Forecast evaluation with Stata
United Kingdom Stata Users Group Conference
London School of Hygiene and Tropical Medicine

Robert Alan Yaffee

September 9, 2010

Acknowledgments

In addition to the people at Stata –particularly, Bill Gould, Vince Wiggins, David Draper, Brian Poi, Robert Guiterrez, Jef Pitblado, Kerry Kammire, Ryman Tiu, Bill Rising, Alan Riley, Alan Acock – who have been very supportive and helpful, I am very grateful to the help and support of my dear friend Ana Timberlake who actually began the Stata Users Group conference in the United Kingdom.

I also would like to thank Sir David Hendry, Jürgen Doornik, Neil Ericsson, Andrew C. Harvey, Siem-Jan Koopman and Sebastien Laurent for their intellectual support and guidance.

I also need to thank David Corbett and Teresa Timberlake, Ashley Dyer, and Noelia Germino for their indispensable logistical support and encouragement.

Robert A. Yaffee, Ph.D.
Silver School of Social Work
New York University
yaffee@nyu.edu
September 2010

The research question

When confronting a research problem or question, we have to decide which approach to use and what kind of model to apply in order to achieve our objectives. Indeed, out standards of evaluation may depend on the kind of knowledge we seek for our purposes. In this approach, we will examine how to assess our forecasts and to select the model based on the evaluation. Are we engaging in exploratory data analysis in order to determine what factors may be of interest

in evaluating situational impact? Are we attempting to study a variety of alternative hypotheses explaining the nature of a relationship between variables of significant theoretical interest? Are we attempting to determine which factors are useful in warning us of an imminent or emerging threat to our security. Or are we endeavoring to evaluate policy options or their implementation? The answer to these questions may provide the focus needed to determine which factors we must apply to evaluate the models we are considering.

Assume that we are moving into an area where little research has been conducted. We engage in exploratory data analysis and for about for factors that may relate to our factors are broadly at work and may be focusing on the solution of a particular problem. We wish to know which series covary so that we may find time series that drive or influence the series of interest. A tentative model may be used for the purpose of exploring a new frontier of knowledge. A dynamic factor analysis could be applied to an environment of many financial series over time to ascertain which time series covary and cobreak in a preliminary attempt to ascertain what are the common forces driving an economy. These models may incorporate many correlated time series, with some series more correlated than others. We may be interested in how wide a theoretical spectrum we are encompassing.

Other models are used to test scientific hypotheses. These models are used to test alternative explanations of particular relationships among these series. We may have several models that we apply. Different models are based on different assumptions about the relationship between our key variables. We wish to compare some models against others, with respect to their forecast accuracy. In this way we may determine which models better capture the regularity that comprises the basis of the forecasts following from these models.

Motivation

Other models are designed for forecasting. After scientific and theoretical issues have been resolved, the purpose of the model may be one of practical warning. These models may contain signature variables that are easier to measure and more stable over time. The parameter constancy provides a more reliable basis for forecasting. Forecasting with such models is generally important as their the purpose is to warn people of an emerging serious situation, or looming danger to life and limb. Forecasting extreme or severe weather systems, health-threatening air quality, and disease diffusion are some prominent examples. Forecasting is not only essential for emergency planning and management, it is also an essential phase of needs assessment, resource allocation and mobilization, policy planning, implementation, and evaluation. However precisely programmatic a protocol may seem, each situation is to some extent different. The physicist Nils Bohr once said, "Prediction is difficult, especially when its about the future." And George E. P. Box is reported to have maintained that All models are wrong, but some are useful (Box and Draper, 1987). Proper preparation for potentially disastrous situations remains a challenge for countries, governments,

or organizations with limited resources.

Michael Clements and Sir David Hendry (2001) suggest some reasons why conditional models, misspecified in unknown ways yield model error. They maintain that a model is an attempt to extract regularities while excluding irregularities from nature. Although modeling and forecasting require covariance stationarity, we live in a nonstationary and changing world (Ibid, 2). Our modeling theory, from which we derive our forecasts, must allow for intermittent structural breaks (Ibid, 2)." The data generating mechanism, from which our time series realization stems, can change over time. Furthermore, some data generating processes change more rapidly than others. Shifts in deterministic factors can cause shifts in equilibrium means over time. In subsequent sections, I will endeavor to clarify nomenclature, discuss study design, and with respect to Box-Jenkins modeling, describe how to assess forecast bias, directional accuracy, out-of-sample and ex ante forecast evaluation using Stata to assess point and interval forecasts. For this presentation, I will deal with forecast evaluation for exponential smoothing and ARIMA models, whereas in later presentations, I will discuss forecast evaluation for ARCH/GARCH, dynamic factor analysis, and state space models. Moreover, I will focus on the forecasts from specific types of forecasts and leave others for treatment at a later time.

Nomenclature

Types of forecasts

Before proceeding further, we need to clarify key concepts of the forecasting nomenclature. Without proper elaboration, some of these concepts may be misinterpreted. There are several classifications of forecasts. One scheme classifies forecasts according to the nature of the forecast horizon, compared to the data available. The forecast horizon is the number of temporal periods over which the forecast is generated and its point of origin provides the basis for common forecast terminology. Another classification of forecasts is by the number of periods forecast at once. One-step-ahead forecasts and multi-step dynamic forecasts are included in this classification. Each of these classifications can use clarification. Among these types of forecasts are *point* forecasts, *interval* forecasts, and *density* that need explanation. Another classification of forecasts is based on the aspect of the forecast being generated. Point forecasts are those which estimate a particular value of the variable being forecast at a particular time in the future. Interval forecasts are predictions that estimate an interval within which the forecast should reside for a particular percentage of the time. There are also density forecasts which yield a probability density distribution of repeated forecasts. In this presentation, we will focus on two types of fixed window forecasts–namely, *ex post* forecasts and *ex ante* forecasts. Rolling window forecasts are deferred for a more detailed treatment at a later time and date. We now attempt to clarify misconceptions about these types of forecasts.

