ECONOMIC FORECASTING

GRAHAM ELLIOTT
ALLAN TIMMERMANN

利得图书馆

sizes or for long horizons. As $(1 - 2h) < 0$ for all h, the first-order effect is to reduce the size of the test statistic.[7]

Monte Carlo evidence in Harvey, Leybourne, and Newbold (1998) shows that their corrections have the desired directional effect on size—the corrections reduce the rejection rate under the null hypothesis by both increasing the critical value and decreasing the value of the test statistic. Size is not controlled for all values of h and all sample sizes, but the distortions are smaller than for the uncorrected t-statistic.

Most of the simulation evidence has been conducted under MSE loss, so d_{t+h} is the difference between two squared terms. Squared errors tend to be quite skewed since they are bounded below at 0, and hence it is unsurprising that small sample tests based on the asymptotic normal distribution are oversized. This effect is exacerbated when the underlying forecast errors are fat tailed. Diebold and Mariano (1995) present simulation evidence that illustrates these effects when the forecast errors are drawn from stable distributions so the Monte Carlo design does not allow estimation error in the construction of the forecasts. Busetti and Marcucci (2013) conduct a Monte Carlo study of the size and power properties of a variety of tests for equal predictive accuracy under squared error loss for nested regression models. They find that the ranking of different tests is quite robust across settings with misspecified models and across different forecast horizons. They also find that highly persistent regressors give rise to a loss in power but do not affect the size of the test.

17.3 COMPARING FORECASTING METHODS: THE GIACOMINI–WHITE APPROACH

Giacomini and White (2006) propose a fundamentally different but highly relevant approach to testing between alternative forecasts. They do not rely on asymptotic results that replace parameter estimates by their probability limits—an approach that basically tests which model is better in population. Instead Giacomini and White (2006) retain the effect that estimation errors have on the forecasts and ask whether two forecasting methods produce the same quality of forecasts (according to the chosen loss function) or if instead one method is better. Their test takes as given an observed sequence of forecasts from the two methods that are being compared and assumes that the parameters of the models are estimated using a rolling window of fixed length. This preserves estimation errors and can be viewed as a sequence of observations on the methods' performance. Tests for equal expected loss as well as tests of orthogonality of forecast errors with respect to all information available when the forecasts were produced can then be performed. The distribution of such tests is approximated under the assumption that the observed sequence of forecasts gets large.

More formally, consider two models that at time t are used to generate one-step-ahead forecasts $f_{1t+1|t}$ and $f_{2t+1|t}$ using a fixed window of ω observations. Each forecast $f_{it+1|t}$ is a function of the data $(z_t, z_{t-1}, \ldots, z_{t-\omega+1})$ and parameter estimates $\hat{\beta}_{1t}, \hat{\beta}_{2t}$ and is denoted by $f_{it+1|t}(\hat{\beta}_{it}), i = 1, 2$ for short. From these forecasts we can construct losses $L(f_{it+1}(\hat{\beta}_{it}), y_{t+1})$ for $t = T_R, \ldots, T - 1$. These can be used

[7] In addition to this test, Harvey, Leybourne, and Newbold (1998) suggest tests based on the median difference instead of the mean difference in losses, and also consider signed rank tests of the difference.

LIBRARY.LEOPOLDS.COM

利得图书馆

to compare the models' finite sample predictive accuracy evaluated at the current parameters, $\hat{\beta}_{1t}$, $\hat{\beta}_{2t}$, through the following null:

$$H_0 : E\left[L(f_{1t+1|t}(\hat{\beta}_{1t}), y_{t+1}) - L(f_{2t+1|t}(\hat{\beta}_{2t}), y_{t+1})\right] = 0. \tag{17.20}$$

It is useful to contrast the null hypothesis in (17.20) with the null of equal predictive accuracy at the population parameters, β_1^*, β_2^*, in the analysis of West (1996) and many subsequent papers:

$$H_0 : E\left[L(f_{1t+1|t}(\beta_1^*), y_{t+1}) - L(f_{2t+1|t}(\beta_2^*), y_{t+1})\right] = 0. \tag{17.21}$$

The null in (17.20) tested by Giacomini and White (2006) is fundamentally different from the null in (17.21). The key difference is that the effect of estimation error does not vanish in (17.20), which assumes a fixed estimation window, whereas it does so for (17.21) which evaluates the expected losses at the (probability) limits of the estimators, $\hat{\beta}$, as the estimation sample gets very large. For example, suppose that the finite-sample bias in the small model due to the omission of relevant predictor variables balances exactly against the reduced effect of estimation error, both measured relative to a larger, unrestricted model. Then, the null hypothesis tested by Giacomini and White should not be rejected. In contrast, the null tested by West should be rejected as the estimation sample expands and the effect of estimation error vanishes.

