Forecasting

by Kevin Kotzé

A significant part of the time series literature considers the ability of a model to predict the future behaviour of a variable with a reasonable degree of accuracy. This objective is important, as most decisions that are taken today are based on what we think will happen in the future. Depending on the decision under consideration, the future can of course be the next minute, day, month, year, etc. For example, a day-trader may be interested in the price of a share in the next minute, hour or day, while the governor of a central bank may be interested in the rate of inflation over the next year.

Forecasting the future is not an easy thing to do, especially when it comes to economic or financial variables that reflect on the complex interactions between individuals, firms and organisations. This has lead to the development of a plethora of models that include large sophisticated mathematical variants, as well as those that are fairly simple. In many instances, we find that the forecasts from these models are comparable and the tools that are used to evaluate them would need to be carefully applied.

When evaluating forecasting models, we need to make use of *ex-post* data (after the fact) that has been realised after the results of the forecast have been generated. In these instances, we would utilise an explicit loss function that may consider whether a forecasting model provides a favourable estimate of the expected future value of a variable. For example, such a loss function would usually look to penalise the error of individual forecasts that arise over a specific period of time.

When comparing the forecasts that have been generated by two models we may also be interested in determining whether or not they are significantly different from one another. Separate tests have been developed for these investigations where the models are either nested or non-nested. The application of these statistics should be an important part of any forecasting exercise.

In addition, there may be a large degree of uncertainty associated with the forecasts of various models, which is usually of interest to economists and financial analysts. If the uncertainty is an inherent property of the variable we are looking to forecast, this might be a good thing. However if it is not, then this could make the forecast less useful. In many respects, different loss functions emphasize either the accuracy of the point estimate or confidence around this point, and both of these topics will be discussed in what follows.

Lastly, many large institutions employ a suite of forecasting models, where the final forecast is some weighted average of the forecasts from many different models. Several empirical studies have suggested that by employing a combination of forecasting models, one is able to reduce instances, where a particularly large forecasting error may arise.

1 Granger causality tests

Like with most other forms of analysis, before we start to construct a forecasting model, the first thing that one would want to do is plot the data. By way of example, consider the data in Figure 1, which suggests that if you are looking to forecast the oil price prior to 1974, when the price was controlled by regulators, then we would need to make use of a model that is able to generate large infrequent changes. However, if you were looking to forecast the price of oil after 1974, when the price of oil was determined by market forces, then one would need to construct a model that is capable of forecasting both small and large changes that arise with a higher frequency.

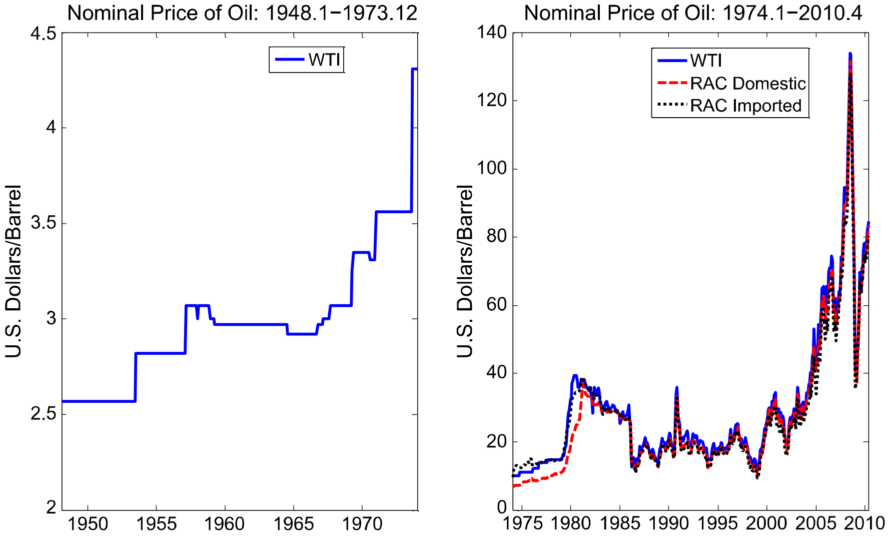


Figure 8.1 The nominal price of crude oil. *Notes*: WTI stands for the West Texas Intermediate price of crude oil and RAC for the U.S. refiners' acquisition cost.

Figure 1: **Oil prices have a clear structural break**

Therefore, when inspecting the data, one would usually also look for changes in the expected mean value of the underlying data-generating process. In addition, it is also a good idea to consider the degree of variability and potential changes in the variability that may have arisen over time. One should also try to detect obvious outliers, before considering how such an outlier may potentially influence the forecasts that would be generated if it is incorporated in the dataset that is used to generate a particular forecast.

After this exercise is complete, we could then determine whether or not we can make use of various variables to forecast future values of a target variable, which we denote y_t. For example, consider that we are at time t and we want to forecast h steps into the future. Hence, we are looking to predict the future value of our target variable at time $t + h$, where h is the forecast horizon. Therefore, we want to generate values for $\mathbb{E}_t \left[y_{t+h} \right]$ given some predictors, which we denote x_t. Where there is linear relationship between these variables we could make use of the regression model:

$$\mathbb{E}_t \left[y_{t+h} \right] = \beta^{\top} x_t$$

where x_t may represent a vector of variables and β would then represent a vector a coefficients. In the simple case, where x_t only contains a single variable, we could perform a Granger causality test to consider whether the value of β is significantly different from zero. To perform such a test we could construct the null for no predictive ability, that could be evaluated against the alternative where the variable has some predictive ability. Hence, we would like to perform the statistical test:

