1-2012

Comment

Dean D. Croushore
University of Richmond, dcrousho@richmond.edu

Real-Time Implications of Forecast-Rationality Tests Based on Multi-Horizon Bounds

Dean Croushore

University of Richmond and Federal Reserve Bank of Philadelphia

Revised Draft of 27 August 2011

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 multi-dimensional answer. I think about forecasts in three dimensions: (1) horizon, (2) sub-sample, 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) finds that the results of tests for bias vary across horizons with no systematic tendency across variables, using individual forecasts from the ASA-NBER survey (now the Survey of Professional Forecasters, SPF). Similarly, Brown and Maital (1981) find 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) find 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 sub-sample. Though researchers seek to find stylized facts, they are thwarted by instabilities in empirical results across sub-samples. Croushore (2010) shows how forecast rationality tests using SPF forecasts change dramatically over time, depending on the starting date and ending date of the sub-sample. For example, Figure

1 shows how the sample ending date affects the results of a rationality test, which is a test that the mean forecast error is zero. The plot shows the p-value testing the null hypothesis that the mean forecast error is zero for different sub-samples. The line labeled *test for bias before break point* shows the p-values for tests using sub-samples 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 sub-samples 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 beginning and ending dates of our samples are often random or occur by happenstance. Suppose the development of the SPF had been delayed five or ten 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 sub-sample. Depending on the exact beginning 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 sub-samples.

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 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 beginning 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 sub-sample period affect the rationality test, but so does the choice of actual. Choosing the initial actual leads to many more sub-samples in which we reject the null hypothesis of no bias than using the other two choices of actuals.

In their paper, "Forecast Rationality Tests Based on Multiple-Horizon Bounds," Patton and Timmermann 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 don't, however, look at the sensitivity of their results to alternative sub-samples.

The Patton-Timmermann paper 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 paper 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 redefinitions and other methodological changes. Alternatively, they must develop a real-time data set with some version of actual data that aren't subject to distortions because of methodological changes, if the data they need aren't conveniently available in an existing real-time data set, such as the Philadelphia Fed's Real-Time Data Set for Macroeconomists (see Croushore-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 paper provides tests that are easy to interpret, because they lend themselves to graphical interpretations. For example, Figure 1 in the Patton-Timmermann paper 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 paper shows plots across horizons of mean squared forecast revisions and the covariance between the forecast and 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 actual should decrease with horizon if the forecasts are optimal, but that is not true for the GDP deflator.

So, the Patton-Timmermann paper 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 paper may not be valid when major benchmark revisions to the data occur. (3) The conclusions are potentially sensitive to the sub-sample choice.

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) use 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 paper 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 sub-sample 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 sub-sample choice than other tests, including the standard Mincer-Zarnowitz test and the test for zero-mean forecast errors.

To conclude, this paper 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 sub-sample dimension.

References

Brown, Bryan W., and Shlomo Maital. "What Do Economists Know? An Empirical Study of Experts' Expectations." *Econometrica* 49 (March 1981), pp. 491–504.

Croushore, Dean. "An Evaluation of Inflation Forecasts from Surveys using Real-Time Data." *B.E. Journal of Macroeconomics: Contributions* (volume 10, issue 1, article 10, 2010).

Croushore, Dean. "Two Dimensions of Forecast Analysis." University of Richmond working paper, May 2011.

Croushore, Dean, and Tom Stark. "A Real-Time Data Set for Macroeconomists." *Journal of Econometrics* 105 (November 2001), pp. 111-130.

Faust, Jon, John H. Rogers, and Jonathan H. Wright. "News and Noise in G-7 GDP Announcements." *Journal of Money, Credit, and Banking* 37 (June 2005), pp. 403–419.

Keane, Michael P., and David E. Runkle. "Testing the Rationality of Price Forecasts: New Evidence From Panel Data." *American Economic Review* 80 (September 1990), pp. 714–735.

Patton, Andrew J., and Allan Timmermann. "Forecast Rationality Tests Based on Multi-Horizon Bounds." *Journal of Business and Economic Statistics*, this issue.

