**June 23, 2014**

The Distribution of Inflation Forecast Errors

By

Ed Gamber

Department of Economics

Lafayette College

Easton PA 18042

gambere@lafayette.edu

Jeff Liebner

Department of Mathematics

Lafayette College

Easton, PA 18042

liebnerj@lafayette.edu

Julie Smith

Department of Economics

Lafayette College

Easton PA 18042

smithjk@lafayette.edu

Abstract

This paper investigates the cross-sectional distribution of inflation forecasts errors over the period 1984 through 2007. Our working hypothesis is that the Fed’s movement toward greater transparency starting in the mid-1990s, likely affected both the distribution of forecast errors and the location of the Fed’s staff forecasts within that distribution. This paper builds on earlier work, which compared Fed forecasts to the mean or median of private sector forecasts by examining the entire distribution of forecast. By examining the entire distribution we are able to compare the forecasting record of particular forecasters against a record comprised of randomly assigned forecasts from the Survey of Professional Forecasters. Since the Fed’s move toward greater transparency beginning in 1994, it’s forecasting record is no longer significantly better than the forecasting record comprised of randomly assigned forecasts.

We thank Pierangelo DePace and the participants of the Southern Economic Association meetings in Tampa, FL for comments and helpful suggestions. We also thank participants at the Wesleyan University seminar for helpful comments. We thank Shannon Nitroy for research assistance. All remaining errors are ours.

1. Introduction

We investigate changes in the cross-sectional distribution of inflation forecast errors. We assess whether the Federal Reserve’s placement within that distribution has shifted over time, and if so, whether such shifts correspond to the Fed’s movement toward greater monetary policy transparency.

Previous studies of the Fed’s forecast performance typically (but not exclusively) compare the Fed’s forecasts with individual forecasts, or the mean or median of a sample of forecasts[[1]](#footnote-1). In this paper we take a more nuanced look at the relative forecasting performance of the Federal Reserve by examining the Fed’s location, and changing location, within the cross-sectional distribution of private-sector forecasts[[2]](#footnote-2).

Previous studies have examined both inflation and output growth forecasts, finding that the Fed’s greatest forecasting advantage is with respect to inflation. Because we are interested in changes in the Fed’s forecasting ability over time, we focus on inflation forecasts. Our measure of inflation is the annualized growth rate of quarterly CPI. By using the CPI inflation rate we are also able to compare the Fed and private sector forecasts to forecasts produced by a variety of “core” measures of inflation.

Our working hypothesis is that the Fed’s movement toward greater transparency starting in the mid-1990s, likely affected both the distribution of forecast errors and the Fed’s location within that distribution. In an earlier paper (Gamber and Smith, 2009) we showed how the Fed’s forecasting superiority has diminished since it began moving toward greater transparency in the mid-1990s. In Gamber, Smith and McNamara (2013) we divided the cross-sectional distribution of SPF forecast errors into quartiles and looked for the location of the Fed within quartiles over various sub-samples. We located a group of forecasters that consistently beat the Fed’s forecasts of output growth and inflation.

The methods we employ in this paper are related to those used by D’Agostino, McQuinn and Whelan (2012). D’Agostino et al. tested whether participants in the Survey of professional forecasters have equal forecasting ability. Using a bootstrap technique that re-assigned forecasts among forecasters they find that the distribution of the forecast accuracy scores by rank are not different than what could be found by randomly assigning forecasts. They concluded, “most of the participants in the Survey of Professional Forecasters appear to have approximately equal forecasting ability.” Their analysis does not assess the performance of individual forecasters. Our analysis does assess the performance of specific forecasts and specific forecasters. In particular we test whether the quality of the Fed’s inflation forecasts are due to random chance, or superior forecasting. We perform this same test on measures of core inflation as well.

We find that the Fed has lost ground against the median SPF forecaster after the movement toward greater transparency beginning in the early 1990s. More specifically, prior to 1994 the Fed’s forecasting record with respect to inflation was significantly better than random chance. Greater monetary policy transparency since 1993 has led to a reduction in the Fed’s forecasting record relative to random chance. With respect to the forecasters in the SPF, the very best SPF forecasters consistently beat the Fed but there is no specific SPF forecaster who consistently beats the Fed. And finally, core inflation measures such as the trimmed mean and median CPI are respectable forecasts but are not consistently better than the Fed.

 Section 2 reviews the literature. Section 3 describes our data and methods. Section 4 presents our results and section 5 concludes.

2. Previous Research

There are several previous studies looking at how the Fed’s Greenbook forecasts compare to forecasts made by the private sector[[3]](#footnote-3). Romer and Romer (2000) compared the Fed’s Greennbook forecasts to several private sector forecasts in order to explain the response of long term interest rates to changes in the federal funds rate target. Using a sample that ends in the early 1990s, they found that the Fed’s Greenbook forecasts for output growth and inflation are more accurate than the private sector’s. Using a slightly longer sample, Sims (2002) came to a similar conclusion, that is, the Fed’s forecasts beat the private sector, especially for inflation. Bernanke and Boivin (2003) and Faust and Wright (2009) found similar results. Using data through 2001, Gamber and Smith (2009) find that the Fed’s relative forecasting advantage for output growth and inflation diminished after the early 1990s.

 The evidence presented in our earlier paper (Gamber and Smith 2009) suggested that the Fed’s move toward greater transparency in the early 1990s likely resulted in a shrinking of the gap between Fed forecast errors and private sector forecast errors. In other words, through increased transparency the Fed would convey information about its intentions through press releases, and other forms of “forward guidance,” leaving less to be inferred from movements in the federal funds rate. Other researchers have looked more directly at the effects of the Fed’s move toward greater transparency. Swanson (2004) found that the Fed’s move toward greater transparency coincided with improved private sector interest rate forecasts. Ehrmann and Fratzscher (2005) found that the Fed’s introduction of policy statements in 1999 led to a change in the way that market participants learn about monetary policy intentions. Blattner, et. al (2008) provide further evidence that increased monetary policy transparency has led to improvements in market participants’ ability to anticipate monetary policy actions. The common theme running through these works is that greater transparency has led to improvements in the forecastability of monetary policy actions. The innovation in our paper is to look at whether the increased transparency has also led to improvements in the forecastability of inflation.

3. Data and Methods

In this paper we use the Consumer Price Index (CPI) as our measure of inflation and examine the quarterly forecast errors. The CPI inflation rate has a long history of being forecasted by both the Survey of Professional Forecasters (SPF) and the Fed. Besides these forecasts from specific forecasters, we also included four simple forecasts (random walk or naïve forecast, the weighted median CPI and trimmed mean CPI both available from the Cleveland Fed and the traditional measure of core inflation the less food and energy inflation rate).

