Mincer, J., and Zarnowitz, V. (1969), "The Evaluation of Economic Forecasts," in *Economic Forecasts and Expectations*, ed. J. Mincer, New York: National Bureau of Economic Research. [2]

Moon, H. R., and Schorfheide, F. (2009), "Estimation with Over Identifying Inequality Moment Conditions," *Journal of Econometrics*, 153, 136–154. [14]

Nordhaus, W. D. (1987), "Forecasting Efficiency: Concepts and Applications," *Review of Economics and Statistics*, 69, 667–674. [2]

Patton, A. J., and Timmermann, A. (2007a), "Properties of Optimal Forecasts under Asymmetric Loss and Nonlinearity," *Journal of Econometrics*, 140, 884–918. [1]

——(2007b), "Testing Forecast Optimality under Unknown Loss," *Journal of the American Statistical Association*, 102, 1172–1184. [14]

—— (2010a), "Monotonicity in Asset Returns; New Tests with Applications to the Term Structure, the CAPM, and Portfolio Sorts," *Journal of Financial Economics*, 98 , 605–625. [3]

——(2010b), "Why do Forecasters Disagree? Lessons from the Term Structure of Cross-sectional Dispersion," *Journal of Monetary Economics*, 57, 803–820. [2]

——(2011), "Predictability of Output Growth and Inflation: A Multi-horizon Survey Approach," *Journal of Business and Economic Statistics*, 29, 397–410. [2]

Pesaran, M. H. (1989), "Consistency of Short-Term and Long-Term Expectations," *Journal of International Money and Finance*, 8, 511–516. [4]

Pesaran, M. H., and Weale, M. (2006), "Survey Expectations," in *Handbook of Economic Forecasting*, eds. G. Elliott, C. W. J. Granger and A. Timmermann, Amsterdam: North Holland, pp. 715–776. [4]

Schmidt, P. (1974), "The Asymptotic Distribution of Forecasts in the Dynamic Simulation of an Econometric Model," *Econometrica*, 42, 303–309. [2]

Sun, H.-J. (1988), "A General Reduction Method for n-Variate Normal Orthant Probability," *Communications in Statistics*, 17(11), 3913–3921. [3]

West, K. D. (1996), "Asymptotic Inference about Predictive Ability," *Econometrica*, 64, 1067–1084. [2]

West, K. D., and McCracken, M. W., "Regression Based Tests of Predictive Ability," *International Economic Review*, 1998, 39, 817–840. [2]

White, H. (2000), "A Reality Check for Data Snooping," *Econometrica*, 68, 1097–1126. [1]

White, H. (2001), *Asymptotic Theory for Econometricians* (2nd ed), San Diego: Academic Press. [8]

Wolak, F. A. (1987), "An Exact Test for Multiple Inequality and Equality Constraints in the Linear Regression Model," *Journal of the American Statistical Association*, 82, 782–793. [1]

——(1989), "Testing Inequality Constraints in Linear Econometric Models," *Journal of Econometrics*, 31, 205–235. [1]

Wooldridge, J. M., and White, H. (1988), "Some Invariance Principles and Central Limit Theorems for Dependent Heterogeneous Processes," *Econometric Theory*, 4, 210–230. [8]

Comment

Dean Croushore
Economics Department, University of Richmond, Richmond, VA 23173-0002, and Federal Reserve Bank of Philadelphia, Philadelphia, PA 19106-1574 (*dcrousho@richmond.edu*)

In the forecasting literature, researchers often seek to determine stylized facts, such as: Are forecasts rational? But forecasts can be characterized in many dimensions and answering the question about whether forecasts are rational may require a multidimensional answer. I think about forecasts in three dimensions: (1) horizon, (2) subsample, and (3) vintage.

One dimension of forecast rationality is the horizon of the forecast. The literature on the rationality of forecasts finds some differences across forecast horizons. Zarnowitz (1985) found that the results of tests for bias vary across horizons with no systematic tendency across variables, using individual forecasts from the ASA-NBER (American Statistical Association–National Bureau of Economic Research) survey (now the Survey of Professional Forecasters, SPF). Similarly, Brown and Maital (1981) found varying bias across horizons for forecasts of variables from the Livingston survey. Generally, the early literature in the 1980s finds many cases of bias in forecasts. However, Keane and Runkle (1990) found convincing evidence of no bias for inflation at short horizons using the individual forecasters in the ASA-NBER survey.

The second dimension of forecast rationality is the subsample. Though researchers seek to find stylized facts, they are thwarted by instabilities in empirical results across subsamples. Croushore (2010) shows how forecast rationality tests using SPF forecasts change dramatically over time, depending on the starting date and ending date of the subsample. For example, Figure 1 shows how the sample ending date affects the results of a rationality test, which is a test that determines whether the mean forecast error is zero. The plot shows the p-value testing the null hypothesis whether the mean forecast error is zero for different subsamples. The line labeled *test for bias before break point* shows the p-values for tests using subsamples that begin in 1971 and end at the date shown on the horizontal axis. The line labeled *test for bias after break point* shows p-values for tests using subsamples that begin at the date shown on the horizontal axis and end at the end of 2008. The idea is that when we look for stylized facts, we are limited by the data available to us. And the starting and ending dates of our samples are often random or occur by happenstance. Suppose the development of the SPF had been delayed 5 or 10 years; then we would have a very different starting date for many of our forecast tests. If the facts we discover are truly stylized facts, then they should not be affected by small changes in the starting or ending dates of our data series. However, a look at Figure 1 suggests that facts about the rationality of SPF inflation forecasts are a function of the subsample. Depending on the exact starting or ending dates of the sample, we reach different conclusions about the rationality of the survey forecasts. Thus, no stylized fact is found that is robust across subsamples.

The third dimension of forecast rationality is the data vintage. Croushore (2011) shows that the results of some forecast rationality tests depend somewhat on the vintage of the data chosen as "actual" to be used to evaluate the accuracy of

Journal of Business & Economic Statistics
January 2012, Vol. 30, No. 1
DOI: 10.1080/07350015.2012.634340

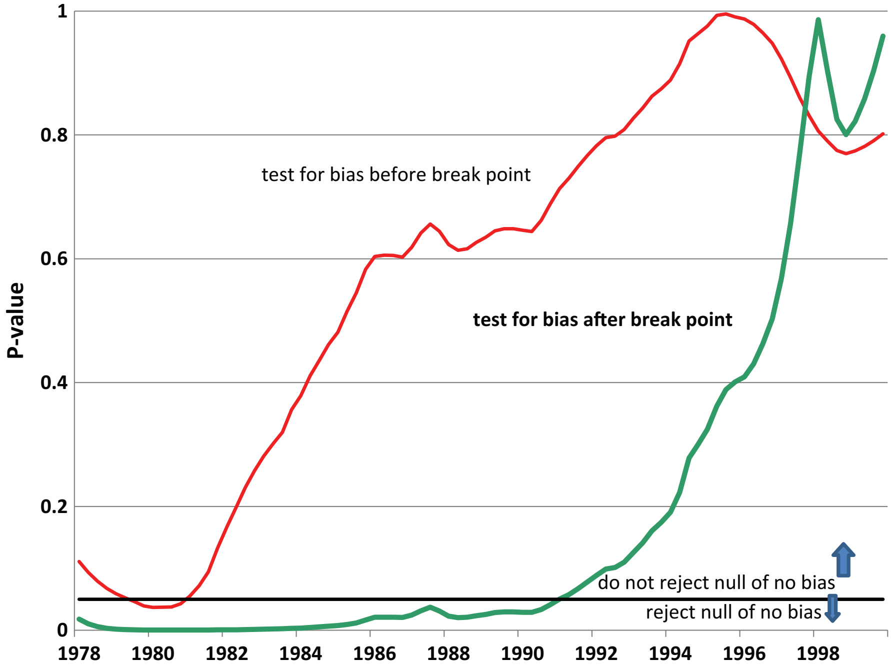


Figure 1. *p*-values for bias at alternative break dates. The plot shows the *p*-value testing the null hypothesis that the mean forecast error is zero for different subsamples. The line labeled *test for bias before break point* shows the *p*-values for tests using subsamples that begin in 1971 and end at the date shown on the horizontal axis. The line labeled *test for bias after break point* shows *p*-values for tests using subsamples that begin at the date shown on the horizontal axis and end at the end of 2008. (Color figure available online.)

forecasts. Many data series are revised for very long periods of time, so how does a researcher choose which measure to use? In the literature, the choices have varied from the release of data two months after the initial release, to the annual revision, to the last vintage before a benchmark revision, to the latest-available data series. But that seemingly innocuous choice may have a large impact on tests for rationality. For example, Figure 2 shows the sensitivity of the zero-mean forecast error test to the choice of both the starting date of the forecast (shown on the horizontal axis) and the choice of variable used as actual (initial, pre-benchmark, or latest-available). Clearly, not only does the subsample period affect the rationality test, but so does the choice of actual. Choosing the initial actual leads to many more subsamples in which we reject the null hypothesis of no bias than using the other two choices of actuals.

In their article, “Forecast Rationality Tests Based on Multiple-Horizon Bounds,” Patton and Timmermann (2011) handle two of the three dimensions of forecast rationality tests: they look across alternative forecast horizons and they develop tests for which choosing an “actual” is not needed. They do not, however, look at the sensitivity of their results to alternative subsamples.

The Patton–Timmermann article accomplishes two main objectives. First, it uses forecasts across alternative horizons, which is valuable because theory implies restrictions on forecasts across different horizons that can be tested. The use of many different horizons avoids issues about choosing which one horizon to analyze. Second, the article develops some tests for which no choice of actual is necessary, which is valuable in avoiding having to choose a vintage of the data to use as actual. Many researchers struggle with this issue. They often use as actuals the latest-available data, which is convenient, but which may be problematic because of re-definitions and other methodological changes. Alternatively, they must develop a real-time dataset with some version of actual data that are not subject to distortions because of methodological changes if the data they need are not conveniently available in an existing real-time dataset, such as the Philadelphia Fed’s Real-Time Data Set for Macroeconomists (see Croushore and Stark 2001). With the Patton–Timmerman tests, no choice of actual is necessary, so researchers avoid having to make this difficult choice. Forecasts, as well as data that will be revised in the future, are treated in a similar manner.

The article provides tests that are easy to interpret, because they lend themselves to graphical interpretations. For example, Figure 1 in the Patton–Timmermann article shows mean squared errors and variances of forecasts from the Greenbook. The sum of the two components should be constant across horizons if the forecasts are optimal, but the graph shows clearly that is not the case. In addition, the variance of the forecasts should increase with horizon if the forecasts are optimal, but that does not hold for the inflation series, as a quick glance at the figure illustrates. Figure 2 in the Patton–Timmermann article shows plots across horizons of mean squared forecast revisions and the covariance between the forecast and the actual (for this test, an actual must be chosen). Mean squared forecast revisions should increase as a function of horizon if the forecasts are optimal, but that is not the case for GDP growth. The covariance between the forecast and the actual should decrease with horizon if the forecasts are optimal, but that is not true for the GDP deflator.

So, the Patton–Timmermann article has many useful features and is the first to provide us with solid analytical results and easy-to-interpret tests. There are three issues about their methods that are worthy of further investigation: (1) The tests may not provide a researcher with the ability to engage in a forecast improvement exercise. (2) The assumptions of the article may not be valid when major benchmark revisions to the data occur. (3) The conclusions are potentially sensitive to the subsample choice.

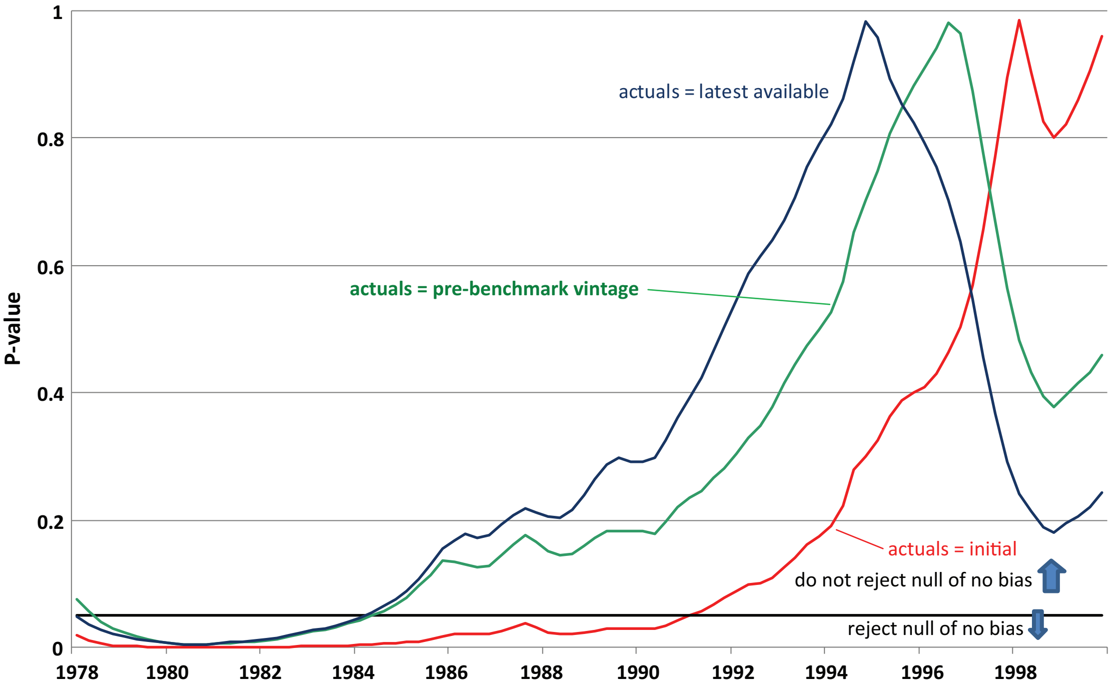


Figure 2. Alternative actuals: p-values for bias after break point. The plot shows the p-value testing the null hypothesis that the mean forecast error is zero for different subsamples and different concepts of actuals. Each line shows p-values for tests using subsamples that begin at the date shown on the horizontal axis and end at the end of 2008. The line labeled $actuals = initial$ shows the p-values for tests using as actuals the value recorded in the initial data release and is the same line as shown in Figure 1. The line labeled $actuals = pre\text{-}benchmark\ vintage$ shows the p-values for tests using as actuals the value recorded in the last vintage before a benchmark revision. The line labeled $actuals = latest\ available$ shows the p-values for tests using as actuals the value recorded in the vintage of May 2011. (Color figure available online.)

The first issue worthy of further investigation is that the tests may not provide a researcher with the ability to engage in a forecast improvement exercise. For example, consider the test discussed earlier for investigating whether mean forecast errors are zero. The mean forecast error is $e_t = x_t^a - x_t^f$, where x_t^a is the actual value and x_t^f is the forecast value. If we run the regression $e_t = \alpha + \varepsilon_t$, we can use the estimated value of α to create an improved forecast: $x_t^i = \hat{\alpha} + x_t^f$, where the improved forecast is x_t^i. Researchers in the 1980s who found bias in forecasts advocated this procedure as a method to reduce forecast errors. Such a test can be used in many different contexts. For example, Faust, Rogers, and Wright (2005), used such a procedure to show how that they can use initial data releases to forecast revisions to GDP in many countries, reducing the mean squared forecast error substantially.

The tests provided by Patton and Timmermann are useful in showing that forecasts are not optimal, but the tests do not lend themselves to forecast improvement possibilities. So, the tests can determine that there is a problem with the forecasts, but provide no guidance about what to do in response. Often in working on forecasts, we observe in-sample predictability of forecast errors, but we are unable to improve the forecasts in a real-time out-of-sample forecast improvement exercise. So, Patton and Timmermann might want to consider how to use their tests to provide guidance to forecasters on how to fix the problems their tests identify.

The second issue worth further investigation is that the assumptions in the article may not be valid under major benchmark revisions to the data. In particular, the monotonicity of mean squared forecast revisions depends on the covariance stationarity of the data series. Under the benchmark revision process, forecast revisions that violate some of the proposed tests could be rational if large benchmark revisions cause a change in the data-generating process. Have such large revisions occurred in practice? It is hard to know for sure, but the Stark plots from Croushore and Stark (2001) are suggestive.

