WORKING PAPERS
RESEARCH DEPARTMENT

WORKING PAPER NO. 06-19
AN EVALUATION OF INFLATION FORECASTS FROM SURVEYS USING REAL-TIME DATA

Dean Croushore

Associate Professor of Economics and Rigsby Fellow
University of Richmond

Visiting Scholar
Federal Reserve Bank of Philadelphia

October 2006

RESEARCH DEPARTMENT, FEDERAL RESERVE BANK OF PHILADELPHIA
Ten Independence Mall, Philadelphia, PA 19106-1574 • www.philadelphiafed.org/econ/index.html

AN EVALUATION OF INFLATION FORECASTS FROM SURVEYS USING REAL-TIME DATA

Dean Croushore

Associate Professor of Economics and Rigsby Fellow
University of Richmond

Visiting Scholar
Federal Reserve Bank of Philadelphia

October 2006

I thank Kundan Kishor, Sean Campbell, and discussants and participants at the 2005 Workshop on Macroeconomic Forecasting, Analysis, and Policy Design with Data Revisions in Montreal, the 2005 Southern Economic Association meetings, the 2006 American Economic Association meetings, and the 2006 Econometric Society Summer meetings for useful comments on this paper. This paper was written in part while the author was a visiting scholar at the Federal Reserve Bank of Philadelphia. The views expressed in this paper are those of the author and do not necessarily represent the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. Amanda Smith provided outstanding research assistance on this project. This paper is available free of charge at http://www.philadelphiafed.org/econ/wps/index.html.

Please send comments to the author at Robins School of Business, 1 Gateway Road, University of Richmond, VA 23173, or e-mail: dcrousho@richmond.edu.

AN EVALUATION OF INFLATION FORECASTS FROM SURVEYS USING REAL-TIME DATA

ABSTRACT

This paper carries out the task of evaluating inflation forecasts from the Livingston Survey and the Survey of Professional Forecasters, using the real-time data set for macroeconomists as a source of real-time data. We examine the magnitude and patterns of revisions to the inflation rate based on the output price index and describe what data to use as "actuals" in evaluating forecasts. We then run tests on the forecasts from the surveys to see how good they are, using a variety of actuals. We find that much of the empirical work from 20 years ago was a misleading guide to the quality of forecasts because of unique events during the earlier sample period. Repeating that empirical work over a longer sample period shows no bias or other problems in the forecasts. The use of real-time data also matters for some key tests on some variables. If a forecaster had used the empirical results from the late 1970s and early 1980s to adjust survey forecasts of inflation, forecast errors would have increased substantially.

AN EVALUATION OF INLATION FORECASTS FROM SURVEYS USING REAL-TIME DATA

Researchers continue to investigate whether forecasts of consumer price inflation from surveys (Michigan, Livingston, and Survey of Professional Forecasters) are biased and inefficient.[1] Arguably, consumer price inflation is not the best measure of inflation because of index construction problems. Better measures of trend inflation come from other variables, such as the GDP deflator. But forecasts of the inflation rate in the GDP deflator are more difficult to evaluate because the past data are revised.

As part of the research program into rational expectations in the early 1980s, economists tested the inflation forecasts based on the GDP deflator from surveys and found a disturbing result: the forecasts exhibited bias and inefficiency. If macroeconomic forecasters had rational expectations, the forecast errors should have had much better properties; instead, the forecasters appeared to make systematic errors. Researchers concluded that perhaps macroeconomic forecasters were irrational or perhaps the surveys were poor measures of inflation expectations. The major consequence was that forecast surveys developed a poor reputation and many researchers ignored them as a source of data on people's expectations.[2]

But perhaps the researchers in the early 1980s were hasty in their denunciation of the surveys. If the researchers were correct, then it should have been a simple task to use their empirical results and provide new and improved forecasts that were better than those sold in the market. The question is, were their results special to the data sample of

---

[1] For a recent example, see Ang, Bekaert, and Wei (2005).

[2] See Maddala (1991) and Thomas (1999) for literature reviews.

the time? Also, were their results possibly an artifact of the data they were using? After all, the GDP deflator is revised substantially. And the sample period in which most of the earlier tests were performed was a time with numerous shocks to relative prices, which were only slowly reflected in the GDP weights. As a result, a real-time analysis of the data is paramount.

This paper is the first to use a variety of real-time data serving as alternative bases for evaluation of forecast accuracy of inflation forecasts based on the output price index. Both the Livingston Survey and the Survey of Professional Forecasters are analyzed. Data revisions are examined to show how they might contribute to different results, depending on the choice of real-time data that are used in the forecast evaluation.

THE DATA

In examining data on inflation and forecasts of inflation, we must account for the noise in high frequency measures of the data. Analysts typically do not care about monthly or quarterly movements of inflation but often analyze it over longer periods, such as one year. Because forecasts are often taken at such a frequency, the focus of this paper is on inflation and inflation forecasts measured over (roughly) a one-year time span.[3]

The real-time data set that we use to evaluate the forecasts is based on the work of Croushore and Stark at the Federal Reserve Bank of Philadelphia.[4] Their real-time data

---

[3] Bryan and Cecchetti (1994) provide a cogent description of the noise in inflation data.

[4] Croushore and Stark (2001) describe the structure of the real-time data set for macroeconomists and evaluate data revisions to some variables. Stark and Croushore (2002) show how data revisions affect

set for macroeconomists collects snapshots of numerous macroeconomic time series data once each quarter since November 1965. Data within any vintage of the data set can be used to show precisely what data were available to a forecaster at any given date. The GDP deflator is one of the variables included within the data set. The timing of the vintages is as of the middle day of the middle month of each quarter.

The only two surveys that span the period from the 1970s to today with forecasts for the GDP deflator are the Livingston Survey and the Survey of Professional Forecasters. The Livingston Survey, which began in the 1940s, collects its forecasts from a wide array of economists, not all of whom have forecasting as their main job.[5] The Survey of Professional Forecasters (SPF), which was known as the ASA-NBER Survey from 1968 to 1990 before it was taken over by the Federal Reserve Bank of Philadelphia, collects its forecasts from economists for whom forecasting is a major part of their jobs.

The Livingston Survey collects economists' forecasts of real output and nominal output (GNP until 1991, GDP since 1992). From these forecasts, we can calculate the implicit forecasts of inflation in the GNP or GDP deflator. Survey forms are sent out about mid-May and mid-November each year and are due back in early June and December. Because the first-quarter values of real output and nominal output are revised in late May each year, we assume that the forecasters knew the revised numbers before making their forecast, so we include those data in our real-time data set. Similarly, we assume the forecasters know the value for the third quarter that is released in late

---

forecasts, while Croushore and Stark (2003) illustrate how data revisions have influenced major macroeconomic research studies. See the Philadelphia Fed's website for the data set at: www.philadelphiafed.org/econ/forecast/reaindex.html.

[5] See Croushore (1997) for details on the Livingston Survey. The survey's data are all available online at: www.philadelphiafed.org/econ/liv/index.html.

November before making their forecasts. This assumption means that the timing of the data is slightly different from that in the real-time data set for macroeconomists, so we collected all the late May and late November values to ensure that we include in our information set all the data available to the survey participants. Because the survey calls for forecasts through the second quarter of the following year (for surveys due in June) and the fourth quarter of the following year (for surveys due in December), the forecasters are really making five-quarter-ahead forecasts. Although the survey itself began in 1946 and forecasts for nominal output have been part of the survey since it began, forecasts for the level of real output did not begin until June 1971. So we begin our sample with that survey. Our sample ends with the survey made in June 2004 because that is the last survey whose one-year-ahead forecasts we can evaluate. To avoid idiosyncratic movements in the forecasts, we examine the median forecast across the forecasters.

The Survey of Professional Forecasters collected forecasts of the GNP deflator from 1968 to 1991, the GDP deflator from 1992 to 1995, and the GDP price index since 1996.[6] The GNP deflator, GDP deflator, and GDP price index behave quite similarly, and there is no apparent break in the inflation series generated from the three different price indexes in either 1992 or 1996. From these forecasts, we can calculate the implicit forecasts of inflation. Survey forms are sent out four times a year after the advance release of the national income and product accounts in late January, April, July, and October and are due back before the data are revised in February, May, August, and

---

[6] See Croushore (1993) for more on the SPF. The survey forecasts can be found online at http://www.philadelphiafed.org/econ/spf/index.html.

November. As a result, the survey forecasts match up exactly with the real-time data set for macroeconomists. The survey calls for forecasts for each quarter for the current and following four quarters, so we can construct an exact four-quarter-ahead forecast. The timing can be seen in the following example: in late January 2005, the national income account data are released and the forecasters know the values of the GDP price index from 1947:Q1 to 2004:Q4. They forecast levels of the GDP price index for 2005:Q1, 2005:Q2, 2005:Q3, 2005:Q4, and 2006:Q1. We examine their one-year-ahead forecasts based on their forecast for 2006:Q1 relative to their forecast for 2005:Q1. Thus, the forecasts span a one-year (four-quarter) period, though it may be relevant to note that the end of their forecast horizon (2006:Q1) is five quarters after the last date for which they observe a realization (2004:Q4). Although the survey itself began in 1968, the early forecasts for the GNP deflator were rounded to the nearest whole number, which causes the forecasts to be quite erratic in the early years of the survey. Because of this, and to analyze the Livingston Survey and SPF forecasts on the same sample period, we look at the SPF forecasts made between 1971:Q1 and 2004:Q2. Our sample ends with the surveys made in 2004:Q2, because that is the last survey whose one-year-ahead forecasts we can evaluate, given realized data through 2005:Q2 (the latest available data when the first draft of this paper was written). As with the Livingston Survey, to avoid idiosyncratic movements in the forecasts, we examine the median forecast across the forecasters.