Data preparation and sample segmentation

Prior to forecasting, the time series that comprise the model have to be, if they are not already stationary, transformed so that they are covariance stationarity. They need to have a stable mean, variance and autocovariance throughout the series. By such a conversion, the only thing that determines the autocorrelation of a series will be the extent of the lag interval measured. The transformation entails detrending either by regression on a trend term or by differencing to change levels to rates. Clements and Hendry suggest that double-differencing (Hendry and Clements, 1999; Castle and Hendry, 2008) may be necessary.

Another phase of the ARIMA diagnosis entails structural break analysis and management. Structural breaks as level shifts and/or outliers should be detected, identified, and modeled. Intercept corrections may be necessary if there are level shifts are discovered in the series. Pulse or path dummies may be required to model the additive outliers or outlier patches.

Moreover, the sample needs to be segmented into an estimation segment and a validation segment. Depending on the size of the time series, a one-fifth to one-half of it is reserved for predictive validation. Ideally, there would be enough observations in the validation sample to provide enough statistical power to statistically determine whether there is significant forecast bias and whether one forecast is statistically significantly better than another. If the validation segment is too small, the research will not have enough power to properly evaluate the forecast against the real data or against another forecast.

Moreover, the researcher needs to be able to distinguish the estimation (historical) from the validation (test) sample. One way to do this is to use Stata's `e(sample)` function. This function = 0 if the sample is not the one on which the estimation occurred and it equals 1 if it is the sample on which the estimation occurs.

I prefer to construct a segment dummy so I can observe the coding when I am proofreading my computer code. I let the segment dummy, `segment = 0` if the sample is the estimation sample and `segment = 1` if the sample is the validation segment.

Ex post forecasts

One of the principal types of forecasts that we address is an *ex post* forecast. We must evaluate our models for predictive validity. To do so, we divide our sample into segments. Forecast evaluation requires at least two segments, although there can be three segments if the model applied is easily or likely to be overfit. The first segment is called the estimation segment or historical segment. The second segment is called the validation or test segment. We estimate the model on the estimation segment and we forecast over the validation segment. *Ex post* forecasts are sometimes referred to as "out-of-sample forecasts." This phrase can be misleading to those unfamiliar with the jargon. According to Hendry, this kind of prediction is really "out-of-estimation-sample" forecast because the forecast horizon begins at the end of the estimation segment and at the begin-

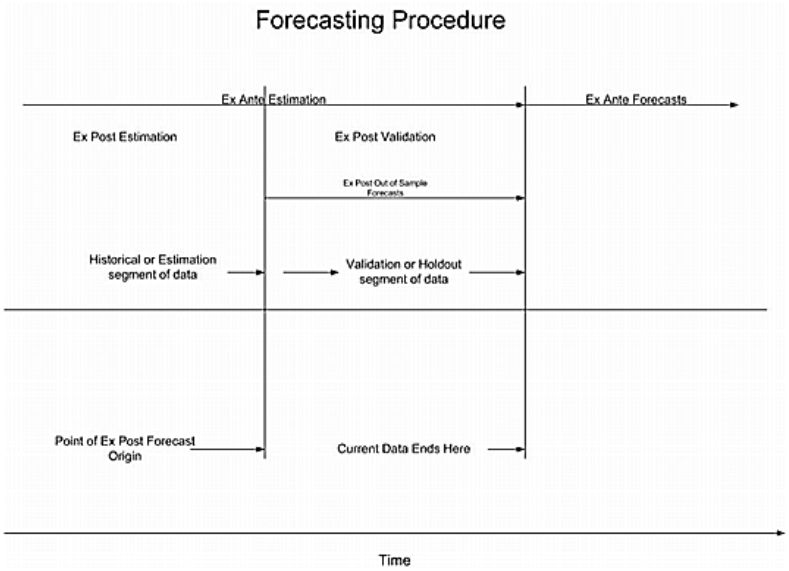

Figure 1: Data preparation for prediction

ning of the validation segment. Forecast evaluation is performed by comparing the *ex post* forecast to the actual data within the validation segment of the data. The difference between them is called the forecast error. An optimal forecast is generally deemed to be one that minimizes the sum of squared errors. The nature of that forecast error is assessed according to some criterion of forecast accuracy. The goodness of fit of the model may be assessed by comparing the estimated to the actual values of the data within the estimation segment, whereas the predictive validity of the models may be compared according to their relative forecast accuracy over the validation segment of the data. An example of an "out-of-estimation-sample" forecast is shown in Figure HS

Ex ante forecasts

Another principal type of forecast is called the *ex ante* forecast. Frequently, we need to forecast beyond the end of the sample data at a particular time. Our point of forecast origin in this type of forecast begins where the actual data cease to exist. Unless we have some conventional "gold standard" of forecast accuracy against which to compare these forecasts, we may have no baseline for comparison at the time of forecast origin. Forecasters customarily generate a "naive" forecast against which to compare this kind of forecast. Makridakis, Wheelwright, and McGee (1983) refer to two kinds of naive forecasts. The "naive" forecast of the first kind is one in which a random walk is extended