For example, Figure 3 shows the Stark plot for the benchmark revision of GDP in examining the key benchmark revision that occurred in January 1996, which was the benchmark revision in which chain weighting was introduced and in which some government purchases were reclassified as investment. The plot shows the demeaned log differences of GDP before and after the benchmark revision of January 1996. It is a plot of $\log[X(t,b)/X(t,a)] - m$, where $X(t,s)$ is the level of X at date t from vintage s, where $s = a$ or $s = b$, $b > a$, and m is the mean of $\log[X(\tau,b)/X(\tau,a)]$ for all the dates that are common to both vintages a and b. The upward trend in the Stark plot means that later data were revised up more than earlier data. But the downward slope at the beginning and end of the sample shows a more complex pattern. This could cause a lack of covariance stationarity across vintages and violate the conditions under which the monotonicity of mean squared forecast revisions is derived. Some work to ensure that this issue is not sufficient to worry about might be in order for data samples that include major benchmark revisions, such as that in 1996.

The third issue worth considering is that the conclusions could be sensitive to subsample choices. This may be worth investigating so that we do not falsely generalize about results based on the overall sample. Potentially, the tests proposed by Patton and Timmermann could be less sensitive to subsample choice

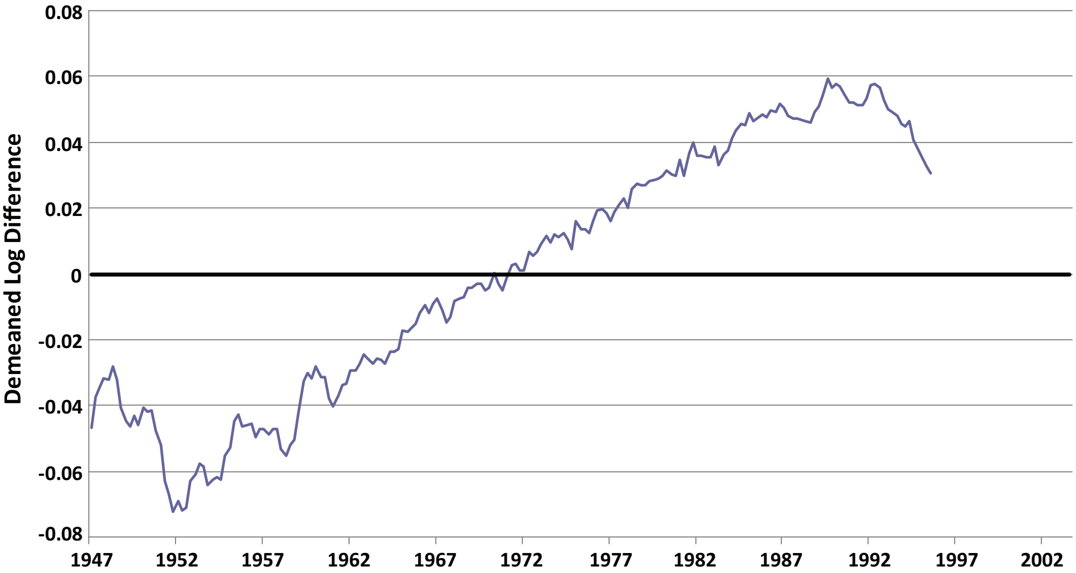


Figure 3. Stark plot across January 1996 benchmark revision. The plot shows the demeaned log differences of GDP before and after the benchmark revision of January 1996. It is a plot of $\log[X(t,b)/X(t,a)] - m$, where $X(t,s)$ is the level of X at date t from vintage s, where $s = a$ or $s = b$, $b > a$, and m is the mean of $\log[X(\tau,b)/X(\tau,a)]$ for all the dates that are common to both vintages a and b. In this plot, $a =$ December 1995 and $b =$ October 1999. (Color figure available online.)

than other tests, including the standard Mincer–Zarnowitz test and the test for zero-mean forecast errors.

To conclude, this article by Patton and Timmermann provides us with an excellent set of tests that can complement much existing research. The tests help us cross two dimensions of forecast rationality: horizon and real-time vintage. They could potentially help as well in the subsample dimension.

[Received April 2011. Revised August 2011.]

REFERENCES

Brown, B. W., and Maital, S. (1981), "What Do Economists Know? An Empirical Study of Experts' Expectations," *Econometrica*, 49, 491–504. [17]

Croushore, D. (2010), "An Evaluation of Inflation Forecasts From Surveys Using Real-Time Data," *B.E. Journal of Macroeconomics: Contributions*, 10, 10. [17]

——— (2011), "Two Dimensions of Forecast Analysis," Working Paper, University of Richmond. [17]

Croushore, D., and Stark, T. (2001), "A Real-Time Data Set for Macroeconomists," *Journal of Econometrics*, 105, 111–130. [18,19]

Faust, J., Rogers, J. H., and Wright, J. H. (2005), "News and Noise in G-7 GDP Announcements," *Journal of Money, Credit, and Banking*, 37, 403–419. [19]

Keane, M. P., and Runkle, D. E. (1990), "Testing the Rationality of Price Forecasts: New Evidence From Panel Data," *American Economic Review*, 80, 714–735. [17]

Patton, A. J., and Timmermann, A. (2011), "Forecast Rationality Tests Based on Multi-Horizon Bounds," *Journal of Business and Economic Statistics*, this issue. [18]

Zarnowitz, V. (1985), "Rational Expectations and Macroeconomic Forecasts," *Journal of Business & Economic Statistics*, 3, 293–311. [17]

Comment

Kajal Lahiri
Department of Economics, University at Albany: SUNY, Albany, NY 12222
(*klahiri@albany.edu*)

1. INTRODUCTION

I enjoyed reading yet another article by Patton and Timmermann (PT hereafter) and feel that it has broken new ground in testing the rationality of a sequence of multi-horizon fixed-target forecasts. Rationality tests are not new in the forecasting literature, but the idea of testing the monotonicity properties of second moment bounds across several horizons is novel and can suggest possible sources of forecasting failure. The basic premise is that since fixed-target forecasts at shorter horizons are based on more information, they should on the average be more accurate than their longer horizon counterparts. The internal consistency properties of squared errors, squared forecasts, squared forecast revisions, and the covariance between the target variable and the forecast revision are tested as inequality

© 2012 American Statistical Association
Journal of Business & Economic Statistics
January 2012, Vol. 30, No. 1
DOI: 10.1080/07350015.2012.634342

constraints across horizons. They also generalize the single-horizon Mincer-Zarnowitz (MZ) unbiasedness test by estimating a univariate regression of the target variable on the longest horizon forecast and all intermediate forecast revisions. Using a Monte Carlo experiment and Greenbook forecasts of four macro variables, PT show that the covariance bound test and the generalized MZ regression using all interim forecast revisions have good power to detect deviations from forecast optimality. I am sure we will be using, extending, and finding caveats with some of the testing proposals suggested in this article for years to come.

2. THEORETICAL CONSIDERATIONS

An important starting point of the article is that for internal consistency of a sequence of optimal forecasts, the variance of forecasts should be a weakly decreasing function of the forecast horizon. This point has been discussed by Isiklar and Lahiri (2007), and was originally the basis for testing whether data revisions are news (and not noise) by Mankiw and Shapiro (1986). In order to highlight the nature of the fixed-target forecast variances, I have plotted in Figure 1 the sequences of monthly real GDP forecasts for 24 target years 1986–2009 from horizons 24 to 1 using consensus forecasts from the Blue Chip surveys. This survey is very well suited to examining the dynamics of forecasts over horizons. The respondents start forecasting in January of the previous year, and their last forecast is reported at the beginning of December of the target year. The shaded bars in the bottom of Figure 1 are the variances of mean forecasts calculated over the target years. Clearly, the variances are nondecreasing functions of horizons and thus the relationship is consistent with rational expectations. Isiklar and Lahiri (2007) explained the relationship by the following logic. Consider $y_t = f_{t,h} + u_{t,h}$ where y_t is the actual GDP growth, $f_{t,h}$ is the h-period ahead forecast ($h = 24, 23, \ldots, 1$) made at time $t - h$, and $u_{t,h}$ denotes the ex-post error associated with this forecast. Since rational expectations imply that $\mathrm{cov}(f_{t,h}, u_{t,h}) = 0$, we have $\mathrm{var}(y_t) = \mathrm{var}(f_{t,h}) + \mathrm{var}(u_{t,h})$, which implies (since for fixed target forecasts, variance of y_t is same for all h) that the variations in forecasts and forecast errors move in opposite directions as the forecast horizon changes. Therefore, as the forecast horizon decreases, the forecast error variability (and therefore the uncertainty) also decreases, but the forecast variability increases. Another way of looking at this increasing variability of forecasts is that as the forecast horizon decreases, more information is absorbed in the forecasts, thus increasing their variability. This information accumulation process can be seen using a simple moving average (MA) data-generating process. Suppose that the actual process has a moving average representation of order q so that $y_t = \mu + \sum_{k=0}^{q} \theta_k \varepsilon_{t-k}$ with $\mathrm{var}(\varepsilon_t) = \sigma^2$. Let $I_{t,h}$ denote the information available at time t-h. Then, the optimal forecast at horizon h will be

$$f_{t,h} \equiv E(y_t | I_{t,h}) = \mu + \sum_{k=h}^{q} \theta_k \varepsilon_{t-k}, \tag{1}$$

and the variance of the forecast is

$$\mathrm{var}(E(y_t | I_{t,h})) = \sigma^2 \sum_{k=h}^{q} \theta_k^2. \tag{2}$$

Similarly, the variance of the forecast when the forecast horizon is $h - 1$ is $\mathrm{var}(E(y_t | I_{t,h-1})) = \sigma^2 \sum_{k=h-1}^{q} \theta_k^2$, so that $\mathrm{var}(f_{t,h-1}) = \mathrm{var}(f_{t,h}) + \theta_{h-1}^2 \sigma^2$.

Thus, when the forecast horizon is very long, that is, several years, the forecasts tend to converge toward the mean of the process, and as information is accumulated, the forecasts change increasing the forecast variability. Figure 1 exhibits this phenomenon very well. Note that for horizons from 24 to 16, the variance seems to remain constant, as was illustrated by Isiklar and Lahiri (2007) for a large number of countries, but with a smaller sample size using the Consensus Survey forecasts. The forecast variability increases because of the variability of the accumulated shocks, that is, $\theta_k \varepsilon_{t-k}$. Therefore, if forecast variability does not change over several long horizons, this may mean that the information acquired at 24 to 16 horizons does not have much impact on the actual value, that is, $|\theta_k|$ is small or equivalently relevant information simply does not exist. Of course, this may also be related to the informational inefficiency of the forecasts. It is possible that even if potentially relevant information over these horizons were available, the forecasters did not incorporate them appropriately causing less than optimal variability in the forecasts. The point here is that due to the nonmonotone arrival and use of information by forecasters at different horizons, the monotonicity properties of second moments like the forecast variance that PT exploit may be less obliging for the detection of forecast suboptimality.

The first difference in the MSE_h provides a measure of the new information content of forecasts when the horizon is h. On the basis of Equation (1), an optimal forecast $f_{t,h}$ satisfies

$$\Delta\mathrm{MSE}(f_{t,h}) \equiv \mathrm{MSE}(f_{t,h+1}) - \mathrm{MSE}(f_{t,h}) = \theta_h^2 \sigma^2, \tag{3}$$

which is equivalent to the information content of the new information in the actual process.

Now let $\tilde{I}_{t,h}$ denote a strict subset of $I_{t,h}$, and $\tilde{f}_{t,h}$ be a suboptimal forecast, which is generated according to

$$\tilde{f}_{t,h} \equiv E(y_t | \tilde{I}_{t,h}) = \tilde{\mu} + \sum_{k=h}^{q} \tilde{\theta}_k \tilde{\varepsilon}_{t-k}, \tag{4}$$

where q denotes the longest forecast horizon at which the first fixed-target forecast is reported—it defines the conditional mean of the actual process when the horizon is q, that is, $\tilde{\mu} = E(y_t | \tilde{I}_{t,q})$; $\tilde{\varepsilon}_{t-h}$ denotes the "news" component used by the forecaster, and $\tilde{\theta}_h$ denotes the impact of this news component as perceived by the forecaster.

For convenience, let us assume that the forecasters observe the news ε_{t-h} correctly, but that their utilization of news is not optimal, so that $\tilde{\theta}_h \neq \theta_h$ and $\tilde{\varepsilon}_{t-h} = \varepsilon_{t-h}$. Thus, we see that the forecast errors follow:

$$y_t - \tilde{f}_{t,h} = (\mu - \tilde{\mu}) + \sum_{k=h}^{q} (\theta_k - \tilde{\theta}_k)\varepsilon_{t-k} + \sum_{k=0}^{h-1} \theta_k \varepsilon_{t-k}, \tag{5}$$

where the first component on the right-hand side (RHS) denotes the bias in the forecast, the second component denotes the error due to inefficiency, and the third component denotes the error due to unforecastable events after the forecast is reported. Calculating mean squared error (MSE) and assuming that sample estimates converge to their population values, we

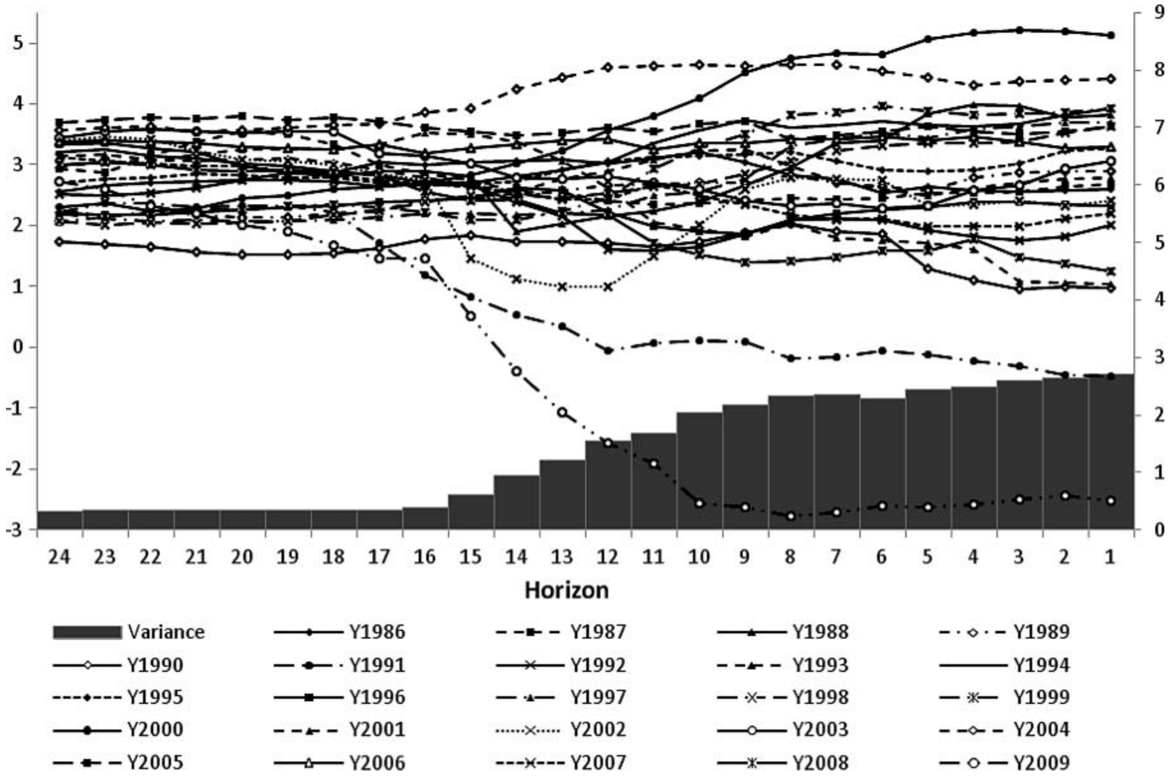


Figure 1. Evolution of fixed target forecasts over horizons and their variances.

get

$$\text{MSE}_h = (\mu - \tilde{\mu})^2 + \sum_{k=h}^{q} (\theta_k - \tilde{\theta}_k)^2 \sigma^2 + \sum_{k=0}^{h-1} \theta_k^2 \sigma^2. \qquad (6)$$

Thus, we find that $\Delta\text{MSE}_h \equiv \text{MSE}_{h+1} - \text{MSE}_h$ is

$$\Delta\text{MSE}_h = \theta_h^2 \sigma^2 - (\theta_h - \tilde{\theta}_h)^2 \sigma^2, \qquad (7)$$

which gives the improvement in forecast content with the new information. The first element on the RHS represents the maximum possible improvement in the quality of forecasts if the information is used efficiently, but the second component represents the mistakes in the utilization of the new information. If the usage of the most recent information $\tilde{\theta}_h$ differs from its optimal value θ_h, the gain from the utilization of new information will decrease and result in excess variability in the forecasts given by the second term. In the special case where $\tilde{\theta}_h = \theta_h$, Equation (7) is equivalent to Equation (3). In this case, ΔMSE_h will measure the content of new information in the actual process, which is simply $\theta_h^2\sigma^2$. Note, however, that a nonnegative MSE differential is compatible with the situation where $\tilde{\theta}_h \neq \theta_h$ for a wide range of parameter values. This underscores the point that the bounds derived by PT are implied by forecast rationality, and hence are not necessary conditions. In other words, if these tests reject the null, we have evidence against forecast rationality, but if the tests do not reject, we cannot say we have evidence in favor of rationality. The issue is whether extant forecast efficiency tests, like those due to Nordhaus (1987) or Davies and Lahiri (1999), would detect forecast irrationality under the latter scenario.