Let us begin our data analysis by looking at plots of the forecasts over time and some measures of the actual values of the inflation rate. For the Livingston Survey, we plot forecasts and actuals based on latest available data (as of August 2005) in Figure 1 from 1971:H1 to 2004:H2. The dates on the horizontal axis represent the dates at which a forecast was made. The corresponding "forecast" point is the forecast for the five-quarter period from the first quarter of the current year to the second quarter of the following year for June surveys, and from the third quarter of the current year to the fourth quarter of the following year for December surveys. For example, the data points shown in the upper panel for 2004:H1 are: (1) the forecast from the June 2004 Livingston Survey for the inflation rate in the GDP price index from 2004:Q1 to 2005:Q2; and (2) the actual inflation rate in the GDP price index based on latest available data (dated August 2005) from 2004:Q1 to 2005:Q2. In our date notation, "H" means "half year"; so, for example, the survey from 2004:H1 means the survey made in the first half of 2004, which was released in June 2004. In the lower panel of Figure 1, the forecast error is shown (defined as the actual value minus the forecast).

Figure 2 shows a similar plot for the Survey of Professional Forecasters. For the SPF, we plot forecasts and actuals based on latest available data (as of August 2005) in Figure 2 from 1971:Q1 to 2004:Q2. As in Figure 1, the dates on the horizontal axis represent the dates at which a forecast was made. The corresponding "forecast" point is the forecast for the four-quarter period from the date shown on the horizontal axis; for example, the data points shown in the upper panel for 2004:Q2 are: (1) the forecast from the May 2004 SPF for the inflation rate in the GDP price index from 2004:Q2 to 2005:Q2; and (2) the actual inflation rate in the GDP price index based on latest available data (dated August 2005) from 2004:Q2 to 2005:Q2. In our date notation, "Q" means "quarter"; so, for example, the survey from 2004:Q2 means the survey made in the second quarter of 2004, which was released in May 2004. In the lower panel of Figure 2, the forecast error is shown (defined as the actual value minus the forecast).

The figures for both the Livingston Survey and the SPF have four features in common: (1) inflation rose much higher than the forecasters thought it would after the oil-price shocks of the 1970s (more so in the Livingston Survey than in the SPF); (2) the forecasters were slow to reduce expected inflation in the early 1980s and their forecast errors were negative for a time (more so in the SPF than in the Livingston Survey); (3) forecast errors were fairly close to zero from the mid-1980s to about 2000; however, the SPF forecasts were persistently too high by a small amount in most of the 1990s;[7] and (4) the upswing in inflation in the past few years caught forecasters by surprise.

REAL-TIME DATA ISSUES

Before we examine the quality of the forecasts, we must tackle the difficult issue of what to use as actuals for calculating forecast errors. In the discussion above, we based our analysis solely on the latest available data (as of August 2005), which is what is typically done in the forecasting literature. But forecasters are quite unlikely to have anticipated the extent of data revisions to the price index that would not occur for many years in the future. More likely, they made their forecasts anticipating the same methods of data construction being used contemporaneously by the government statistical agencies.

How big are the revisions to the data on the price index for output? In the real-time data set we can consider a variety of actuals, including the value recorded one-year

---

[7] For more on this issue, see Croushore (1998b).

after the initial release, the value recorded three years after the initial release, the last value recorded before a new benchmark revision occurs (a concept that maintains a consistent method by which the government calculates growth rates, including the same base year), and the value recorded in the latest available data (as of August 2005). How different are these alternative concepts of actuals? And how large are the consequent data revisions?

Figure 3 shows all four concepts for actuals that we use in this paper for the four-quarter inflation rate. For each date shown on the horizontal axis, the actual value is defined as the inflation rate from four quarters ago to that date. In the figure, it is hard to see big differences across the vintages, but the differences are as large as 1.4 percent, which could lead to somewhat different forecast errors. Also, there are periods with persistent differences between the actuals, as in 1971 to 1973, 1979 to 1980, 1989 to 1991, and 1994 to 1998.

Revisions from initial release to each of the actuals also vary substantially, as the four panels of Figure 4 show.[8] The four panels all have the same scale, so you can observe the relative size of the revisions to the data and plot both quarterly data (so you can observe when shocks to the one-quarter inflation rate occur) and four-quarter data (our central object of interest). The revisions based on one-year later data in panel a and the pre-benchmark revisions in panel c are fairly similar in size, although the largest revisions are quite different in magnitude. But the three-year revisions in panel b and the revisions from initial to latest available in panel d are much larger and more volatile. A

[8] Aruoba finds that the annual inflation rate in the output deflator has a nonzero mean revision, though his concept of revision is slightly different from the one we use here, since he does not look at revisions caused by benchmark revisions.

number of revisions in the four-quarter inflation rate exceed 1 percent. Histograms describing the distribution of the revisions (Figure 5, panels a, b, c, and d) show the range of revisions and their relative size and probability. With such large revisions, tests of forecasts might well be affected, depending on what actuals are chosen to use in the evaluation.

Because the scaling in Figures 3 and 4 have a wide range to accommodate all the data, it is hard to discern how a particular number was revised. But for any observation date, it is possible to track the inflation rate as it is successively revised across all the vintages since its initial release. This is done in Figure 6 for the inflation rate between 1973:Q1 and 1974:Q1. The inflation rate for this period was initially released as 8.4 percent in the vintage of May 1974, revised to 9.1 percent in August 1974, then back down to 8.3 percent in February 1976. It bounces around a bit before being revised down to 7.5 percent in February 1981. By February 1986 it is revised back up to 8.2 percent, bounces around a bit more over time, then is cut to 7.2 percent (its historical low point) in the data vintage of February 1996, which is the first vintage with chain weighting. Minor redefinitions since then cause the number to fluctuate slightly, and in the latest available data set (August 2005), it is 7.6 percent. With so many fluctuations in the measure of the inflation rate, it is clear that the result of any statistical method that evaluates a forecast made in 1973:Q1 for the coming year is going to depend significantly on what is chosen to serve as the actual value of the variable.

Even in the longer run, revisions to the inflation rate are significant. To see this, consider the average inflation rate over five-year intervals, measured at different vintage dates. In Table 1, we consider vintage dates for every pre-benchmark vintage and show the five-year average inflation rate. Across vintages, the five-year average inflation rate changes by as much as 0.7 percentage point. Thus, the value of the inflation rate for substantial periods of time is not precisely measured.

ARE THE FORECASTS BIASED?

In the literature on testing forecasts for accuracy, a key test is one to see if the forecasts are biased. A forecast is biased if forecasts differ systematically from their realized values. We examine bias in forecasts by regressing the actuals on the forecasts and testing the null hypothesis that the constant term equals zero and the slope coefficient equals one.[9] If the null hypothesis is rejected, then the forecasts are biased. We will examine regressions of this type, using our four alternative definitions of actuals to see if the test results are sensitive to the choice of actuals.

A glance at the early years shown in Figures 1 and 2 suggests that the forecasts are biased. Because many of the studies of bias in the survey data were undertaken in the early 1980s, during the period in which the theory of rational expectations was being tested empirically, it is clear why the tests suggested that the survey forecasts were not rational. Scatter plots of the data from both surveys allow us to eyeball the bias in the surveys, as shown in Figure 7. In the period from 1971 to 1981, there is a clear tendency in both surveys for the forecasts to be too low (points to the left of the 45-degree line) relative to actuals. After that, however, from 1982 to 2004 the forecasts are much better

[9] Alternative tests for bias include calculating the forecast error (defined as actual minus forecast) and testing whether the mean over time of the forecast errors is zero, or performing a sign test, which tests for a zero median. These tests are performed in the next section.

in both surveys, with a slight tendency in the SPF for the forecasts of inflation to be too high.

To examine the bias more formally, we run the bias test, based on the regression:

$$\pi_t = \alpha + \beta \pi_t^f + \varepsilon_t , \tag{1}$$

where π_t is the actual inflation rate and π_t^f is the forecast at each date t. If the forecasts are not biased, we should estimate $\hat{\alpha} = 0$ and $\hat{\beta} = 1$, as first suggested by Theil (1966). Webb (1987), however, has challenged this view of bias, arguing that even if we reject this joint hypothesis, data revisions, coefficients that change over time, and peso-type problems may prevent someone from using the results of Equation (1) to make better forecasts.

Most studies have focused on bias tests for inflation forecasts by looking at the consumer price index. Only a few studies have examined the forecasts of the output price index as we do here. A number of papers investigate bias in the inflation forecasts but suffer from a failure to account for the nature of real-time data, including Dua and Ray (1992), Hafer and Hein (1985), Rhim et al. (1994), and Vanderhoff (1984). The most notable study that accounts for data revisions and finds bias in the forecasts is Zarnowitz (1985), who finds bias in the SPF forecasts from 1968 to 1979, using pre-benchmark vintages as actuals. By contrast Keane and Runkle (1990) find evidence against bias using individual data, rather than the survey average data that Zarnowitz used. Following up on Zarnowitz, Baghestani and Kianian (1993) found the SPF output deflator forecasts biased from 1981 to 1991, though they used the advance NIPA release as actuals.