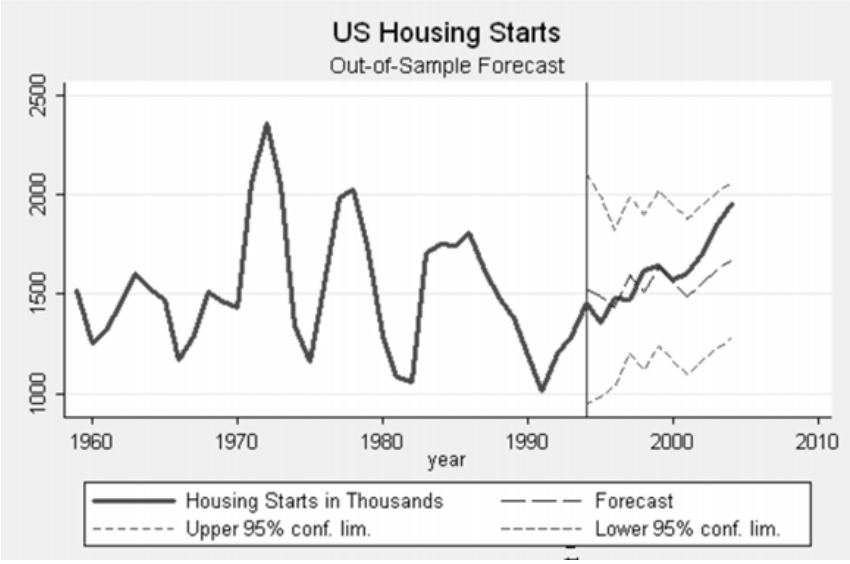

Figure 2: An *ex post* forecast of US Housing Starts (in thousands) 1959-2005

from the last value of the variable being forecast, whereas a "naive" forecast of the second kind is a deseasonalized extension of the variable being forecast as a basis for comparison (Ibid). Those who do *ex ante* forecasting, usually wait until the real data are collected and retrospectively evaluate their earlier forecast. They then compare empirical evaluation against the naive approach used earlier. Depending upon whether we employing an ex post forecast or an ex ante forecast, our forecast horizon will begin at a different point of forecast origin, (Pindyck and Rubinfeld, 1997) as shown in Figure 1. An example of an *ex ante* unconditional forecasts of U.S. imports after 2005 is shown in Figure 3

Forecast evaluation is a means by which predictive validity of models are assessed. It is customary to evaluate models for omnibus goodness of fit and other criteria for the purpose of model comparison and evaluation. Without such an assessment model selection would be impossible when there is no unique solution for a particular model.

One-step-ahead forecasts

Another classification of forecasts pertains to the type of algorithm applied. If the forecasts are uncorrelated with one another and they are generated so that each point projected is an iterated projection of values, then an iterated projection of values is a form of a one-step-ahead forecast. Each forecast of this type assumes that the previous forecast is actual data upon which the one-step ahead forecast is based. Because this forecast builds on previous optimal estimates this type of forecast can generally be more accurate than a multi-

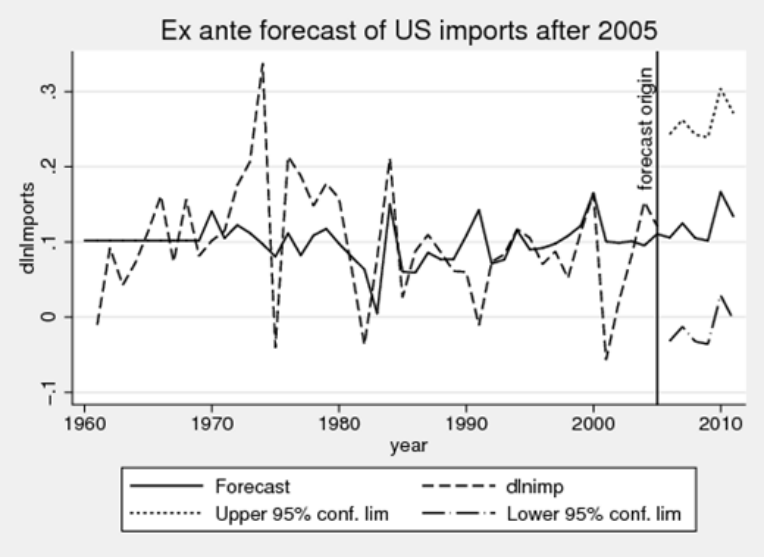

Figure 3: A unconditional *ex ante* forecast of US imports after 2005

step dynamic forecast. The state space models generally perform this kind of forecast with the Kalman filter. An example of this iterated projection is shown in Figure 4.

Static forecasts

Static models, such as extrapolations, presume that the world does not change. The ability to apply an extrapolative forecast implies a stability of the situation. With an evolving world, more complex models that permit dynamics are required. In the short run, the models must accommodate observable dynamics. In the relatively long run, the process is one of a kind of dynamic equilibrium or relatively static. There are some exceptions of course. Static regressions contain contemporaneously interdependent endogenous series and predictive series, although their lagged values may exhibit no relationship, yet these phenomenon are unusual(Hendry, 1995).