$$\begin{aligned} H_0 &: \beta = 0 \\ H_1 &: \beta \neq 0 \end{aligned}$$

To complete this test we need to compare the estimated value of $\hat{\beta}$ against the restricted value of this coefficient under the null hypothesis, which is zero. Therefore, we are interested in the difference, $(\hat{\beta} - 0)$. To determine if this result is significantly different from zero, we need to consider the variability of the underlying data. If there is a large amount of variability in the underlying data then we may conclude that a larger value of $(\hat{\beta} - 0)$ may be insignificantly different from zero, while if there is a small amount of variability in the data then only small values for $(\hat{\beta} - 0)$ would be insignificantly different from zero. Therefore, we use construct the t-test as follows:

$$t_{\beta} = \frac{\left(\hat{\beta} - 0\right)}{\mathsf{std}(\hat{\beta})}$$

where we can reject the null hypothesis when the value for the t-statistic is large, which would suggest that the model has predictive power. The calculated values from this statistic are compared against the critical values from a normal distribution, where we are able to reject the null hypothesis within a 95% confidence interval when, $|t_{\beta}| \leq 1.96$. Alternatively, we could consider the p-values for the t-test, where if the p-value is 0.05 then you are on the boundary of rejecting at the 95% confidence interval of a normal distribution.

This statistic is conditional on the value of h, as the model could be useful when predicting one-step ahead, but it may not be useful when trying to forecast twelve-steps ahead. In addition, when we have several predictors, then we would need to construct an F-test to consider the joint predictive ability of a group of variables.

To calculate an appropriate value for the denominator in the t-test we make use of the error term in the regression model. For example, consider the regression model:

$$\mathbb{E}_t \left[y_{t+h} \right] = \beta^{\top} x_t + \varepsilon_{t+h}, \quad \text{where} \quad \varepsilon_{t+h} \sim \mathsf{i.i.d.}\mathcal{N}(0,1)$$

In those cases where we make use of lags of the dependent variable to forecast forward, such as the case of an autoregressive model, it would be unlikely that shocks to the variable would be independent. To illustrate this, consider the AR(1) that may be used for successive h-step-ahead forecasts:

$$\begin{aligned} \mathbb{E}_t \left[y_{t+1} \right] &= \phi_1 y_t + \varepsilon_{t+1} \\ \mathbb{E}_t \left[y_{t+2} \right] &= \phi_1 y_{t+1} + \varepsilon_{t+2} = \phi_1 \left(\phi y_t + \varepsilon_{t+1} \right) + \varepsilon_{t+2} \\ \mathbb{E}_t \left[y_{t+3} \right] &= \phi_1 y_{t+2} + \varepsilon_{t+3} = \phi_1 \left(\phi y_{t+1} + \varepsilon_{t+2} \right) + \varepsilon_{t+3} \\ \vdots &= \vdots \end{aligned}$$

In this case, the values for ε_{t+h} represent the forecast errors. This error term would not be independent and hence we need to correct the denominator in the above t-test, since future errors may be dependent on previous errors. In addition, the larger is h the more serially correlated the error term. Therefore, we need to make use of Newey and West (1987), heteroskedasticity and autocorrelation consistent (HAC) estimate of the variance.

Note also that when making use of these tests, to consider whether lagged values of x_t Granger cause y_t, it would not necessarily be an indicator that x_t would have useful predictive power. This case is shown in Meese and Rogoff (1983), where they note that although the interest rate differential $i_{t+h} - i^*_{t+h}$ Granger causes a change in the exchange rate, $s_{t+h} - s_t$, it does not necessarily perform well when used in an out-of-sample forecast evaluation exercise, where we only have information about i_t and i^*_t and not the future values for these variables. In addition, another reason for not making use of in-sample Granger causality statistics for the evaluation of forecasts is that they would encourage over-fitting, where we make use of data that can only describe a particular portion of the known dataset.

2 Out-of-sample forecasting notation

To describe the procedure for constructing forecasts, we firstly need to introduce some notation. By way of example, if we have quarterly data and want to obtain forecasts over the next eight quarters (i.e. two years), then we would want to generate eight successive forecasts. Therefore, we would want to generate a number of h-step ahead forecasts, where $h = \{1, 2, \ldots, 8\}$. The end of the forecasting horizon may be represented by H, where in this case, $H = 8$.

Since we would be unable to evaluate the performance of a model from a single forecast, we would usually want to cosider how the model performs over a number of successive forecasts. In such a case we make use of an in-sample portion of size R and an out-of-sample portion of size $P + H$, where P refers to number of predictions. Note that the parameter values are estimated on a sample of data that is not used to test the accuracy of the predictions. In the machine learning literature the in-sample portion of data would refer to the *training dataset* and the out-of-sample portion would refer to the *testing dataset*. In what follows, the principles that we are going to discuss would also apply to those that are found in the machine learning literature. Therefore, to generate the first forecast for $\mathbb{E}_t [y_{t+1}]$ we would make use of the data from period R to generate a forecast for period $R + 1$. In certain instances, we will refer to the *information set*, which is denoted I_t. This term refers to the information, relating to the values of a variable, that is available at a particular point in time. Therefore, in most cases I_t would refer to information pertaining to all the past realised values of a particular variable or variables.

In an out-of-sample evaluation of a model, we would want to compare the forecasts against future realisations of the data. To complete this exercise the researcher would divide the available sample size of $T + H$ into an in-sample portion of size R and an out-of-sample portion of size $P + H$.