While the use of ΔMSE_h provides an estimate of the improvement in forecasting performance at horizon h in an ex-post sense, a similar measure can be constructed based solely on forecasts without using the actual data on the target variable—a point emphasized by PT. Notice that, based on Equation (1), the optimal forecast revision $r_{t,h} \equiv f_{t,h} - f_{t,h+1}$ is nothing but $r_{t,h} = \theta_h \varepsilon_{t-h}$. In the suboptimal case of Equation (4), we have the forecast revision process $r_{t,h} = \tilde{\theta}_h \varepsilon_{t-h}$. Calculating the mean squared revisions (MSRs) at horizon h and taking the probability limit, we get $\text{MSR}_h = p\lim_T \frac{1}{T} \sum_{t=1}^{T} r_{t,h}^2 = \tilde{\theta}_h^2 \sigma^2$, which provides a measure for the reaction of the forecasters to news. But since forecasters react to news based on their perceptions of the importance of the news, this measure can be seen as the content of the new information as perceived by the forecasters in real time. Note the clear difference between ΔMSE_h and MSR_h. While the former is driven by the forecast errors, the latter has nothing to do with the actual process or the outcomes. But both of the measures should give the same values if the survey forecasts are optimal. This result was originally pointed out by Isiklar and Lahiri (2007).

The difference between MSR_h and ΔMSE_h may provide important behavioral characteristics of the forecasters such as over or underreaction to news at a specific forecast horizon. MSR_h can be seen as a measure of how forecasters interpret the importance of news at a specific horizon, and ΔMSE_h can be seen as the “prize” they get as a result of revising their forecasts. Suppose that forecasters make large revisions at horizon h, but the performance of the forecasts does not improve much at that horizon, then one may conjecture that the forecasters react excessively to the news. To see this more clearly, simple algebra yields $\text{MSR}_h - \Delta\text{MSE}_h = 2(\tilde{\theta}_h^2 - \theta_h \tilde{\theta}_h)\sigma^2$, which is positive

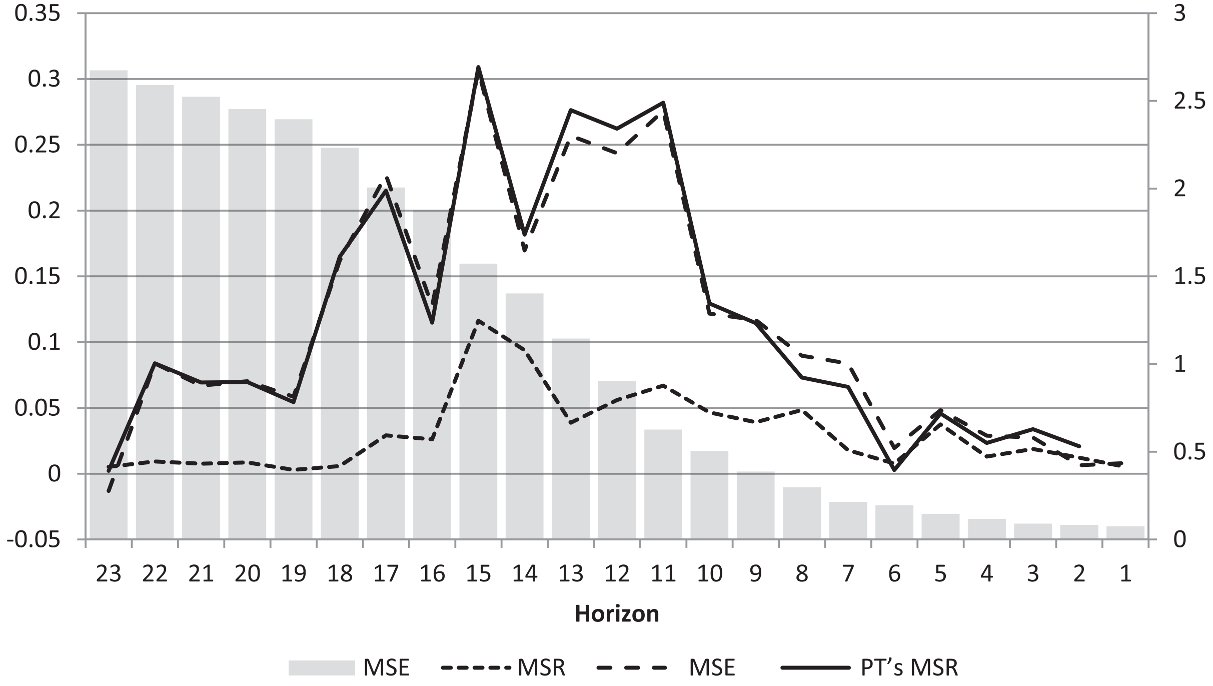


Figure 2. MSE, MSF, and their differentials, Blue Chip Consensus Forecasts 1986–2009.

when $\tilde{\theta}_h^2 > \theta_h \tilde{\theta}_h$. This is the same as the condition $|\tilde{\theta}_h| > |\theta_h|$. But $|\tilde{\theta}_h| > |\theta_h|$ is equivalent to overreaction to the news when the horizon is h. Thus, forecast optimality can be tested by the equality between MSR_h and $\Delta\mathrm{MSE}_h$. This concept is the basis of the amended Nordhaus-type rationality test suggested by Lahiri and Sheng (2008) in which forecast error is regressed on the latest forecast revision for testing its significance. Note that PT's bounds test on the MSRs is defined slightly differently as the difference between $f_{t,1} - f_{t,\,m}$ and $f_{t,1} - f_{t,\,m-1}$ where m is an intermediate horizon.

3. ADDITIONAL SURVEY EVIDENCE

In order to visualize the differences in MSR_h as we defined above, $\Delta\mathrm{MSE}_h$ and PT's MSR, we plotted these values in Figure 2, calculated from the Blue Chip consensus forecasts over 1986–2009. The actual values are the first available real-time data obtained from the Philadelphia Fed's real-time database. First note that PT's MSR and $\Delta\mathrm{MSE}_h$ are almost identical at most horizons because when the shortest horizon forecast is very close to the actual value, as is actually the case, the two measures will be very similar. Second, both are nonnegative at all horizons suggesting forecast rationality by the PT criteria. However, we find a substantial wedge between MSR_h and $\Delta\mathrm{MSE}_h$ particularly in the middle horizons that suggests substantial underreaction to news, and hence inefficiency. Thus, PT's MSE and MSR differential bounds conditions are not stringent enough to detect inefficiency in the Blue Chip consensus forecasts. This is consistent with the evidence they report. Note, however, with the same Blue Chip data series over 1977–2009, the Nordhaus test readily rejects rationality over multiple horizons with adjusted R^2 in excess of 0.20 and the coefficient on the lagged forecast revision around 0.58. This result is valid over different sample periods 1977–2009 and 1986–2009, over all horizons and also over horizons between 16 and 6. Note that even though PT did not consider the Nordhaus test in their experiments, their bounds test that the variance of the forecast revision should not exceed twice the covariance between the forecast revision and the actual value is effectively the Nordhaus test in disguise because, given that the longer horizon forecasts have larger MSEs than shorter horizon forecasts, this particular PT condition is derived using the Nordhaus condition that forecast errors should be uncorrelated with forecast revisions under forecast efficiency (see their proof of Corollary 4 in the appendix).

We also experimented with the extended MZ regression using consensus Blue Chip real GDP forecasts from 1977 to 2009. Data on horizons 16 through 7 are available throughout the Blue Chip sample. PT's univariate optimal revision regression generalizing the MZ regression rejected the null that the intercept is zero and that the coefficients of the horizon 16 forecast and the series of intermediate forecast revisions are one with the p-value of 0.07. But all individual MZ regressions accepted the unbiasedness hypothesis with p-values in excess of 0.5. This result is very similar to what PT found with Greenbook forecasts on real GDP. However, the high multicollinearity between successive revisions tends to make this regression highly unstable, particularly when forecasts on a large number of horizons are available. Thus, one should be careful while using this test—the conclusions using this test may depend on the horizons included in the extended regression.

I also used another rich survey panel dataset—the U.S. Survey of Professional forecasters (SPF)—over 1968Q4–2011Q1 using three primitive forecasts of individuals (ID nos 40, 65, and 85) each having more than 100 quarters of participation, and also all forecasters who participated at least 10 times yielding a total of 425 forecasters in the "all" group. I used real GDP forecasts for six available horizons—beginning with the current quarter. Various forecast statistics are reported in Table 1. The actual

Table 1. Real GDP forecast error statistics for SPF data

	Forecaster ID no. 40			Forecaster ID no. 65			Forecaster ID no. 85			All		
Horizon quarter	MSE	MSR	ΔMSE	MSE	MSR	ΔMSE	MSE	MSR	ΔMSE	MSE	MSR	ΔMSE
1	7.295	2.997	0.788	4.447	5.645	4.62	4.572	2.283	2.66	7.083	7.203	3.793
2	8.083	1.765	3.442	9.067	2.962	2.066	7.232	1.789	−0.789	10.876	6.650	2.444
3	11.525	0.766	1.147	11.133	2.927	0.474	6.443	1.326	−0.148	13.320	6.696	3.543
4	12.672	1.495	−4.129	11.607	4.934	3.69	6.295	2.142	5.057	16.863	5.601	−0.065
5	8.543	-	-	15.297	-	-	11.352	-	-	16.798	-	-

GDP values are again the real-time figures released one month after the end of the quarter. Here, we find very few negative ΔMSE_h or MSR values that would suggest inefficiency. Only the MSE differential for forecaster 40 between quarter 5 and quarter 4 is substantially negative, suggesting forecast suboptimality. However, this evidence of inconsistency can be a result of the arrival of the current year's real GDP value for predicting the first quarter GDP growth for the next year. More generally, relevant information regarding different target values may arrive at different times in a nonmonotone manner, and as a result, the relative forecast accuracy over horizons may not be smooth. PT's approach of pooling all horizons together to test forecast rationality can mask this important horizon-specific heterogeneity in forecast efficiency. In other words, forecasts may be efficient at certain horizons but not at others.

In Figure 3, we have plotted total sum of squares of forecast revisions (defined as $S_h^t = \sum_{i=1}^{N_h} \sum_{t=1}^{t_h^i} \frac{(r_{t,h}^i - \bar{\bar{r}}_h)^2}{\sum_{i=1}^{N_h} T_h^i}$, where i refers to the ith forecaster, $r_{t,h}^i = f_{t,h}^i - f_{t,h+1}^i$, $\bar{\bar{r}}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} \frac{1}{T_h^i} \sum_{t=1}^{T_h^i} r_{t,h}^i$, and N_h and T_h^i denote the available observations) at horizons from 23 to 1 to illustrate that the maximum amount of revisions take place in the middle horizons, and interestingly, the maximum amount of underreaction to news takes place at these horizons too. See Lahiri and Sheng (2010) for additional evidence on this point. These results mean that the efforts put forward by forecasters to produce serious forecasts vary by horizons, and the process begins seriously at around 16-month horizon. Thus, forecasting efforts and the resulting efficiency are also conditioned by the institutions' requirements under which the forecasters operate. To truly understand the forecasting inefficiencies, the demand side of the forecasting market should also be considered, in addition to the schedule of official data announcements.

There are a few more such negative (though small) MSE differentials in Table 1. PT also found a similar result with respect to Greenbook real GDP forecasts; see also Clements et al. (2007). We find that the MSE differentials and MSRs are quite different particularly in the middle horizons and the latter tend to underestimate the former, suggesting underreaction to new information. Following PT, we also calculated MSR differentials between $f_{t,2} - f_{t,4}$ and $f_{t,2} - f_{t,3}$, and $f_{t,3} - f_{t,5}$ and $f_{t,3} - f_{t,4}$; and between $f_{t,4} - f_{t,6}$ and $f_{t,4} - f_{t,5}$ for the three long-standing forecasters and also for the "all' group using disaggregate data. In none of the cases did we find any evidence of negative MSR differentials and thus we fail to detect any indication of irrationality based on this MSR criterion. However, the Nordhaus test and the regressions of forecast errors on forecast revisions, *a la* Lahiri and Sheng (2008, 2010), readily detected deviations from rationality in most cases.

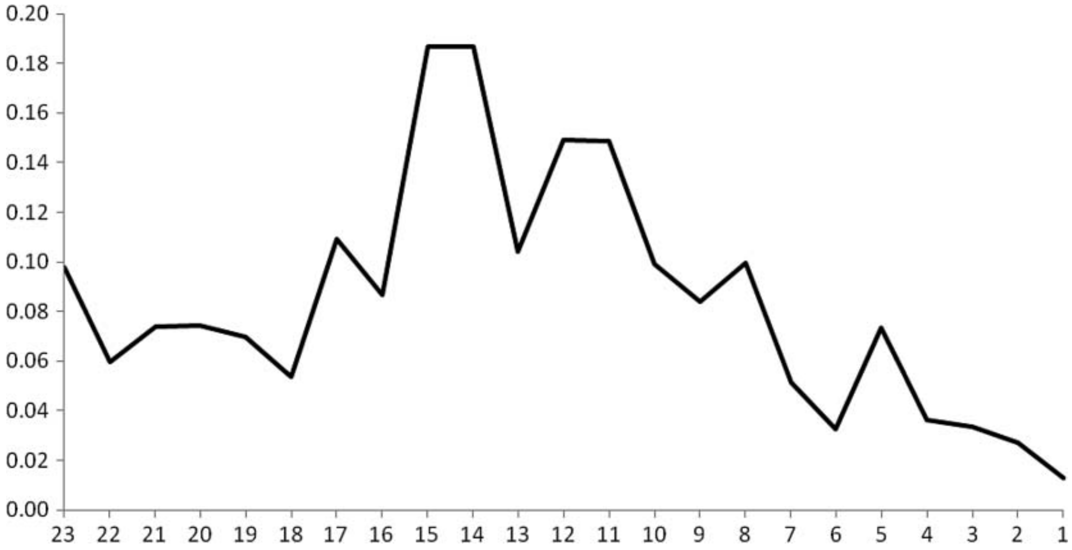


Figure 3. Total sum of squares in GDP forecast revisions during 1986–2009: Blue Chip Surveys.

4. CONCLUSION

I find the ideas put forward in this article to rigorously test the optimality of forecasts across horizons interesting and the efforts to implement the tests quite commendable. Despite their mathematical elegance, these derived bounds are implied by forecast rationality, and hence are not necessary conditions. Thus, if the tests do not reject the null, we cannot say we have evidence in favor of rationality. For instance, even though the forecast variances and MSEs are observed to be weakly decreasing functions of the forecast horizon, the underlying forecasts can be easily be still inefficient. By using Blue Chip and SPF survey forecasts, we found, like PT, that some of the bounds tests proposed in the paper are not very powerful to detect suboptimality in instances where the extant Nordhaus test readily identifies it. However, their new extended MZ test based on a regression of the target variable on the long-horizon forecast and the sequence of interim forecast revisions works well, provided the multicollinearity problem does not become serious. We have argued that by testing the equality of MSE differentials with mean square forecast revisions, one can also examine forecast rationality over multiple horizons. In order to truly understand the pathways through which forecasts fail to satisfy forecast optimality, we have also to consider the demand side of the forecasting market and the schedule of official data announcements. For instance, there is evidence that forecasters record maximum suboptimality at horizons where they also make maximum forecast revisions. The observed suboptimality of Greenbook forecasts that PT found cannot be understood unless the institutional requirements of such forecast are appreciated. Nevertheless, the importance of testing the joint implications of forecast rationality across multiple horizons when such information is available as proposed by the authors must be appreciated.

ACKNOWLEDGMENTS

The author thanks Antony Davies, Gultikin Isiklar, Huaming Peng, Xuguang Sheng, and Yongchen Zhao for many useful discussions and research assistance, and to the co-editors for making a number of helpful comments.

