f_d(0)$.

Small-Sample Modification of DM Test

- HLN (1997) :use an approximately unbiased estimate of the variance of the mean loss differential
- Forecast accuracy is measured in terms of mean squared prediction error

Small-Sample Modification of DM Test ... continue

- H-step ahead forecast errors are assumed to have zero autocorrelations at order h and beyond

- Small-sample modification

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

t-distribution with T-1 $d.f.$

Applications

- Predictability of nominal exchange rates (Mark 1995)
- Comparing predictive ability of flexible-specification, fixed-specification, linear and nonlinear econometric models of macroeconomic variables (Swanson & White 1997)

Applications

- Predictive ability with cointegrated variables (Corradi, Swanson & Olivetti 2001)
- Predictive ability in the presence of structural breaks (Clark & McCracken 2003)
- Forecast comparison of volatility models versus GARCH (1,1) – (Hansen & Lunde 2005)

Applications

- Forecast comparison of volatility models versus GARCH (1,1) – (Hansen & Lunde 2005)

A Multivariate Test

- $k+1$ alternative forecast methods, $k \geq 1$.
- $H_0 : E\,g(e_{1t}) = E\,g(e_{2t}) = ... = E\,g(e_{k+1,t})$ or, equivalently, $E\,g(e_{1t}) = E\,g(e_{2t}) \wedge E\,g(e_{2t}) = E\,g(e_{3t}) \wedge ... \wedge E\,g(e_{kt}) = E\,g(e_{k+1,t})$, where $\wedge$ denotes logical 'and'.
- The null hypothesis of *equal predictive accuracy* is $E\,\mathbf{d}_t = \mathbf{0}$ for all t, where $\mathbf{d}_t = (d_{1t}, ..., d_{kt})'$ and $d_{jt} = g(e_{jt}) - g(e_{j+1,t})$, $j = 1, ..., k$.
- Under fairly weak conditions $T\,\bar{\mathbf{d}}'\hat{\Omega}^{-1}\bar{\mathbf{d}}$ is asymptotically χ^2_k under H_0.
- $\bar{\mathbf{d}}$ is the sample mean of the vector loss differential series $\mathbf{d}_1, ..., \mathbf{d}_T$ and $\hat{\Omega}$ is a consistent estimator of the asymptotic variance of $\sqrt{T}\bar{\mathbf{d}}$.
- For simplicity, it is assumed that $\{\mathbf{d}_t\}$ is a vector MA(q) process such that $\lim_{T\to\infty} \mathrm{Var}\,(\sqrt{T}\bar{\mathbf{d}}) = \Gamma(0) + \sum_{h=1}^{q} \left[\Gamma(h) + \Gamma'(h)\right] = \Omega$.
- $\Gamma(h)$ is the autocovariance matrix of $\{\mathbf{d}_t\}$ at lag h.

Invariance and Bias

Remark 2: Any reordering of the alternative forecasting methods potentially alters the appearance of $\mathbf{d}_t$.

Proposition

For any two vectors of loss differentials based on two distinct orderings, $\mathbf{d}_t$ and $\mathbf{d}_t^*$ say, there exists a nonsingular matrix $\mathbf{B}$ such that $\mathbf{B}\mathbf{d}_t = \mathbf{d}_t^*$.

Consequently, the proposed Wald statistic $S_1 = T\,\bar{\mathbf{d}}'\hat{\Omega}^{-1}\bar{\mathbf{d}}$ is unaffected by reordering.

Remark 3: In our multivariate extension of the Diebold-Mariano test we estimate $\mathrm{Var}\,\bar{\mathbf{d}}$ by $T^{-1}\hat{\Omega}$. In finite samples, however, $T^{-1}\hat{\Omega}$ is biased.

Two Modified Tests

The finite-sample correction of HLN for the DM test extends to our multivariate setting, and it can be improved in terms of the order of the error in its underlying approximation. This results in two modified tests with potentially better finite-sample properties than S_1.

- The HLN modification of S_1:

$$S_2 = \frac{T - 1 - 2q + T^{-1}q(q+1)}{T} S_1,$$

- The on S_2 'improved' modification of S_1:

$$S_3 = \frac{T - 1 - 2q - T^{-1}q(q+1)}{T} S_1.$$

Monte Carlo Setup

- Consider the simple case with three alternative methods and Gaussian prediction errors, that are to be compared in terms of mean prediction error. That is, we test $H_0 : E\, e_{1t} = E\, e_{2t} = E\, e_{3t}$.
- We consider q-dependent, contemporaneously correlated forecast errors where q is either 0, 1 or 2.
- The Gaussian vector MA(q) forecast errors are given by