In testing forecasts over a four-quarter (SPF) or five-quarter horizon (Livingston), we face the issue of overlapping observations. Because the forecasts span a longer period than the sampling frequency, any shock affects the actuals for several consecutive periods. For example, an oil-price shock in 1973:Q2 affects any measurement of actuals that includes that quarter and therefore the forecast errors for forecasts made over the period, including 1973:Q2. For the SPF, this means that the forecast errors from surveys taken in 1972:Q2, 1972:Q3, 1972:Q4, 1973:Q1, and 1973:Q2 are all correlated; for the Livingston Survey, correlation occurs among forecast errors from surveys taken in 1972:H1, 1972:H2, and 1973:H1. To allow for these overlapping observations, we must either cut the SPF sample into five pieces (taking every fifth observation) and the Livingston Survey into three pieces, or adjust the covariance matrix using methods suggested by Brown and Maital (1981), using the method of Hansen and Hodrick (1980), perhaps as modified by Newey and West (1987) to guarantee a positive definite covariance matrix.

Table 2 shows the results of the test for bias for the Livingston Survey and the SPF, with both alternative methods of accounting for overlapping observations and with two different actuals, latest available and pre-benchmark. The test is run for both the first 11 years of the sample (1971 to 1981) and for the full sample. The coefficient estimates for Equation (1) are shown along with the $\bar{R}^2$, the Durbin-Watson statistic, and the p-value for the joint hypothesis test of a zero constant and slope = 1. When we use the sample with overlapping observations, the p-value is based on a chi-squared test using the Newey-West method for adjusting the covariance matrix. When we split the sample to avoid the problem of overlapping observations, the p-value is based on an F test. A p-value of less than 0.05 indicates a rejection of the null hypothesis of unbiasedness.

For the Livingston Survey, shown in panel a, using latest available data on actuals, we reject the null hypothesis of unbiased forecasts for the 1971 to 1981 period; but over the full sample, we do not reject the null. We get the same result no matter how we deal with the overlapping-observations problem with actuals being the latest available data. This result is weakened somewhat when we use pre-benchmark vintages as actuals, in which case the adjustment of the covariance matrix to deal with the overlapping-observations problem on the early sample leads to no rejection of the null hypothesis. But for the three split samples, we continue to reject the null hypothesis for the 1971 to 1981 period. For the SPF, however, the survey looks more unbiased. The only rejection that we have is for the overlapping-observations sample for the 1971 to 1981 period; in all other cases, we do not reject the null. Looking at other actuals (one year later and three years later), the pattern in terms of rejections of the null hypothesis is exactly like that for pre-benchmark actuals for both surveys, so those results are not reported here.

FORECAST-IMPROVEMENT EXERCISES

The next question we seek to answer is: if you observed the pattern of past forecast errors, could you have used the knowledge to make better forecasts? Consider trying to improve on the forecasts in the following way. Run the bias regression in Equation (1), estimate $\hat{\alpha}$ and $\hat{\beta}$, then create a new and improved forecast, π_t^i:

$$\pi_t^i = \hat{\alpha} + \hat{\beta}\pi_t^f . \tag{2}$$

Those who argued in the early 1980s that the forecasts were irrational suggested that this approach would have led forecasters to have much smaller forecast errors than in fact they had. But suppose we had followed their advice over time. How big would the subsequent forecast errors be? And would following this advice lead to a lower root-mean-squared forecast error (RMSFE)?

Running this experiment, using real-time data, leads to the results shown in Table 3. We try the forecast-improvement exercise in three different ways. The "full sample" results use all the forecasts starting in 1971 in a rolling fashion. We begin with the forecasts made in 1983:H1 for the Livingston Survey and 1978:Q1 for the SPF, to ensure that our regression has at least 20 observations. We run the regression given by Equation (1), use the estimates of $\hat{\alpha}$ and $\hat{\beta}$, then create a new and improved forecast, π_t^i according to Equation (2). Then we step forward one period, add one more data point to the sample, then rerun Equation (1) and form a new and improved forecast from Equation (2). For each date, we collect just the one-year-ahead forecast formed at that date. We proceed in this fashion through the end of the sample. Then, we calculate the root-mean-squared error (RMSE) of the forecasts made by the forecast-improvement method and compare them with the RMSE of the survey itself.

The columns labeled "10-year Window" and "5-year Window" repeat the exercise described above. But instead of using all the forecasts going back to 1971, they use just 10 years of forecasts and five years of forecasts, respectively, ignoring the early forecasts. The idea here is that the degree of bias in the forecasts may have changed over time, so this method allows us to ignore the older forecasts in attempting to improve later forecasts.

The results show that the use of Equation (2) to improve the survey results is not very fruitful. The RMSE for any attempt at forecast improvement is higher than the RMSE for the original survey. Interestingly enough, the root-mean-squared-forecast error for the original survey is not much affected by whether we use as actuals the one-quarter-later value or the one-year-later value.

The results are somewhat sensitive to the sample period chosen. However, if the sample begins earlier, there are fewer observations in the early part of the sample period, so the attempt at forecast improvement is thwarted by additional sampling uncertainty. If the sample begins later, the surveys are more accurate, and there is less to improve upon, as the period of substantial bias in the survey drops out of the sample.

ADDITIONAL TESTS FOR FORECAST ACCURACY

Diebold and Lopez (1996) suggest a variety of tests that forecasts should pass, other than bias tests.

Sign Test. The sign test is based on the hypothesis that forecast errors are independent with a zero median. As a result, the number of positive observations in the sample has a binomial distribution. The sign test is valid only for samples that do not have overlapping observations, so we cut the Livingston sample into three parts and the SPF sample into five parts to run the test.

The results are shown in Table 4. Choosing a significance level of 0.05, we reject the null for a sample split into n sub-samples at a significance level of $0.05/n$. More precisely, if we have n independent samples, for each sample the relevant significance level is α, where α is given by the equation: $1 - (1 - \alpha)^n = 0.05$. For $n = 3$, $\alpha = 0.0170$; for $n = 5$, $\alpha = 0.0102$. If the p-value shown in the right-hand column is less than this, we would reject the null hypothesis.

In no case do we reject the null hypothesis that the forecast errors are independent with a zero median. Thus the forecasts appear to pass the sign test, at least for using latest available data as actuals and pre-benchmark data as actuals.

Wilcoxon Signed-Rank Test. Under the same null hypothesis of independent errors with a zero median, the Wilcoxon signed-rank test also requires a symmetric distribution, thus accounting for the relative magnitudes of the forecast errors. The test statistic is the sum of the ranks of the absolute values of the positive forecast errors, where the forecast errors are ranked in increasing order, which is distributed as a standard normal.

Table 5 shows the results, with p-values reported in the last column. The null hypothesis is never rejected for any of the subsamples for both surveys.

Zero-Mean Test. Good forecasts should be such that the mean forecast error is zero. Again, as Table 6 shows, the forecasts pass the test easily, as the mean of each forecast is quite close to zero, whether we use latest available actuals or pre-benchmark actuals.

Dufour Test. The Dufour test is a bit more sophisticated than the sign test and the Wilcoxon signed-rank test, but it is based on the same principle. It follows the same structure as the Wilcoxon test but is applied to the product of successive forecast errors, thus testing whether the forecast errors are white noise.

The results are shown in Table 7. Again, in no case do we reject the null hypothesis. Thus, we cannot reject the hypothesis that the forecast errors are white noise.

Pearce Test. Perhaps the most convincing study of rational expectations in the 1970s was Pearce (1979). Pearce made the observation that, looking at CPI data, if you were to simply estimate an ARIMA model using standard Box-Jenkins techniques, you would create better forecasts than the Livingston Survey had. This test was so simple that it convinced even the most diehard supporter of rational expectations that something was wrong with the survey forecasts. But, again, Pearce's sample period happened to be fairly short and was one in which inflation generally rose unexpectedly. With more than 20 more years of data, is Pearce's result still valid? Croushore (1998a) argues that it is not for the case of the CPI.[10] We can run a similar exercise for the output price index.

Pearce assumed that the model appropriate for inflation was an IMA(1,1) process.[11] We run one set of experiments based on that process, and another in which we assume an AR process that is determined by calculating SIC values period by period, thus allowing the model to change over time. We also run two permutations, one with real-time data and one with latest available data, to see how much real-time data matters for creating such forecasts, as in Stark and Croushore (2002).

The results of the exercise are shown in Table 8. There is little support for the view that a simple ARIMA model can do better than the survey forecasts. The best evidence in favor of the ARIMA model comes from using latest available data rather than real-time data, and only for the SPF when latest available data are used as actuals. Thus, there is no way that a forecaster in real time could have used an ARIMA model to improve on the survey forecasts.

---

[10] Ang, Bekaert, and Wei (2005) perform a similar test with the same result.

[11] See Stock and Watson (2006) for a detailed analysis of forecasting inflation using an IMA(1,1) process.

CONCLUSION

The Livingston Survey and the Survey of Professional Forecasters developed poor reputations because of the systematic pattern of forecast errors found in the 1970s. Using basic statistical tests, researchers found that the forecast errors from the surveys failed to pass a number of basic tests, most importantly the Pearce test. But when we look at a much longer sample of data, which goes beyond the years in which movements of inflation were dominated by oil-price shocks and bad monetary policy, we find that the inflation forecasts pass those statistical tests convincingly. In addition, the evaluation of forecast errors depends in part on the choice of actuals, with actuals taken to be latest available data providing the least favorable evaluation of the forecasts.

The bottom line is: if you want a good measure of inflation expectations, you should use the forecasts from the SPF or Livingston surveys.

REFERENCES

Ang, Andrew, Geert Bekaert, and Min Wei. "Do Macro Variables, Asset Markets, or Surveys Forecast Inflation Better?" Working Paper, July 2005.

Aruoba, S. Boragan. "Data Revisions Are Not Well Behaved." Manuscript, University of Maryland, December 2004.