Dynamic forecasts

Some people may be inclined to think that a dynamic forecast is any forecast generated from time-series data. These models must take the lag structure into account and may have variables to measure. There may be measurement error in the variables and this may change over time. They may entail adjustment to rational expectations. They may involve leading indicators. They may vary about a dynamic equilibrium. In so doing, they distinguish this kind of fore-

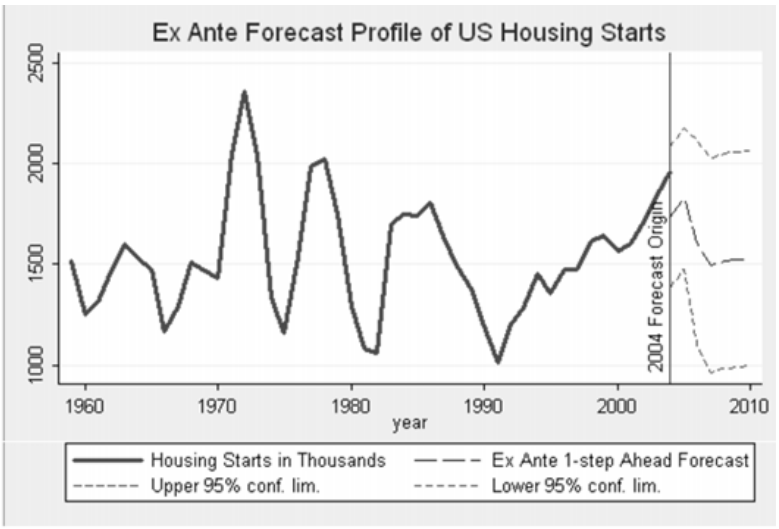

Figure 4: A 6 year one-step-ahead forecast of U.S. housing starts (in thousands) from 2004

cast from a static forecast, which might be generated prediction intervals from a cross-sectional regression. Another view of a dynamic forecast is a multi-step-ahead forecast that can be distinguished from a one-step-ahead forecast. Whereas the one-step-ahead forecast is an iterated projection of uncorrelated forecasts, generated by an ARIMA `predict` command, with the Kalman filter, the multi-step dynamic forecast with `prais` has been considered a simultaneous projection h-steps based only on the data prior to the point of forecast origin. Because such dynamic prediction is based only on the data prior to the forecast origin, it can be considerably less accurate than the iterative one-step ahead projection of the Kalman filter. Dynamic forecasts do not build upon incremental projections that could enhance its accuracy. Consequently, this type of forecast may not be as accurate as the one-step-ahead forecast. Others refer to it as another types of forecast (Stata Time Series Reference Manual, Release 11).

Rolling origin forecasts

There are two kinds of rolling origin forecasts. A rolling origin forecast can have a fixed length window. This approach is often applied when there are unusual end-effects in a time series being forecast. The end-effects can seriously bias a forecast. Because end effects are not common throughout the whole time series, a fixed-length window may be rolled along the time path, period by period, until it reaches the end of the data. The point of forecast origin is therefore incrementally moved forward until the window cannot move further

without being compressed. For each movement of the window, the measures of forecast accuracy are computed and stored. After the rolling ceases, the average of those computed forecast accuracy measures is calculated and this is used to assess the general forecast accuracy. The advantage of this approach is that unusual outliers or outlier patches or level shifts do not wreak havoc on the whole assessment. The other kind of rolling origin forecast is one where the window has a variable length. The origin can move either forward or backward but the window is either respectively shortened or widened over the remaining time periods. The advantage of this form of recursive least squares estimation is that the overall evaluation becomes less dependent upon the length of the forecast horizon. But a detailed discussion of evaluation for these kinds of forecasting is beyond the scope of this introductory exposition.

Unconditional and conditional forecasting

Predictability is necessary but not sufficient for forecastability (Ibid, 8). Forecastability requires not only a systematic relationship but a knowledge of how the information set enters the conditional density of the data generating process (Ibid, 8). Misspecification can arise there due to hidden or unanticipated correlations with excluded or unknown variables but they may also come about due to unanticipated changes in variables over the forecast horizon. An important distinction should be made between unconditional and conditional forecasts. Sometimes smoothers are used to perform simple and quick forecasts. Sometimes an ARIMA model is based on a single series. These forecasts are not conditional on exogenous time series that influence them. They are unconditional forecasts. Other forecasts have a time series regression framework. An endogenous variable may be influence by proximate, indirect, or direct effects associated with it. These forecasts depend on predetermined or weakly exogenous variable to constitute the model from which the forecast is generated. These types of forecasts entail preliminary prediction of all such exogenous or weakly exogenous variables over the forecast horizon, before the formulated model can be applied to generate the forecast of the endogenous variable over that forecast horizon. The preparation for the conditional forecast is the preliminary forecast of all of the predetermined or weakly exogenous variables over the forecast horizon. In Figure 5, we observe a conditional forecast of an error correction model of personal consumption expenditures as a function of per capita disposable income.