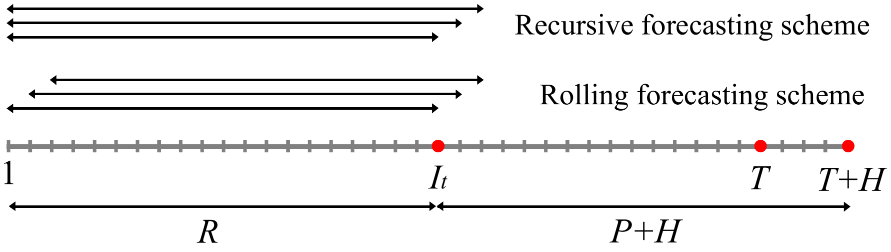


Figure 2: **Notation for different forecasting schemes**

The notation that is used in forecasting exercises differs slightly to what we had previously, as the entire sample of realised values is represented by $T+H$. When looking to perform an out-of-sample forecasting evaluation one would usually generate a number of successive forecasts, where the information set would include an additional single observation when generating successive forecasts, as the positioning of I_t moves towards T. When the variable under consideration is y_t, we would generate the initial h-step-ahead forecast that we denote y^f_{R+h}. The value for this forecast would then be compared to the realised value of this time series at period y_{R+h} to generate the forecast error:

$$e^f_{R+h} = y_{R+h} - y^f_{R+h}$$

To consider whether or not the model provides accurate forecasts, we mimic what the forecaster would have done in real time, by estimating the model over various points of time to generate a sequence of forecast errors. Therefore, after estimating the model with the use of data for the in-sample period R to calculate the initial forecast errors, the in-sample period would then increase to $R+1$. The model parameters are then re-estimated to generate a new value for the $\hat{\beta}$ coefficients, which are used to generate the new set of forecasts for the period y^f_{R+1+h}. In this case the calculation of the forecast errors would be given by $e^f_{R+1+h} = y_{R+1+h} - y^f_{R+1+h}$. This procedure would continue until the last estimation, which takes place at time T to generate forecasts up until period $T+H$.

When using a recursive scheme, the initial observation in the in-sample period is fixed at the first observation of the data, while the rolling-window scheme would maintain a fixed number of observations for the in-sample period. Therefore, a recursive estimation scheme would use all available information (from period $t=0$ to I_t), while in a the rolling scheme the first observation in the in-sample period would increase by a single observation prior to the successive estimation of the model parameters. The rolling-window scheme is usually preferred when there are potential structural breaks that arise during the in-sample period, which would influence the estimates of $\hat{\beta}$, while the recursive scheme would usually produce more accurate forecasts when the variable is relatively stable over time.

For example, assume that our dataset extends from 1995q1 to 2014q4, and we want to generate a number of one-step ahead forecasts over the out-of-sample period 2010q1 to 2014q4. This would imply that R would be 2009q4, and the initial in-sample period would be 1995q1 to 2009q4, which would be used to generate a forecast for 2010q1. After calculating the forecast error, the revised in-sample period for a recursive scheme would span 1995q1 to 2010q1. When using a rolling window scheme the in-sample period would be 1995q2 to 2010q1. Either of these methods could be used to generate a forecast for 2010q2, which would be used to calculate a second forecast error. This procedure continues until we have full set of forecast errors. After using a particular model in an out-of-sample forecasting exercise we would have generated a total of P forecast errors for each horizon, h. Therefore, we have a sequence of forecast errors that may be expressed as follows:

$$\left\{ e^f_{t+h} \right\}^T_{t=R}$$

In most cases we would want to evaluate both the short-term and long-term forecasting performance of a model. Hence, if we want to consider the forecasting performance of a model that is used to generate one- to H-step ahead forecasts, we would usually make use of a matrix for capturing the forecasting errors, where each of the vectors for the h-step-ahead forecasting errors are placed in a separate column.

$$\boldsymbol{e}_H = \left\{ \begin{array}{cccc} e^f_{R+h}(h=1) & e^f_{R+h}(h=2) & \ldots & e^f_{R+h}(h=H) \\ e^f_{R+1+h}(h=1) & e^f_{R+1+h}(h=2) & \ldots & e^f_{R+1+h}(h=H) \\ \vdots & \vdots & \ddots & \vdots \\ e^f_{T+h}(h=1) & e^f_{T+h}(h=2) & \ldots & e^f_{T+h}(h=H) \end{array} \right\} \tag{2.1}$$

Note that the respective columns or rows in the matrix would represent a time series variable. For example, the first column would represent the one-step ahead forecasts errors over the out-of-sample period. This would imply that this matrix of forecasting errors is of dimension $(P \times H)$ and we will see that this representation will be particularly useful when calculating the respective statistics that are used to evaluate the performance of different forecasting models.

Before moving on it is perhaps worth noting that these forecasting experiments are usually termed *pseudo* (or quasi) out-of-sample forecasting experiments, where the use of the *pseudo* term would infer:

- When conducting the evaluation we have at our disposal all the out-of-sample information, which are the realisations of the observed variable. This information could have been used in the formulation and specification of the model (i.e. we see that the out-of-sample portion of the dataset exhibits regime-switching behaviour and as a result we decide to make use of a regime-switching model).
- In economics, and specifically in macroeconomics, most data series are heavily revised over time. For example, when the statistical agency of a given country releases a value for gross domestic product for the first time, this number will typically only be an estimate of the true value of economic output. During subsequent periods, this estimate will be subject to a number of revisions as the agency obtains more information about what actually occurred over that period. Databases with true real-time data exist for some countries and where such data is utilised in a forecasting experiment, we refer to it as a *real-time* out-of-sample forecasting evaluation.

Of course, doing a real-time out-of-sample forecasting experiment is subject to the same flaw as described in the first bullet, which is of particular importance when the model has undergone significant revision over time, using information that may not have been available at the time that pertained to the initial in-sample period.

Examples of dataset vintages that may be used to evaluate forecasts at the time when the realised value was first released are describd in Croushore and Stark (2001), and you will find such data for the United States at: https://www.philadelphiafed.org/research-and-data/real-time-center/real-time-data/ (https://www.philadelphiafed.org/research-and-data/real-time-center/real-time-data/). I’ve compiled similar datasets for South Africa, which are stored on my GitHub and GitLab repositories. See either https://github.com/KevinKotze (https://github.com/KevinKotze) or https://gitlab.com/KevinKotze (https://gitlab.com/KevinKotze).