The SPF is an unbalanced panel—survey participants enter and exit (and often re-enter again). It is likely that some forecasters have lower (higher) errors simply because they forecasted during period when forecasting was relatively easy (hard). We use a normalization discussed below to take account of the changing degree of difficulty in forecasting over time.

For the Fed’s forecasts we use the Philadelphia Fed’s Greenbook forecast dataset, which realigns the Fed’s forecasts to correspond in time with the SPF quarterly forecasts. All of our data are quarterly annualized rates of change. The full sample is 1984:I through 2007:IV. We chose 1984 as the starting date in order to focus on the period since the onset of the Great Moderation. The end date of 2007 is based on the availability of the Greenbook forecasts.

We use two normalization methods. First, we follow D’Agostino et al. (2012) by normalizing forecast errors to account for the degree of forecasting difficulty. Specifically, we divide each forecaster’s squared forecast error by the cross-sectional mean squared error (MSE)[[4]](#footnote-4). Second, we normalize by dividing each forecaster’s squared forecast error by a three-quarter moving average of the best fitting ARMA model.

We also follow D’Agostino et al. by computing a score for each forecaster. The score is the average over time of each forecaster’s normalized squared forecast error. Together, the normalization and the score allow us to compare forecasters with various (continuous or intermittent) forecasting records.

We begin by examining the summary statistics of the raw (non-normalized) forecast errors to see whether we can detect any change in the distribution before and after the beginning of the Fed’s movement toward greater transparency (February 1994). The summary statistics of the raw errors appear in appendix table 1. From this table, we see that the means, variances, amount of skewness and kurtosis appear to differ across the pre and post-1994 samples. We next undertake a more formal analysis of changes in the distribution of forecast errors using the Kolmogorov-Smirnov (KS) test, which allows us to test whether the two samples come from the same distribution.[[5]](#footnote-5) We conducted the KS test on every possible sample split in the dataset (with a 15% window). The null hypothesis is that there is no difference in the two distributions. All p-values, at every sample split were highly significant. In the table below we report the dates at which the sample split produced the lowest p-value.

|  |  |
| --- | --- |
| Forecast horizon (quarters ahead) | Date of lowest p-value |
| 0 | 1995.2 |
| 1 | 1995.1 |
| 2 | 1996.3 |
| 3 | 1995.4 |
| 4 | 2000.2, 2001.4 and 2004.1 |

For the zero to four-quarter horizons, the KS test indicates that there is a break in the mid-1990s. For the 4 quarter ahead horizon there is a local minimum in the mid-1990s but the global minimums occur in the early 2000s.

2. Forecaster Score Rankings

 We computed forecaster scores for each forecaster as described in the previous section. Tables 1 through 6 present the forecaster scores and rankings for various forecasters and forecasts as well as the results of a bootstrapping exercise to determine whether specific forecasters and forecasts are better than random chance. Overall the results suggest that the Fed is no longer a superior forecaster to SPF forecasters and simple core inflation forecasts. Splitting the sample in 1994 when the Fed increased its transparency does allow us to discern some interesting changes in the behavior in inflation forecasts.

 In Table 1A with the full sample (1984 to 2007), we use the normalization method of D’Agostino et al. We find that the forecasting advantage of the Fed is weak. Looking at the current quarter the Fed is significantly better than all core inflation measures (inflation less food and energy, median inflation, and trimmed mean inflation). The Fed is worse than the best SPF forecaster chosen from those forecasters that had forecast at least five times. Interestingly, the Fed is equivalent to the median SPF forecaster. Beyond the current quarter, the Fed ranks in the middle of the pack and is statistically similar to the forecasts from simple lagged core inflation measures. As a benchmark we include the naïve forecast (random walk) and find that the Fed is still statistically superior to this forecast of inflation suggesting that removing some noise (either by statistical procedure or modeling) from the headline inflation process is needed to obtain an adequate inflation forecast. The median SPF forecasters ranks higher than the Fed and other simple inflation forecasts and is statistically better than the Fed when comparing the scores.

We next directly test whether any of these forecasters or forecasts are better than those assigned by random chance.  For each forecaster, we consider all of the quarters in which that particular forecaster made a prediction.  We then take the forecast errors from the
all of the forecasters who made a prediction in that quarter and randomly reassign them with replacement.  After doing this for each of the quarters in which a particular forecaster made predictions, we recalculate the forecaster's score.  After finding 999 new scores in this manner, we compare the forecaster's actual score to these simulated scores.  To determine if the forecasters perform better than random chance, we find the percentage of these scores that are equal to or less than the observed score.  This p-value represents the probability of a particular forecaster obtaining the observed score by random chance.  In Table 1B, we present these p-values. With the exception of the one-quarter ahead time horizon the Fed’s score is significant at the 0.10 level, meaning that the Fed is a better forecaster than forecasts chosen randomly from the SPF.  The median of the SPF forecasts is also better than random chance in all five time horizons forecast at the 0.005 level.  Finally, the median CPI and trimmed mean CPI are better forecasts starting at the two quarter ahead time horizon indicating that these measures tend to provide solid forecasts at slightly longer time horizons.

In addition, using the D’Agostino et al. normalization we break the sample at the time the Fed increased transparency in 1994. The early sample runs from 1984-1993 and the later sample runs from 1994-2007 in Tables 2A and 3A, respectively. Comparing these two sub-samples results is important because the comparison provides information about changes in the abilities of the Fed, SPF participants and simple lagged inflation measures to provide useful information about future inflation. First, examining Table 2A and 3A, we see that the Fed maintains its superior ability to forecast the current quarter. The Fed is equivalent to the median CPI and trimmed mean CPI starting one quarter ahead during both sample periods. The best SPF forecast still beats the Fed but again it is not a consistent forecaster so its practical use is limited. The most important difference is that the move to greater transparency seems to have allowed the median SPF forecast to gain ground on the Fed in term of forecasting ability. Keep in mind that because we are normalizing we are already accounting for the fact that it became easier to forecast so our results supports earlier research by Gamber and Smith (2009) and Liu and Smith (2013).

Looking at how the forecasts do against random assignment in Tables 2B and 3B we again see the result that transparency has made the Fed less superior. The Fed falls below the 10% threshold in all horizons except the current quarter whereas the median SPF is above the 10% threshold. Somewhat disappointingly, there is no clear-cut winner between the median CPI and trimmed mean CPI which means looking at one of those measures alone may not provide adequate information about future inflation.

The D'Agostino's normalization method is useful in that it accounts for the variation in forecasting difficulty from quarter to quarter; however, using a more mechanical normalization method might provide different insight into the degree to which forecasters changed over the sample. We use an ARMA (Autoregressive-moving average) model. We find the best fitting ARMA model for each quarter and normalize each quarter by a three quarter moving average, a technique similar to the one used in Gamber, Smith and Weiss (2011).