Zarnowitz, Victor. "Rational Expectations and Macroeconomic Forecasts." *Journal of Business & Economic Statistics* 3 (October 1985), pp. 293–311.

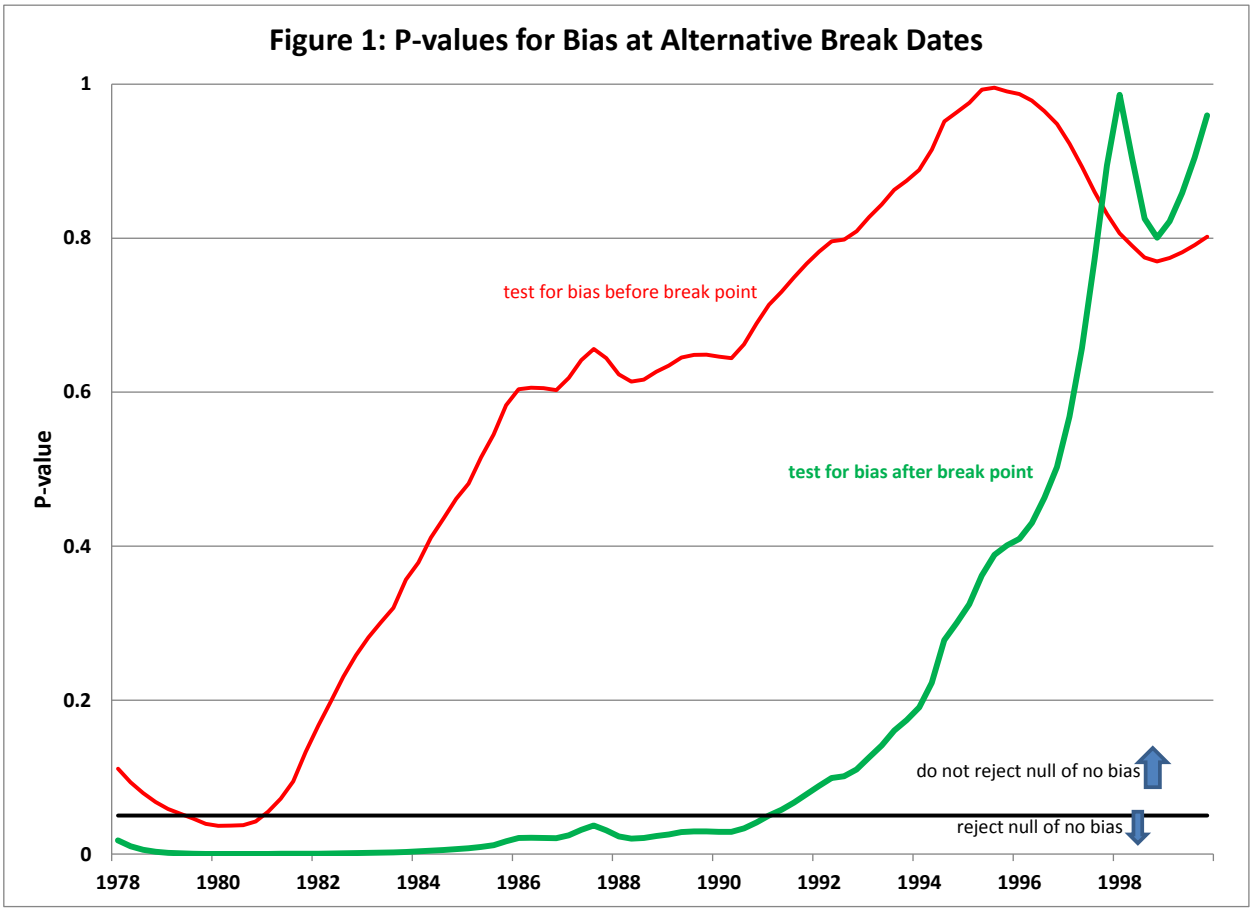

Figure 1: P-values for Bias at Alternative Break Dates

The plot shows the p-value testing the null hypothesis that the mean forecast error is zero for different sub-samples. The line labeled *test for bias before break point* shows the p-values for tests using sub-samples 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 sub-samples that begin at the date shown on the horizontal axis and end at the end of 2008.