Baghestani, Hamid M., and Amin M. Kianian. "On the Rationality of US Macroeconomic Forecasts: Evidence from a Panel of Professional Forecasters." *Applied Economics* 25 (1993), pp. 869–878.

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

Bryan, Michael F., and Stephen G. Cecchetti, "Measuring Core Inflation," in N. Gregory Mankiw, ed., *Monetary Policy*. Chicago: University of Chicago Press, 1994, pp. 195-215.

Croushore, Dean. "Introducing: The Survey of Professional Forecasters," Federal Reserve Bank of Philadelphia *Business Review* (November/December 1993), pp. 3-15.

Croushore, Dean. "The Livingston Survey: Still Useful After All These Years," Federal Reserve Bank of Philadelphia *Business Review*, March/April 1997, pp. 15-27.

Croushore, Dean. "Evaluating Inflation Forecasts," Federal Reserve Bank of Philadelphia Working Paper 98-14, June 1998a.

Croushore, Dean. "Low Inflation: The Surprise of the 1990s," Federal Reserve Bank of Philadelphia *Business Review*, July/August 1998b, pp. 3-12.

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

Croushore, Dean, and Tom Stark, "A Real-Time Data Set for Macroeconomists: Does the Data Vintage Matter?" *Review of Economics and Statistics* 85 (August 2003), pp. 605–617.

Diebold, Francis X., and Jose A. Lopez. "Forecast Evaluation and Combination," in G.S. Maddala and C.R. Rao, eds., *Handbook of Statistics*. Amsterdam: North Holland, 1996, pp. 241-68.

Dua, Pami, and Subash C. Ray. "ARIMA Models of the Price Level: An Assessment of the Multilevel Adaptive Learning Process in the USA." *Journal of Forecasting* 11 (September 1992), pp. 507–516.

Hafer, R. W., and Scott E. Hein. "On the Accuracy of Time-Series, Interest Rate, and Survey Forecasts of Inflation" *Journal of Business* 58 (October 1985), pp. 377–398.

Hansen, Lars-Peter, and Robert J. Hodrick. "Foreign Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis," *Journal of Political Economy* 88 (October 1980), pp. 829-53.

Keane, Michael P., and David E. Runkle. "Testing the Rationality of Price Forecasts: New Evidence From Panel Data," *American Economic Review* 80 (1990), pp. 714-35.

Maddala, G.S. "Survey Data on Expectations: What Have We Learnt?" in Marc Nerlove, ed., *Issues in Contemporary Economics, vol. II. Aspects of*

Macroeconomics and Econometrics. New York: New York University Press, 1991.

Newey, Whitney K., and Kenneth D. West. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55 (May 1987), pp. 703-8.

Pearce, Douglas K. "Comparing Survey and Rational Measures of Expected Inflation," *Journal of Money, Credit and Banking* 11 (November 1979), pp. 447-56.

Rhim, Jong C., Mohammed F. Khayum, and Timothy J. Schibik. "Composite Forecasts of Inflation: An Improvement in Forecasting Performance." *Journal of Economics and Finance* 18 (Fall 1994), pp. 275–286.

Stark, Tom, and Dean Croushore. "Forecasting with a Real-Time Data Set for Macroeconomists." *Journal of Macroeconomics* 24 (December 2002), pp. 507–31.

Stock, James H., and Mark W. Watson. "Why Has U.S. Inflation Become Harder to Forecast?" manuscript, September 2006.

Theil, Henri. *Applied Economic Forecasting*. Amsterdam: North Holland, 1966.

Thomas, Lloyd B., Jr. "Survey Measures of Expected U.S. Inflation," *Journal of Economic Perspectives* 13 (Fall 1999), pp. 125–144.

Vanderhoff, James. "A 'Rational' Explanation for 'Irrational' Forecasts of Inflation." *Journal of Monetary Economics* 13 (May 1984), pp. 387–392.

Webb, Roy H. "The Irrelevance of Tests for Bias in Series of Macroeconomic Forecasts," Federal Reserve Bank of Richmond *Economic Review* (November/December 1987), pp. 3-9.

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

Table 1
Average Inflation Rate of Over Five Years
For Pre-Benchmark Vintages
Annualized percentage points

Vintage Year:	‘75	‘80	‘85	‘91	‘95	’99	’03	‘05
Period								
49Q4 to 54Q4	2.6	2.7	2.7	2.5	2.4	2.6	2.5	2.6
54Q4 to 59Q4	2.6	2.6	2.6	2.9	2.9	2.4	2.5	2.5
59Q4 to 64Q4	1.4	1.5	1.5	1.6	1.6	1.3	1.3	1.3
64Q4 to 69Q4	3.6	3.9	3.9	4.1	4.1	3.7	3.7	3.7
69Q4 to 74Q4	6.3	6.5	6.2	6.8	6.5	6.3	6.3	6.3
74Q4 to 79Q4	NA	7.1	7.0	7.5	7.7	7.2	7.1	7.1
79Q4 to 84Q4	NA	NA	6.1	6.1	6.4	6.2	6.0	6.0
84Q4 to 89Q4	NA	NA	NA	3.3	3.6	3.4	3.1	3.0
89Q4 to 94Q4	NA	NA	NA	NA	2.9	3.1	2.8	2.7
94Q4 to 99Q4	NA	NA	NA	NA	NA	NA	1.7	1.6
99Q4 to 04Q4	NA	NA	NA	NA	NA	NA	NA	2.3

This table shows the inflation rates over the five year periods shown in the first column for each pre-benchmark vintage shown in the column header.

Table 2a
Test for Bias, Livingston Survey

Sample period	$\hat{\alpha}$	$\hat{\beta}$	$\bar{R}^2$	D.W.	*p-value*
Actuals=Latest					
Overlapping Observations					
First subsample:					
71H1-81H2	4.160	0.643	0.37	0.44	.012
	(1.643)	(0.189)			
Full sample:					
71H1-04H1	-0.625	1.100	0.73	0.29	.460
	(0.512)	(0.113)			
Split-Sample Results					
First subsample:					
71H1-81H2	3.160	0.752	0.29	1.89	.001
	(3.050)	(0.379)			
71H1-04H1	-0.954	1.168	0.75	1.01	.140
	(0.820)	(0.142)			
Second subsample:					
71H1-81H2	4.403	0.621	0.52	2.41	.000
	(1.898)	(0.225)			
71H1-04H1	-0.374	1.052	0.76	0.93	.720
	(0.766)	(0.129)			
Third Subsample:					
71H1-81H2	4.947	0.555	0.11	2.58	.000
	(3.397)	(0.420)			
71H1-04H1	-0.592	1.089	0.66	0.90	.493
	(0.962)	(0.167)			

Table 2a (continued)

Sample period	$\hat{\alpha}$	$\hat{\beta}$	$\bar{R}^2$	D.W.	*p-value*
Actuals=Pre-Benchmark Vintage					
Overlapping Observations					
First subsample:					
71H1-81H2	3.664	0.716	0.29	0.43	.121
	(2.379)	(0.253)			
Full sample:					
71H1-04H1	-0.770	1.143	0.71	0.35	.319
	(0.510)	(0.118)			
Split-Sample Results					
First subsample:					
71H1-81H2	2.464	0.811	0.23	1.93	.010
	(3.681)	(0.458)			
71H1-04H1	-0.881	1.150	0.72	1.43	.197
	(0.868)	(0.150)			
Second subsample:					
71H1-81H2	4.189	0.677	0.32	2.56	.000
	(2.899)	(0.344)			
71H1-04H1	-0.610	1.113	0.73	1.47	.354
	(0.869)	(0.147)			
Third Subsample:					
71H1-81H2	4.410	0.660	0.08	2.66	.000
	(4.358)	(0.539)			
71H1-04H1	-0.848	1.172	0.66	1.33	.137
	(1.042)	(0.181)			

Note: standard errors are in parentheses.

Table 2b
Test for Bias, SPF

Sample period	$\hat{\alpha}$	$\hat{\beta}$	$\bar{R}^2$	D.W.	*p-value*
Actuals=Latest Available					
Overlapping Observations					
First subsample:					
71Q1-81Q4	4.327	0.486	0.20	0.18	.012
	(1.604)	(0.229)			
Full sample:					
71Q1-04Q2	-0.200	1.036	0.64	0.13	.866
	(0.422)	(0.114)			
Split-Sample Results					
First subsample:					
71Q1-81Q4	3.922	0.537	0.24	2.15	.135
	(1.950)	(0.297)			
71Q1-04Q2	-0.164	0.997	0.67	0.96	.826
	(0.656)	(0.139)			
Second subsample:					
71Q1-81Q4	4.834	0.392	0.10	2.05	.065
	(1.854)	(0.288)			
71Q1-04Q2	-0.028	0.989	0.62	1.06	.965
	(0.693)	(0.148)			
Third Subsample:					
71Q1-81Q4	4.366	0.485	0.08	2.18	.162
	(2.438)	(0.377)			
71Q1-04Q2	-0.256	1.048	0.62	1.14	.941
	(0.736)	(0.158)			
Fourth Subsample:					
71Q1-81Q4	3.722	0.612	0.15	1.88	.140
	(2.437)	(0.390)			
71Q1-04Q2	-0.357	1.106	0.63	0.89	.795
	(0.742)	(0.165)			
Fifth Subsample:					
71Q1-81Q4	4.589	0.435	0.04	1.95	.141
	(2.420)	(0.383)			
71Q1-04Q2	-0.238	1.051	0.62	0.87	.946
	(0.724)	(0.159)			

Table 2b (continued)