Considering the full sample in Table 4A, we see that the rank of the Fed and the median SPF forecaster has declined. The Fed maintains its forecasting advantage against core inflation measures in the current quarter and is indistinguishable at all other horizons as it was under the D’Agostino et al. normalization. Only the best SPF forecaster is superior to the Fed and once again the forecaster that is superior to the Fed varies by time horizon. Therefore, there is not just one SPF forecaster that we can follow to beat the Fed.

Instead of examining the p-value from a t-test in the bootstrap exercise to gauge if the forecast is better or worse than random chance we use the Wilcoxon sign test, which does not require the normality assumption that the t-test does. The method that we used earlier does not work well under the ARMA normalization because some of the ARMA models produce extremely accurate forecast during some quarters. When the ARMA models forecast errors are close to zero the forecaster’s score tend to infinity and therefore skewing the ranking of actual scores.

The Wilcoxon sign test is a non-parametric test that compares pairs of observations from two populations without a normality requirement. Here, one population is a certain forecaster or forecast and the other is the ARMA model. The null hypothesis for the two-sided test is that there is no difference in medians. Thus, the p-value is the probability of observing the results of the data, given that the null hypothesis is true. In this context, the null hypothesis is that either the ARMA or the other forecast is equally likely to have a lower score than the other.

Examining Table 4C, in the current quarter a p-value of 0.919 indicates that there is a 91.9% chance of seeing the actual results given that the Fed and the ARMA model are equally good at predicting that quarter. So, it is likely that the two forecasters have about the same ability. On the other hand, at the four-quarter ahead horizon a p-value of .052 indicates that it is likely that Fed and the ARMA are not equally good at predicting that time horizon. For the median SPF, we find that the median SPF forecaster is likely to have equal forecasting ability as the ARMA model. For the median CPI inflation measure both in the current horizon and the four-quarter ahead horizon there are differences in ability. We need to examine the one-sided test to figure out which forecast is better when we reject the null hypothesis of equal forecasting ability from the two-sided Wilcoxon sign test.

For the one-sided test, the null hypothesis is that one population's score is higher than the other (which one depends on the direction of the test), and the p-value is the probability of obtaining results as extreme or more extreme than what was observed given the null hypothesis is true. Examining the Fed’s forecast in the four-quarter ahead horizon a p-value of 0.026 indicates that there is a 2.6% chance of seeing the observed results given that the ARMA scores lower than the Fed. Thus, we would reject the null, and conclude that the Fed is better or as good as the ARMA in that quarter. Additionally, for the current quarter median CPI we obtain a p-value of 0.999 indicating that there is a 99.9% chance of observing the results given the ARMA score is lower than the median CPI. Thus we would fail to reject the null that the median is better than the ARMA in that quarter.

Finally, in Table 4C the proportion indicates the fraction of quarters that the given forecast had a lower score than the ARMA model did in the given horizon. We can see that the Fed is a better forecaster than the ARMA model starts around 50% (49%) and rises to 60% (60.4%) in the four quarter-ahead horizon. Interestingly, the SPF proportion starts higher than the Fed (53.1%) but ends lower than the Fed (57.3%) which may indicate that the Fed can beat the ARMA model at long time horizons more frequently because it knows the future path of monetary policy.

In Tables 5A and 6A, we examine the changes in the results under the ARMA normalization when splitting the sample at 1994. The results are similar to those presented for the whole sample. The main different to note is that the best SPF forecaster is equivalent to the Fed in the pre-1994 sample and superior to the Fed after 1994. Otherwise the Fed is equivalent to the core inflation measures and median SPF forecasters in horizons 1 to 4; the Fed maintains its superiority against those measures in the current quarter. Also, Tables 5C and 6C, provide similar results of the Wilcoxon sign test.

Finally, much of the previous literature describes how the Fed has lost its superiority to the median SPF forecaster[[6]](#footnote-6). Table 7 directly compares the ability of the Fed and the median SPF forecaster under the Wilcoxon sign test. We find that the Fed again has lost its superiority. The p-values for the two sample splits in that chart are all greater than 10% except in the pre-1994 current quarter sample. In that particular case the Fed is likely to be better than the median SPF forecaster but that is the only indication of the superiority of the Fed.

**Conclusion**

 This paper investigated the cross-sectional distribution of inflation forecasts errors over the period 1984 through 2007. We build on previous work by D’Agostino et al. to examine the Fed’s forecasting record relative to other forecasters and forecasts before and after the Fed’s move to greater transparency in 1994. We find, that the Fed has lost most of its forecasting superiority since 1994. However, there is no SPF forecaster who is consistently better than the Fed. And furthermore, there is no core measure of inflation that is consistently better than the Fed. Overall the results indicate that for the CPI, that these other forecasters and forecast of inflation provide similar information to that which is available in the Greenbook forecasts.

**References**

Bai, Jushan and Pierre Perron, Estimating and Testing Linear Models with Multiple Structural Changes, Econometrica, vol. 66, no. 1, pp. 47-78, 1998.

Bernanke, Ben S. and Jean Boivin, 2003 “Monetary Policy in a data-rich environment,” *Journal of Monetary Economics*, vol. 50, pp. 525-46.

Blattner, Tobias, Marco Catenaro, Michael Ehrmann, Rolf Strauch and Jarkko Turunen, “The Predictability of Monetary Policy,” European Central Bank Occasional Paper, no. 83, March 2008.

Campbell, Sean D. “Macroeconomic Volatility, Predictability and Uncertainty in the Great Moderation: Evidence from the Survey of Professional Forecasters,” *Journal of Business and Economic Statistics*, vol. 25, 2007, pp. 191-200.

Ehrmann, Michael and Marcel Fratzscher, “Transparency, Disclosure and the Federal Reserve, European Central Bank Working Paper no 457, March 2005.

D’ Agostino, Anonello and Karl Whelan, “Federal Reserve Information During the Great Moderation,” *Journal of the European Economic Association*, vol. 6, issue 2-3, April-May 2008, pp. 609-20.

D’ Agostino, Anonello, Kieran McQuinn, and Karl Whelan, “Are Some Forecasters

Really Better Than Others?” *Journal of Money, Credit and Banking*, Vo.l. 44, No. 4,

June 2012, pp. 715-32.

Faust, Jon and Jonathan Wright, 2009 “Comparing Greenbook and Reduced Form Forecasting Using a Large Real Time Data Set,” *Journal of Business and Economic Statistics*, vol. 27, no. 4, pp. 468-79.