3 Forecasting with random walk models

To consider if the forecasts from a particular model are any good, we may want to compare them to those of a random-walk model. Forecasts from such a model may be expressed as follows:

$$\mathbb{E}_t \left[y_{t+h} \right] = y_t$$

where the forecast errors that may be summarised by the sequence:

$$\left\{ e^{rw}_{t+h} \right\}^{T}_{t=R}$$

Hence, if the random-walk model provides superior forecasts it would imply that we should rather make use of the current value of a time series variable to predict the future of this variable, rather than use a competing model forecast.

4 Forecasting with autoregressive models

With an autoregressive model, we are able to relate the current value of a process to previous values. For example, we may specify an AR(1) model as follows,

$$y_t = \phi_1 y_{t-1} + \varepsilon_t \tag{4.1}$$

where the residual is Gaussian white noise, $\varepsilon_t \sim \mathsf{i.i.d.}\, \mathcal{N}(0, \sigma^2)$. In this case we can simply iterating the model forward over a number of steps to calculate future values of a process,

$$\begin{aligned} y_{t+1} &= \phi_1 y_t + \varepsilon_{t+1} \\ y_{t+2} &= \phi_1 y_{t+1} + \varepsilon_{t+2} \\ \vdots &= \vdots \\ y_{t+H} &= \phi_1 y_{t+H-1} + \varepsilon_{t+H} \end{aligned}$$

This expression may be summarise, by inserting the first line into the second line to relate successive future values to the current observed realisation. Hence, for $h = 2$ we have,

$$\begin{aligned} y_{t+2} &= \phi_1 (\phi_1 y_t + \varepsilon_{t+1}) + \varepsilon_{t+2} \\ &= \phi_1^2 y_t + \phi_1 \varepsilon_{t+1} + \varepsilon_{t+2}, \end{aligned}$$

and after doing this recursively for $h = 3$ we have,

$$\begin{aligned} y_{t+3} &= \phi_1 \left(\phi_1 \left(\phi_1 y_t + \varepsilon_{t+1} \right) + \varepsilon_{t+2} \right) + \varepsilon_{t+3} \\ &= \phi_1^3 y_t + \phi_1^2 \varepsilon_{t+1} + \phi_1 \varepsilon_{t+2} + \varepsilon_{t+3}, \end{aligned}$$

This gives rise to a pattern, where after sorting terms and simplifying,

$$y_{t+h} = \phi_1^h y_t + \sum_{i=0}^{h-1} \phi_1^i \varepsilon_{t+h-i} \tag{4.2}$$

Thus y_{t+h} is a function of y_t, which represents the available information set, I_t, that contains information about all past realised observations of this variable. In this case, the actual observed (i.e. realised) future values of y_{t+h} will also contain the effects of future shocks. Unfortunately, information about the future realised values of these shocks is not included in I_t. Hence, this part of the future realised value of y_{t+h} may give rise to an irreducible forecasting error.

4.1 Point forecasts for autoregressive models

To compute the point forecast of y_{t+h}, we take the conditional expectation of $\mathbb{E}_t\left[y_{t+h}|I_t\right]$. In the case of the simple AR(1) in (4.1), this is equivalent to $\mathbb{E}_t\left[y_{t+h}|y_t\right]$. In addition, since all the future error terms in this model have an expected future mean value of zero, we can use (4.2) to calculate the one-step ahead forecast $\mathbb{E}_t\left[y_{t+1}|y_t\right] = \phi_1 y_t$. Similarly, the two-step ahead forecast would be derived from $\mathbb{E}_t\left[y_{t+2}|y_t\right] = \phi_1^2 y_t$, and the more general expression may then be given as,

$$\mathbb{E}_t\left[y_{t+h}|y_t\right] = \phi_1^h y_t \tag{4.3}$$

This expression is also occasionally referred to as a predictor. If the variable that we are seeking to forecast is described by a *stable* AR(1) (i.e. it is assumed that $|\phi_1| < 1$), and the structure does not include a constant, as in (4.1), then the term $\phi_1^h y_t$ would tend towards zero as the forecast horizon increases. Hence,

$$\mathbb{E}_t\left[y_{t+h}|y_t\right] \rightarrow 0 \qquad \text{when } \; h \rightarrow \infty$$

Therefore, the effect of shocks that may be contained in y_t would dissipate, as the forecast horizon increases.

4.2 Autoregressive models with an intercept

If we assume that an intercept is included to the stable AR(1) equation, such that,

$$y_t = \mu + \phi_1 y_{t-1} + \varepsilon_t \tag{4.4}$$

Using the same recursions as above, we are able to derive the expression,

$$\mathbb{E}_t\left[y_{t+h}|y_t\right] = (1 + \phi_1 + \phi_1^2 + \ldots + \phi_1^{h-1})\mu + \phi_1^h y_t \tag{4.5}$$

Hence, the one-step ahead forecast, where $h = 1$, is simply $\mathbb{E}_t\left[y_{t+h}|y_t\right] = \mu + \phi_1 y_t$. Similarly, the two-step ahead forecast, where $h = 2$, is $\mathbb{E}_t\left[y_{t+2}|y_t\right] = (1 + \phi_1)\mu + \phi_1^2 y_t$. Therefore, when h goes to infinity, we see that the forecast converges on the unconditional mean of the process,

$$\mathbb{E}_t\left[y_{t+h}|y_t\right] \longrightarrow \frac{\mu}{1 - \phi_1} \qquad \text{when } h \rightarrow \infty \tag{4.6}$$

This follows from the summation formula for an infinite sequence, $(1 + \phi_1^1 + \ldots + \phi_1^\infty) = 1/(1 - \phi_1)$.