Sample period	$\hat{\alpha}$	$\hat{\beta}$	$\bar{R}^2$	D.W.	*p-value*
Actuals=Pre-Benchmark Vintage					
Overlapping Observations					
First subsample:					
71Q1-81Q4	4.110	0.527	0.15	0.17	.110
	(2.254)	(0.309)			
Full sample:					
71Q1-04Q2	-0.301	1.069	0.61	0.15	.798
	(0.450)	(0.122)			
Split-Sample Results					
First subsample:					
71Q1-81Q4	4.090	0.532	0.09	2.22	.245
	(2.667)	(0.406)			
71Q1-04Q2	-0.335	1.043	0.63	1.41	.864
	(0.749)	(0.158)			
Second subsample:					
71Q1-81Q4	4.324	0.482	0.07	1.96	.196
	(2.477)	(0.384)			
71Q1-04Q2	-0.167	1.029	0.60	1.39	.975
	(0.765)	(0.164)			
Third Subsample:					
71Q1-81Q4	3.591	0.624	0.08	2.18	.314
	(3.078)	(0.476)			
71Q1-04Q2	-0.427	1.101	0.60	1.48	.846
	(0.817)	(0.174)			
Fourth Subsample:					
71Q1-81Q4	3.707	0.587	0.08	1.80	.252
	(2.793)	(0.447)			
71Q1-04Q2	-0.357	1.103	0.62	1.11	.814
	(0.757)	(0.757)			
Fifth Subsample:					
71Q1-81Q4	4.858	0.405	–0.05	1.85	.250
	(3.207)	(0.507)			
71Q1-04Q2	-0.258	1.076	0.57	1.16	.904
	(0.825)	(0.181)			

Note: standard errors are in parentheses.

Table 3

RMSFEs for Forecast-Improvement Exercises

Survey	Period	Original Survey	Attempts to Improve on Survey: Full Sample	10-year Window	5-year Window
Actuals = One-Year Later					
Livingston	1983:H1–2002:H2	0.94	1.56	1.62	1.47
SPF	1978:Q1–2003:Q2	1.09	1.66	1.76	1.69
Actuals = One-Quarter Later					
Livingston	1983:H1–2002:H2	0.98	1.52	1.60	1.45
SPF	1978:Q1–2003:Q2	1.11	1.67	1.71	1.82

Table 4
Sign Test

Survey	Period	Subsample	N	Reject null?	p-value
Actuals=Latest					
Livingston	‘71H1-’04H1	1	23	no	.1444
		2	22	no	.6698
		3	22	no	.3938
SPF	‘71Q1-’04Q2	1	26	no	.4328
		2	27	no	.1779
		3	27	no	.3359
		4	27	no	.8474
		5	27	no	.8474
Actuals=Pre-Benchmark					
Livingston	‘71H1-’04H1	1	23	no	.2971
		2	22	no	.2008
		3	22	no	.3938
SPF	‘71Q1-’04Q2	1	26	no	.1166
		2	27	no	.1779
		3	27	no	.1779
		4	27	no	.8474
		5	27	no	.5637

Table 5
Wilcoxon Signed-Rank Test

Survey	Period	Subsample	N	Reject null?	p-value
Actuals=Latest					
Livingston	‘71H1-’04H1	1	23	no	.4115
		2	22	no	.4651
		3	22	no	.5057
SPF	‘71Q1-’04Q2	1	26	no	.2376
		2	27	no	.3366
		3	27	no	.4279
		4	27	no	.7186
		5	27	no	.3488
Actuals=Pre-Benchmark					
Livingston	‘71H1-’04H1	1	23	no	.5034
		2	22	no	.3382
		3	22	no	.4264
SPF	‘71Q1-‘04Q2	1	26	no	.1742
		2	27	no	.2488
		3	27	no	.3488
		4	27	no	.5971
		5	27	no	.4004

Table 6
Zero-Mean Test

Survey	Period	Subsample	N	Reject null?	p-value
Actuals=Latest					
Livingston	‘71H1-’04H1	1	23	no	.8308
		2	22	no	.7708
		3	22	no	.7525
SPF	‘71Q1-’04Q2	1	26	no	.9791
		2	27	no	.9562
		3	27	no	.7492
		4	27	no	.7070
		5	27	no	.7070
Actuals=Pre-Benchmark					
Livingston	‘71H1-’04H1	1	23	no	.8308
		2	22	no	.7708
		3	22	no	.9157
SPF	‘71Q1-’04Q2	1	26	no	.5319
		2	27	no	.7976
		3	27	no	.3135
		4	27	no	.8208
		5	27	no	.8738

Table 7
Dufour Test

Survey	Period	Subsample	N	Reject null?	p-value
Actuals=Latest					
Livingston	‘71H1-’04H1	1	22	no	.1080
		2	21	no	.0496
		3	21	no	.0420
SPF	‘71Q1-’04Q2	1	25	no	.3533
		2	26	no	.3337
		3	26	no	.4237
		4	26	no	.3158
		5	26	no	.4237
Actuals=Pre-Benchmark					
Livingston	‘71H1-’04H1	1	22	no	.3720
		2	21	no	.0918
		3	21	no	.0792
SPF	‘71Q1-’04Q2	1	25	no	.3819
		2	26	no	.5850
		3	26	no	.5850
		4	26	no	.6752
		5	26	no	.5506

Table 8
Pearce Test
Root Mean Squared Forecast Errors

		Real-Time Data		Latest Available Data	
Sample period	Survey	IMA(1,1)	SIC	IMA(1,1)	SIC
Survey of Professional Forecasters					
Actuals=Latest					
1971:Q1–1981:Q1	2.237	2.360	3.286	2.249	3.032
1971:Q1–2004:Q2	1.538	1.589	1.998	1.507	1.819
Actuals=Pre-Benchmark					
1971:Q1–1981:Q1	2.584	2.771	3.578	2.667	3.325
1971:Q1–2004:Q2	1.691	1.786	2.147	1.719	1.980
Livingston Survey					
Actuals=Latest					
1971:H1–1981:H1	2.380	3.196	4.980	3.133	4.611
1971:H1–2003:H2	1.696	1.968	2.880	1.899	2.661
Actuals=Pre-Benchmark					
1971:H1–1981:H1	2.795	3.639	5.316	3.592	4.946
1971:H1–2003:H2	1.860	2.222	3.070	2.171	2.861

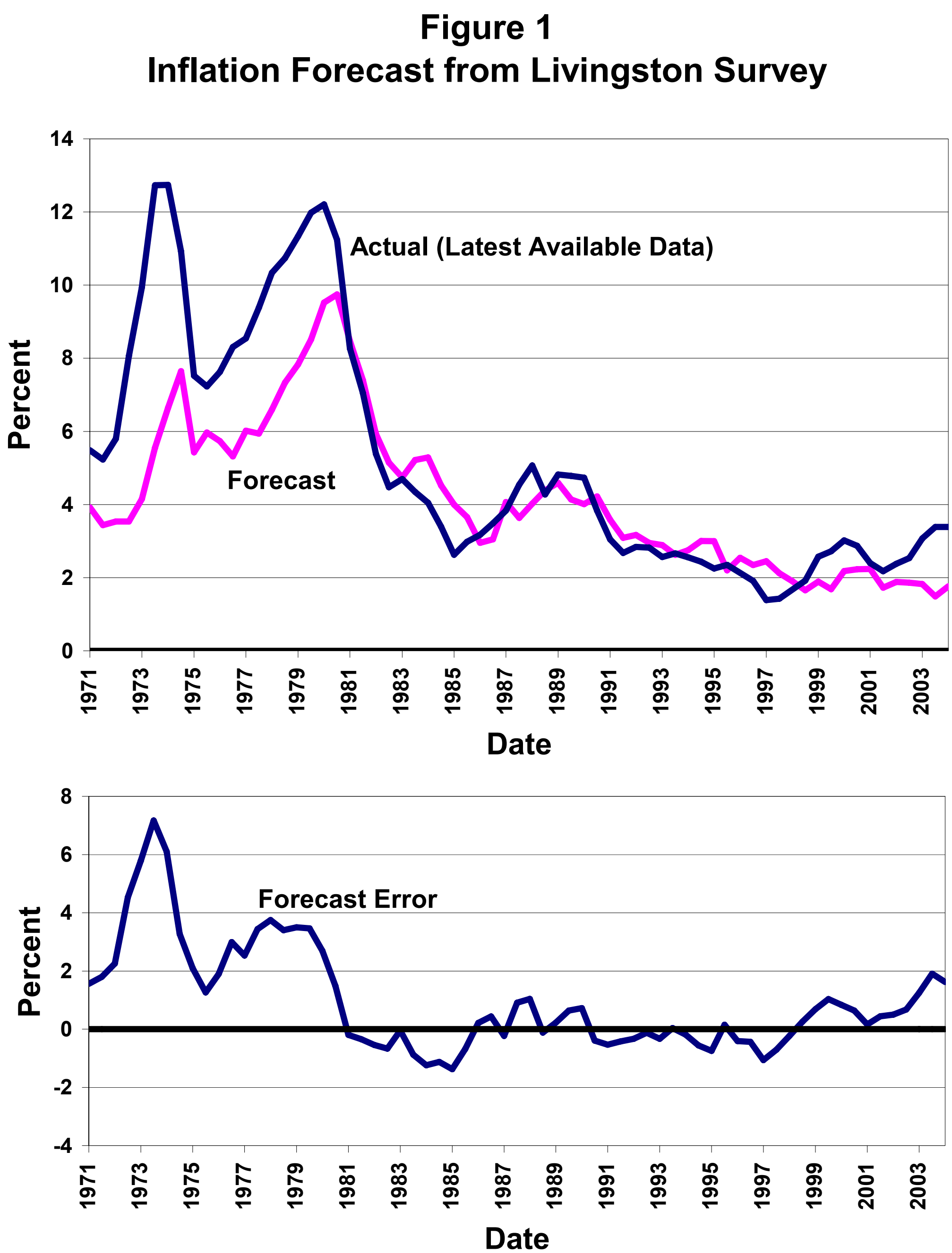


Figure 1: Inflation Forecast from Livingston Survey
These charts show, in the top panel, the inflation forecast from the Livingston Survey from 1971:H1 to 2004:H1, compared with actual values based on latest available data; and in the bottom panel, the resulting forecast error.