Fischer, Henning, Marta García-Bárzana, Peter Tillmann and Peter Winker, 2012, Evaluating FOMC forecast ranges: an interval data approach, Joint Discussion Paper Series in Economics, Universities of Aachenm Grieβen, Göttingenm Kassel, Marburg and Siegen, no. 130-2012.

Gamber, E. N., and J. K. Smith. “Are the Fed’s Inflation Forecasts Still Superior to the

Private Sector’s?” *Journal of Macroeconomics*, vol. 31, 2009, pp. 240-51.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, D. McNamara, “Where is the Fed in the Distribution of

Forecasts,” *Journal of Policy Modeling*, vol. 36, issue 2, March-April 2014, pp. 296-312.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, M. Weiss, "Forecast errors before and during the Great

Moderation." *Journal of Economics and Business* 63.4 (2011): 278-289.

Gavin, W.T. (2003). FOMC Forecasts: Is All the Information in the Central Tendency?.

Federal Reserve Bank of St. Louis Review 85 (3), 27–46.

Gavin, W.T. and R.J. Mandal (2003). Evaluating FOMC forecasts. *International Journal*

*of Forecasting* 19 (4), 655–667.

Gavin, W.T. and G. Pande (2008). FOMC Consensus Forecasts. Federal Reserve Bank of

St. Louis Review 90 (3, Part 1), 149–164.

Liu, D. and J. K. Smith. “Inflation forecasts and core inflation: Where is the information

on future inflation?” *Quarterly Review of Economics and Finance*, forthcoming 2013.

<http://dx.doi.org/10.1016/j.qref.2013.07.006>

Romer, Chistine D., and David H. Romer, 2000, “Federal Reserve Information and the Behavior of Interest Rates,” *The American Economic Review*,” vol. 90, no. 3, pp. 429-57.

Romer, Chistine D., and David H. Romer (2008). The FOMC versus the Staff: Where Can MonetaryPolicymakers Add Value? *American Economic Review* 98 (2), 230–235.

Romer, David (2010). A New Data Set on Monetary Policy: The Economic Forecasts of

Individual Members of the FOMC. *Journal of Money, Credit and Banking* 42 (5), 951–

957.

Sims, Christopher A., 2002, “The Role of Models and Probabilities in the Monetary Policy Process,” Brookings Papers on Economic Activity , Vol. 2002, No. 2, pp. 1-40.

Swanson, Eric T. 2004, Federal Reserve Transparency and Financial Market Forecasts of Short-Term Interest Rates, Board of Governors of the Federal Reserve System Finance and Economics Discussion Series, number 2004-06.

**Table 1A: Score and Rank of forecasts or forecasters Full sample (1984-2007) D’Agostino et al. Normalization**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fore-cast | Actual Score(h=0) | Score ratio and Significance1(h=0) | Actual Score(h=1) | Score ratio and Significance(h=1) | Actual Score(h=2) | Score ratio and Significance(h=2) | Actual Score(h=3) | Score ratio and Significance(h=3) | Actual Score(h=4) | Score ratio and Significance(h=4) |
| Rank2 (of 138) | Rank | Rank | Rank | Rank |
| Fed | 0.417 | - | 0.854 | - | 0.837 | - | 0.807 | - | 0.830 | - |
| 10 | 68 | 60 | 60 | 70 |
| CPILFE | 1.447 | 3.470\*\* | 1.122 | 1.31 | 0.965 | 1.153 | 0.993 | 1.230 | 1.019 | 1.228 |
| 118 | 101 | 93 | 85 | 96 |
| Median CPI | 1.255 | 3.010\*\* | 0.847 | 0.99 | 0.808 | 0.965 | 0.822 | 1.012 | 0.793 | 0.955 |
| 111 | 64 | 53 | 63 | 53 |
| Trimmed Mean CPI | 0.936 | 2.245\*\* | 0.865 | 1.01 | 0.781 | 0.933 | 0.811 | 1.005 | 0.930 | 1.120 |
| 83 | 69 | 48 | 61 | 84 |
| Naive | 2.380 | 5.707\*\* | 1.556 | 1.822\* | 1.821 | 2.176\*\* | 1.858 | 2.302\*\* | 1.563 | 1.883\*\* |
| 138 | 125 | 130 | 131 | 125 |
| Best SPF | 0.153ID=541 | 0.367\* | 0.276ID=19 | 0.323\*\* | 0.353ID=406 | 0.422\*\* | 0.264ID=430 | 0.327\*\* | 0.020ID=444 | 0.241\*\* |
| 1 | 1 | 1 | 1 | 1 |
| Median SPF | 0.375 | 0.899 | 0.571 | 0.669\*\* | 0.639 | 0.763\* | 0.619 | 0.767\*\* | 0.612 | 0.737\*\* |
| 6 | 13 | 19 | 20 | 19 |

1: Significance indicates whether the given forecaster's score in the given horizon was significantly different from the Fed's in that horizon.

* \* indicates significance at the .05 level
* \*\* indicates significance at the .01 level
* Tests were conducted for only those who made 5 or more forecasts in the given horizon.
* Actual scores were calculated using the entire sample.

2: Rank of the given forecaster among those who made 5 or more forecasts in the given horizon. Rank is based on the forecaster's score. All horizons have the same number of forecasters as the current period unless otherwise noted.

**Table 1B: Rank relative to random chance Full sample (1984-2007) D’Agostino et al. Normalization**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Fore-cast | P-value3 (h=0) | P-value (h=1) | P-Value (h=2) | P-value(h=3) | P-value(h=4) |
| Fed | 0.001 | 0.114 | 0.068 | 0.030 | 0.071 |
| CPILFE | 0.995 | 0.864 | 0.395 | 0.505 | 0.614 |
| Median CPI | 0.954 | 0.100 | 0.046 | 0.042 | 0.020 |
| Trimmed Mean CPI | 0.338 | 0.138 | 0.017 | 0.046 | 0.305 |
| Naive | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Best SPF | 0.001 | 0.028 | 0.025 | 0.008 | 0.045 |
| Median SPF | 0.001 | 0.004 | 0.004 | 0.003 | 0.001 |

3: The p-value is calculated by ordering the actual score of the given forecast and the 999 scores simulated by the bootstrap exercise. The location or rank of the actual score within this ordering is divided by 1000 to obtain the p-value.