4.3 Higher-order autoregressive models

For completeness, the case of an AR(p) model with an intercept is provided below,

$$y_t = \mu + \sum_{i=1}^{p} \phi_p y_{t-i} + \varepsilon_t \tag{4.7}$$

where p is the maximum lag length, and the residual is Gaussian white noise. For simplicity, we take the conditional expectation at each forecast horizon when computing the forecast recursively.

$$\begin{aligned}
\mathbb{E}_t\left[y_{t+1}|y_t\right] &= \mu + \phi_1 y_t + \phi_2 y_{t-1} + \ldots + \phi_{p+1} y_{t-p+1} \\
\mathbb{E}_t\left[y_{t+2}|y_t\right] &= \mu + \phi_1 \mathbb{E}_t\left[y_{t+1}|y_t\right] + \phi_2 y_t + \ldots + \phi_{p+2} y_{t-p+2} \\
\mathbb{E}_t\left[y_{t+3}|y_t\right] &= \mu + \phi_1 \mathbb{E}_t\left[y_{t+2}|y_t\right] + \phi_2 \mathbb{E}_t\left[y_{t+1}|y_t\right] + \phi_3 y_t + \ldots + \phi_{p+3} y_{t-p+3} \\
\vdots &= \vdots \\
\mathbb{E}_t\left[y_{t+h}|y_t\right] &= \mu + \phi_1 \mathbb{E}_t\left[y_{t-1+h}|y_t\right] + \phi_2 \mathbb{E}_t\left[y_{t-2+h}|y_t\right] + \ldots + \phi_p \mathbb{E}_t\left[y_{t-p+h}|y_t\right] \\
&\quad + \phi_h y_t + \ldots + \phi_{p+h} y_{t-p+h}
\end{aligned} \tag{4.8}$$

Where the last line assumes that $p > h$.

4.4 Forecast errors in autoregressive models

As has been noted in (4.9), the forecast error, e^f_{R+h}, in period $t+h$ would be calculated from the expression:

$$e^f_{R+h} = y_{R+h} - y^f_{R+h} \tag{4.9}$$

where y_{R+h} is the *ex-post* actual realisation of the value for respective variable. Making use of this expression and the recursive formulas that are provided in (4.2) and (4.3), we can calculate the forecast error at different forecast horizons,

$$\begin{aligned}
e^f_{R+1} &= y_{R+1} - y^f_{R+1} = (\phi_1 y_R + \varepsilon_{R+1}) - \phi_1 y_R = \varepsilon_{R+1} \\
e^f_{R+2} &= y_{R+2} - y^f_{R+2} = (\phi_1^2 y_R + \phi_1 \varepsilon_{R+1} + \varepsilon_{R+2}) - \phi_1^2 y_R = \phi_1 \varepsilon_{R+1} + \varepsilon_{R+2} \\
\vdots &= \vdots
\end{aligned} \tag{4.10}$$

where at horizon h,

$$\begin{aligned}
e^f_{R+H} = y_{R+H} - y^f_{R+H} &= \left(\phi_1^H y_t + \sum_{i=0}^{H-1} \phi_1^i \varepsilon_{t+H-i}\right) - \phi_1^H y_t \\
&= \sum_{i=0}^{H-1} \phi_1^i \varepsilon_{R+H-i}
\end{aligned} \tag{4.11}$$

This result suggests that the forecast errors are just the coefficients of the MA representation of the AR(1) process if the process is stationary (and the MA representation exists). When we assume that the errors in autoregressive model are Gaussian white noise, the expected value of all future realisations of the error, as derived in (4.11) are zero. Therefore, when we assume that $\mathbb{E}_t\left[\varepsilon_{t+h}|I_t\right] = 0$, it would be the case that

$$\mathbb{E}_t \left[e^f_{t+h} \right] = \mathbb{E}_t \left[y_{t+h} - y^f_{t+h} \right] = \mathbb{E}_t \left[y_{t+h} \right] - \mathbb{E}_t \left[y^f_{t+h} \right] = 0 \tag{4.12}$$

If this is the case then it would imply that the predictor is unbiased.

4.5 Example of autoregressive forecasts

To illustrate the workings of the way in which one is able to construct a forecast from an autoregressive model, we can work with hypothetical AR(1) and AR(2) examples. In this case we assume that the coefficients are as follows:

- AR(1) with $\mu = 0.4$ and $\phi_1 = 0.7$
- AR(2) with $\mu = 0.3$, $\phi_1 = 0.6$ and $\phi_2 = 0.1$.

For both processes we assumed that they start off from the same initial situation, where $y_t = 2$ and $y_{t-1} = 1.5$. Using the formulae from (4.5) for the AR(1) model and (4.8) for the AR(2) model, the forecasts for horizons $h = 1, h = 2, h = 3, h = 5$ and $h = 10$ for both processes are reported in Table 1.

These results suggest that the forecasts converge on the unconditional mean of the processes as the forecast horizon increases, where for the AR(1) process, $0.4/(1 - 0.7) = 1.3$. Similarly, for the AR(2) process, $0.3/(1 - 0.6 - 0.1) = 1$.

	Forecast horizon				
	1-step	**2-step**	**3-step**	**5-step**	**10-step**
AR(1)	1.8	1.66	1.56	1.45	1.35
AR(2)	1.65	1.49	1.36	1.19	1.04

Table 1: **AR forecasts**

5 Forecast evaluation

To evaluate the performance of a forecast we may consider the use of various different statistics. For example, we may wish to derive an estimate of the bias that is associated with the forecast to determine whether or not we are consistently underestimating or overestimating the expected future mean value. In addition, when comparing the forecasting performance of a number of different models, we would usually want to evaluate a number of successive h-step ahead forecasts that have been generated over time. After generating the forecasts we could then sum up the forecasting errors to determine which model provides the smallest total forecasting error. However, as we would not wish for the positive and negative values of the forecast errors to cancel each other out, we may wish to take the square or absolute value of these errors. This has given rise to a number of different loss functions, which we denote $\mathscr{L}_{t+h}$ for the forecast that is h-steps ahead.

When seeking to evaluate a forecast, we also need give due consideration to the particular loss function that would be relevant to the problem at hand. Figure 3 presents examples of three different loss functions, where no penalty is imposed on errors that have a value of zero, while the penalty increases at different rates for errors that are different from zero.