[Received June 2011. Revised July 2011.]

REFERENCES

Clements, C. P., Joutz, F., and Stekler, H. (2007), "An Evaluation of Forecasts of the Federal Reserve: A Pooled Approach," *Journal of Applied Econometrics*, 22, 121–136. [24]

Davies, A., and Lahiri, K. (1999), "Re-examining the Rational Expectations Hypothesis Using Panel Data on Multi-period Forecasts," in *Analysis of Panels and Limited Dependent Variables*, eds. C. Hsiao, K. Lahiri, L. F. Lee, and H. Pesaran, Cambridge, U.K.: Cambridge University Press, pp. 226–254. [22]

Isiklar, G., and Lahiri, K. (2007), "How Far Ahead can We Forecast? Evidence from Cross-country Surveys," *International Journal of Forecasting*, 23, 167–187. [21,22]

Lahiri, K., and Sheng, X. (2008), "Evolution of Forecast Disagreement in a Bayesian Learning Model," *Journal of Econometrics*, 144, 325–340. [23,24]

Lahiri, K., and Sheng, X. (2010), "Learning and Heterogeneity in GDP and Inflation Forecasts," *International Journal of Forecasting*, 26, 265–292. [24]

Mankiw, N. G., and Shapiro, M. D. (1986), "News or Noise? An Analysis of GNP Revisions," *Survey of Current Business*, 66, 20–25. [21]

Nordhaus, W. (1987), Forecasting Efficiency; Concepts and Applications, *Review of Economics and Statistics*, 69, 667–674. [22]

Comment

Barbara Rossi
Department of Economics, 204 Social Science Building, Duke University, Durham, NC 27708, ICREA, and University Pompeu (*brossi@econ.duke.edu*)

Patton and Timmermann (2011) propose new and creative forecast rationality tests based on multi-horizon restrictions. The novelty is to consider the implications of forecast rationality jointly across the horizons. They focus on testing implications of forecast rationality such as the fact that the mean squared forecast error should be increasing with the forecast horizon (Diebold 2001; Patton and Timmermann 2007) and that the mean squared forecast should be decreasing with the horizon. They also consider new regression tests of forecast rationality that use the complete set of forecasts across all horizons in a univariate regression, which they refer to as the "optimal revision regression" tests. One of the advantages of the proposed procedures is that they do not require researchers to observe the target variable, which sometimes is not clearly available. In fact, Patton and Timmermann (2011) show that both their inequality results as well as the "optimal revision regression" test hold when the short horizon forecast is used in place of the target variable. Their work is an excellent contribution to the literature.

The main objective of this comment is to check the robustness of forecast rationality tests to the presence of instabilities. The existence of instabilities in the relative forecasting performance of competing models is well known [see Giacomini and Rossi (2010) and Rossi and Sekhposyan (2010), among others; Rossi (2011) provide a survey of the existing literature on forecasting in unstable environments]. First, we show heuristic empirical evidence of time variation in the rolling estimates of the coefficients of forecast rationality regressions. We then use fluctuation rationality tests, proposed by Rossi and Sekhposyan (2011), to test for forecast rationality, while, at the same time, being robust to instabilities. We also consider a version of Patton and Timmermann's (2010) optimal revision

© 2012 American Statistical Association
Journal of Business & Economic Statistics
January 2012, Vol. 30, No. 1
DOI: 10.1080/07350015.2012.634343

Table 1. Full-sample Mincer and Zarnowitz (1969) forecast rationality tests

	GDP deflator inflation			Output growth		
h	α	β	Joint	α	β	Joint
0	1.1600 (0.5599)	5.2183 (0.0736)	8.9164 (0.0116)	3.9022 (0.1421)	2.2612 (0.3228)	4.2239 (0.1210)
1	0.5753 (0.7500)	0.3029 (0.8594)	4.5608 (0.1022)	13.5417 (0.0011)	17.8056 (0.0001)	18.0331 (0.0001)
2	0.5275 (0.7681)	0.5369 (0.7646)	9.3500 (0.0093)	8.3291 (0.0155)	8.5530 (0.0139)	9.0285 (0.0110)
3	0.0277 (0.9862)	1.4636 (0.4810)	8.4619 (0.0145)	6.5810 (0.0372)	4.0920 (0.1292)	7.2051 (0.0273)
4	0.1992 (0.9052)	3.6437 (0.1617)	12.2100 (0.0022)	1.6620 (0.4356)	0.7996 (0.6705)	3.0507 (0.2175)
5	2.2151 (0.3304)	9.1047 (0.0105)	14.8061 (0.0006)	0.1852 (0.9116)	0.0045 (0.9977)	1.5906 (0.4515)

NOTE: Full-sample Mincer and Zarnowitz (1969) regression, Equation (1). P-values based on HAC robust estimates (with bandwidth equal to 3) for testing $\alpha = 0$ (column labeled "α"), $\beta = 1$ (column labeled "β "), and both $\alpha = 0$ and $\beta = 1$ (column labeled "Joint") in parentheses.

regression test robust to instabilities, which we will refer to as the "fluctuation revision" test. Finally, we discuss the empirical evidence.

We focus on the same data as in Patton and Timmermann (2010), which include the Federal Reserve "Greenbook" forecasts of quarter-over-quarter rates of change in GDP and the GDP deflator. The data are from Faust and Wright (2009), starting in 1980 and ending in 2002.

First, we consider the typical "full-sample" Mincer and Zarnowitz (1969) forecast rationality test. Let the target value to be forecasted at time t using information up to time $t-h$ be y_t and let the forecast be denoted by $y_{t|t-h}$. The Mincer and Zarnowitz (1969) regression is as follows:

$$y_t = \alpha + \beta y_{t|t-h} + \varepsilon_{t,h}, \quad t = 1, \ldots, P, \tag{1}$$

where P is the number of out-of-sample forecasts, h is the forecast horizon, and $\varepsilon_{t,h}$ is the residual. If the forecasts are unbiased, the constant α should be statistically insignificantly different from zero; if the forecasts are optimal, the slope β should be statistically insignificantly different from unity. The null hypothesis of forecast rationality is that $\alpha = 0$ and $\beta = 1$, jointly. Table 1 reports the results. The table shows that forecast rationality is rejected at the 5% significance level for the GDP deflator inflation at most horizons, and it is rejected at horizons 1–3 for GDP growth.

However, the estimates of α and β may not be stable over time. The presence of instability is a serious concern, since it would imply that typical forecast rationality tests are invalid. See Rossi (2005) for an intuitive discussion of why full-sample tests are invalid in the presence of instabilities. To provide informal evidence, Figure 1 reports estimates of α and β in rolling regressions, using a window of 60 out-of-sample forecast observations. The x-axis is the time of the latest forecast included in the rolling regression sample. Figure 1(A) shows that the estimates of α and β in regression (1) for the GDP deflator forecasts are quite unstable over time: α is closer to 0 and β is closer to 1 in the late 1990s than in the mid-1990s. Similarly, Figure 1(B) shows that parameter estimates for GDP growth forecasts are also quite unstable over time. This evidence is only suggestive, though, since it ignores parameter estimation uncertainty.

In what follows, we will consider formal tests to investigate whether the empirical evidence in favor of the rejection of rationality in the Greenbook forecasts may depend on the sample period. We use the Fluctuation Rationality test developed by Rossi and Sekhposyan (2011), which is designed to test forecast rationality in unstable environments. Consider the general regression

$$v_t = g'_{t-h} \cdot \theta + \eta_{t,h}, \quad t = 1, \ldots, P, \tag{2}$$

where θ is a $(k \times 1)$ parameter vector, v_t is the realized variable, g_{t-h} is a $(p \times 1)$ vector of variables known at time $t-h$, and $\eta_{t,h}$ is the residual. Equation (2) corresponds to Equation (1) for $\theta = [\alpha, \beta]'$, $v_t = y_t$, and $g_{t-h} = [1, y_{t|t-h}]$. Consider the following rolling regression. Let $\widehat{\theta}_t$ be the parameter estimate in regression (2) computed over centered rolling windows of size $m = 60$. That is, consider estimating regression (2) using data from $t - m/2$ up to $t + m/2 - 1$, for $t = m/2, \ldots, P - m/2 + 1$. Also, let the Wald test in the corresponding regressions be defined as

$$\mathcal{W}_{t,m} = \left(\widehat{\theta}_t - \theta_0\right)' \widehat{V}_{\theta,t}^{-1} \left(\widehat{\theta}_t - \theta_0\right), \quad \text{for } t = m/2, \ldots, P - m/2 + 1, \tag{3}$$

where $\widehat{V}_{\theta,t}$ is a Heteroscedasticity and Autocorrelation robust (HAC) estimator of the asymptotic variance of the parameter estimates in the rolling windows, see Newey and West (1987). Rossi and Sekhposyan (2011) define the Fluctuation Rationality test as

$$\sup_t \mathcal{W}_{t,m}, \quad \text{for } t = m/2, \ldots, P - m/2 + 1. \tag{4}$$

The test rejects the null hypothesis $H_0 : E(\widehat{\theta}_t) = \theta_0$ for all $t = m/2, \ldots, P - m/2 + 1$ if $\max_t \mathcal{W}_{t,m} > \kappa_{\alpha,k}$, where $\kappa_{\alpha,k}$ are the critical values at the $100\alpha\%$ significance level. The critical values at 5% are reported in table 1 of Rossi and Sekhposyan (2011) for various values of $\mu = [m/P]$ and the number of restrictions, k.

A simple, two-sided t-ratio test on the sth parameter, $\theta_0^{(s)}$, can be obtained as $(\widehat{\theta}_t^{(s)} - \theta_0^{(s)})\, \widehat{V}_{\theta^{(s)},t}^{-1/2}$, where $\widehat{V}_{\theta^{(s)},t}$ is an element in

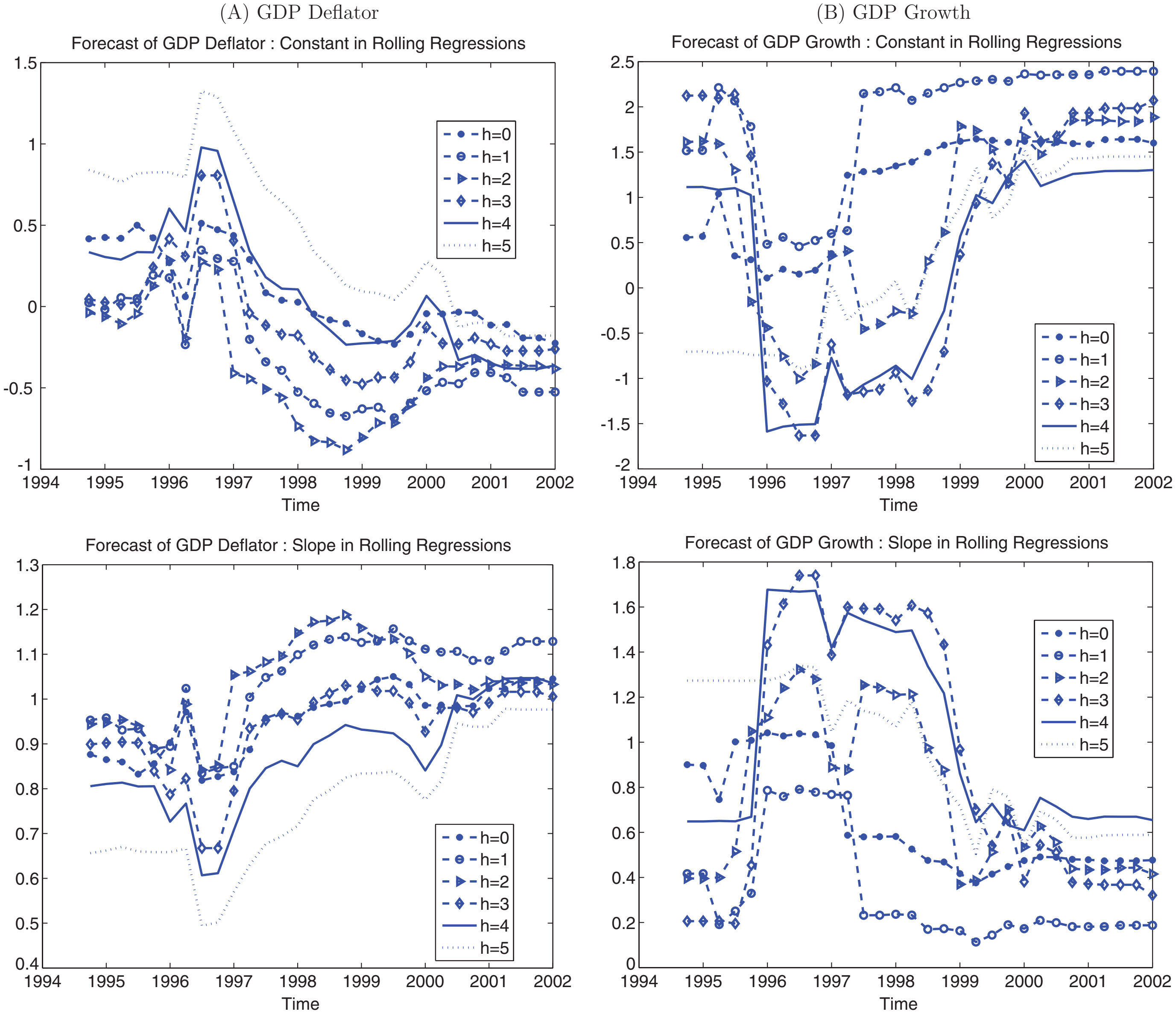


Figure 1. Rolling estimates of parameters of α and β in the Mincer and Zarnowitz (1969) regressions for various horizons (h). (Color figure available online.)

the sth row and sth column of $\widehat{V}_{\theta,t}$. We reject the null hypothesis $H_0 : E(\widehat{\theta}_t^{(s)}) = \theta_0^{(s)}$ for all $t = m/2, \ldots, P - m/2 + 1$ at the 100α% significance level if $\max_t |(\widehat{\theta}_t^{(s)} - \theta_0^{(s)})\widehat{V}_{\theta^{(s)},t}^{-1/2}| > \overline{\kappa}_\alpha$, where $\overline{\kappa}_\alpha$ are the critical values provided by Giacomini and Rossi (2010).

Figure 2 shows fluctuation rationality results for the Mincer and Zarnowitz (1969) regressions for the following cases: for testing $H_0 : \alpha = 0$ and $\beta = 1$ jointly for each horizon (first row); for testing $H_0 : \alpha = 0$ (second row); and for testing $H_0 : \beta = 1$ (third row). The figure also plots 95% confidence bands. The former is a one-sided test, whereas the latter are two-sided t-tests. Specifically, the Mincer and Zarnowitz's regression is $W_{t,m} = (\widehat{\theta}_t - \theta_0)' \, \widehat{V}_{\theta,t}^{-1}(\widehat{\theta}_t - \theta_0)$ for $t = m/2, \ldots, P - m/2 + 1$ and $\theta_0 = [0, 1]'$ (first row); the fluctuation rationality test on the constant is $W_{t,m}^{\alpha} = \widehat{\alpha}_t' \; \widehat{V}_{\alpha,t}^{-1/2}$ (second row) and that on the slope is $W_{t,m}^{\beta} = (\widehat{\beta}_t - 1) \; \widehat{V}_{\beta,t}^{-1/2}$ (third row), where $\widehat{V}_{\alpha,t}$ and $\widehat{V}_{\beta,t}$ are the diagonal elements of $\widehat{V}_{\theta,t}$. The figure shows that forecast rationality is rejected for horizons 0, 2, and 5 for the GDP deflator, and for horizon 1 for the GDP growth.

The framework discussed above can also be generalized to develop a version of the Patton and Timmermann's (2010) optimal revision regression test (implemented with a proxy) robust to instability. We refer to this test as the "fluctuation revision" test. The fluctuation revision test is defined as in Equation (4), where $\mathcal{W}_{t,m}$ is defined by Equation (3), $v_t = y_t$, $g_{t-h} = [1, y_{t|t-h_H}, d_{t|h_1,h_2}, \ldots, d_{t|h_{H-1},h_H}]$, H is the maximum forecast horizon, $d_{t|h_{H-1},h_H}$ denotes the forecast revision between horizons h_{H-1} and h_H, and $\theta_0 = [0, 1, \ldots, 1]'$. Figure 3 shows the test statistic, $\mathcal{W}_{t,m}$, over time; the dotted lines report the 5% critical value. According to the figures, the implications of forecast rationality considered by Patton and Timmermann (2010) are not rejected for the GDP deflator, whereas they are rejected for GDP growth (mainly in the late 1990s).