$$\begin{pmatrix} e_{1t} \\ e_{2t} \\ e_{3t} \end{pmatrix} = \begin{pmatrix} \frac{1+\theta_1 B+\theta_2 B^2}{\sqrt{1+\theta_1^2+\theta_2^2}} & 0 & 0 \\ 0 & \frac{1+\theta_1 B+\theta_2 B^2}{\sqrt{1+\theta_1^2+\theta_2^2}} & 0 \\ 0 & 0 & \frac{1+\theta_1 B+\theta_2 B^2}{\sqrt{1+\theta_1^2+\theta_2^2}} \end{pmatrix} \begin{pmatrix} v_{1t} \\ v_{2t} \\ v_{3t} \end{pmatrix},$$

where $\mathbf{v}_t \sim N_3(\mathbf{0}, \mathbf{R})$ and

$$\mathbf{R} = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}.$$

By construction, $\mathbf{e}_t \sim N_3(\mathbf{0}, \mathbf{R})$.

- The loss differential vector at time t is given by $\mathbf{d}_t = \begin{pmatrix} d_{1t} \\ d_{2t} \end{pmatrix} = \begin{pmatrix} e_{1t} - e_{2t} \\ e_{2t} - e_{3t} \end{pmatrix}$.

Table 1: The table reports the percentage of the simulations in which the *true* null of equal mean forecast errors is rejected at the asymptotic 10% critical value. The model generating the Gaussian forecast errors takes the form of a trivariate MA(2) process. The parameters θ_1 and θ_2 denote the MA coefficients for the three forecast errors and control the serial correlation. The parameter ρ controls the contemporaneous correlation between the forecast errors. S_1 and S_2 denote, respectively, the original and HLN versions of the proposed test. S_3 is the improved version of S_2. The presented results are based on 1 000 000 Monte Carlo replications.

	$\rho = 0$				$\rho = 0.5$				$\rho = 0.9$			
Test	$T = 50$	$T = 100$	$T = 200$	$T = 400$	$T = 50$	$T = 100$	$T = 200$	$T = 400$	$T = 50$	$T = 100$	$T = 200$	$T = 400$
						$\theta_1 = \theta_2 = 0$						
S_1	11.96	10.99	10.48	10.27	11.99	10.99	10.50	10.24	12.03	11.00	10.51	10.27
S_2	11.49	10.75	10.37	10.21	11.51	10.76	10.38	10.18	11.54	10.76	10.39	10.22
S_3	11.49	10.75	10.37	10.21	11.51	10.76	10.38	10.18	11.54	10.76	10.39	10.22
						$\theta_1 = 0.5, \theta_2 = 0$						
S_1	15.22	12.54	11.22	10.63	15.23	12.55	11.24	10.63	15.29	12.52	11.26	10.66
S_2	13.67	11.79	10.87	10.46	13.67	11.81	10.89	10.45	13.72	11.79	10.91	10.48
S_3	13.63	11.78	10.86	10.46	13.62	11.80	10.88	10.45	13.67	11.78	10.91	10.48
						$\theta_1 = 0.9, \theta_2 = 0$						
S_1	15.09	12.50	11.21	10.63	15.10	12.50	11.21	10.62	15.16	12.47	11.25	10.64
S_2	13.55	11.74	10.84	10.45	13.57	11.75	10.86	10.44	13.61	11.73	10.89	10.47
S_3	13.50	11.73	10.83	10.45	13.52	11.74	10.85	10.44	13.56	11.72	10.89	10.47
						$\theta_1 = \theta_2 = 0.5$						
S_1	18.80	14.21	12.03	11.04	18.81	14.23	12.05	11.04	18.84	14.20	12.07	11.06
S_2	16.05	12.94	11.42	10.74	16.08	12.95	11.43	10.74	16.14	12.93	11.46	10.76
S_3	15.91	12.91	11.41	10.74	15.95	12.92	11.42	10.74	16.00	12.90	11.45	10.76
						$\theta_1 = \theta_2 = 0.9$						
S_1	18.79	14.22	12.03	11.04	18.81	14.23	12.05	11.04	18.85	14.21	12.07	11.05
S_2	16.03	12.95	11.42	10.74	16.10	12.95	11.43	10.73	16.12	12.93	11.45	10.76
S_3	15.89	12.92	11.42	10.74	15.96	12.92	11.43	10.73	15.98	12.90	11.45	10.76

Multivariate Case
Monte Carlo Results

- The proposed test can be oversized in moderate samples
- The test benefits noticeably from the finite-sample correction, even in moderately large samples
- However, the finite-sample correction provides only a partial adjustment

Multivariate Case Follow-up Work

- Consider alternative types of weak stationarity
- Extensions to
  - Panel data (Pesaran)
  - High frequency data
  - Qualitative and limited dependent variable
  - Semiparametric approaches
- Compare with White / Hansen's data snooping reality test
- Relation to Ken West's test for predictive ability
- Semiparametric approaches to multivariate tests of forecasting performance
- Power considerations

The End