**Table 2A: Score and Rank of forecasts or forecasters Sample (1984-1993) D’Agostino et al. Normalization**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fore-cast | Actual Score(h=0) | Score ratio and Significance1 | Actual Score(h=1) | Score ratio and Significance | Actual Score(h=2) | Score ratio and Significance | Actual Score(h=3) | Score ratio and Significance | Actual Score(h=4) | Score ratio and Significance |
|  Rank2(of 70) | Rank | Rank | Rank | Rank |
| Fed | 0.360 | - | 0.803 | - | 0.628 | - | 0.565 | - | 0.577 | - |
| 5 | 36 | 20 | 17 | 18 |
| CPILFE | 1.738 | 4.828\*\* | 1.215 | 1.513 | 1.104 | 1.758\* | 1.060 | 1.876\*\* | 0.999 | 1.731\* |
| 65 | 55 | 51 | 47 | 43 |
| Median CPI | 1.217 | 3.381\*\* | 0.764 | 0.951 | 0.837 | 1.333 | 0.809 | 1.432\* | 0.789 | 1.367 |
| 57 | 34 | 37 | 30 | 33 |
| Trimmed Mean CPI | 1.021 | 2.836\*\* | 0.842 | 1.049 | 0.766 | 1.220 | 0.824 | 1.458 | 0.786 | 1.362 |
| 48 | 38 | 32 | 31 | 31 |
| Naive | 2.433 | 6.758\*\* | 1.533 | 1.909 | 2.037 | 3.244\* | 1.711 | 3.030 | 1.664 | 2.884 |
| 71 | 64 | 69 | 68 | 65 |
| Best SPF | 0.281ID=406 | 0.78 | 0.296ID=442 | 0.369\* | 0.257ID=431 | 0.409\*\* | 0.264ID=430 | 0.467\* | 0.265ID=411 | 0.459\* |
| 1 | 1 | 1 | 1 | 1 |
| Median SPF | 0.310 | 0.861 | 0.477 | 0.594 | 0.582 | 0.927 | 0.562 | 0.995 | 0.558 | 0.967 |
| 3 | 11 | 16 | 16 | 14 |

1: Significance indicates whether the given forecaster's score in the given horizon was significantly different from the Fed's in that horizon.

* \* indicates significance at the .05 level
* \*\* indicates significance at the .01 level
* Tests were conducted for only those who made 5 or more forecasts in the given horizon.
* Actual scores were calculated using the entire sample.

2: Rank of the given forecaster among those who made 5 or more forecasts in the given horizon. Rank is based on the forecaster's score. All horizons have the same number of forecasters as the current period unless otherwise noted.

**Table 2B: Rank relative to random chance Sample (1984-1993) D’Agostino et al. Normalization**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Fore-cast | P-value3 (h=0) | P-value (h=1) | P-Value (h=2) | P-value(h=3) | P-value(h=4) |
| Fed | 0.001 | 0.170 | 0.006 | 0.002 | 0.003 |
| CPILFE | 0.996 | 0.837 | 0.722 | 0.655 | 0.534 |
| Median CPI | 0.817 | 0.084 | 0.210 | 0.159 | 0.117 |
| Trimmed Mean CPI | 0.554 | 0.232 | 0.093 | 0.181 | 0.118 |
| Naive | 1.000 | 0.987 | 1.000 | 0.996 | 0.998 |
| Best SPF | 0.042 | 0.047 | 0.009 | 0.003 | 0.001 |
| Median SPF | 0.001 | 0.001 | 0.002 | 0.003 | 0.003 |

3: The p-value is calculated by ordering the actual score of the given forecast and the 999 scores simulated by the bootstrap exercise. The location or rank of the actual score within this ordering is divided by 1000 to obtain the p-value.

**Table 3A: Score and Rank of forecasts or forecasters Sample (1994-2007) D’Agostino et al. Normalization**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fore-cast | Actual Score(h=0) | Score ratio and Significance1 | Actual Score(h=1) | Score ratio and Significance | Actual Score(h=2) | Score ratio and Significance | Actual Score(h=3) | Score ratio and Significance | Actual Score(h=4) | Score ratio and Significance |
|  Rank2(of 97) | Rank | Rank | Rank | Rank |
| Fed | 0.457 | - | 0.892 | - | 0.984 | - | 0.979 | - | 1.011 | - |
| 8 | 53 | 66 | 64 | 67 |
| CPILFE | 1.247 | 2.729\*\* | 1.051 | 1.178 | 0.875 | 0.889 | 0.967 | 0.988 | 1.044 | 1.033 |
| 77 | 71 | 49 | 62 | 70 |
| Median CPI | 1.300 | 2.845\*\* | 0.925 | 1.040 | 0.798 | 0.811 | 0.846 | 0.864 | 0.814 | 0.805 |
| 78 | 56 | 33 | 43 | 35 |
| Trimmed Mean CPI | 0.888 | 1.943\*\* | 0.879 | 0.990 | 0.787 | 0.800 | 0.787 | 0.804 | 1.017 | 1.006 |
| 54 | 50 | 31 | 34 | 68 |
| Naive | 2.346 | 5.133\*\* | 1.573 | 1.763 | 1.668 | 1.695\* | 1.964 | 2.006\*\* | 1.488 | 1.472 |
| 97 | 90 | 91 | 95 | 89 |
| Best SPF | 0.152ID=541 | 0.333\* | 0.330ID=455 | 0.370\*\* | 0.440ID=501 | 0.447\*\* | 0.359ID=538 | 0.367\*\* | 0.314ID=416 | 0.311\*\* |
| 1 | 1 | 1 | 1 | 1 |
| Median SPF | 0.42 | 0.919\*\* | 0.679 | 0.761\*\* | 0.651 | 0.662\*\* | 0.658 | 0.672\*\* | 0.650 | 0.643\*\* |
| 6 | 12 | 16 | 14 | 13 |

1: Significance indicates whether the given forecaster's score in the given horizon was significantly different from the Fed's in that horizon.

* \* indicates significance at the .05 level
* \*\* indicates significance at the .01 level
* Tests were conducted for only those who made 5 or more forecasts in the given horizon.
* Actual scores were calculated using the entire sample.

2: Rank of the given forecaster among those who made 5 or more forecasts in the given horizon. Rank is based on the forecaster's score. All horizons have the same number of forecasters as the current period unless otherwise noted.

**Table 3B: Rank relative to random chance Sample (1994-2007) D’Agostino et al. Normalization**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Fore-cast | P-value3(h=0) | P-value (h=1) | P-Value (h=2) | P-value(h=3) | P-value(h=4) |
| Fed | 0.001 | 0.279 | 0.475 | 0.476 | 0.569 |
| CPILFE | 0.896 | 0.634 | 0.219 | 0.449 | 0.673 |
| Median CPI | 0.929 | 0.377 | 0.054 | 0.150 | 0.082 |
| Trimmed Mean CPI | 0.333 | 0.225 | 0.055 | 0.047 | 0.613 |
| Naive | 1.000 | 0.998 | 1.000 | 1.000 | 0.997 |
| Best SPF | 0.001 | 0.009 | 0.036 | 0.012 | 0.076 |
| Median SPF | 0.001 | 0.005 | 0.008 | 0.002 | 0.001 |

3: The p-value is calculated by ordering the actual score of the given forecast and the 999 scores simulated by the bootstrap exercise. The location or rank of the actual score within this ordering is divided by 1000 to obtain the p-value.