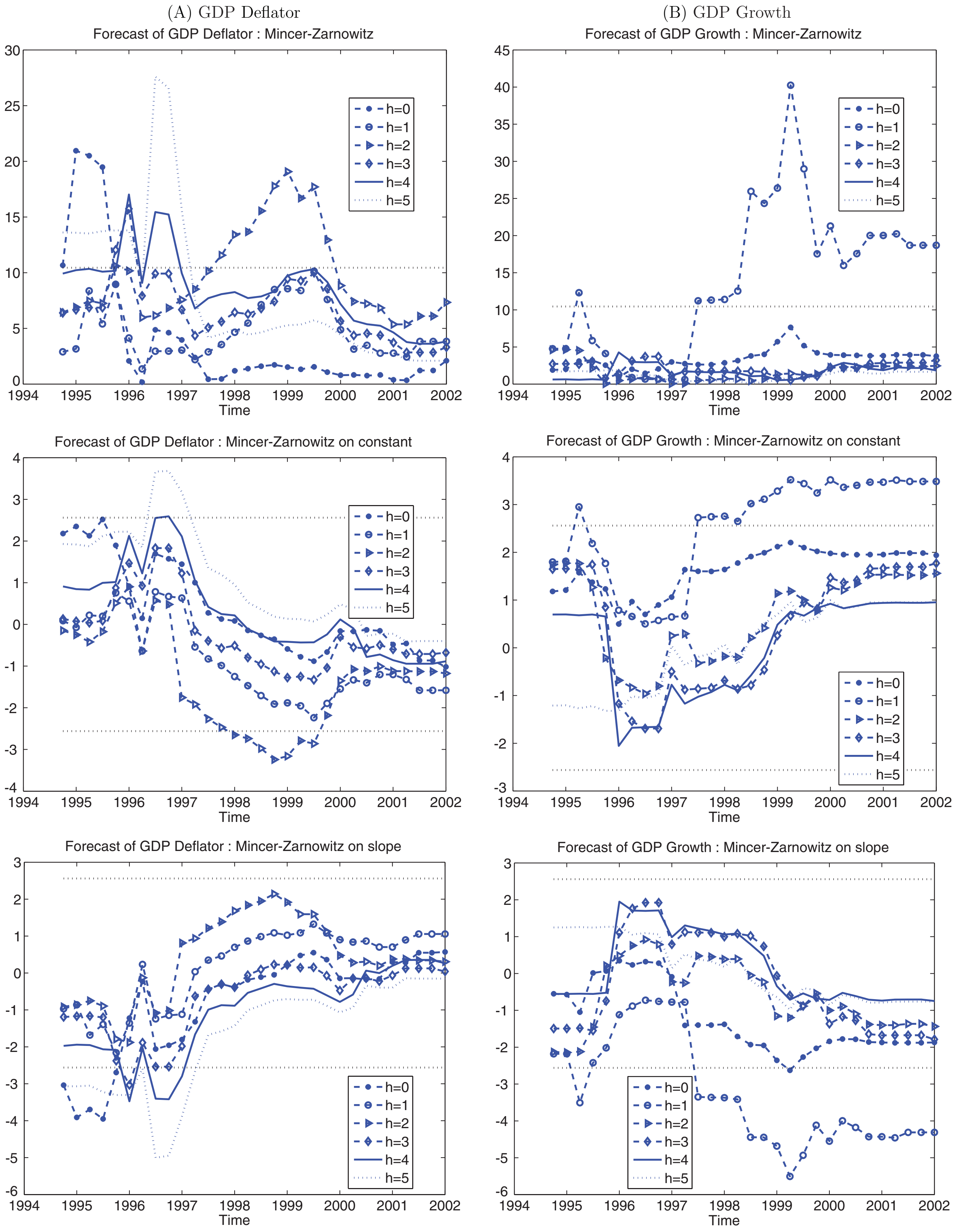
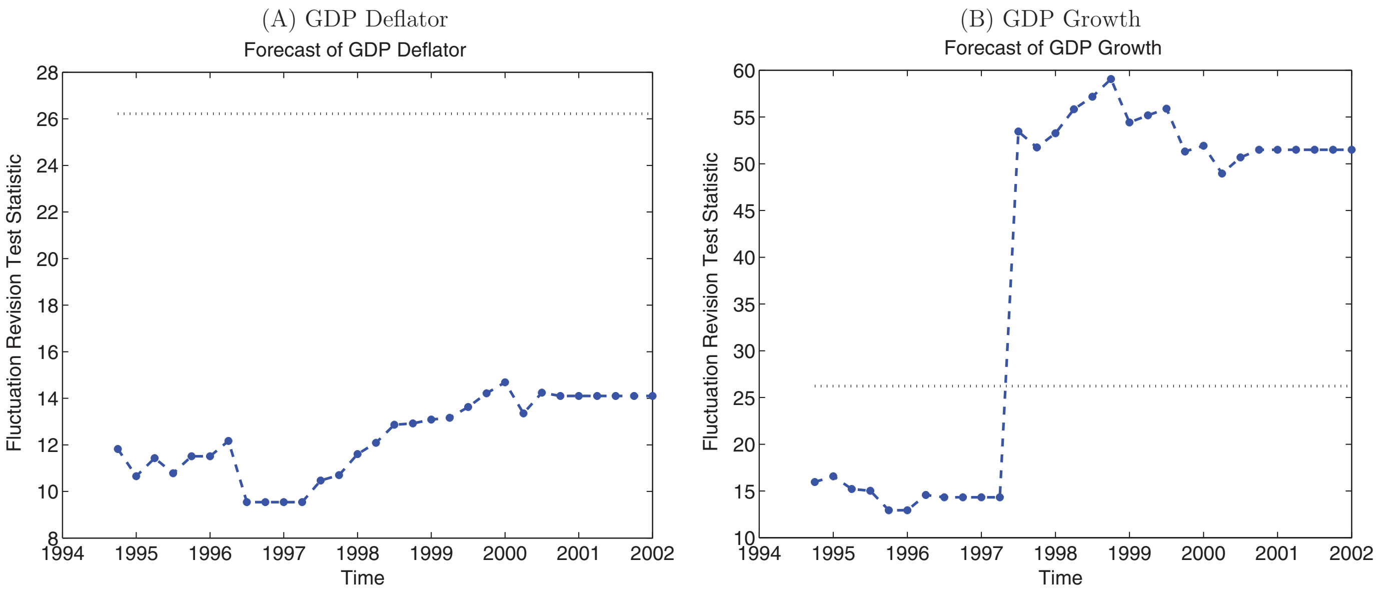


Figure 2. Fluctuation rationality tests for the Mincer and Zarnowitz (1969) regressions for the following cases: for testing $H_0 : \alpha = 0$ and $\beta = 1$ jointly for each horizon (first row); for testing $H_0 : \alpha = 0$ (second row); and for testing $H_0 : \beta = 1$ (third row). In particular, the figure reports Equation (3) together with 95% confidence bands (dotted lines). (Color figure available online.)


(A) GDP Deflator

(B) GDP Growth

Figure 3. Fluctuation revision test over time (line with stars) for GDP deflator data (left-hand side panel) and GDP growth (right-hand side panel), together with the critical value (dotted line). (Color figure available online.)

To conclude, we found empirical evidence in favor of instabilities in the parameters of forecasting rationality regressions. Studying such instabilities might provide useful information as to when rejections of forecast rationality occurred, as well as their possible economic causes.

ACKNOWLEDGMENTS

The author is grateful to A. Patton and J. Wright for sharing their codes and data, and to K. Hirano and J. Wright for organizing the special JBES session at the American Economic Association (AEA).

[Received March 2011. Accepted August 2011.]

REFERENCES

Diebold, F. X. (2001), *Elements of Forecasting* (2nd ed.), Cincinnati, OH: South-Western College Publishing. [25]

Faust, J., and Wright, J. (2009), "Comparing Greenbook and Reduced Form Forecasts using a Large Realtime Dataset," *Journal of Business and Economic Statistics*, 27, 468–479. [26]

Giacomini, R., and Rossi, B. (2010), "Forecast Comparisons in Unstable Environments," *Journal of Applied Econometrics*, 25, 595–620. [27]

Mincer, J., and Zarnowitz, V. (1969), "The Evaluation of Economic Forecasts," in *Economic Forecasts and Expectations*, ed. J. A. Mincer, New York: National Bureau of Economic Research, pp. 81–111. [26]

Newey, W., and West, K. (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708. [26]

Patton, A., and Timmermann, A. (2007), "Properties of Optimal Forecasts Under Asymmetric Loss and Nonlinearity," *Journal of Econometrics*, 140, 884–918. [25]

Patton, A., and Timmermann, A. (2012), "Forecast Rationality Tests Based on Multi-Horizon Bounds," *Journal of Business and Economic Statistics*, forthcoming. [25]

Rossi, B. (2005), "Optimal Tests for Nested Model Selection with Underlying Parameter Instabilities," *Econometric Theory*, 21, 962–990. [26]

Rossi, B. (2011), "Forecasting in Unstable Environments," in *Handbook of Forecasting*, Vol. 2, eds. G. Elliott and A. Timmermann, North Holland: Elsevier. Forthcoming. [25]

Rossi, B., and Sekhposyan, T. (2010), "Have Models' Forecasting Performance Changed Over Time, and When?," *International Journal of Forecasting*, 26, 808–835. [25]

Rossi, B., and Sekhposyan, T. (2011), "Forecast Optimality Tests in the Presence of Instabilities," Mimeo, Duke University. [25]

Comment

Lennart Hoogerheide
Department of Econometrics and Tinbergen Institute, Vrije Universiteit Amsterdam, The Netherlands
(*l.f.hoogerheide@vu.nl*)

Francesco Ravazzolo
Norges Bank, Oslo, Norway (*francesco.ravazzolo@norges-bank.no*)

Herman K. van Dijk
Econometric Institute and Tinbergen Institute, Erasmus University Rotterdam and Department of Econometrics,
Vrije Universiteit Amsterdam, The Netherlands (*hkvandijk@ese.eur.nl*)

Forecast rationality under squared error loss implies various bounds on second moments of the forecasts across different horizons. For example, the mean squared forecast error should be nondecreasing in the horizon. Patton and Timmermann (2011) propose rationality tests based on such restrictions, including interesting new tests that can be conducted *without* having data on the target variable; that is, these tests can be performed by checking only the "internal consistency" of the "term structure" of forecasts.

One of their novel tests that is easily implemented and that performs well in Monte Carlo simulations (in the sense that the actual size is equal to the nominal size and that the power is high) considers the hypothesis of optimal forecast revision in the context of a linear regression of the most recent forecast on the long-horizon forecast and the sequence of interim forecast revisions. That is, it considers the following regression:

$$\hat{Y}_{t|t-1} = \tilde{\alpha} + \tilde{\beta}_H \, \hat{Y}_{t|t-H} + \sum_{j=2}^{H-1} \tilde{\beta}_j \left(\hat{Y}_{t|t-j} - \hat{Y}_{t|t-j-1}\right) + v_t, \quad (1)$$

where the null hypothesis of "rationality" or "optimal revision" corresponds to the hypothesis

$$H_0 : \tilde{\alpha} = 0 \cap \tilde{\beta}_2 = \ldots = \tilde{\beta}_H = 1. \quad (2)$$

Note that the time of the variable to be predicted is "fixed" at time t, while the regressors are the forecasts for this time t "running backward," made at time $t-1$ to $t-H$.

For a simple interpretation of the hypothesis, we rewrite the optimal revision regression (1) as

$$\hat{Y}_{t|t-1} - \hat{Y}_{t|t-2} = \tilde{\alpha} + \tilde{\gamma}_H \, \hat{Y}_{t|t-H} + \sum_{j=2}^{H-1} \tilde{\gamma}_j \left(\hat{Y}_{t|t-j} - \hat{Y}_{t|t-j-1}\right) + v_t, \quad (3)$$

with $\tilde{\gamma}_h \equiv \tilde{\beta}_h - 1$ $(h = 2, \ldots, H)$. In (3) the null hypothesis of "rationality" or "optimal revision" obviously corresponds to the hypothesis

$$H_0 : \tilde{\alpha} = 0 \cap \tilde{\gamma}_2 = \ldots = \tilde{\gamma}_H = 0. \quad (4)$$

One of the attractive properties of this test proposed by Patton and Timmermann (2011) is that it has a clear intuitive interpretation: under the null hypothesis of "*no expected forecast correction*" the last update of the forecast, $\hat{Y}_{t|t-1} - \hat{Y}_{t|t-2}$, does not need to *correct* a bias of $\hat{Y}_{t|t-2}$ $(\tilde{\alpha} = 0)$, or the previous updates $\hat{Y}_{t|t-j} - \hat{Y}_{t|t-j-1}$ ($\tilde{\gamma}_j = 0$ for $j = 2, \ldots, H-1$), or the long-horizon forecast $\hat{Y}_{t|t-H}$ $(\tilde{\gamma}_H = 0)$.

In our comment we address several points. Our main point is to exploit the fact that no actually observed target variable is required and to extend the analysis of Patton and Timmermann to the case of risk measures such as value-at-risk and expected shortfall for which we never observe the *true* value. The tests can also be used for volatility or variance measures.

Consider the following example in which the target variable evolves according to a stationary AR(2) process:

$$Y_t = \phi_0 + \phi_1 \, Y_{t-1} + \phi_2 \, Y_{t-2} + \varepsilon_t, \qquad \varepsilon_t \sim \text{iid}\, N(0, \sigma^2), \quad (5)$$

with $\phi_0 = 0$, $\phi_1 = 0.5$, and $\sigma^2 = 1$. For ϕ_2 we consider several values: $\phi_2 = 0.0, 0.1, 0.2, 0.3$. We estimate a simple AR(1) model, Equation (5) with $\phi_2 = 0$. We simulate 1,000 datasets of 1,500 observations, where the first 1,000 in-sample observations are used for ordinary least squares (OLS) estimation of the parameters $\theta = (\phi_0, \phi_1, \sigma^2)'$ and the last 500 out-of-sample observations are used for evaluation of value-at-risk forecasts. Define $\text{VaR}^{95\%}_{t|t-h}$ as the 5% quantile of the predicted distribution of Y_t at time $t-h$ $(h = 1, 2, \ldots)$:

$$\text{VaR}^{95\%}_{t|t-h} = \hat{Y}_{t|t-h} + \hat{\sigma}\, \Phi^{-1}(0.05)$$

with

$$\hat{Y}_{t|t-h} = \hat{\phi}_0 \, \frac{1 - \hat{\phi}_1^h}{1 - \hat{\phi}_1} + \hat{\phi}_1^h \; Y_{t-h}.$$

These $\text{VaR}^{95\%}_{t|t-h}$ take the role of $\hat{Y}_{t|t-h}$ in (3), which thus becomes

$$\text{VaR}^{95\%}_{t|t-1} - \text{VaR}^{95\%}_{t|t-2} = \tilde{\alpha} + \tilde{\gamma}_H \, \text{VaR}^{95\%}_{t|t-H} + \sum_{j=2}^{H-1} \tilde{\gamma}_j \left(\text{VaR}^{95\%}_{t|t-j} - \text{VaR}^{95\%}_{t|t-j-1}\right) + v_t. \quad (6)$$

© 2012 American Statistical Association
Journal of Business & Economic Statistics
January 2012, Vol. 30, No. 1
DOI: 10.1080/07350015.2012.634348

Table 1. Estimated AR(1) model for simulated data from the AR(2) model

	95% VaR (or 99% VaR)				95% VaR				99% VaR			
	optimal forecast revision ($H = 3$)				UC		CC		UC		CC	
ϕ_2	H_0 : model is correct including estimated parameters		H_0 : model is correct allowing for estimation error in parameters		H_0 : model is correct including estimated parameters							
0.0	0.116	(0.010)	0.046	(0.007)	0.080	(0.009)	0.084	(0.009)	0.081	(0.009)	0.088	(0.009)
0.1	0.490	(0.016)	0.315	(0.015)	0.101	(0.010)	0.090	(0.009)	0.077	(0.008)	0.085	(0.009)
0.2	0.979	(0.005)	0.934	(0.008)	0.105	(0.010)	0.113	(0.010)	0.078	(0.008)	0.080	(0.009)
0.3	1.000	(0.000)	0.999	(0.001)	0.148	(0.011)	0.161	(0.012)	0.097	(0.009)	0.098	(0.009)

NOTE: Percentage of rejections (size or power) at 5% nominal size in the optimal forecast revision regression test, and tests for UC and CC of value-at-risk forecasts. Results are computed for 1,000 simulated datasets. Numerical standard errors are given within parentheses.