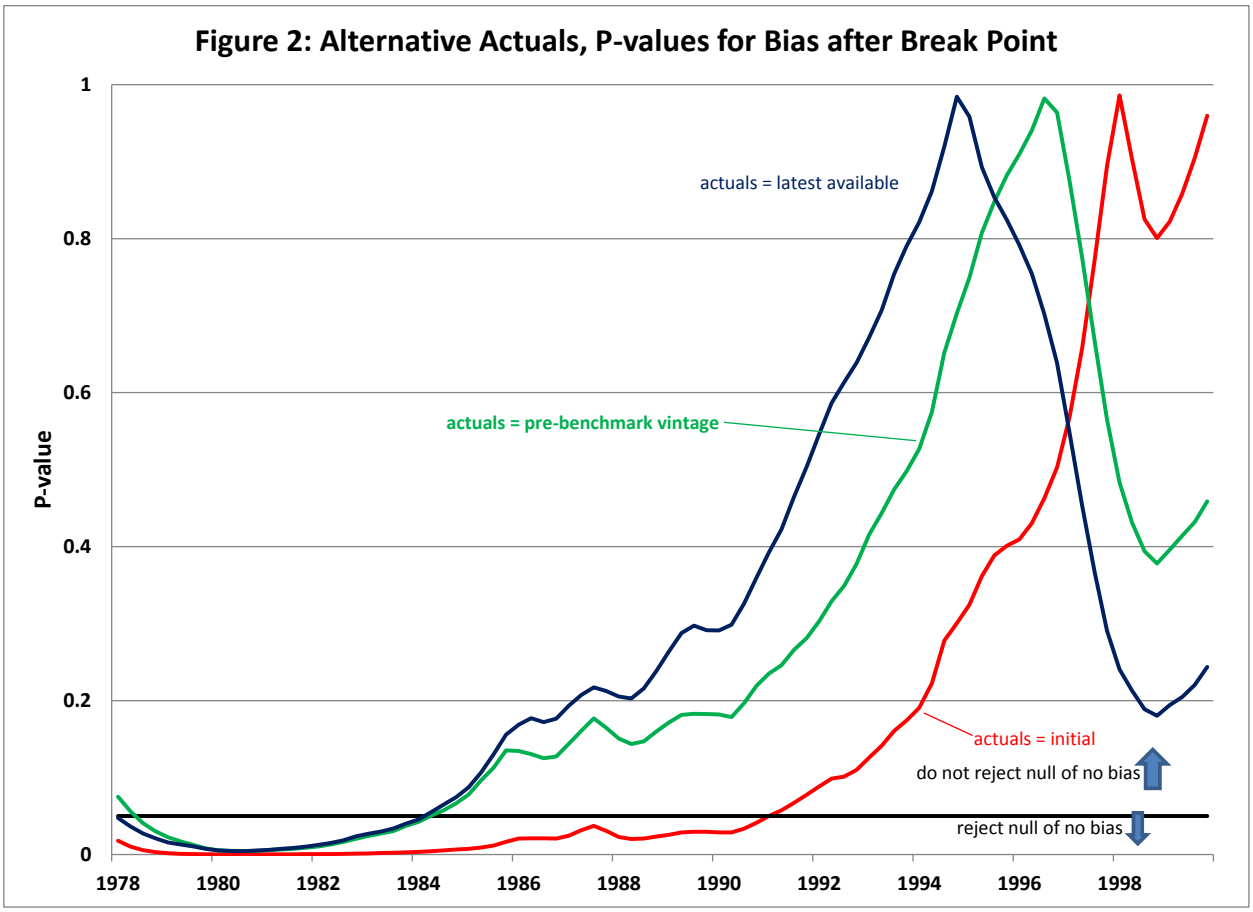

Figure 2: Alternative Actuals: P-values for Bias after Break Point

The plot shows the p-value testing the null hypothesis that the mean forecast error is zero for different sub-samples and different concepts of actuals. Each line shows p-values for tests using sub-samples 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 shown in Figure 1. The line labeled *actuals = pre-benchmark vintage* shows the p-values for tests using as actuals the value recorded in the last vintage before a benchmark revision. And the line labeled *actuals = latest available* shows the p-values for tests using as actuals the value recorded in the vintage of May 2011.