**Table 4A: Score and Rank of forecasts or forecasters Full sample (1984-2007) ARMA Normalization**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fore-cast | Actual score(h=0) | Score ratio & significance1 (h=0) | Actual score(h=1) | Score ratio & significance (h=1) | Actual score(h=2) | Score ratio & significance (h=2) | Actual score(h=3) | Score ratio & significance (h=3) | Actual score(h=4) | Score ratio & significance (h=4) |
|  Rank2(of 138) | Rank | Rank | Rank | Rank |
| Fed | 1.486 | - | 1.538 | - | 1.921 | - | 4.26 | - | 1.049 | - |
| 30 | 95 | 111 | 111 | 80 |
| CPILFE | 6.555 | 4.411\*\* | 1.611 | 1.047 | 1.881 | 0.979 | 6.9 | 1.620 | 1.222 | 1.165 |
| 129 | 101 | 107 | 124 | 90 |
| Median CPI | 5.193 | 3.494\*\* | 1.308 | 0.850 | 1.291 | 0.672 | 3.96 | 0.930 | 0.911 | 0.868 |
| 119 | 76 | 68 | 109 | 62 |
| Trimmed Mean CPI | 4.977 | 3.349\*\* | 1.356 | 0.882 | 1.454 | 0.757 | 6.24 | 1.465 | 1.279 | 1.219 |
| 117 | 82 | 86 | 121 | 93 |
| Naive | 5.777 | 3.888\*\* | 1.354 | 0.880 | 2.573 | 1.339 | 5.18 | 1.216 | 1.947 | 1.856 |
| 126 | 81 | 126 | 117 | 126 |
| Best SPF |  0.397ID=549 | 0.267\*\* | 0.005ID=19 | 0.003\*\* | 0.220ID=86 | 0.115\*\* | 0.196ID=430 | 0.046 | 0.173 ID=444 | 0.165\*\* |
| 1 | 1 | 1 | 1 | 1 |
| Median SPF | 2.242 | 1.509 | 1.101 | 0.716 | 1.256 | 0.654 | 2.99 | 0.702 | 0.846 | 0.806 |
| 57 | 57 | 65 | 92 | 56 |

1: Significance indicates whether the given forecaster's score in the given horizon was significantly different from the Fed's in that horizon.

* \* indicates significance at the .05 level
* \*\* indicates significance at the .01 level
* Tests were conducted for only those who made 5 or more forecasts in the given horizon.
* Actual scores were calculated using the entire sample.

2: Rank of the given forecaster among those who made 5 or more forecasts in the given horizon. Rank is based on the forecaster's score. All horizons have the same number of forecasters as the current period unless otherwise noted.

**Table 4B: Rank relative to random chance Full Sample (1984-2007) ARMA Normalization**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Fore-cast | P-value3(h=0) | P-value (h=1) | P-value (h=2) | P-value (h=3) | P-value (h=4) |
| Fed | 0.999 | 0.999 | 1.000 | 1.000 | 0.668 |
| CPILFE | 1.000 | 1.000 | 1.000 | 1.000 | 0.968 |
| Median CPI | 1.000 | 0.984 | 0.988 | 1.000 | 0.231 |
| Trimmed Mean CPI | 1.000 | 0.996 | 1.000 | 1.000 | 0.982 |
| Naive | 1.000 | 0.992 | 1.000 | 1.000 | 1.000 |
| Best SPF | 0.103 | 0.001 | 0.002 | 0.001 | 0.001 |
| Median SPF | 1.000 | 0.805 | 0.981 | 1.000 | 0.093 |

3: The p-value is calculated by ordering the actual score of the given forecast and the 999 scores simulated by the bootstrap exercise. The location or rank of the actual score within this ordering is divided by 1000 to obtain the p-value.

**Table 4C: Wilcoxon Sign Test Results Full Sample (1984-2007) ARMA Normalization**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Forecast | P-value | h=0 | h=1 | h=2 | h=3 | h=4 |
| Fed |  |  |  |  |  |  |
| One sided | 0.620 | 0.380 | 0.305 | 0.131 | 0.026 |
| Proportion1 | 0.490 | 0.521 | 0.531 | 0.563 | 0.604 |
| CPILFE |  |  |  |  |  |  |
| One sided | 1.000 | 0.821 | 0.092 | 0.305 | 0.459 |
| Proportion  | 0.302 | 0.458 | 0.573 | 0.531 | 0.510 |
| Median CPI |  |  |  |  |  |  |
| One sided | 0.999 | 0.179 | 0.380 | 0.305 | 0.026 |
| Proportion | 0.354 | 0.552 | 0.521 | 0.610 | 0.604 |
| Trimmed Mean CPI |  |  |  |  |  |  |
| One sided | 0.995 | 0.541 | 0.131 | 0.238 | 0.131 |
| Proportion  | 0.375 | 0.500 | 0.563 | 0.542 | 0.563 |
| Naive |  |  |  |  |  |  |
| One sided | 0.999 | 0.041 | 0.380 | 0.762 | 0.009 |
| Proportion  | 0.344 | 0.594 | 0.521 | 0.469 | 0.625 |
| Median SPF |  |  |  |  |  |  |
| One sided | 0.305 | 0.541 | 0.238 | 0.131 | 0.092 |
| Proportion | 0.531 | 0.500 | 0.542 | 0.563 | 0.573 |

1: Proportion of quarters that forecast's errors are smaller than the ARMA model errors.

**Table 5A: Score and Rank of forecasts or forecasters Sample (1984-1993) ARMA Normalization**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fore-cast | Actual score(h=0) | Score ratio & significance1 (h=0) | Actual score(h=1) | Score ratio & significance (h=1) | Actual score(h=2) | Score ratio & significance (h=2) | Actual score(h=3) | Score ratio & significance (h=3) | Actual score(h=4) | Score ratio & significance (h=4) |
| Rank2(of 70) | Rank | Rank | Rank | Rank |
| Fed | 1.193 | - | 0.862 | - | 0.931 | - | 0.737 | - | 0.409 | - |
| 13 | 36 | 36 | 29 | 21 |
| CPILFE | 8.322 | 6.976\* | 1.446 | 1.677 | 1.294 | 1.390 | 1.272 | 1.726 | 0.771 | 1.885\* |
| 67 | 61 | 51 | 55 | 53 |
| Median CPI | 6.267 | 5.253\* | 0.957 | 1.110 | 0.777 | 0.835 | 1.281 | 1.738 | 0.621 | 1.518 |
| 64 | 41 | 27 | 56 | 44 |
| Trimmed Mean CPI | 6.027 | 5.052\* | 1.136 | 1.318 | 0.675 | 0.725 | 1.201 | 1.630 | 0.651 | 1.592 |
| 63 | 49 | 22 | 54 | 47 |
| Naive | 7.220 | 6.052 | 1.231 | 1.428 | 1.609 | 1.728 | 1.412 | 1.916 | 0.935 | 2.286 |
| 66 | 51 | 62 | 58 | 62 |
| Best SPF | 0.413ID=406 | 0.346 |  0.047ID=19 | 0.054\*\* | 0.140ID=431 | 0.150\*\* | 0.667ID=411 | 0.905\*\* |  0.092ID=442 | 0.225\*\* |
| 1 | 1 | 1 | 1 | 1 |
| Median SPF | 2.055 | 1.723 | 0.722 | 0.838 | 0.794 | 0.853 | 0.823 | 1.117 | 0.491 | 1.200 |
| 35 | 28 | 28 | 34 | 27 |