Our null hypothesis is not

$$H_0\text{: forecast rationality or optimality under squared error loss} \qquad (7)$$

but

$$H_0\text{: the estimated model for VaR prediction is correct.} \qquad (8)$$

That is, we use the test regression (6) without requiring the assumption of squared error loss. The price for this is that, to the best of our knowledge, one generally has to use simulation from the assumed model to generate the distribution of the F-statistic for the null hypothesis in Equation (4). However, for the AR(1) model with iid $N(0, \sigma^2)$ errors, the errors in (6) are given by $v_t = \phi_1 \varepsilon_{t-1} \sim \text{iid}N(0, \phi_1^2 \sigma^2)$, so that under H_0 the F-statistic has its standard F-distribution. Since $\hat{\sigma}\ \Phi^{-1}(0.05)$ is constant, applying the optimal revision regression test to $\text{VaR}^{95\%}_{t|t-h}$ amounts to the test for $\hat{Y}_{t|t-h}$.

Results for the test (with $H = 3$) are presented in the first column of Table 1. Even if the AR(1) model is true ($\phi_2 = 0.0$), the percentage of rejections (at a nominal size of 5%) is 11.6% (with a numerical standard error of 1.0%). The obvious reason is that there are errors in the parameter estimates. The Monte Carlo simulation by Patton and Timmermann (2011) (with a nominal size of 10%) does not suffer from this phenomenon, as they assume that the process *and its parameter values* are known to forecasters.

If we want to test for the validity of the model, taking into account the presence of errors in parameter estimates, then we must adapt (i.e., increase) the critical value. We propose the following method:

Procedure for optimal revision testing taking into account errors in parameter estimates

Step 1. Compute parameter estimates $\hat{\theta}$ in model for observed time series y (e.g., AR(1) model with $\theta = (\phi_0, \phi_1, \sigma^2)'$); generate forecasts $(1, 2, \ldots, H$ steps ahead); compute F-statistic $F(y)$ in optimal revision regression.

Step 2. Simulate N (e.g., $N = 1{,}000$) data series $y^{(i)}$ $(i = 1, \ldots, N)$—with the same number of observations as those of the observed time series y—from the estimated model with parameters $\hat{\theta}$.

Step 3. Compute parameter estimates $\hat{\theta}^{(i)}$ for each simulated data series $y^{(i)}$ $(i = 1, \ldots, N)$.

Step 4. Generate forecasts $(1, 2, \ldots, H$ steps ahead) for each estimated model with parameters $\hat{\theta}^{(i)}$ and data $y^{(i)}$ $(i = 1, \ldots, N)$.

Step 5. Compute the F-statistic $F(y^{(i)})$ $(i = 1, \ldots, N)$ in optimal revision regression for each set of forecasts from step 4.

Step 6. Compare $F(y)$ with the desired percentile of the sample of *F-statistics under* H_0 $F(y^{(i)})$ $(i = 1, \ldots, N)$ from step 5.

Results for this adapted test (with $H = 3$) are in the second column of Table 1. For $\phi_2 = 0$ the percentage of rejections (at a nominal size of 5%) is 4.6% (with a numerical standard error of 0.7%), so that we have no evidence that the size is distorted.

In order to assess the power of the test, we compare the performance to the well-known unconditional coverage (UC) and conditional coverage (CC) tests for the 95% and 99% value-at-risk; see Kupiec (1995) and Christoffersen (1998). In this example, for the optimal revision regression the test results are the same for each $100(1-\alpha)\%$ value-at-risk with $\alpha \in (0, 1)$. The percentage of rejections for $\phi_2 = 0.1,\ 0.2$, and 0.3 is clearly larger for the optimal revision regression than for the UC and CC tests. Intuitively, this makes sense, since the optimal revision regression uses a large set of forecasts for multiple horizons, whereas the UC and CC tests are only based on the limited information in the set of 0/1 variables that indicate whether the predicted value-at-risk is exceeded by the actual observation. The nominal size for the UC and CC tests is chosen somewhat larger than 5%, as the discrete distributions of the test statistics do not allow for an exact nominal size of 5%. The nominal size is 5.4% and 5.0% (5.3% and 6.4%) in the UC and CC tests for the 95% VaR (99% VaR). The critical values for the UC and CC tests are computed by simulating 100,000 series of iid 0/1 variables under H_0, as the asymptotically valid χ^2 distributions may be rather poor approximations in finite samples, especially for the CC test.

Next, consider the example in which the target variable evolves according to a stationary ARCH(2) process

$$Y_t = \varepsilon_t \sqrt{\sigma_t^2}, \qquad \varepsilon_t \sim \text{iid } N(0,1),$$
$$\sigma_t^2 = \phi_0 + \phi_1 Y_{t-1}^2 + \phi_2 Y_{t-2}^2, \tag{9}$$

with $\phi_0 = 0.5$ and $\phi_1 = 0.5$. For ϕ_2 we consider several values: $\phi_2 = 0.0, 0.1, 0.2, 0.3$. We estimate a simple ARCH(1) model, Equation (9) with $\phi_2 = 0$. Again, we simulate 1,000 datasets of 1,500 observations, where the first 1,000 in-sample observations are used for estimation of the parameters ϕ_0, ϕ_1 and the last 500 out-of-sample observations are used for evaluation of value-at-risk forecasts:

$$\text{VaR}_{t|t-h}^{95\%} = \sqrt{\hat{\sigma}_{t|t-h}^2}\, \Phi^{-1}(0.05)$$

with

$$\hat{\sigma}_{t|t-h}^2 = \hat{\phi}_0 \frac{1 - \hat{\phi}_1^h}{1 - \hat{\phi}_1} + \hat{\phi}_1^h Y_{t-h}^2.$$

Applying the optimal revision regression test to $\text{VaR}_{t|t-h}^{95\%}$ (or any other $100(1-\alpha)\%$ VaR with $\alpha \in (0, 0.5)$) amounts to the test for the standard deviation $\sqrt{\hat{\sigma}_{t|t-h}^2}$. In this case we cannot even use the critical value from the standard F-distribution for the first, "strict" optimal revision test (of validity of the model including the parameter values) for two reasons. First, the regressors in (6) may even have small explanatory power for the regressand if the model is correct. For example, in the ARCH(1) model the regressors have no explanatory power for the regressand in test regression (3) for the variance $\hat{\sigma}_{t|t-h}^2$, but since the VaR is proportional to the standard deviation $\sqrt{\hat{\sigma}_{t|t-h}^2}$ this is not necessarily true. Second, the errors v_t in the optimal revision regression (6) can be substantially non-Gaussian, having a negatively skewed and fat-tailed distribution. The histogram in the top panel of Figure 1 shows the negative skewness of the distribution of the errors v_t for one dataset simulated from the ARCH(1) model. This skewness is caused by the negative skewness of the distribution of the regressand $(\text{VaR}_{t|t-1}^{95\%} - \text{VaR}_{t|t-2}^{95\%})$ in (6); the latter is illustrated by the histogram in the middle panel. The bottom panel shows the reason for the negative skewness of $(\text{VaR}_{t|t-1}^{95\%} - \text{VaR}_{t|t-2}^{95\%})$: $\text{VaR}_{t|t-2}^{95\%}$ is more "moderate" than $\text{VaR}_{t|t-1}^{95\%}$, since $\text{VaR}_{t|t-2}^{95\%}$ is closer to the unconditional VaR. Therefore, $\text{VaR}_{t|t-1}^{95\%}$ is sometimes much more negative than $\text{VaR}_{t|t-2}^{95\%}$, whereas it is often slightly less negative. The result is a distribution of $(\text{VaR}_{t|t-1}^{95\%} - \text{VaR}_{t|t-2}^{95\%})$ that has a positive mode and substantially negative skewness. The small differences between the histograms of the errors v_t and the dependent variable $(\text{VaR}_{t|t-1}^{95\%} - \text{VaR}_{t|t-2}^{95\%})$ reflect that the regressors in (6) have small explanatory power for the regressand, even though the ARCH(1) model is correct. For these reasons, the actual size may be much larger than the nominal size if we would use the critical value from the F-distribution (e.g., an actual size larger than 50% for a nominal size of 5%). Therefore, we require simulation for the critical value in both versions of the optimal revision test. There is also heteroscedasticity for which we use weighted least squares (WLS), assuming $\text{var}(v_t)$ proportional to $\text{var}(y_{t-1})$ (which seems to be a usable approximation). The aim of WLS is to increase the power of the test; the computation of the critical value by simulation already takes care of the size.

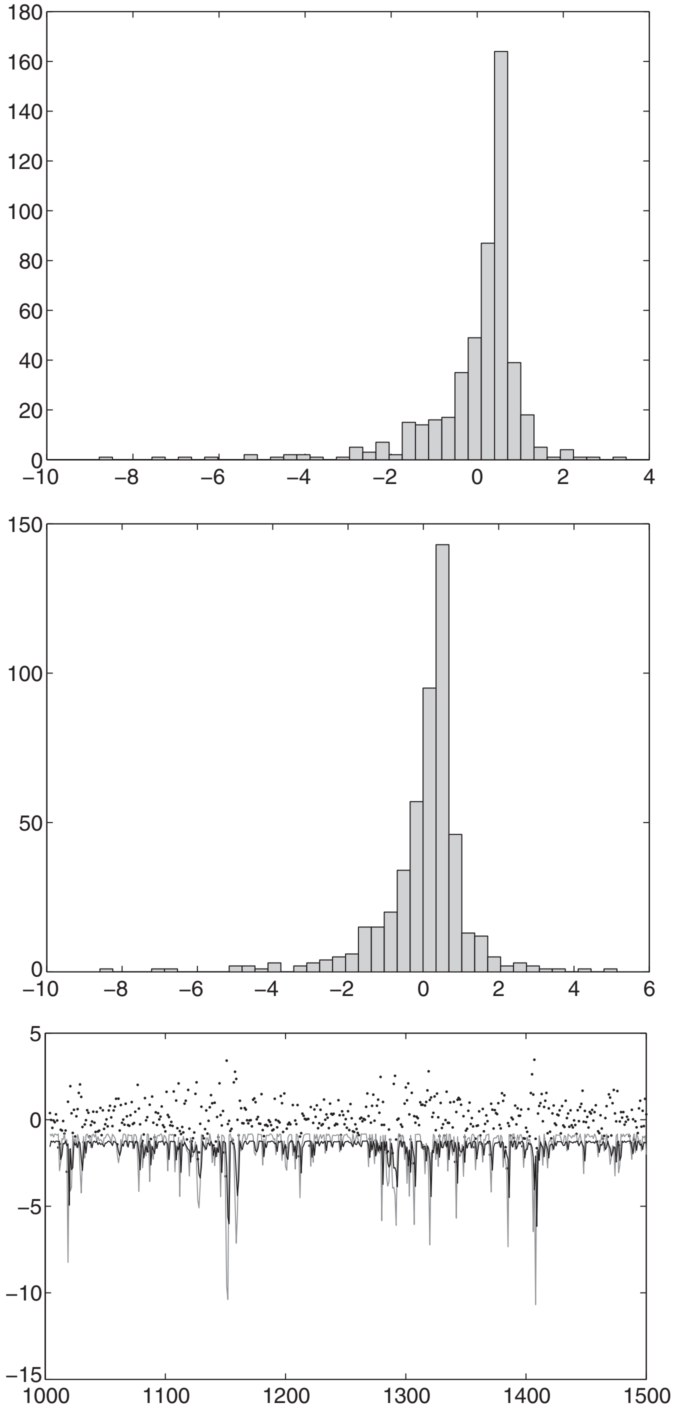


Figure 1. Simulated dataset from the ARCH(1) model: histograms of error terms v_t (top panel) and regressand $(\text{VaR}_{t|t-1}^{95\%} - \text{VaR}_{t|t-2}^{95\%})$ (middle panel) in optimal revision test regression (6); graph of simulated data y_t in out-of-sample period (dots), together with $\text{VaR}_{t|t-1}^{95\%}$ (gray line) and $\text{VaR}_{t|t-2}^{95\%}$ (black line) (bottom panel).

In the first test (of validity of the model including the parameter values) we perform the procedure without step 3, using the "true" parameters $\hat{\theta}$ (instead of $\hat{\theta}^{(i)}$) of our simulated data series in steps 4 and 5. Results are presented in Table 2. Again, the percentage of rejections of the first optimal revision test is larger

Table 2. Estimated ARCH(1) model for simulated data from the ARCH(2) model

	95% VaR (or 99% VaR)				95% VaR				99% VaR			
	optimal forecast revision ($H = 3$)				UC		CC		UC		CC	
ϕ_2	H_0 : model is correct including estimated parameters		H_0 : model is correct allowing for estimation error in parameters		H_0 : model is correct including estimated parameters							
0.0	0.081	(0.009)	0.049	(0.007)	0.087	(0.009)	0.068	(0.008)	0.074	(0.008)	0.084	(0.009)
0.1	0.175	(0.012)	0.111	(0.010)	0.104	(0.010)	0.084	(0.009)	0.083	(0.009)	0.092	(0.009)
0.2	0.386	(0.015)	0.295	(0.014)	0.145	(0.011)	0.112	(0.010)	0.103	(0.010)	0.111	(0.010)
0.3	0.495	(0.020)	0.438	(0.016)	0.197	(0.013)	0.160	(0.012)	0.127	(0.011)	0.132	(0.011)

NOTE: Percentage of rejections (size or power) at 5% nominal size in the optimal forecast revision regression test, and tests for UC and CC of value-at-risk forecasts. Results are computed for 1,000 simulated datasets. Numerical standard errors are given within parentheses.

than 5% for $\phi_2 = 0$, reflecting the effect of errors in parameter estimates. For the second optimal revision test, we do not have evidence that the actual size deviates from 5%. The optimal revision test again has greater power than the UC and CC tests.

In the optimal revision regression test in the AR(1) model a very wrong value of $\hat{\sigma}$ cannot be detected, since the value of $\hat{\sigma}$ does not affect the F-statistic. The UC and CC tests can detect this, which stresses that the optimal revision regression test should preferably be used *in addition to different tests*.

Finally, we discuss two other points. First, if we apply the optimal revision regression test to an *in-sample* window for which the model has been estimated, then a "generated regressor/regressand problem" implies that the F-statistic does not have the standard F-distribution under H_0, even if the errors v_t are normally distributed. For example, in the AR(1) model we have

$$\hat{Y}_{t|t-1} = \hat{\phi}_0 + \hat{\phi}_1\, Y_{t-1},$$

$$\hat{Y}_{t|t-2} = \hat{\phi}_0 \left(1 + \hat{\phi}_1\right) + \hat{\phi}_1^2\, Y_{t-2},$$

$$\hat{Y}_{t|t-1} - \hat{Y}_{t|t-2} = \hat{\phi}_1 \left(Y_{t-1} - \hat{\phi}_0 - \hat{\phi}_1\, Y_{t-2}\right).$$

That is, $\hat{Y}_{t|t-1} - \hat{Y}_{t|t-2}$ equals $\hat{\phi}_1$ times the OLS residual, which is obviously perpendicular to the AR(1) model's regressors, the constant term 1, and Y_{t-2}, if we estimate the optimal revision regression (with $H = 2$) for the same window as the parameters ϕ_0, ϕ_1. Then the estimated coefficients ($\hat{\alpha}$ and $\hat{\gamma}_2$) and F-statistic are exactly equal to 0 for any data series. This reflects that in general the critical values should be smaller if one applies the optimal revision regression test to an in-sample window (or a window that has overlap with an in-sample window).

Bayesian inference may be a useful alternative for testing inequalities of (co)variances or coefficients, which is the focus of alternative tests proposed by Patton and Timmermann (2011), especially for small or moderate data samples. Advantages are that no asymptotic approximations need to be used, and that one does not require "complicated" asymptotic distributions under H_0. A disadvantage is that one needs an explicit model for the distribution, but this may anyway be required for reliable inference in finite samples. We intend to investigate this possibility in further research, simulating from the involved (possibly highly nonelliptical) target distributions by the methods of Hoogerheide, Kaashoek, and Van Dijk (2007), Hoogerheide and Van Dijk (2010), and Hoogerheide, Opschoor, and Van Dijk (2011).