Measuring forecast accuracy

Simple non-causal methods may forecast more accurately than more complex causal models. E.g., some exponential smoothers have outperformed more complex models in the international M-3 forecasting competition. One of the models that performed superbly in the M-3 competition in 2000 was the "Theta model

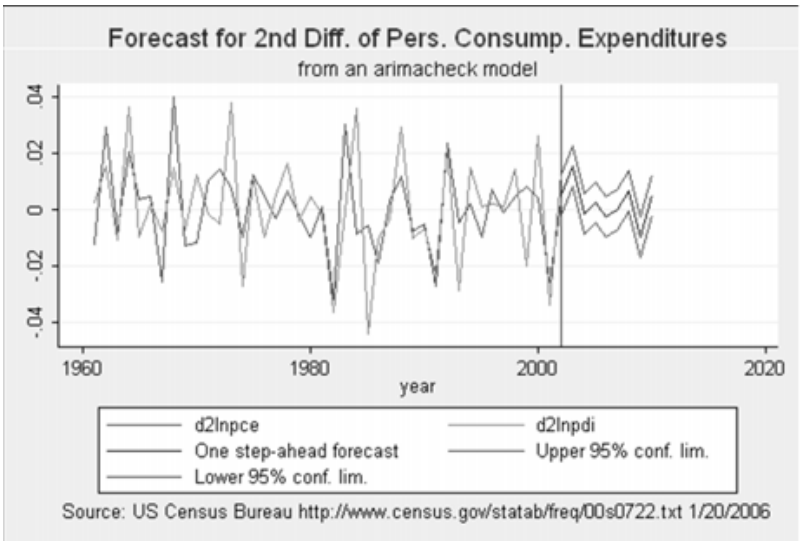

Figure 5: A conditional forecast of personal consumption expenditures as a function of personal disposable income

(Assimakopoulos and Nikolopoulos, 2000)." The theta model, in its simplest form, is merely the average of the simple exponential smoother and a linear regression line. Figure 6 displays the theta model forecast in comparison to the raw data of life expectancy of all races in the United States over the years shown on the horizontal axis.

For a model to be well-specified, it should optimize the encompassment of the driving forces within a system of time series. When the system is not as closed as the models imply, there may be unanticipated level shifts or impacts from changes in other fields owing to innovations, discoveries, interactions and correlations with legislative, administrative, or policy changes. Sir David Hendry, Graham Mizon, Neil Ericsson and others have referred to this part of a model as the extent to which the model "encompasses" the theory under consideration (Hendry and Mizon, 2005 ; Ericsson, 1994). Hendry has also maintained in his writings that econometric models should fulfill the model assumptions that indicate proper specification. The model should exhibit linear functional form, no structural breaks, no significant residual autocorrelation or ARCH effects, residual homoskedasticity, residual normality, as well as parameter constancy. For a model to be theoretically and statistically congruent, it has to fulfill these assumptions. Failure to pass the assumptions would detract from the validity of the model.

The forecast error is defined as

$$\hat{e}_{t+h} = (\hat{y}_{t+h} - y_{t+h}) \tag{1}$$

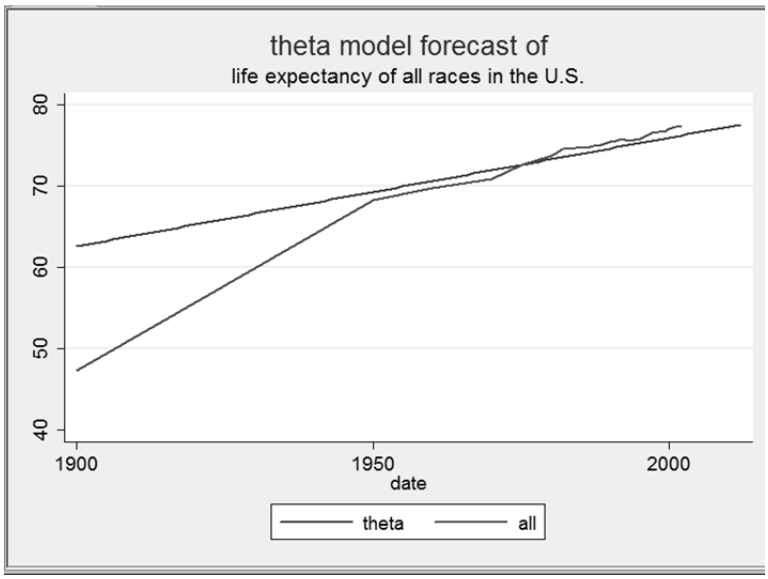

Figure 6: A Theta model forecast of life expectancy of all races in the U.S.

which means that the sum of squared forecast error (SSFE) is

$$SSFE = (\hat{y}_{t+h} - y_{t+h})^2 \tag{2}$$

where h = the number of time periods over which forecasting is performed.

If we divide the number of time periods over which the forecast is generated into the sum of squared forecast errors, we obtain the forecast error variance, otherwise called the mean square forecast error (MSFE).

$$MSFE = SSFE/h \quad = \frac{\sum_{T}^{T+H}(\hat{y}_{t+h} - y_{t+h})^2}{H} \tag{3}$$

where H = the total number of time periods over which forecasting is performed.

Alternatively, we could take use the root mean square forecast error (RMSFE), which is simply the square root of the mean square forecast error.

$$RMSFE = \sqrt{MSFE} \tag{4}$$

Although the forecast competition has been described as a tournament to ascertain what model can obtain a minimum forecast error variance, the use of the forecast error variance or its square root as a criterion of forecast accuracy has been criticized. Scott Armstrong notes that the forecast error variance is vulnerable to outlier inflation, owing to the squaring of the difference between

the actual and the forecast. If the series being forecast contains outliers, other measures of forecast accuracy that are not so inflatable by the presence of outliers would be preferred. One such measure is the absolute error, sum of absolute errors, or mean absolute deviation, mean absolute percentage error, or one of the various forms of the symmetric mean absolute percentage error.