Conversely, when comparing nested models, (17.20) can set a higher standard relative to tests such as (17.21) since the large model is now required to outperform the small model by a margin big enough to make up for the greater effect that estimation error has on the large model's forecasting performance.

Giacomini and White (2006) establish that, under a finite estimation window,

$$T_P^{-1/2} \sum_{t=T_R}^{T-1} \left[\Delta L_{t+1}(\hat{\beta}_{1t}, \hat{\beta}_{2t}, y_{t+1}) - E[\Delta L]\right] \rightarrow^d \mathrm{N}(0, \tilde{S}_y(0)),$$

where $\Delta L_{t+1}(\hat{\beta}_{1t}, \hat{\beta}_{2t}, y_{t+1}) = L(f_{1t+1|t}(\hat{\beta}_{1t}), y_{t+1}) - L(f_{2t+1|t}(\hat{\beta}_{2t}), y_{t+1})$ measures the differential loss, while the variance $\tilde{S}_y(0)$ is given by

$$\tilde{S}_y(0) = \lim_{T_P \rightarrow \infty} \mathrm{Var}\left(T_P^{-1/2} \sum_{t=T_R}^{T-1} \left[L(f_{1t+1|t}(\hat{\beta}_{1t}), y_{t+1}) - L(f_{2t+1|t}(\hat{\beta}_{2t}), y_{t+1}) - E[\Delta L_{t+1}]\right]\right).$$

Note that this result is obtained in the limit for $T_P \rightarrow \infty$ but without the assumption that T_R expands asymptotically. Hence, Giacomini and White (2006) do not need to make assumptions such as positive-definiteness of Ω and so their approach allows for both nested and nonnested comparisons. Moreover, a much wider class of forecast methods that do not necessarily fit in the mold of the analysis of West (1996) can be considered, including Bayesian, nonlinear, and nonparametric models, as well as forecasts based on a variety of nonstandard estimators. In each case it is important that the effect of estimation error does not vanish so that $\tilde{S}_y(0)$ does not become degenerate.

LIBRARY.LEOPOLDS.COM

利得图书馆

17.3.1 Conditional Test of Forecasting Performance

Giacomini and White (2006) introduce conditional tests for predictive accuracy that are conditional on current information, Z_t. For these tests the null hypothesis in (17.20) is altered to

$$E\left[\Delta L_{t+1}(\hat{\beta}_{1t}, \hat{\beta}_{2t}, y_{t+1})|Z_t\right] = 0, \tag{17.22}$$

where $Z_t = \{z_1, \ldots, z_t\}$ is the information set available at time t. For a single-period forecast horizon, $h = 1$, the null is that the loss difference is a martingale difference sequence with respect to Z_t. For longer horizons, the null implies that information available at time, t, is not correlated with the difference in the losses.

The hypothesis in (17.22) is interesting from an economic point of view since it allows us to test whether certain forecasting models are better in some economic states than others. For example, Henkel, Martin, and Nardari (2011) find that stock returns are predictable during economic recessions but not during expansions. This could be tested by letting the conditioning information include a recession indicator, $\mathbb{1}(\text{NBER}_t = 1)$, that equals 1 if the NBER views period t as a recession, and otherwise equals 0, although the NBER indicator is not available in real time.