Summarizing, Patton, and Timmermann (2011) have proposed a set of interesting and useful tests for forecast rationality or optimality under squared error loss, including an easily implemented test based on a regression that only involves (long-horizon and short-horizon) forecasts and no observations on the target variable. We have discussed an extension, a simulation-based procedure that takes into account the presence of errors in parameter estimates. This procedure can also be applied in the field of "backtesting" models for value-at-risk. Applications to simple AR and ARCH time series models show that its power in detecting certain misspecifications is larger than the power of well-known tests for correct UC and CC.

[Received June 2011. Accepted September 2011.]

REFERENCES

Christoffersen, P. F. (1998), "Evaluating Interval Forecasts," *International Economic Review*, 39 (4), 841–862. [31]

Hoogerheide, L. F., Kaashoek, J. F., and Van Dijk, H. K. (2007), "On the Shape of Posterior Densities and Credible Sets in Instrumental Variable Regression Models With Reduced Rank: an Application of Flexible Sampling Methods Using Neural Networks," *Journal of Econometrics*, 139 (1), 154–180. [33]

Hoogerheide, L. F., Opschoor, A., and Van Dijk, H. K. (2011), "A Class of Adaptive EM-based Importance Sampling Algorithms for Efficient and Robust Posterior and Predictive Simulation," Tinbergen Institute Discussion Paper TI 2011-004/4. [33]

Hoogerheide, L. F., and Van Dijk, H. K. (2010), "Bayesian Forecasting of Value at Risk and Expected Shortfall Using Adaptive Importance Sampling," *International Journal of Forecasting*, 26 (2), 231–247. [33]

Kupiec, P. H. (1995), "Techniques for Verifying the Accuracy of Risk Measurement Models," *Journal of Derivatives*, 3, 73–84. [31]

Patton, A. J., and Timmermann, A. (2011), "Forecast Rationality Tests Based on Multi-Horizon Bounds," *Journal of Business & Economic Statistics*, 30 (1), 1–17. [30]

Comment

Kenneth D. West
Department of Economics, University of Wisconsin (kdwest@wisc.edu)

This comment proposes joint tests that are applicable when the forecaster supplies forecasts for several horizons, for a given variable. An example would be to predict quarterly GDP growth one, two, three, and four-quarters ahead, producing a times series of quarterly observations on one-, two-, three-, and four-quarter ahead forecasts. The tests focus on the target date, comparing short- and long-horizon forecasts of the same object. Many of the tests are inequality based. The one that I will use as an illustration in these comments is based on a familiar idea: because of mean reversion, the variance of a forecast of a stationary variable falls with the horizon (technically, is weakly decreasing in horizon). Specifically:

$$\begin{aligned}&\text{var(one step ahead forecast)} \geq \text{var(two step ahead forecast)}\\&\quad\geq \cdots \geq \text{var}(H \text{ step ahead forecast}). \qquad (1)\end{aligned}$$

The authors examine Equation (1) and their other tests in simulations, finding that the tests generally work reasonably well.

The key contribution of the article is to develop tests of forecasting models that *only* require data on forecasts, and do not require data on realizations. Such tests allow one to sidestep debates about whether forecasters are trying to predict current or final vintages of data that are subject to revision. In addition, by focusing on the target date, this "fixed event" approach is more robust to changes in regime than are approaches that focus on the date of prediction.

I will give the article high praise by stating that the idea of using only data from forecasts is obvious. My praise is sincere: ideas that are obvious once said, but nonetheless are only now being said, are ideas with lots of potential.

Having stated my overall high opinion of the article, I will take the remainder of my limited space to indicate some points of possible disagreement. I am not sure that my interpretation of the tests aligns very well with that of the authors.

Let me raise three questions. The first two are related. 1) How powerful are the tests in detecting the possibility that the forecaster is using a single internally consistent, though misspecified, model across all horizons? 2) How do we interpret the tests if forecasts are constructed using the "direct" rather than "iterated" methods, or more generally if we concede at the outset that our forecasting model is misspecified? My answer to these two questions is that perhaps the tests in this article are better described as testing whether an internally consistent forecasting method is being used, rather than as testing for flaws in our forecasts. A final unrelated question is: 3) How does error in estimation of parameters required to make forecasts affect the performance of the tests? My answer to this question is that we do not know, and that means we do not have reliable evidence on how well these tests will work in practice.

1. Some of the paper's inequalities hold—possibly weakly, as equalities—if an AR(1) model is used, whether or not the AR(1) model is correct. For example, suppose forecasts are generated from a zero mean AR(1) with parameter φ, with $|\varphi| < 1$. Let Y_t denote the stationary variable under study, which may or may not follow an AR(1); my only assumption is that Y_t is stationary. Then, the one step ahead forecast is of course φY_t, the two step ahead forecast is $\varphi^2 Y_t$, and so on. For this AR(1) model, the inequalities in my Equation (1) are

$$\text{var}(\varphi Y_{t-1}) \geq \text{var}(\varphi^2 Y_{t-2}) \geq \cdots \geq \text{var}\big(\varphi^H Y_{t-H}\big). \qquad (2)$$

Note that since $|\varphi| < 1$, these inequalities hold regardless of whether or not Y_t is generated by an AR(1) model.

Let me hold off on further discussion until after I make my second point.

2. To think about how the test (1) deals with iterated direct forecasts, let us work through an example with a specific DGP. Suppose that Y_t follows an MA(2) of the following form:

$$Y_t = \varepsilon_t + \theta\varepsilon_{t-2}, \quad |\theta| < 1,\ \theta \neq 0, \quad \varepsilon_t \sim \text{white noise}. \qquad (3)$$

Note that lag 2 appears in the DGP, while lag 1 does not.

Let us contrast iterated and direct forecasts, supposing that in each case the forecaster uses not the correct MA(2) model but a model that relies on a single lag of Y_t. Thus, we are using a misspecified model for our forecasts. The iterated forecast generates AR(1) forecasts. Because Y_t is uncorrelated with Y_{t-1}, the AR(1) coefficient φ is 0. This means that for the iterated method the inequalities in (1) are trivially satisfied (with equality).

Now consider the direct forecast, in which the h step ahead forecast of Y_t is generated by projecting Y_t onto Y_{t-h}. The one step ahead forecast is again zero. For two step ahead forecasts, standard projection arguments yield $Y_{t|t-2} = [\theta/(1+\theta^2)]Y_{t-2}$. It is easily seen that the monotonicity in (1) fails:

$$\begin{aligned}&\text{var(one step ahead direct forecast)} = 0 < [\theta/(1+\theta^2)]^2\text{var}(Y_t)\\&\quad= \text{var(two step ahead direct forecast)}. \qquad (4)\end{aligned}$$

Since we are using a misspecified model, the direct method is of course preferable to the iterated method by a mean squared error criterion. Specifically, the two methods have identical forecasts and forecast errors one step ahead, but for two step ahead forecasts, the variance of the forecast error for the direct method is strictly smaller: var(two step ahead direct forecast error) < var(two step ahead iterated forecast error). Yet, the test (1) rejects when the direct method is used and does not reject when

© 2012 American Statistical Association
Journal of Business & Economic Statistics
January 2012, Vol. 30, No. 1
DOI: 10.1080/07350015.2012.634350

the iterated method is used. [Of course, as the authors note, the rejection under the direct method indicates that one can do better. If the variance of the two step ahead forecast error is smaller than that of the one step ahead, one can form the one step ahead forecast by reusing the two step ahead forecast and then the inequality will hold (weakly, as equality). But the fact remains that the test rejects under the direct but not iterated method.]

Let me now discuss my points 1 and 2. My point 1 was that Equation (1) holds by construction if forecasts are generated by an AR(1) model. But that is not true of all the article's tests. Indeed, the article gives a counterexample, in Section 4.3.2. In this counterexample, the test used is one that examines whether forecast error variances are weakly increasing. The DGP is an AR(2). The parameters of the AR(2) are not, however, likely to characterize economic data, because, as the authors state, the autocorrelation at lag 2 is distinctly larger than that at lag 1. The authors do not present other counterexamples. In terms of my point 2, despite my simple example in Equation (3), there are plausible scenarios in which one or more of the authors' tests will not reject when the direct method is being used.

Thus, my AR(1) and MA(2) examples are nothing but examples with uncertain generality. It is unclear how illustrative these examples are about the behavior of this article's tests when those tests are applied to plausible forecasts and plausible DGPs. I do think, however, that we need to be open to the possibility that for a reasonable class of models, the tests in this article are better described as testing whether the iterated method, or some other internally consistent forecasting method, is being used, rather than as testing for flaws in our forecasts.

3. My final comment relates to the simulations in the article. These all assume that forecasts are made from population parameters. In practice, economic forecasts overwhelmingly rely on estimated regression parameters rather than population parameters. This might seem a pedantic point, but in fact performance of tests involving forecasts often is wildly sensitive to the use of estimated rather than population parameters. Simulations in West and McCracken (1998), for example, find that in univariate Mincer-Zarnowitz regressions, nominal 5% tests sometimes have actual size as large as 45% when estimated regression parameters are used for prediction. I thus take the simulation evidence in this article with a large grain of salt. I also note that adjustments to produce better size are available for some conventional tests (again see West 1996; West and McCracken 1998; Clark and West 2007; Hubrich and West 2011). I hope that the authors will think about developing similar adjustments for their tests.

ACKNOWLEDGMENT

The author thanks the National Science Foundation for financial support.

[Received April 2011. Accepted June 2011.]

REFERENCES

Clark, T. E., and West, K. D. (2007), "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics*, 138(1), 291–311. [35]

Hubrich, K., and West, K. D. (2010), "Forecast Comparisons for Small Nested Model Sets," *Journal of Applied Econometrics*, 25(4), 574–594. [35]

West, K. D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica*, 64, 1067–1084. [35]

West, K. D., and McCracken, M. W. (1998), "Regression Based Tests of Predictive Ability," *International Economic Review*, 39, 817–840. [35]

Rejoinder

Andrew J. Patton
Department of Economics, Duke University, 213 Social Sciences Building, Box 90097, Durham, NC 27708
(*andrew.patton@duke.edu*)

Allan Timmermann
University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093 (*atimmerm@ucsd.edu*)

1. INTRODUCTION

We thank our discussants for their stimulating and thoughtful comments on our article. Each of the five discussions raises an interesting set of points, and across all of them a few key themes arise. The following sections discuss each of these topics in turn.

2. INSTABILITIES IN THE DATA GENERATING PROCESS

Croushore and Rossi both raise the question of how tests of forecast rationality fare in the presence of instabilities in the data generating process for the target variable or the forecasts. Croushore presents empirical results showing that the statistical significance of forecast bias varies greatly depending on the choice of start and end dates for the sample period. Rossi outlines interesting work from Rossi and Sekhposyan (2011), which allows the researcher to conduct tests of forecast rationality that are robust to the presence of such instabilities, by computing test statistics across a range of subsamples, and then obtaining a critical value for some function (e.g., the sup) of these test statistics. We view this as a very useful extension since the type of macroeconomic variables often considered in forecast rationality tests are indeed subject to breaks, notably inflation. By Proposition 3, instability in the target variable should not affect the multi-horizon bounds we propose, provided that forecasters' information set is expanding through time and the fixed-event setup is adopted. However, the finite sample properties of the bounds tests could well be affected by such instability, in which case an approach like that of Rossi and Sekhposyan could be useful.

Croushore also notes that some of the tests could wrongly reject the null of forecast rationality due to changes in the definition of the target variable through time. As he acknowledges, this is less of an issue when the test is based solely on the forecasts, at least if changes to the definition of the target variable happen after the target date. If it happens midstream, that is, while forecasts for a given target date are still being produced, the results could of course be affected. This corresponds to changing the definition of the target variable at some date, $t - h_M$, say, from Y_t to $\dddot{Y}_t$, in which case the variances of the optimal long-term and short-term forecasts, $\hat{Y}^*_{t|t-h_L} = E[\dddot{Y}_t|\mathcal{F}_t]$ and $Y^*_{t|t-h_L} = E[Y_t|\mathcal{F}_t]$, respectively, (for $h_S < h_M < h_L$) need not satisfy the mean squared forecast bounds of the article since they apply to different variables. This can equivalently be thought of as a violation of the assumption of an expanding information set as time progresses, which is one of the key assumptions underpinning the proposed tests.

3. PARAMETER ESTIMATION ERROR

Hoogerheide, Ravazzolo, and van Dijk present an interesting application of our "optimal revision regression" [see Equation (20) of the main article] to a Value-at-Risk forecasting application. This application exploits the fact that no data on the target variable are required, a desirable feature in volatility or value-at-risk applications, where the target variable is latent. In a simulation study for this application, these authors document that the test is (somewhat) oversized when parameter estimation error (PEE) is ignored (size of 0.12 or 0.08 for a nominal 0.05 test). West also raises concerns about the finite sample size of our rationality tests in the presence of PEE, and notes that in other applications PEE can lead to size distortions.

Hoogerheide et al. propose a simulation-based method for obtaining critical values that take into account parameter estimation error, while West provides references for asymptotic adjustments for PEE. In applications where the forecast evaluator is also the forecast producer, and so simulation-based and asymptotic adjustments are feasible, we agree that this is an improvement over simply ignoring PEE. For such cases, it is also feasible to construct critical values for the second moment bounds that account for the estimation scheme. However, we note that survey or expert forecasts, such as those in our empirical application, are typically based both on (unknown) econometric models and also on subjective judgment. In such applications, it is not clear how PEE should be handled.

We should note that PEE may affect the critical values for certain tests, but more important is whether such tests are even applicable. For example, while Corollary 1 states that the mean squared forecast error must be weakly increasing in the forecast horizon, in the presence of PEE it has long been known that this monotonicity property may break down, see Hoque, Magnus, and Pesaran (1988), Theorem 2, and Magnus and Pesaran (1989), Theorem 1, for example, in the case of an AR(1) model. Thus, a term structure of mean squared error (MSE) values that is not weakly increasing need *not* be indicative of

© 2012 American Statistical Association
Journal of Business & Economic Statistics
January 2012, Vol. 30, No. 1
DOI: 10.1080/07350015.2012.634354

Table 1. Monotonicity bounds and regression results in the presence of parameter estimation error in an AR(1) model

	$\phi = 0.1$			$\phi = 0.5$			$\phi = 0.95$		
Estimation sample size:	25	50	100	25	50	100	25	50	100
Tests									
Inc MSE	·	·	√	·	√	√	√	√	√
Dec Cov	·	√	√	√	√	√	√	√	√
Cov bound	√	√	√	√	√	√	√	√	√
Dec MSF	√	√	√	√	√	√	·	√	√
Inc MSFR	√	√	√	√	√	√	·	√	√
Dec Cov, with proxy	√	√	√	√	√	√	√	√	√
Cov bound, with proxy	√	√	√	√	√	√	√	√	√
MZ on short horizon	·	·	·	·	·	·	·	·	·
Univar opt. revision regr.	·	·	·	·	·	·	·	·	·
Univar opt. revision regr., with proxy	·	·	·	·	·	·	·	·	·
Univar MZ, Bonferroni	·	·	·	·	·	·	·	·	·
Univar MZ, Bonferroni, with proxy	·	·	·	·	·	·	·	·	·
Vector MZ	·	·	·	·	·	·	·	·	·
Vector MZ, with proxy	·	·	·	·	·	·	·	·	·

NOTE: This table presents results on whether a given monotonicity bound or regression parameter result based on forecast optimality in the absence of estimation error continues to hold in the presence of estimation error. If the result holds it is marked with "√," otherwise it is marked with "·". The model considered is a simple AR(1) with an intercept, $y_t = \mu + \phi y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim N(0, 1)$. Three in-sample estimation samples are considered (25, 50, and 100 observations) and three levels of persistence are considered $(\phi = 0.1, 0.5, 0.95)$.

forecast suboptimality, rather it may simply reflect PEE. Furthermore, the regression parameter restrictions (such as those from the Mincer–Zarnowitz regression or the "optimal revision regression," MZ and ORR) may also fail to hold in the presence of PEE. For example, a finite sample bias in the estimation of the lag coefficients in AR models is well known, and this will inevitably lead to deviations of the population MZ and ORR parameters from their hypothesized values. In general, as the regression-based tests are based on *equality* restrictions rather than *in*equality restrictions, one might suspect that these are more sensitive to the presence of PEE.