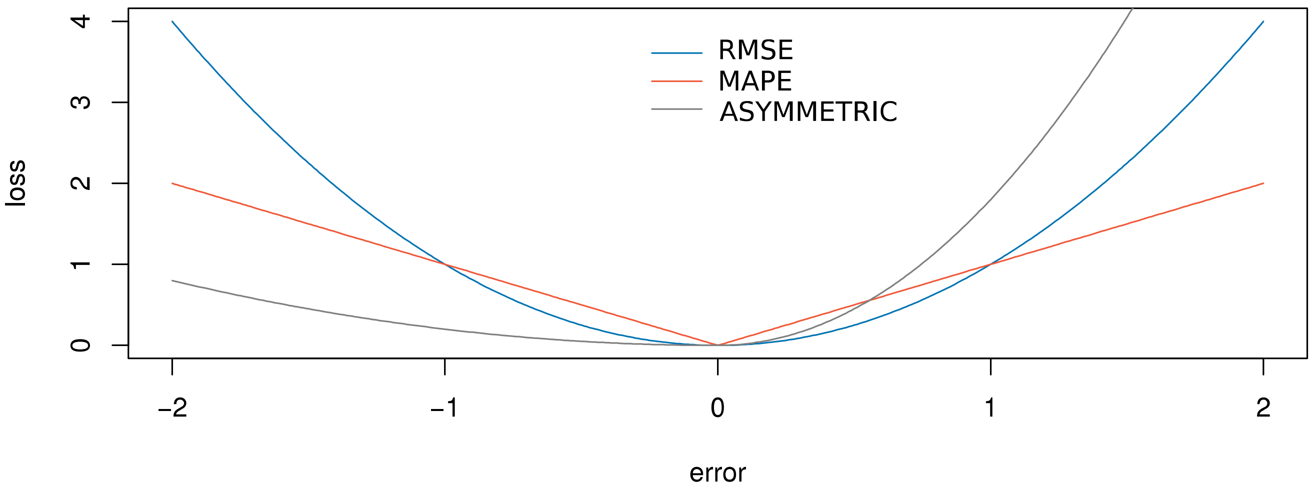


Figure 3: **Loss functions for forecast evaluation**

The most popular of these loss functions, is a squared loss function that is used in the root-mean squared error (RMSE). In this case we work with the square of the forecasting error, which implies that we are working with a quadratic function that places a larger penalty on large errors. In addition, if we are not that concerned with large forecasting errors then we could work with the absolute value of the forecasting error, which is used in the calculation of the mean absolute percentage error (MAPE). This loss function is linear and does not penalise large errors to the same extent as the RMSE.

There are also cases where an asymmetric loss function would be appropriate. For example, consider the case where an over-prediction results in a more unsatisfactory outcome than an under-prediction. Such an instance may arise in several financial applications that make use of various forms of asymmetric loss functions. In addition, we would also want to consider the degree to which we are concerned by outliers in the forecasting error as this would also influence our choice of loss function.

While the literature contains many different loss functions, we consider three of the most popular in what follows. The model that produces the smallest value for the loss function is usually termed the most efficient model. In general we would want to make use of the model that provides an unbiased and efficient forecast.

5.1 Bias

To measure whether the forecast errors are centred around zero we can take the simple average of the forecast errors, to provide an estimate of the forecast bias, which is the expected value of the forecast error. This statistic may be calculated as follows:

$$\mathbb{E}_t \left[e^f_{t+h} \right] = \frac{1}{P} \sum_{t=R}^{T} e^f_{t+h} \tag{5.1}$$

In the case where we have generated a matrix for the forecast errors, as in (2.1), then we could derive an estimate of the bias for each h-step ahead after deriving the mean for each of the column vectors.

In most cases, the value of the estimated bias that is closest to zero is preferred. If the forecast is biased, the model would have either made consistent mistakes of either predicting too high or too low a value. Alternatively, we may have made a few particularly large forecasting errors that skew the result.

Note that, this would not be the only statistic of interest as a model that has made both very large over-predictions along with very large under-predictions could still provide a bias that is equal to zero.

5.2 Root mean squared error (RMSE)

The most widely used method for evaluating forecasts is the root mean squared error (RMSE), which is usually calculate for each h-step ahead forecast. It is calculated by firstly taking the sum of the square of all the forecasting errors for a particular h-step ahead forecast. Thereafter, we take the mean of these values before taking the square root of this value. Therefore, where we have a column vector of out-of-sample forecast errors for a particular h-step ahead forecast, we could calculate the RMSE as,

$$\text{RMSE}_h = \sqrt{\frac{1}{P} \sum_{t=R}^{T} \left(e^f_{t+h} \right)^2} \tag{5.2}$$

As such, the RMSE is a symmetric loss function, where forecasts that are either too high or too low are weighted equally. Naturally, smaller forecast errors are considered to be better than larger ones, and as such a low RMSE value indicates better forecasting performance.

Time	**Outcome**	**AR(1)**			**AR(2)**		
		y^f_{t+1}	e^f_{t+1}	$\{e^f_{t+1}\}^2$	y^f_{t+1}	e^f_{t+1}	$\{e^f_{t+1}\}^2$
t-2	2.00						
t-1	1.5						
t	2						
t+1	1.8	1.8	0	0	1.65	0.15	0.02
t+2	1.5	1.66	-0.16	0.03	1.58	-0.08	0.01
t+3	1.2	1.45	-0.25	0.06	1.38	-0.18	0.03

Time	Outcome		AR(1)			AR(2)	
t+4	1.4	1.24	0.16	0.03	1.17	0.23	0.05
t+5	1.6	1.38	0.22	0.05	1.26	0.34	0.12
	$\text{BIAS}_{h=1}$		-0.01			0.09	
	$\text{RMSE}_{h=1}$			0.18			0.21

Table 2: **Calculating the BIAS and RMSE for one-step ahead forecast**

Continuing with the two autoregressive examples introduced in Table 1, where the AR(1) model is given as $y_t = 0.4 + 0.7 y_{t-1} + \varepsilon_t$, and the AR(2) is given as $y_t = 0.3 + 0.6 y_{t-1} + 0.1 y_{t-2} + \varepsilon_t$, we have calculated the bias and the RMSE in Table 2. In this example, the forecasting horizon is set to $h = 1$, and we have assumed some given outcomes that are provided. Note that for the initial forecast, the values are equivalent to those in Table 1. However, since we now have realised outcomes for subsequent periods, we use these to generate the subsequent forecasts.