Forecast bias

To determine whether there is significant forecast bias, we perform a nonparametric Wilcoxon signed-rank test between the actual data within the validation segment and the forecast. The nonparametric test allows the observations to be correlated and it does not require parametric assumptions about the nature of the forecast distribution. This permits evaluation of forecast financial series that have non-Gaussian fat-tailed distributions (a t distribution with the degrees of freedom less than six) or survival or extreme value distributions, which are notoriously non-Gaussian. The drawback is that the power efficiency of this test is only about 95% of that of a paired t-test. Ideally, the forecast horizon will be long enough to provide the necessary power to detect a small to medium effect size here.

The Stata command to perform this test is `signrank actual = forecast`. If we were to employ a Diebold-Mariano test we would be relying on some parametric assumptions.

Theil-Mincer-Zarnowitz test for Weak rationality of forecast

Were we to insist on a joint test of no significant constant and a slope of unity, we could use a Newey West regression of the actual data on the forecast. If the data generating process were correctly modeled, there would be no statistically significant slope and there would be a regression coefficient equal to unity. Jacob Mincer and Victor Zarnowitz proposed a regression model, shown in Eq. 5, in a paper entitled "The Evaluation of Economic Forecasts" a test for the weak rationality of the forecast (B. Hansen lecture notes, 2010). In this case, the actual data within the forecast horizon, is regressed upon the forecast in that same segment, the value of the constant (c) should equal zero and the value of the regression coefficient (b) should equal one for the condition of weak forecast rationality to exist. When the actual data was regressed on the Forecast, the constant had to be equal to zero and the regression coefficient had to be equal to unity for this weak rationality of the forecast to hold. We allow for some heteroskedasticity and residual autocorrelation to be corrected by the Newey-West autocorrelation correction to the Halbert White heteroskedastically consistent variance estimator.

$$Actual_t = c + b * Forecast_t \tag{5}$$

Directional accuracy

Sometimes people are interested in the directional accuracy of the endogenous series with respect to an exogenous series. Correlations test this covariance. A Pearson correlation coefficient and a Spearman's Rho rank-order correlation to test the monotonicity of the two series are given. The latter provides a relative scale from minus one to plus one, indicating the level of monotonicity obtained between these two series.

Absolute measures of forecast accuracy

In addition to the measures based on squared errors, we have a number of measures that are based on absolute error that are measured over the out-of-estimation sample forecast horizon. Among these is the sum of absolute errors (SAE), shown in Eq. 6, the mean absolute error (MAE), in Eq. 7, the mean absolute percentage error (MAPE), displayed in Eq. 9, and the median absolute percentage error (MdAPE), which is the middle observation of the rank-ordered absolute percentage error (APE) over the forecast horizon, if h is an odd number, or, if h is an even number, the MdAPE is the average of the h/2 and the h/2 + 1 observation of those rank-ordered APEs.

$$SAE = \sum_{t=T}^{T+h} |Actual_{t+h} - Forecast_{t+h}| \tag{6}$$

where T=point of forecast origin (h=0).

$$MAE = \frac{\sum_{t=T}^{T+h} |Actual_{t+h} - Forecast_{t+h}|}{h} \tag{7}$$

$$APE = 100 \sum_{t=T}^{T+h} |(Actual_{t+h} - Forecast_{t+h})/(Actual_{t+h})| \tag{8}$$

$$MAPE = (100)/(h) \sum_{t=T}^{T+h} |(Actual_{t+h} - Forecast_{t+h})/(Actual_{t+h})| \tag{9}$$

Each of these measures has their relative advantages and disadvantages. The SAE and MAE have no scale by which they can be relatively measured. The MAPE is somewhat scale dependent in that when forecasting very low values or integers—such as a one or two– the size of the measure is easily inflated to 100% or more. Therefore, when using the MAPE, it is important to accompany it with the MAE or an MSFE to provide a sense of balance. Because of this inflation, several scholars have proposed a symmetric MAPE (sMAPE).

The earliest version of the symmetric MAPE uses the addition of the forecast and the actual in the denominator, rather than just the actual. (O'Conner and Lawrence, 1993). The disadvantage is that this summation in the denominator places the symmetric MAPE not on a scale of 0 to 100, but on a scale of 0 to 200 (sMAPE1) For this reason, Michele Hibon and Spyros Makridakis (2000) proposed a revised version of the symmetric MAPE (sMAPE2), whereby we avoid the scale inflation with small numbers and restrict the fluctuations between -200 to 200 instead of a lack of any boundaries.

$$sMAPE1 = \frac{(100)}{(h)} \sum_{t=T}^{T+h} |(Actual_{t+h} - Forecast_{t+h})|/(Actual_{t+h} + forecast_{t+h}) \tag{10}$$

$$sMAPE2 = (200)/(h) \sum_{t=T}^{T+h} |(Actual_{t+h} - Forecast_{t+h})/(Actual_{t+h} + forecast_{t+h})| \tag{11}$$

The second version appears to be one of the preferred ones. It is the one that is included in these programs.