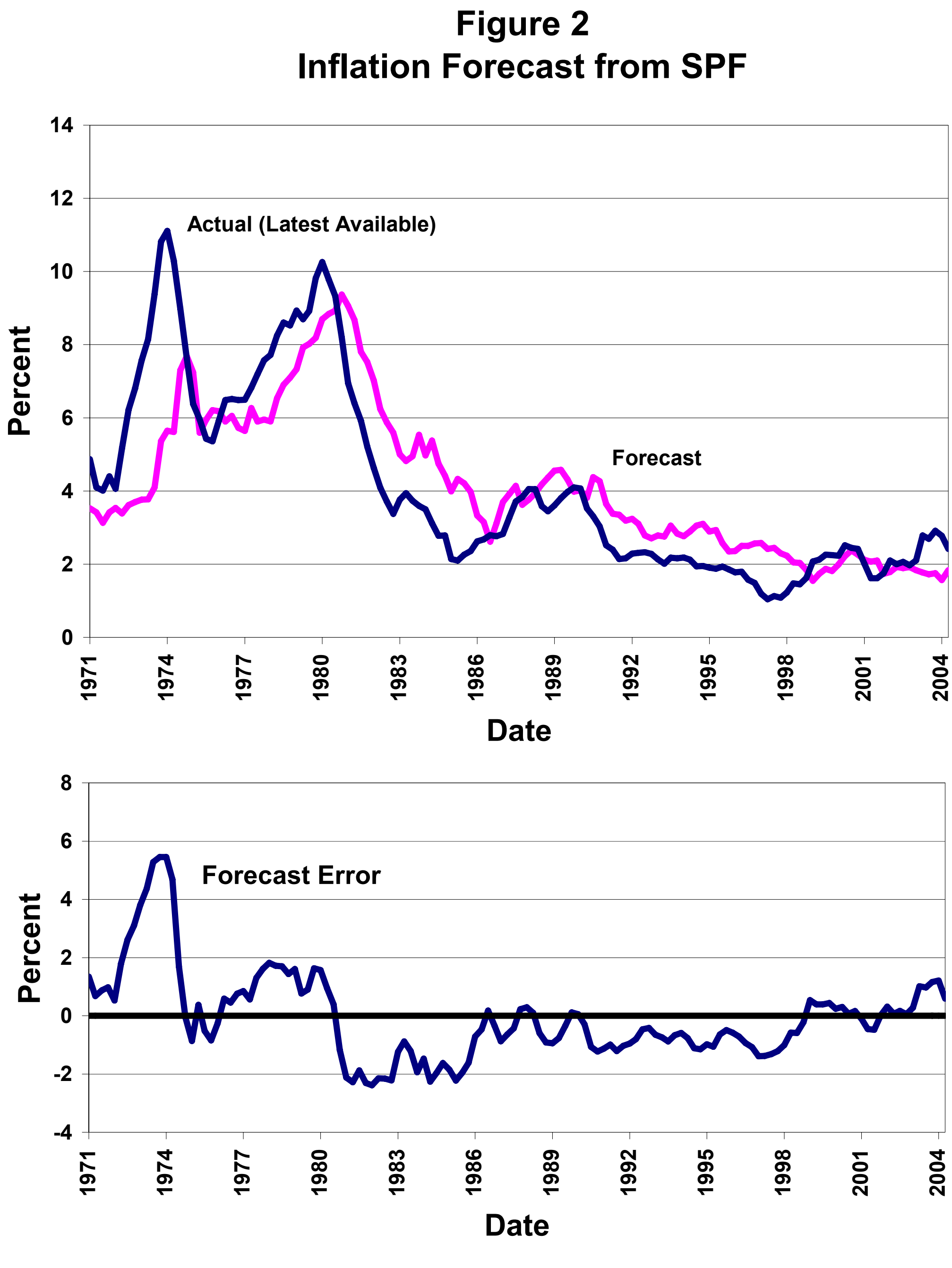


Figure 2: Inflation Forecast from SPF
These charts show, in the top panel, the inflation forecast from the SPF from 1971:Q1 to 2004:Q2, compared with actual values based on latest available data; and in the bottom panel, the resulting forecast error.