With only five forecast errors to evaluate, the estimates of both the bias and the RMSE are of little use and are merely included to show how easy it is to compute these results. In most real-world applied evaluations one would usually make use of a larger set of out-of-sample forecasts.

5.2.1 Decomposing the RMSE

It is possible to show that the RMSE has two sources of potential errors, which may be illustrated in the following example. Where a one-step ahead forecast error is generated by an AR(1) model, $y_t = \mu + \phi_1 y_{t-1} + \varepsilon_t$, we would obtain a value for the predicted value, y^f_{t+1}. This forecast would relate to the value of the process that would be generated by $\hat{y}_{t+1} = \hat{\mu} + \hat{\phi}_1 y_t + \varepsilon_{t+1}$. At period t, the values for $\hat{\mu}$, $\hat{\phi}_1$, and ε_{t+1} are unknown. Therefore, we could express the forecast error as,

$$e^f_{t+1} = y_{t+1} - y^f_{t+1} = \varepsilon_{t+1} + \left[(\mu - \hat{\mu}) + \left(\phi_1 - \hat{\phi}_1 \right) y_t \right] \tag{5.3}$$

Note that while we would expect ε_{t+1} to take on a value that is close to zero, after we have realised subsequent values of the process, this would not be the case on each and every occasion. This expression suggests that the forecast error may be due to an unexpected shock, which may be termed an irreducible error and changes in the parameter values of the model. Using (5.3) to compute the mean squared-error, we get:

$$\mathbb{E}_t \left[\left(y_{t+1} - y^f_{t+1} \right)^2 \right] = \sigma^2_{\varepsilon,t+1} + \mathsf{var} \left[(\mu - \hat{\mu}) + \left(\phi_1 - \hat{\phi}_1 \right) y_t \right] \tag{5.4}$$

This expression suggests that the mean squared-error is composed of two parts. The first pertains to the part that is a function of the error term and the other is a function of possible changes in the coefficient values. These two parts give rise to uncertainty relating to the shock or error and the parameter uncertainty, which is given by the variance of the parameter estimation discrepancies in

(5.4). If we assume that the parameters in the estimated autoregressive model are identical to their population counterparts, the last term in the square brackets of (5.4) would equal zero. In this case the mean squared error is then equal to σ_{ε}^{2}, which is equivalent to the error in the regression model. When working with macroeconomic models, these parameters would be termed *deep* or *structural* parameters.

5.2.2 Comparing out-of-sample and in-sample fit

When performing a forecasting evaluation we evaluate the *out-of-sample* fit of the model. In contrast we could also consider the *in-sample* fit of the model, which may be summarised by the coefficient of determination, information criteria, and various other measures.

In most cases, it is not always the case that a model that has a good in-sample fit will ensure that the out-of-sample performance of the model is suitable. The intuition behind this is provided by (5.4), where a good in-sample fit (as suggested by a high R^2) may be obtained by including a large number of regressors. When performing a out-of-sample evaluation, this would typically lower the estimate of σ_{ε}^{2}, but it would also usually increase the estimation uncertainty, which would increase the last part of (5.4).

Thus, when trying to evaluate the appropriateness of a statistical model, it is usually a good idea to assess both the in-sample and out-of-sample performance of the model.

5.3 Mean absolute error (MAE) and mean absolute percentage error (MAPE)

As an alternative to the quadratic loss function, which is used in the RMSE, one could make use of linear loss function, such as the mean absolute error. This statistic is calculated as:

$$\mathrm{MAE}_h = \frac{1}{P}\sum_{t=R}^{T}\left|e_{t+h}^{f}\right|$$

Note that the RMSE would impose larger penalties on extreme outliers (or very large forecast errors) as a result of the quadratic loss function. The mean absolute percentage error has been used in a number of practical applications as provides an intuitive result and may be expressed in percentage terms. It may be calculated as follows:

$$\mathrm{MAPE}_h = \frac{1}{P}\sum_{t=R}^{T}\left|\frac{y_{t+h} - y_{t+h}^{f}}{y_{t+h}}\right|$$

6 Comparing different forecasts

Most institutions make use of a suite of forecasting models that may be used to generate respective forecasts for a specific variable. To determine which of these is responsible for the most accurate forecast, we can compare the value of our preferred loss function for each of the models, and simply choose the model that is associated with the best score (i.e. choose the model with the lowest RMSE or bias). Or alternatively, we could use a particular loss function that could be used to evaluate the forecasts from competing models, after we have generated the sequence of forecast errors for each model. Consider for example the case where we have two sequences of forecast errors that are produced by a linear regression model, $\left\{ e^{f}_{t+h} \right\}^{T}_{t=R}$, and a random-walk model, $\left\{ e^{rw}_{t+h} \right\}^{T}_{t=R}$.

To identify the model that produces the most accurate estimate we could use a quadratic loss function, where we firstly take of the sum of the square of the individual forecasts errors from each model before calculating the difference:

$$\Delta \mathscr{L}_{t+h} = \sum_{t=R}^{T} \left(e^{f}_{t+h} \right)^2 - \sum_{t=R}^{T} \left(e^{rw}_{t+h} \right)^2$$

If the result is positive then it would suggest that the model that was used to generate e^{f}_{t+h} is inferior. However, if this difference is negligible then we may want to suggest that the models have equal predictive ability.