1: Significance indicates whether the given forecaster's score in the given horizon was significantly different from the Fed's in that horizon.

* \* indicates significance at the .05 level
* \*\* indicates significance at the .01 level
* Tests were conducted for only those who made 5 or more forecasts in the given horizon.
* Actual scores were calculated using the entire sample.

2: Rank of the given forecaster among those who made 5 or more forecasts in the given horizon. Rank is based on the forecaster's score. All horizons have the same number of forecasters as the current period unless otherwise noted.

**Table 5B: Rank relative to random chance Sample (1984-1993) ARMA Normalization**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Fore-cast | P-value3(h=0) | P-value (h=1) | P-value (h=2) | P-value (h=3) | P-value (h=4) |
| Fed | 0.824 | 0.265 | 0.401 | 0.068 | 0.001 |
| CPILFE | 1.000 | 0.971 | 0.934 | 0.905 | 0.107 |
| Median CPI | 1.000 | 0.458 | 0.102 | 0.927 | 0.010 |
| Trimmed Mean CPI | 1.000 | 0.780 | 0.025 | 0.837 | 0.018 |
| Naive | 1.000 | 0.889 | 0.998 | 0.979 | 0.426 |
| Best SPF | 0.110 | 0.001 | 0.001 | 0.001 | 0.001 |
| Median SPF | 1.000 | 0.061 | 0.125 | 0.188 | 0.001 |

3: The p-value is calculated by ordering the actual score of the given forecast and the 999 scores simulated by the bootstrap exercise. The location or rank of the actual score within this ordering is divided by 1000 to obtain the p-value.

**Table 5C: Wilcoxon Sign Test Results Sample (1984-1993) ARMA Normalization**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Forecast | P-value | h=0 | h=1 | h=2 | h=3 | h=4 |
| Fed |  |  |  |  |  |  |
| One sided | 0.215 | 0.040 | 0.077 | 0.019 | 0.003 |
| Proportion1 | 0.575 | 0.650 | 0.625 | 0.675 | 0.725 |
| CPILFE |  |  |  |  |  |  |
| One sided | 0.997 | 0.437 | 0.040 | 0.134 | 0.134 |
| Proportion  | 0.300 | 0.525 | 0.650 | 0.600 | 0.600 |
| Median CPI |  |  |  |  |  |  |
| One sided | 0.960 | 0.134 | 0.040 | 0.215 | 0.000 |
| Proportion | 0.375 | 0.600 | 0.650 | 0.575 | 0.775 |
| Trimmed Mean CPI |  |  |  |  |  |  |
| One sided | 0.923 | 0.215 | 0.040 | 0.318 | 0.003 |
| Proportion  | 0.400 | 0.575 | 0.650 | 0.550 | 0.725 |
| Naive |  |  |  |  |  |  |
| One sided | 0.960 | 0.134 | 0.134 | 0.682 | 0.019 |
| Proportion  | 0.375 | 0.600 | 0.600 | 0.475 | 0.675 |
| Median SPF |  |  |  |  |  |  |
| One sided | 0.134 | 0.318 | 0.040 | 0.077 | 0.040 |
| Proportion | 0.600 | 0.550 | 0.650 | 0.625 | 0.650 |

1: Proportion of quarters that forecast's errors are smaller than the ARMA model errors.

**Table 6A: Score and Rank of forecasts or forecasters Sample (1994-2007) ARMA Normalization**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fore-cast | Actual score(h=0) | Score ratio & significance1 (h=0) | Actual score(h=1) | Score ratio & significance (h=1) | Actual score(h=2) | Score ratio & significance (h=2) | Actual score(h=3) | Score ratio & significance (h=3) | Actual score(h=4) | Score ratio & significance (h=4) |
|  Rank2(of 97) | Rank | Rank | Rank | Rank |
| Fed | 1.695 | - | 2.020 | - | 2.628 | - | 6.771 | - | 1.506 | - |
| 21 | 76 | 84 | 77 | 57 |
| CPILFE | 5.430 | 3.204\*\* | 1.732 | 0.857 | 2.295 | 0.873 | 10.870 | 1.605 | 1.549 | 1.029 |
| 89 | 66 | 79 | 86 | 62 |
| Median CPI | 4.407 | 2.600\*\* | 1.586 | 0.785 | 1.644 | 0.626 | 5.931 | 0.876 | 1.133 | 0.752 |
| 79 | 58 | 51 | 74 | 41 |
| Trimmed Mean CPI | 4.286 | 2.529\* | 1.529 | 0.757 | 1.998 | 0.760 | 9.790 | 1.446 | 1.700 | 1.129 |
| 74 | 54 | 69 | 83 | 71 |
| Naive | 4.746 | 2.800\*\* | 1.441 | 0.713 | 3.261 | 1.241 | 7.864 | 1.161 | 2.670 | 1.773 |
| 84 | 49 | 94 | 81 | 93 |
| Best SPF | 0.397ID=549 | 0.234\*\* | 0.322ID=455 | 0.159\* | 0.298ID=533 | 0.113\*\* | 0.371ID=454 | 0.055 | 0.109ID=416 | 0.072\*\* |
| 1 | 1 | 1 | 1 | 1 |
| Median SPF | 2.376 | 1.402 | 1.372 | 0.679 | 1.586 | 0.604 | 4.536 | 0.670 | 1.099 | 0.730 |
| 35 | 45 | 48 | 67 | 37 |

1: Significance indicates whether the given forecaster's score in the given horizon was significantly different from the Fed's in that horizon.

* \* indicates significance at the .05 level
* \*\* indicates significance at the .01 level
* Tests were conducted for only those who made 5 or more forecasts in the given horizon.
* Actual scores were calculated using the entire sample.