Figure 3
Alternative Actuals for the Inflation Rate

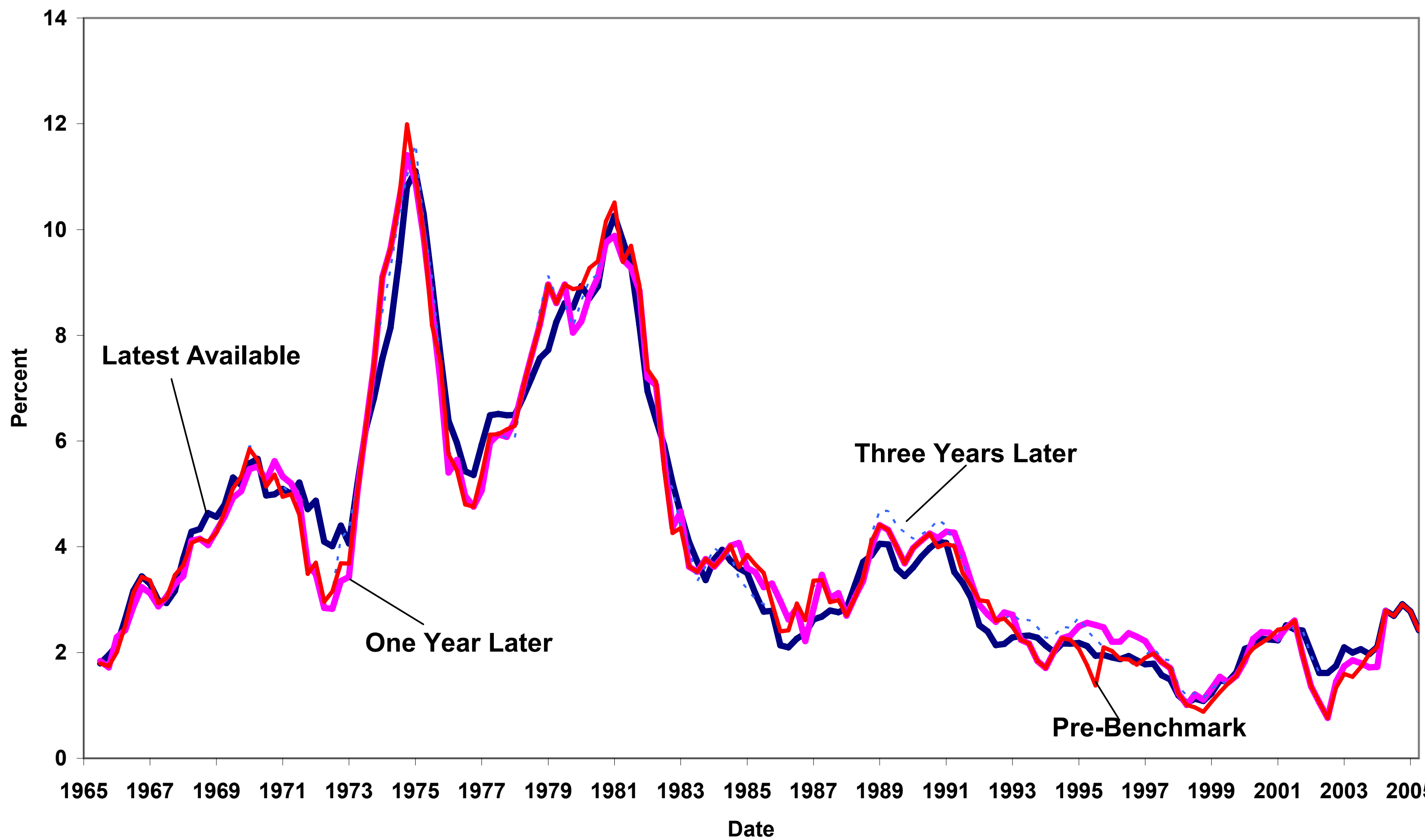


Figure 4, panel a
Revisions to Inflation Rate
Initial to One Year Later

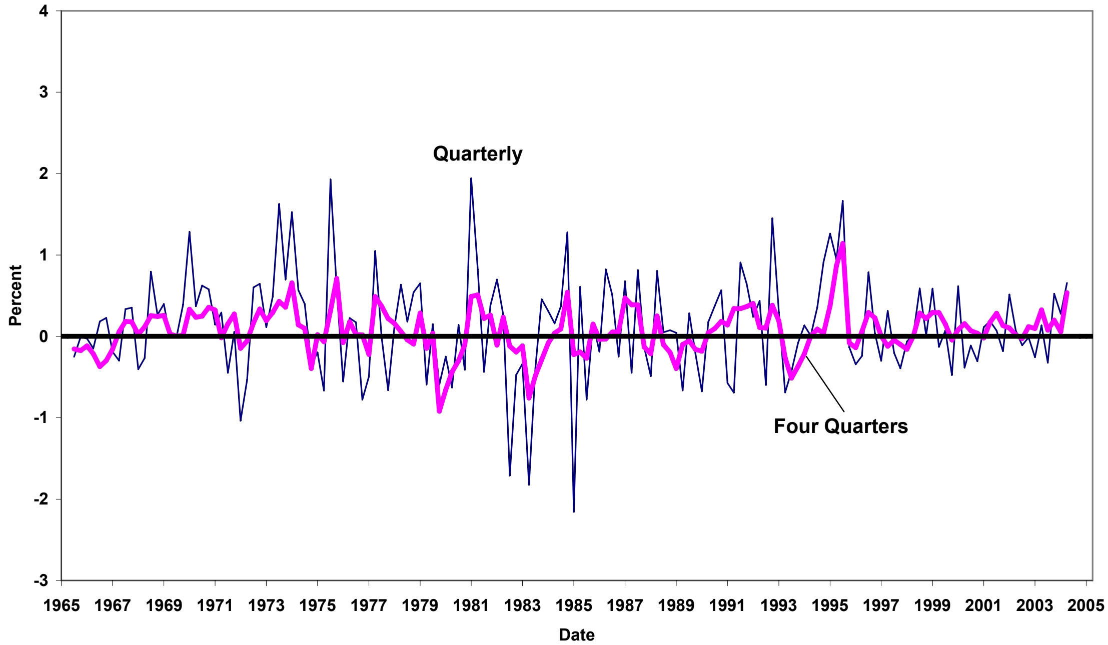


Figure 4, panel b
Revisions to Inflation Rate
Initial to Three Years Later

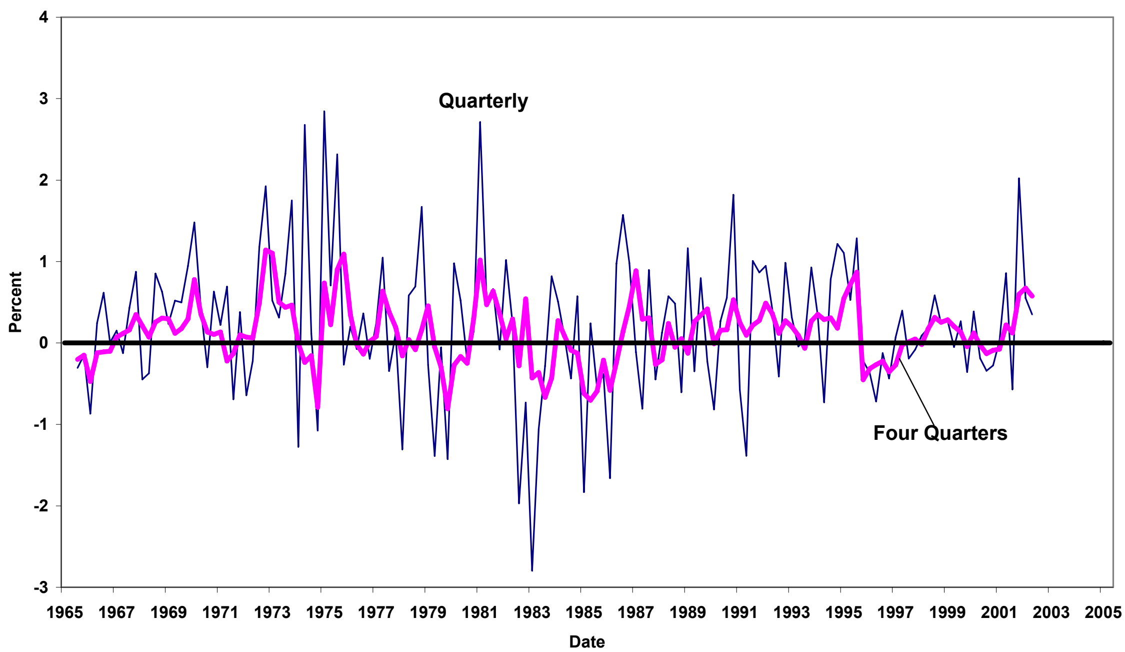


Figure 4, panel c
Revisions to Inflation Rate
Initial to Pre-Benchmark

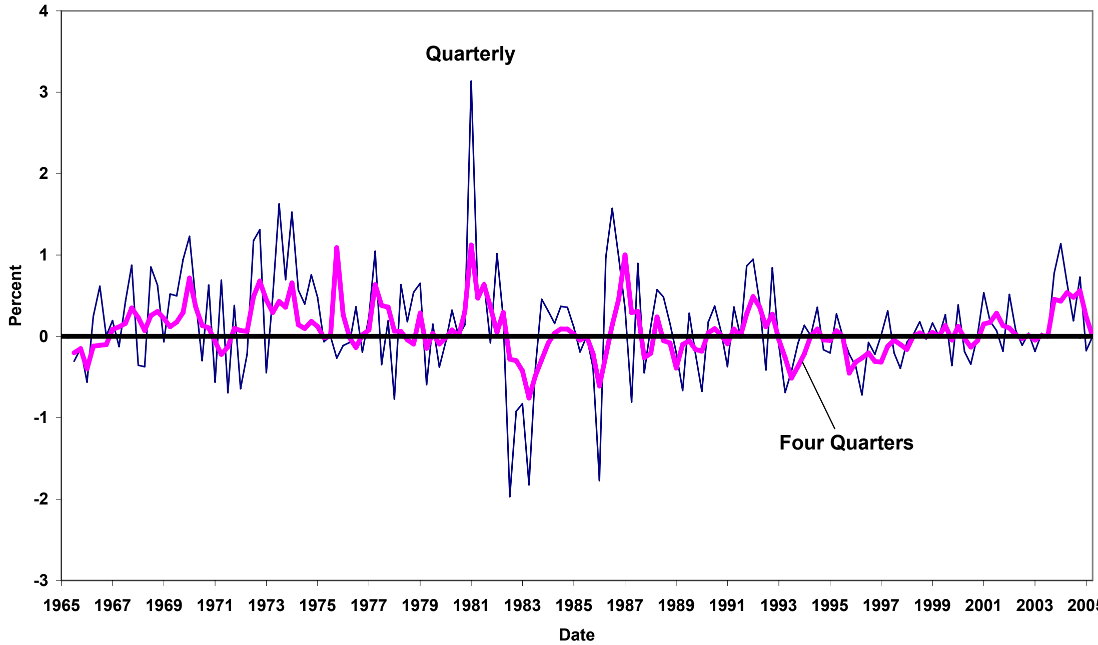


Figure 4, panel d
Revisions to Inflation Rate
Initial to Latest Available

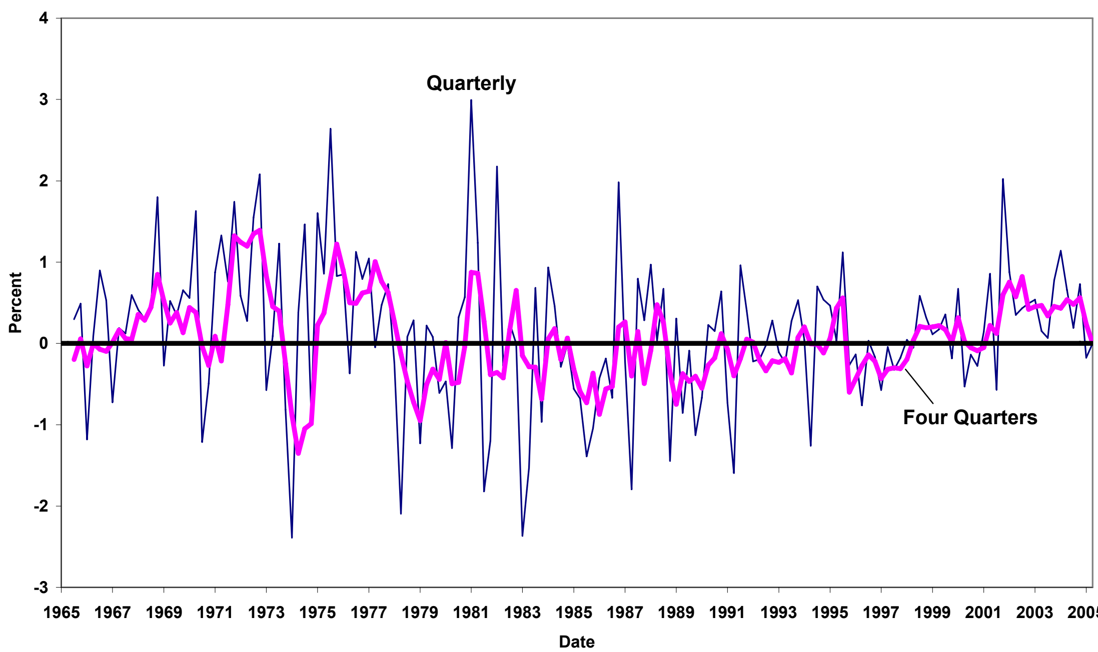
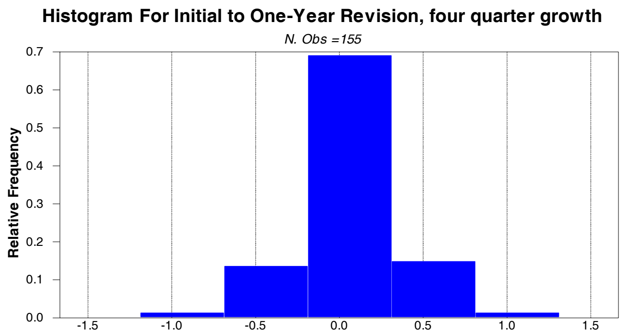
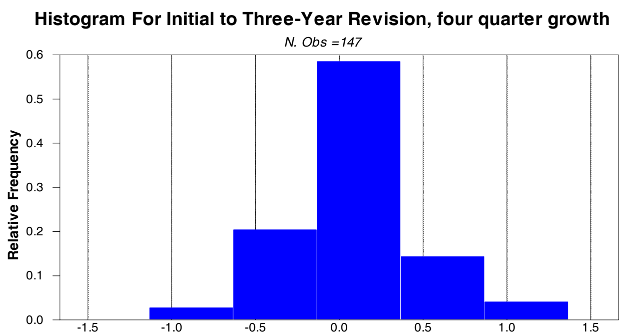
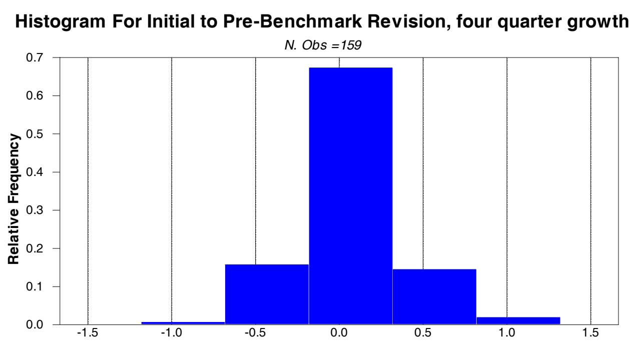