To make the conditional null hypothesis in (17.22) operational, we need to choose a set of test functions, v_t, which are functions of data available at the time the forecast is made, i.e., functions of Z_t. Letting v_t be a $(q \times 1)$ vector, we can test the moment restriction in (17.22) using a standard GMM quadratic form,

$$\text{GW}(T_P) = T_P \left(T_P^{-1} \sum_{t=T_R}^{T-1} v_t \Delta L_{t+1} \right)' W^{-1} \left(T_P^{-1} \sum_{t=T_R}^{T-1} v_t \Delta L_{t+1} \right), \tag{17.23}$$

where W is the optimal weight matrix for the GMM problem and $\Delta L_{t+1} = L(f_{1t+1|t}(\hat{\beta}_{1t}), y_{t+1}) - L(f_{2t+1|t}(\hat{\beta}_{2t}), y_{t+1})$. Under relatively mild and standard conditions, Giacomini and White show that this statistic has a limiting χ^2_q distribution under the null hypothesis.[8]

An interesting aspect of this test is that different choices for the window length in the rolling regressions, ω, change what is being tested. The reason is that if we change the window length, we also alter $\hat{\beta}_{1t}$, $\hat{\beta}_{2t}$ and thus the sequence of forecasts, $\{f_{1t+1|t}(\hat{\beta}_{1t}), f_{2t+1|t}(\hat{\beta}_{2t})\}$, and the null hypothesis. The upshot of this is that with the same models but different window lengths, forecasters might find that the tests yield different results. However, this property is part of the point of undertaking the test for the forecast method (which includes the choice of estimation window) rather than attempting to learn which forecasting model is best when evaluated at the limit of the parameter estimates.

A failure of finding that the test in (17.23) detects superior performance for a model whose parameters are estimated using a rolling window does not imply, of course, that the same model, with parameters estimated on an expanding window, would not have generated better forecasts. Using a rolling window estimator in such situations can worsen the performance of large models with a greater number of estimated parameters and so can impair the test's ability to identify these models

[8] For example, strict stationarity is not needed for this result.

LIBRARY.LEOPOLDS.COM

利得图书馆

as being superior relative to more parsimonious models with fewer estimated parameters, even if the large model is the best specification. For example, forecasts based on rolling window estimation with 10 years of observations might lead to rejections of the large model, while the large model could be preferred with a rolling estimation window of 20 years of observations.

17.4 COMPARING FORECASTING PERFORMANCE ACROSS NESTED MODELS

When comparing the finite-sample performance of two nested models, estimation error can cause the large model to produce less precise forecasts—generate higher MSE values—than the small model which requires estimation of fewer parameters. The test statistics proposed by McCracken (2000) take this into account. Specifically, the distribution of test statistics that account for estimation error shifts further to the left and in many cases takes on negative values, the greater the number of additional parameters that have to be estimated for the large model.

This property means that a possible outcome of the test of equal predictive accuracy could be to favor a large forecasting model even though this model generates less precise forecasts in a particular finite sample than a smaller model. The logic in such cases is that although the large model underperformed the small model in a finite sample, its performance was not as bad as one would have expected given the additional number of parameters that require estimation by the large model.

From the point of conducting inference about two models, this is a valid point. However, from the perspective of a forecaster who is deciding on which model to use, it seems risky to choose the large model in situations where it is underperforming the smaller model. This holds even if the sample evidence suggests that the larger model eventually will be preferred when enough data are available to estimate its additional parameters with greater precision.

Two approaches have been proposed to address these issues for nested models. One approach suggested by Clark and West (2007) recenters the test statistic in a way that explicitly adjusts the test for the greater effect of parameter estimation error on the large model. The second approach is to directly focus the test on finite-sample performance, as suggested by Giacomini and White (2006). We first explain how recursive parameter estimation induces standard evaluation test statistics to follow nonstandard distributions and next describe these approaches.

17.4.1 Complications Arising from Nested Models

West (1996) established that two nonnested models' relative predictive accuracy will asymptotically be normally distributed, albeit with standard errors that need to account for estimation error. When models are nested, this result, or the results of West and McCracken (1998), need no longer hold. Under the null that the additional predictors of the larger model have zero coefficients and so are irrelevant, the forecasts from the large and small model will be identical. In the limit as the effect of estimation error vanishes, the standard error of the difference between the two models' forecast performance will therefore be 0. Test statistics based on differential MSE performance will therefore have nonstandard limiting distributions.

LIBRARY.LEOPOLDS.COM