A Relative Measure of forecast accuracy

Henri Theil developed a relative measure of forecast accuracy. The measure is designed to compare the forecast error variance to that of a naive forecast or that of a random walk. This measure is particularly useful for *ex ante* forecasts, which frequently have to be compared to an artificial baseline. In 1983 Richard Meese and Ken Rogoff, reported in the International Journal of Economics, that many econometric structural models did not forecast as accurately as a random walk (cited in Hansen notes, 2010). Henri Theil's U, shown in Eq. 12 formed a ratio of the proportional error of the forecast to that of the naive forecast or random walk extension of the last value carried over the forecast horizon. Models with values less than one were deemed respectable whereas those with U values greater than one were not so respectable.

$$Theil's\ U = \sqrt{\frac{\sum_{t=T}^{T+h}((Actual_{t+h} - Forecast_{t+h})/(Actual_{t+h}))^2}{\sum_{t=T}^{T+h}((Naive1_{t+h} - forecast_{t+h})/(Actual_{t+h}))^2}} \tag{12}$$

Forecast evaluation modules

oforeval.ado for *ex post* forecasts

Version 2.0.0. of oforeval.ado has been written by this author to provide an out-of-sample ARIMA forecast evaluation over the forecast horizon specified by the user. A help file is also available, called oforeval.hlp. Both of these files should be stored in the c : $/ado/personal/o$ (where / represents a backslash in the pathname of the) directory on the hard drive of the user's computer or wherever he deposits his personal ado programs.

Preprocessing

1)The user constructs a segment variable, to be coded 0 during the estimation segment and 1 during the forecast evaluation.
2)The user performs his ARIMA analysis over the estimation segment.
For example, `arima dlnwpi if segment == 0, ar(13) ma(8)`
3)The user performs his forecasting over the validation segment of the data. He gives it the name "oforecast."
for example, `predict oforecast if segment == 1`

Command syntax

`oforeval arg1 arg2 arg3`
where
`arg1` = the variable name of the variable being forecast
after transformations.
`arg2` = the observation number representing point of forecast origin.
`arg3` = highest order of autocorrelation used in command.

Examples:

```
oforeval ds12.air 122 0
oforeval dlnwpi 116 3
```

(continued on the next page)

oforeval output

```
. predict oforecast if segment==1
(option xb assumed; predicted values)
. oforeval dlnwpi 116 4
```

```
       Out of estimation sample forecast evaluation
                     of dlnwpi
                Date:  23 Nov 2010
                Time:  22:37:35
```

```
                Forecast bias tests


   Paired t-test over forecast horizon

       t  =                              -0.5898
       p-value =                         0.5716

   Signrank test of difference between Forecast and Actual

       Signrank Z        =               -0.5331
       Probability > |Z| =               0.2970
```

```
Theil-Mincer-Zarnowitz Test of weak forecast rationality
        using a  Newey - West  regression of
        Actual = cons + B*Forecast

       Joint test of cons=0 and B=1
           F value =                     11.0174
           df   =                            2
           p-value =                     0.0069
```

(continued on the next page)

```
            Measures of directional accuracy

Pearson correlation between
forecast and actual  =                      -0.1230

Spearman rank correlation
         between forecast & actual    =    -0.1833
         p-value of Spearman Rho      =    0.6368

             Measures of forecast accuracy

Sum of Squared Errors            =         0.0013
Mean Square Forecast Error       =         0.0000
Root Mean Square Forecast Error  =         0.0001
Sum of Absolute Errors           =         0.0935
Mean Absolute Error              =         0.0104
Mean Absolute Percentage Error   =         171.8847
Median Absolute Percentage Error =         70.8047
Symmetric Mean absolute percentage
           error v.1          =            109.7530
Length of Forecast Horizon  (h)  =          8
```

(continued on the next page)

```
        Theil's U test of Forecast Accuracy

    Forecast Accuracy compared to Naive 1 forecast

          Theil's U   =                    0.8031
```

```
note bene:  Theil's U Scoring

  Theil's U = 1 when the naive method is as good as the
     forecasting technique being evaluated
  Theil's U > 1: There is no point in using a formal
     forecasting technique because naive method
     generates better results.
  Theil's U < 1: The Forecasting Technique being used is
     better than the naive method.
  Naive method 1 is using the last value carried
     forward as a forecast.

 Caveat: Statistical tests performed over short
 forecast horizons may lack sufficient statistical power
 to detect real differences between forecasts and
 actual data.  Power analysis may be in order under
 such circumstances.

Acknowledgements:
 I want to thank Kerry Kammire and Brian Poi
 for their helpful programming advice and assistance.
 Other references can be found in the help file.

     Author: Robert Alan Yaffee
     New York University
     robert.yaffee@nyu.edu
```

foreval.ado for *ex ante* forecasts

Version 1.0.0. of foreval.ado has been written by this author to provide an *ex ante* ARIMA forecast evaluation based on a naive1 extension of the last value carried forward. This is essentially a random walk extension of that value to provide a basis of comparison for a forecast projected beyond on the of the available data. A help file is also available, called foreval.hlp. Both of these files should be stored in the $c:/ado/personal/f$ directory, where the / may represent a backslash on the computer hard drive of the user or wherever his personal ado programs reside.

Preprocessing

1)The user constructs a segment variable, to be coded 0 during the estimation segment and 1 during the forecast evaluation.

2)The user decides how far into the future he intends to forecast.
3)The user performs his ARIMA analysis over the estimation segment.
For example, `arima dlnwpi if segment == 0, ar(1 3) ma(8)`
4)The user extends the date-time variable over the intended forecast horizon with: `tsappend, add(h)`
where h = the number of temporal periods over which he will forecast.
5) If other predictor series or event-dummy variables serve as regressors in the model, their need to be forecast over the planned forecast
horizon at this juncture in the protocol, before the endogenous series can be forecast.
6)The user performs his forecasting over the validation segment of the data. He gives it the name "oforecast."
for example, `predict forecast if segment == 1`

Command syntax

`foreval arg1 arg2 arg3`
where
`arg1` = the variable name of the variable being forecast after transformations.
`arg2` = the observation number representing point of forecast origin.
`arg3` = highest order of autocorrelation used in command.