2: Rank of the given forecaster among those who made 5 or more forecasts in the given horizon. Rank is based on the forecaster's score. All horizons have the same number of forecasters as the current period unless otherwise noted.

**Table 6B: Rank relative to random chance Sample (1984-1993) ARMA Normalization**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Fore-cast | P-value3(h=0) | P-value (h=1) | P-value (h=2) | P-value (h=3) | P-value (h=4) |
| Fed | 0.998 | 1.000 | 1.000 | 1.000 | 0.998 |
| CPILFE | 1.000 | 1.000 | 1.000 | 1.000 | 0.997 |
| Median CPI | 1.000 | 0.997 | 1.000 | 1.000 | 0.834 |
| Trimmed Mean CPI | 1.000 | 0.995 | 1.000 | 1.000 | 0.999 |
| Naive | 1.000 | 0.994 | 1.000 | 1.000 | 1.000 |
| Best SPF | 0.088 | 0.007 | 0.019 | 0.073 | 0.008 |
| Median SPF | 1.000 | 0.975 | 0.997 | 1.000 | 0.781 |

**Table 6C: Wilcoxon Sign Test Results Sample (1994-2007) ARMA Normalization**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Forecast | P-value | h=0 | h=1 | h=2 | h=3 | h=4 |
| Fed |  |  |  |  |  |  |
| One sided | 0.886 | 0.886 | 0.748 | 0.656 | 0.447 |
| Proportion1 | 0.429 | 0.429 | 0.464 | 0.482 | 0.518 |
| CPILFE |  |  |  |  |  |  |
| One sided | 0.995 | 0.930 | 0.447 | 0.748 | 0.656 |
| Proportion  | 0.339 | 0.411 | 0.518 | 0.464 | 0.482 |
| Median CPI |  |  |  |  |  |  |
| One sided | 0.989 | 0.447 | 0.825 | 0.656 | 0.825 |
| Proportion | 0.357 | 0.518 | 0.446 | 0.482 | 0.446 |
| Trimmed Mean CPI |  |  |  |  |  |  |
| One sided | 0.959 | 0.825 | 0.656 | 0.344 | 0.825 |
| Proportion  | 0.393 | 0.446 | 0.482 | 0.536 | 0.446 |
| Naive |  |  |  |  |  |  |
| One sided | 0.998 | 0.114 | 0.748 | 0.748 | 0.114 |
| Proportion  | 0.321 | 0.589 | 0.464 | 0.464 | 0.589 |
| Median SPF |  |  |  |  |  |  |
| One sided | 0.656 | 0.748 | 0.748 | 0.447 | 0.447 |
| Proportion | 0.482 | 0.464 | 0.464 | 0.518 | 0.518 |

1: Proportion of quarters that forecast's errors are smaller than the ARMA model errors.

**Table 7: Wilcoxon Sign Test Results Full Sample (1984-2007) Fed vs. median SPF tests**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Forecast | P-value | h=0 | h=1 | h=2 | h=3 | h=4 |
| Full sample |  |  |  |  |  |  |
| One sided | 0.620 | 0.853 | 0.893 | 0.235 | 0.177 |
| Proportion1  | 0.490 | 0.427 | 0.427 | 0.531 | 0.542 |
| Pre 1994 |  |  |  |  |  |  |
| One sided | 0.024 | 0.566 | 0.168 | 0.434 | 0.682 |
| Proportion  | 0.625 | 0.45 | 0.58 | 0.475 | 0.48 |
| Post 1994 |  |  |  |  |  |  |
| One sided | 0.656 | 0.661 | 0.916 | 0.248 | 0.446 |
| Proportion | 0.482 | 0.45 | 0.393 | 0.536 | 0.5 |

1: Proportion of quarters that Fed forecasts errors are smaller than the median SPF forecast errors.

Appendix Table 1: Summary statistics on raw forecast errors

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Horizon | Summary Statistic | Mean | Variance | Skewness | Kurtosis |
| h=0 | Pre-94 | 1.32 | 10.44 | 6.24 | 49.71 |
| Post-94 | 1.77 | 13.76 | 5.08 | 37.1 |
| Full Sample | 1.61 | 12.64 | 5.41 | 40.61 |
| h=1 | Pre-94 | 2.25 | 26.31 | 4.95 | 29.79 |
| Post-94 | 2.62 | 26.00 | 5.16 | 41.15 |
| Full Sample | 2.49 | 26.08 | 5.07 | 36.95 |
| h=2 | Pre-94 | 2.52 | 32.29 | 5.13 | 30.96 |
| Post-94 | 2.61 | 18.39 | 3.98 | 29.41 |
| Full Sample | 2.58 | 23.13 | 4.74 | 33.01 |
| h=3 | Pre-94 | 3.16 | 54.88 | 5.08 | 28.94 |
| Post-94 | 3.03 | 29.64 | 8.61 | 171.66 |
| Full Sample | 3.08 | 38.26 | 6.85 | 88.48 |
| h=4 | Pre-94 | 3.35 | 55.83 | 4.86 | 26.65 |
| Post-94 | 6.01 | 403.05 | 6.12 | 38.05 |
| Full Sample | 5.1 | 286.03 | 7.02 | 52.46 |

1. Romer and Romer (2000), Sims, (2002), Gamber and Smith (2009). [↑](#footnote-ref-1)
2. Throughout this paper we use Fed forecast to refer to the forecasts produced by the Staff of the Board of Governors in preparation for the Greenbook. [↑](#footnote-ref-2)
3. Several other studies assess the forecasting record of the FOMC forecasts. See Fischer, et. al (2012), Gavin (2003), Gavin and Mandal (2003), Gavin and Pande (2008), Romer and Romer (2008) and Romer (2010). [↑](#footnote-ref-3)
4. D’Agostino et al., 2012, equation 3, p. 719 the normalized squared forecaster error is  where  is forecaster *i*’s squared forecast error at time *t*, is forecaster *i*’s normalized squared error forecast at time *t*,  is the number of forecasters at time t. [↑](#footnote-ref-4)
5. The KS test is a non-parametric test that does not rely on linear regression as does the Bai-Perron test (see Bai and Perron, 1998). Most importantly, Bai-Perron assumes that the error terms are iid normally distributed with mean 0 and fixed standard deviation (at least between proposed breaks). The KS test (when applied to before vs. after a break) tests whether the distribution has changed.  The Bai-Perron tests whether the parameters of the model have changed under the maintained assumption that the model is correctly specified. Thus, the KS test allows us to test for any sort of change. [↑](#footnote-ref-5)
6. See for example, Campbell (2007), D’Agostino and Whelan (2008), Gamber and Smith (2009, 2014). [↑](#footnote-ref-6)