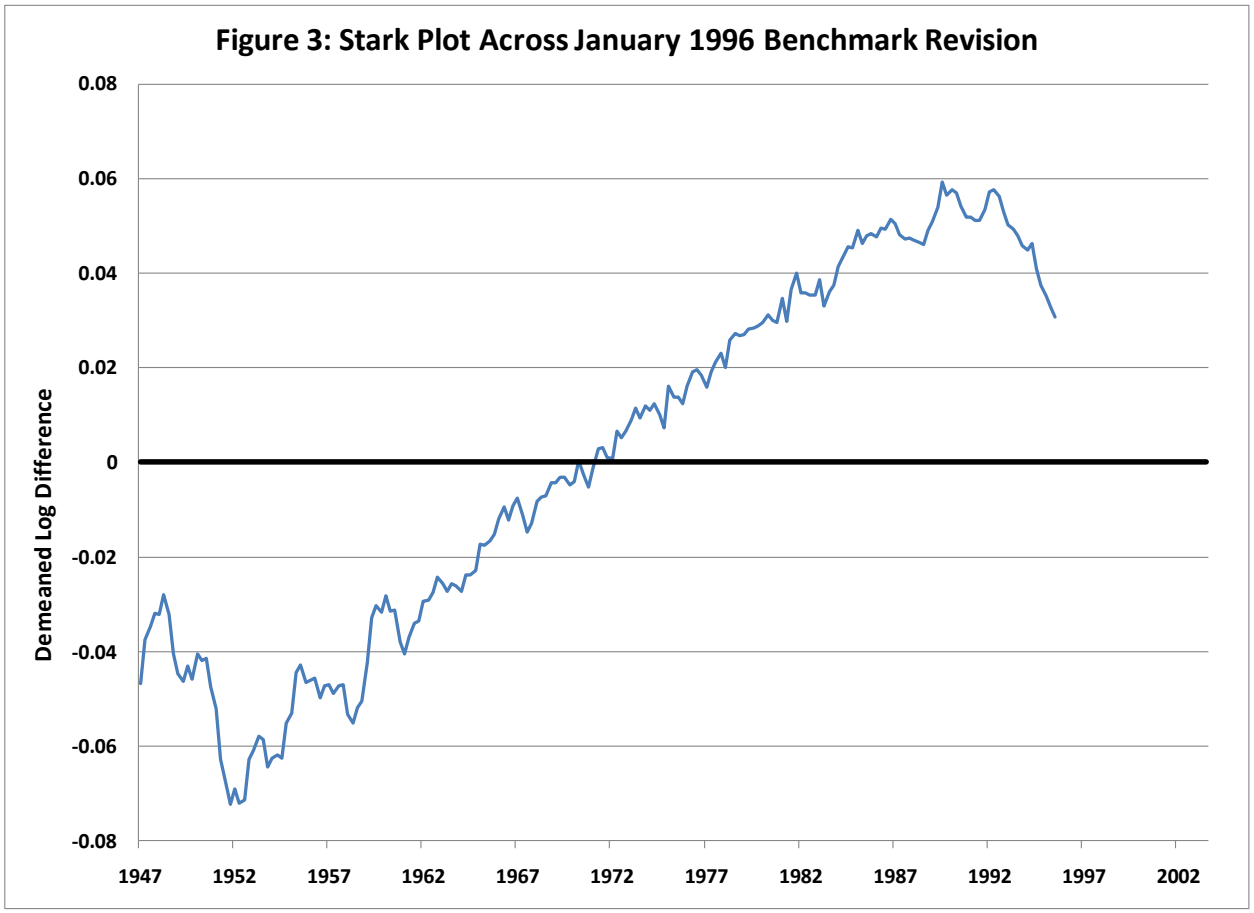

Figure 3: Stark Plot Across January 1996 Benchmark Revision

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.