To study whether this is an issue in applications similar to those considered in our Monte Carlo study, we conducted the following study: consider a simple AR(1) with intercept as in Section 5 of the article, with autoregressive coefficient $\phi \in \{0.1, 0.5, 0.95\}$. We consider rolling window estimation of this model based on $R \in \{25, 50, 100\}$ observations. (The results for $R = 1000$ are very close to the $R = 100$ case and so are not reported.) Table 1 records a "√" if the monotonicity results established in the article hold across forecasting horizons $h = 1, 2, \ldots, 8$ in the presence of estimation error, and records a "·" if this is violated. This table shows that for estimation sample sizes of 100 or more, all of the monotonicity and bounds results in the article continue to hold, at least for this popular data generating process. For an estimation sample of size 50, consistent with the results in Magnus and Pesaran (1989, Table 1), all results in the paper hold, except for a single result (MSE) for the least predictable case $(\phi = 0.1)$. It is only when the sample size is 25 observations that we see scattered evidence of PEE leading to breakdowns in the monotonicity results of the article. This is in contrast with the regression-based results, all of which fail to hold in the presence of PEE, even with large estimation samples due to the finite sample biases in the parameter estimates. This confirms the intuition above that tests based on inequality restrictions are likely to be less sensitive to PEE than those based on equality restrictions.

In Table 2, we present the results of a simulation study of the finite sample size of the proposed tests in the presence of PEE. These results can be compared to those in Table 1 of the article for $H = 8$ and zero measurement error. The results confirm that the bounds-based tests (in the top seven rows) are largely robust to PEE. It is only when persistence is low $(\phi = 0.1)$ and the sample size is very short (25 observations) that we see some evidence that these tests are oversized. The regression-based tests, on the other hand, are all grossly oversized, which is consistent with the theoretical results in Table 1, and with the findings reported in Table 1 by Hoogerheide et al., which showed that the parameter restrictions that hold under the absence of PEE do not, in fact, hold in general.

The results in Tables 1 and 2 are specific to the simple AR(1) data generating process (DGP). The effect of PEE will generally depend on both the complexity of the model used to compute forecasts, larger effects likely associated with large-scale multivariate models, and on the true DGP, both of which are typically unknown. It will also depend on the estimation method and the scheme used to update the parameter estimates (fixed, rolling, or expanding window). Another issue is that most results are limited to forecasts based on plug-in estimators which ignore the effect of estimation error and so could be improved, for example, through a bias-adjustment procedure, in a way that depends on the forecast horizon.

4. FINITE SAMPLE PERFORMANCE

Because samples used to evaluate forecast rationality are typically quite small, we agree with the observation made by Croushore that forecast rationality tests can be quite sensitive to the sample at hand. Figure 1 in Croushore's analysis

Table 2. Monte Carlo simulation of size of tests of forecast optimality in the presence of parameter estimation error in an AR(1) model

	$\phi = 0.1$			$\phi = 0.5$			$\phi = 0.95$		
Estimation sample:	25	50	100	25	50	100	25	50	100
Tests									
Inc MSE	23.3	20.4	18.4	3.7	4.2	4.3	0.0	0.0	0.0
Dec Cov	6.4	10.1	17.4	1.5	2.4	6.8	0.3	0.3	0.3
Cov bound	24.5	17.3	18.4	0.3	0.7	1.6	5.5	4.6	4.2
Dec MSF	0.6	2.2	8.4	0.2	1.1	5.5	3.2	3.1	3.3
Inc MSFR	0.0	0.5	1.9	0.4	2.0	3.1	0.0	0.0	0.0
Dec Cov, with proxy	0.2	2.1	8.8	0.2	1.7	6.2	0.2	0.3	0.3
Cov bound, with proxy	2.3	6.2	12.4	0.8	2.4	5.1	5.4	4.9	4.5
MZ on short horizon	94.0	58.8	33.2	22.5	4.1	5.9	38.7	15.4	14.6
Univar opt. rev. regr.	61.6	36.4	26.5	35.8	24.0	21.9	36.6	23.9	15.8
Univar opt. rev. regr., with proxy	31.1	26.1	23.8	50.4	32.8	26.3	45.2	26.6	17.9
Univar MZ, Bonf.	85.5	56.8	38.4	83.9	64.0	53.9	80.7	60.5	37.6
Univar MZ, Bonf., with proxy	53.1	41.6	43.0	81.0	64.1	24.5	79.7	60.0	36.9
Vector MZ	99.8	97.3	92.4	97.8	93.9	89.3	73.2	57.8	37.2
Vector MZ, with proxy	57.8	54.3	67.5	80.8	83.2	87.3	67.5	48.5	30.0
Bonf, using actuals	40.1	22.7	21.9	17.5	10.8	9.6	18.1	8.3	6.1
Bonf, using forecasts only	12.5	10.2	12.1	25.2	13.4	11.5	21.4	8.6	6.0
Bonf, all tests	34.5	17.0	18.7	21.2	11.1	9.0	18.2	6.3	4.9

NOTES: This table presents the outcome of 1,000 Monte Carlo simulations of the size of various forecast optimality tests, corresponding to Table 1 of the article. Data are generated by a first-order autoregressive process with persistence parameter ϕ of 0.1, 0.5, or 0.95. Three rolling window estimation samples are considered (25, 50, and 100 observations). The maximum forecast horizon is assumed to be eight periods. The simulations assume an out-of-sample size of 100 observations and a nominal size of 10%. The inequality tests are based on the Wolak (1989) test and use simulated critical values based on a mixture of chi-squared variables. Tests labeled "with proxy" refer to cases where the one-period forecast is used in place of the predicted variable.

illustrates that the finite sample power of the test is an important consideration. For the graph with a fixed end date (2006), the length of the evaluation sample expands as one moves to the left along the x-axis, thereby increasing the power of the test and the probability of detecting any biases that may be present in the forecasts. This does not, however, help explain why the evidence against the null of no bias is weaker to the right in the graph that conditions on the starting date (1971) and expands the evaluation window as time moves forward. In this case, the test in fact rejects short evaluation windows.

In principle, using information on several forecast horizons should help improve power. For example, the optimal revision regression

$$Y_t = \alpha + \beta_H \hat{Y}_{t|t-h_H} + \sum_{j=1}^{H-1} \beta_j d_{t|h_j,h_{j+1}} + u_t, \tag{1}$$

tests for biases in the individual forecast revisions as well as in the long-run forecast. This allows us to decompose any biases into the individual forecast horizons, which is more difficult if only a single forecast horizon is considered. For comparison, the regressions proposed in Nordhaus (1987) and Lahiri and Sheng (2008)

$$d_{t|h_S,h_M} = \beta d_{t|h_M,h_L} + u_t \quad \text{or} \tag{2}$$

$$e_{t|t-h_S} = \gamma d_{t|h_S,h_L} + u_t, \tag{3}$$

are equality tests of covariances for a given horizon although they exploit multi-horizon information through use of forecast revisions.

The empirical work reported by Lahiri suggested that forecasters revise their forecasts in a nonsmooth way as a function of the forecast horizon which can cause problems such as ill-behaved regressors in the optimal revision regression. This observation suggests that one could increase the finite sample power of the tests by carefully selecting a subset of horizons and focusing on the associated moments.

One way to study the finite sample performance of the new forecast rationality tests is to use the bootstrap reality check of White (2000) or the refinement proposed by Hansen (2005), rather than rely on asymptotical critical values. Table 3 illustrates the outcome of 1,000 Monte Carlo simulations used to study the size and power of a subset of the variance bounds. Compared with the tests based on the Wolak method, these tests are more undersized under the White approach, but slightly less so using the Hansen refinement when $H = 8$. The power of the tests is comparable across the three methods, with a slightly better performance under the asymptotical critical values. The bootstrap approach is likely to outperform by a greater margin as the number of forecast horizons, H, gets larger.

5. INTERPRETATION OF THE TESTS

Lahiri and West both raise the point that the tests proposed in our article will fail to detect certain deviations from rationality, that is, they are not consistent tests. West makes this point in the context of a mis specified AR(1) forecasting model applied to a MA(2) data-generating process, while Lahiri considers a correctly specified MA(q) model based on incorrect parameters. This is a valid concern, and is one that applies not only to our bound-based tests but also to the familiar Mincer–Zarnowitz tests, forecast error autocorrelation tests, our proposed optimal revision regression, etc. Consistent regression-based tests of

Table 3. Comparison of asymptotic- and bootstrap-based tests

	Size			Equal noise			Increasing noise		
	Asymp	White	Hansen	Asymp	White	Hansen	Asymp	White	Hansen
Tests									
Inc MSE	5.2	0.3	6.3	14.4	8.6	11.3	0.4	1.3	1.7
Dec Cov	4.7	0.0	6.2	13.8	8.6	11.6	11.2	10.5	10.3
Cov bound	0.0	1.5	1.3	78.9	74.8	76.0	91.6	93.3	88.9
Dec MSF	0.7	0.1	2.4	18.2	8.7	12.3	100.0	56.0	61.2
Inc MSFR	4.4	0.0	6.9	16.7	9.8	12.5	0.0	0.5	0.6
Dec Cov, with proxy	6.0	0.1	5.0	15.5	11.8	14.7	13.9	13.0	12.0
Cov bound, with proxy	0.0	5.9	5.1	99.2	98.7	98.5	100.0	100.0	100.0

NOTES. This table presents the outcome of 1,000 Monte Carlo simulations of the size and power of various forecast optimality tests, corresponding to Tables 1 and 2 of the article. All tests have a nominal size of 10%. The forecast horizon is assumed to be eight periods, the level of measurement error is assumed to be "medium," and the out-of-sample period is assumed to be 100 observations. The first column of each panel presents results based on the Wolak (1989) test and uses simulated critical values based on a mixture of chi-squared variables. These values are repeated from column 5 of Table 1 and columns 5 and 11 of Table 2 in the article (reported here for ease of comparison). The second and third columns present results based on a bootstrap implementation based on White (2000) and Hansen (2005), respectively.

forecast rationality were considered by Corradi and Swanson (2002), though to the best of our knowledge the multi-horizon equivalent of the tests in that article are not available in the literature. Moreover, practice tests with power against generic alternatives can be difficult to interpret and are unlikely to have much power in any given direction, given the short evaluation samples that are typically available.

While we acknowledge that each of our bounds has power only against certain deviations from rationality, when used as a *suite* of tests they cover a variety of deviations. Further, our simulation study suggests that combining these tests via a Bonferroni approach does not lead to an overly conservative test, and provides power in a (finite) variety of directions.

5.1 Improving Imperfect Forecasts

One of the benefits of using tests with power in a specific direction is that they offer clues to how suboptimal forecasts can be improved. Croushore raises the point that unlike more standard regression-based tests of forecast rationality, the bounds tests proposed in our article do not naturally provide an "improved" forecast as an output. For example, consider a standard Mincer–Zarnowitz regression

$$Y_t = \beta_0^h + \beta_1^h \hat{Y}_{t|t-h} + u_{t|t-h}. \tag{4}$$

If the test that $\left[\beta_0^h, \beta_1^h\right] = [0, 1]$ is rejected, one can immediately use $\hat{Y}^c_{t|t-h} \equiv \hat{\beta}_0^h + \hat{\beta}_1^h \hat{Y}_{t|t-h}$ as a (linear) bias-corrected forecast. Along similar lines, our proposed optimal revision regression in Equation (1) may also be used to generate an improved forecast

$$\hat{Y}_t^c = \hat{\alpha} + \hat{\beta}_H \hat{Y}_{t|t-h_H} + \sum_{j=1}^{H-1} \hat{\beta}_j d_{t|h_j, h_{j+1}}. \tag{5}$$

It is true that the bounds tests we propose do not immediately lead to an improved forecast. However, considering what types of mis-specification will be detected by the various bounds tests does offer the possibility of improving the forecasts. We next illustrate how the rationality tests can be interpreted using two examples of suboptimal forecasts.

5.2 The Lazy Forecaster

Consider the case of a lazy forecaster, who, in constructing a short-horizon forecast, $\tilde{Y}_{t|t-h_S}$, does not update his long-horizon forecast, $\tilde{Y}_{t|t-h_L}$, with relevant information, and hides this lack of updating by adding a small amount of zero-mean, independent noise to the long-horizon forecast. In that case

$$\tilde{Y}_{t|t-h_S} = \tilde{Y}_{t|t-h_L} + u_{t-h_S},\ u_{t-h_S} \sim \text{iid}\left(0, \sigma_u^2\right). \tag{6}$$

We then have

$$\begin{aligned} V\left[e_{t|t-h_S}\right] &= V\left[Y_t - \tilde{Y}_{t|t-h_L} - u_{t-h_S}\right] = V\left[e_{t|t-h_L}\right] \\ &\quad + V\left[u_{t-h_S}\right] > V\left[e_{t|t-h_L}\right]. \end{aligned} \tag{7}$$

Thus, such a forecaster will be revealed via a violation of the MSE bounds. This forecaster will also violate the bound on the variance of the forecast revision

$$\begin{aligned} V\left[\tilde{Y}_{t|t-h_S} - \tilde{Y}_{t|t-h_L}\right] &= V\left[u_{t-h_S}\right] > \text{cov}\left[\tilde{Y}_{t|t-h_S} - \tilde{Y}_{t|t-h_L}, Y_t\right] \\ &= \text{cov}\left[u_{t-h_S}, Y_t\right] = 0. \end{aligned} \tag{8}$$

As shown in Corollary 4, the variance of the optimal forecast revision is bounded above by twice its covariance with the target variable. When the forecast is "updated" with pure noise, this bound is violated and in this case the long-horizon forecast is better than the short-horizon forecast, even when the latter is available.

5.3 Overreacting to News

As a second example, consider a forecaster who overreacts to new information about the target variable. We will model this as a forecast given by

$$\hat{Y}_{t|t-h} = \hat{Y}^*_{t|t-h-1} + \gamma d^*_{t|h,h+1}, \qquad \gamma \geq 1. \tag{9}$$

When $\gamma = 1$, we have $\hat{Y}_{t|t-h} = \hat{Y}^*_{t|t-h-1} + d^*_{t|h,h+1} = \hat{Y}^*_{t|t-h}$, and thus the forecast is optimal, but when $\gamma > 1$ the forecaster updates her forecast by more than what is optimal. Consider now what this does to the observed forecast revisions

$$d_{t|h,h+1} = \hat{Y}_{t|t-h} - \hat{Y}_{t|t-h-1} = \gamma d^*_{t|h,h+1} + (1-\gamma) d^*_{t|h+1,h+2}. \tag{10}$$

Notice that unlike optimal forecast revisions, which are uncorrelated, these forecast revisions will be negatively correlated

$$\begin{aligned}\operatorname{cov}\left[d_{t|h,h+1}, d_{t|h+1,h+2}\right] &= \gamma(1-\gamma)V\left[d^{*}_{t|h+1,h+2}\right] \\ &\leq 0 \text{ for } \gamma \geq 1 \text{ (with equality when } \gamma = 1). \end{aligned} \tag{11}$$

This implies that mean squared forecast revisions (MSFR) can violate the weakly increasing property established in Corollary 1

$$\begin{aligned} V\left[d_{t|h,h+2}\right] &= V\left[d_{t|h,h+1} + d_{t|h+1,h+2}\right] \\ &= V\left[d_{t|h,h+1}\right] + V\left[d_{t|h+1,h+2}\right] \\ &\quad + 2\operatorname{cov}\left[d_{t|h,h+1}, d_{t|h+1,h+2}\right] \\ &\lesseqqgtr V\left[d_{t|h,h+1}\right]. \end{aligned} \tag{12}$$

Only when $\gamma = 1$ does the third term drop out and we are ensured that the left-hand side is weakly greater than the right-hand side. Thus, a violation of the MSFR condition, as found in our empirical application for GDP growth forecasts, illustrated in Figure 2, may indicate overreaction to news arriving between the release dates of forecasts.

ADDITIONAL REFERENCES

Corradi, V., and Swanson, N. R. (2002), "A Consistent Test for Nonlinear Out of Sample Predictive Accuracy," *Journal of Econometrics*, 110, 353–381. [39]

Hoque, A., Magnus, J.R., and Pesaran, B. (1988), "The Exact Multi-Period Mean-Square Forecast Error for the First-Order Autoregressive Model," *Journal of Econometrics*, 39, 237–246. [36]

Lahiri, K., and Sheng, X. (2008), "Evolution of Forecast Disagreement in a Bayesian Learning Model," *Journal of Econometrics*, 144, 325–340. [38]

Magnus, J.R., and Pesaran, B. (1989), "The Exact Multi-Period Mean-Square Forecast Error for the First-Order Autoregressive Model With an Intercept," *Journal of Econometrics*, 42, 157–179. [36]

Rossi, B., and Sekhposyan, T. (2011), "Forecast Optimality Tests in the Presence of Instabilities," Mimeo, Duke University. [36]