Examples:

```
foreval ds12.air 122 0
foreval dlnwpi 116 3
```

(continued on the next page)

Output

```
. foreval dlnwpi 124 3
```

```
        Ex ante ARIMA forecast evaluation

                Date: 24 Nov 2010
                Time: 08:50:05
          Evaluations cover forecast horizon only
```

```
      Test of Forecast Bias against naive 1 baseline

        Paired t-test
          t    =       -1.355
          df   =           11
          p-value =     0.203

        Wilcoxon paired rank test

         Signrank Z   =       1.334
         Probability > |z| =  0.909


          Caveat: appears missing if forecast = straight line
```

```
       Theil's test of weak forecast rationality

       Newey-West Regression joint test
         of constant=0 and regression coefficient
         of actual data on forecast =  1

          F    =          43.9
          df   =          1
          p > |F| =       0.000

       If either slope or constant or both are omitted,
          the omission may be due to collinearity where
          both are approximately equal to zero.
```

(continued on the next page)

```
      Measures of Forecast Accuracy

         Sum of Squared Errors           (SSE)   =   0.001
         Mean Square Forecast Error      (MSFE)  =   0.000
         Root Mean Square Forecast Error (RMSFE) =   0.008
         Mean Absolute Error             (MAE )  =   0.007
         Mean Absolute Percentage Error  (MAPE)  =   8.809%
         Symmetric MAPE version one      (SMAPE1)=   9.790
         Symmetric MAPE version two      (SMAPE2)=  16.761
         Median Absolute Percentage Error(MdAPE) =  46.658%
         Forecast horizon =              (h)     =   12 obs

Caveats:
1) Forecast is compared to random walk extension of
   variable being forecast.
2) if h < 30 statistical power
   of forecast evaluation will be limited.
```

Directions for further development

In this paper we only address the forecast evaluation for ARIMA models. In later versions, we address issues of structural breaks, and present programs for ARCH/GARCH forecast evaluation as well as those for dynamic factor analysis and state space models.

(continued on the next page)

References

Armstrong, S.,(2000). Evaluating Forecast Methods in Armstrong, S.(ed.) Principles of Forecasting, Boston: Kluwer Publishers, 460-472.

Assimakopoulos, V. and Nikolopoulos, K. (2000). The Theta model: A decomposition approach to forecasting. International Journal of Forecasting, (16), 521-530.

Baum, C. (2009) An introduction to Stata programming, College Station, TX: Stata Press.

Box, George E. P.; Norman R. Draper (1987). Empirical Model-Building and Response Surfaces. Wiley. 424, 688.
This quote can be found at http://en.wikiquote.org/wiki/George Box.

Cameron, A.C. and Trivedi, P.K. (2009). Microeconometrics using Stata. College Station, TX: Stata Press, chapters 4, 10, & 13.

Castle, J. and Hendry, D. F. (2008). ISF2008 presentation in Nice France.

Castle, J. and Shephard, N. (2009). The Methodology and Practice of Econometrics. Oxford, UK: Oxford University Press,

Clements, M.P. and Hendry, D.F. (2001) Forecasting Non-stationary Economic Time Series, Cambridge, MA: MIT Press, xxv, 4-35.

Doornik, J. A. and Hendry, D. F. (2004). Econometric Modeling with PcGive, Vol I, London, UK: Timberlake Consultants, Ltd., other information criteria formulae are found here.

Ericsson, N.R. (1994). Testing Exogeneity: An Introduction. Oxford, UK: Oxford University Press, 30.

Hamilton, James D. (1994). Time Series Analysis, Princeton, NJ: Princeton University Press, 71-116.

Hansen, B. Lecture notes on world wide web:
http://www.ssc.wisc.edu/ bhansen/390/390Lecture23.pdf, Autumn 2010.

Hendry, D. F. (1995). Dynamic Econometrics. Oxford, UK: Oxford University Press., 17-19, 233.
Hendry, D.F. and Mizon, G.E. 2005 Forecasting in the Presence of Structural Breaks and Policy Regime Shifts. In Andrews, DWK and Stock, JH (eds), Identification and Inference for Econometric Models: Essays in Honor of Thomas

Rothenberg (pp. 480-502). Cambridge: Cambridge University Press.

Hibon, M. and Makridakis, S. (2000) The M-3 Competition: results, conclusions, and implications. International Journal of Forecasting,
(16), 461-476.

Makridakis, S., Wheelwright, S. and McGee, V. (1983). Forecasting: Methods and Applications, Inc.New York, NY: John Wiley & Sons, 47-50.

OConnor, M. and Lawrence, M., 1998. Judgemental forecasting and the use of available information. In: Wright, G. and Goodwin, P. Editors, 1998. Forecasting With Judgment Wiley, Chichester, 6590.

Pindyck, R. and Rubinfeld, D. (1997) Economic Models and Economic Forecasts. McGraw Hill/Irwin: New York, NY., 203.

Stata Release 11 Time Series Reference Manual (2009). College Station, TX: Stata Press, Inc., 63-64.