Figure 5: Histograms Showing the Distribution of Data Revisions







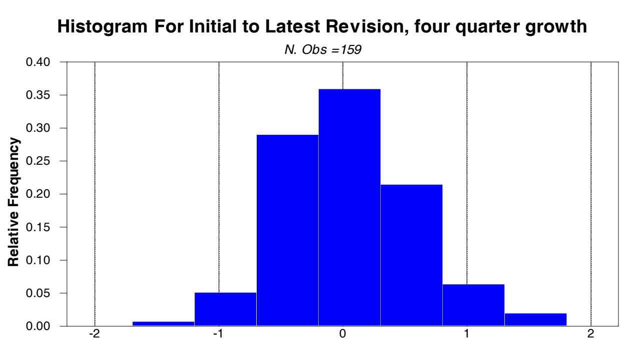


Figure 6
Inflation Rate from 1973Q1 to 1974Q1
(as viewed from the perspective of 126 different vintages)

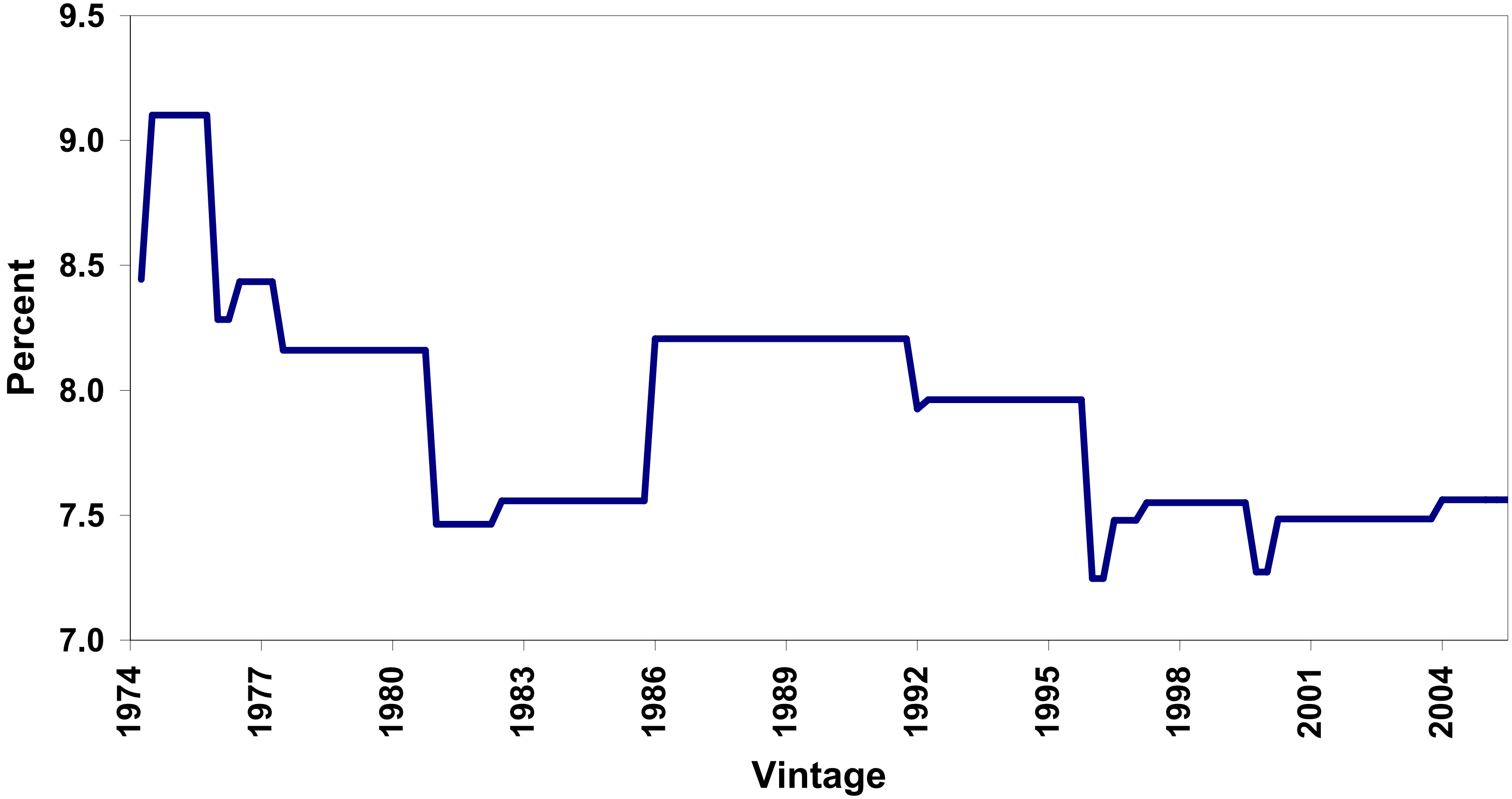


Figure 7, panel a
Forecasts Versus Latest-Available Actuals
Livingston Survey: 1971:H1 to 1981:H2

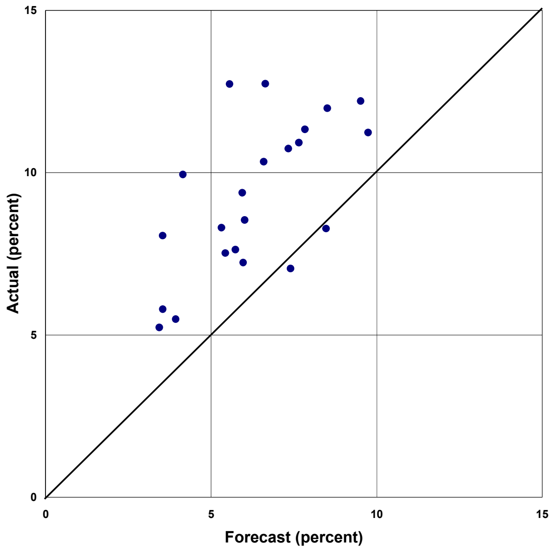


Figure 7, panel b
Forecasts Versus Latest-Available Actuals
SPF: 1971:Q1 to 1981:Q4

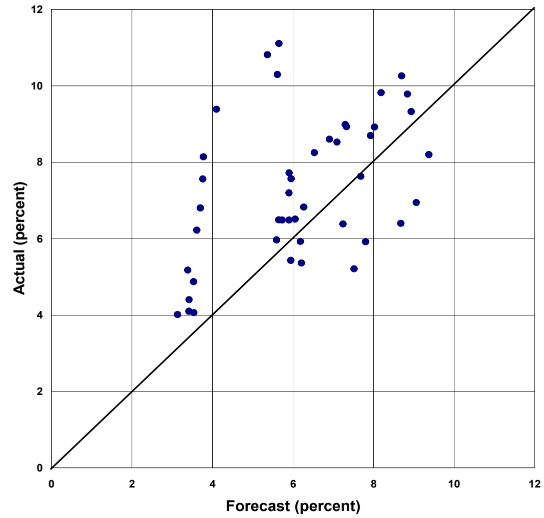


Figure 7, panel c
Forecasts Versus Latest-Available Actuals
Livingston Survey: 1971:H1 to 2004:H1

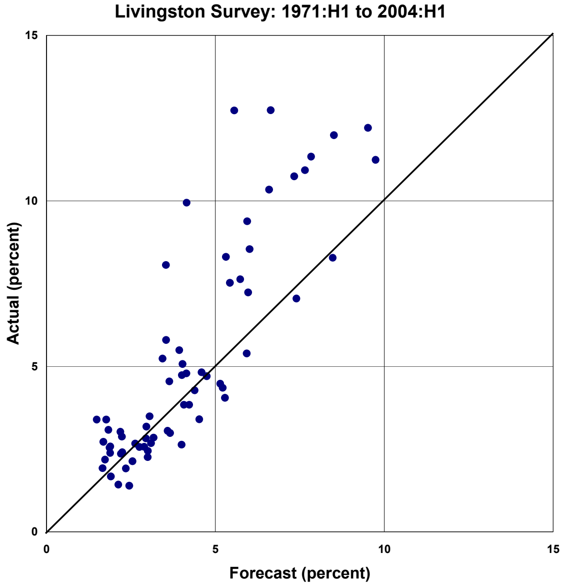


Figure 7, panel d
Forecasts Versus Pre-Benchmark Actuals
SPF: 1971:Q1 to 2004:Q2