6.1 Test for equal forecast ability

To formally test whether a model has equal predictive ability we make use of a t-test, where the null hypothesis is that the expected difference in the forecast errors is not different from zero. Therefore,

$$\begin{aligned} H_0 \; &: \; \mathbb{E}_{T+H} \left[\sum_{t=R}^{T} \left(e^{f}_{t+h} \right)^2 - \sum_{t=R}^{T} \left(e^{rw}_{t+h} \right)^2 \right] = 0 \\ \therefore \; H_0 \; &: \; \mathbb{E}_{T+H} \left[\Delta \mathscr{L}_{t+h} \right] = 0 \end{aligned}$$

To perform this test we can take the sequence of values for $\Delta \mathscr{L}_{t+h}$ and regress it on a constant, c, in the following regression. In this case the $\hat{\beta}$ coefficient could be subjected to a t-test to see if it is significantly different from zero. This regression model would take the form:

$$\Delta \mathscr{L}_{t+h} = \hat{\beta} c + \epsilon_{t+h}$$

and the t-test would be expressed as:

$$t_{\beta} = \frac{\hat{\beta} - 0}{\mathsf{std} \left(\Delta \mathscr{L}_{t+h} \right)}$$

Note that in the above the case the estimated value of $\hat{\beta}$ would be equal to the average value of the difference in the loss function.

$$\hat{\beta} = \frac{1}{P} \Delta \mathscr{L}_{t+h} = \Delta \bar{\mathscr{L}}_{t+h}$$

This rather intuitive test was initially described in Diebold and Mariano (1995), while West (1996) describes the use of a similar test. This statistic employs the hypothesis:

$$H_0 : \hat{\beta} = 0 \quad \text{vs} \quad H_1 : \hat{\beta} \neq 0 \tag{6.1}$$

In this case the null hypothesis implies no significant difference in forecasting performance. Therefore, if $\hat{\beta} = 0$, it will be the case that $\mathbb{E}_t \left[\Delta \mathscr{L}_{t+h} \right] = \mathbb{E}_t \left[\epsilon_t \right] = 0$ under the standard ordinary least squares assumptions. As such, a rejection of the null would imply that the forecasting performance of the two models are significantly different from one another (at some given level of significance). Diebold and Mariano (1995) show that the distribution of the difference in the loss function converges to the distribution of the normal distribution when the number of forecasts that we have is sufficiently large (i.e. more than 100 observations). This implies that when $|t_{DM}| \leq 1.96$ we are unable to reject the null when working with a 95% confidence interval. In the usual manner, when this statistic is higher in absolute value than the critical value, then we can reject the null hypothesis.

From a practical perspective, this statistic could also be calculated as follows:

$$t_{DM} = \frac{\hat{\beta} - 0}{\mathsf{std}\left(\Delta \bar{\mathscr{L}}_{t+h} \right)} = \frac{\hat{\beta} - 0}{\sqrt{\mathsf{var}\left(\Delta \bar{\mathscr{L}}_{t+h} \right)}} = \frac{\hat{\beta} - 0}{\sqrt{\mathsf{var}\left(\frac{1}{P} \sum_{t=R}^{T} \Delta \mathscr{L}_{t+h} \right)}}$$

$$\therefore t_{DM} = \frac{\hat{\beta} - 0}{\sqrt{\mathsf{var}\left(\sum_{t=R}^{T} \Delta \mathscr{L}_{t+h} \right)}} \sqrt{P}$$

When performing this test it is worth noting that when $h > 1$, the values for $\Delta \mathscr{L}_{t+h}$ and the residuals ϵ_t would usually be serially correlated. Therefore, one would usually employ methods for HAC standard errors, which are discussed in Newey and West (1987). For a more detailed treatment of how to compare different forecasts, the interested reader should consult West (2006), which provides an informative summary.

An important limitation of the Diebold and Mariano (1995) statistic may be considered in light of the above example, where the regression model makes use of parameter estimates to generate predictions, while the random-walk model does not. Typically the variability of the loss that is based on a parameter estimate will be greater than the loss that is not based on a parameter estimate. The reason for this may be illustrated by the fact that the loss that is calculated from the linear regression is generated as:

$$\epsilon_{t+h} = y_{t+h} - \hat{\phi}^{\top} x_t$$

which would be more variable if there is a large degree of parameter uncertainty relating to the value of $\hat{\phi}$. Hence, such a model would usually over-estimate the degree of variability that is associated with the losses. The effect of the estimation error on the degree of variability in the forecast error is considered in West (1996), where he compares the forecasting errors for one model that includes estimation errors, which is denoted as M_1, and another that does not, which is denoted M_2.

$$M_1 : \mathscr{L}_T^{\hat{\phi}} = \frac{1}{P} \sum_{t=R}^{T} \left\{ y_{t+h} - \hat{\phi}^{\top} x_t \right\}^2$$

$$M_2 : \mathscr{L}_T^{\phi} = \frac{1}{P} \sum_{t=R}^{T} \left\{ y_{t+h} - \phi^{\top} x_t \right\}^2$$

Thereafter, to consider the distribution of $\hat{\phi}$, he makes use of a mean value expansion of $\hat{\phi}$ around ϕ to show that the parameter estimation decreases when R increases for fixed values of P. Such an expansion could take the form

$$\Delta \bar{\mathscr{L}}_T^{\hat{\phi}} \approx \Delta \bar{\mathscr{L}}_T^{\phi} + \frac{\partial \bar{\mathscr{L}}_T^{\phi}}{\partial \left(\phi\right)} \left[\left(\hat{\phi} - \phi\right) \sqrt{R} \right] \sqrt{\frac{P}{R}}$$

where $\frac{\partial \bar{\mathscr{L}}_T^{\phi}}{\partial (\phi)} \left[\left(\hat{\phi} - \phi\right) \sqrt{R} \right] \sqrt{\frac{P}{R}}$ is the component that is due to the estimation error. Therefore, in such cases the Diebold and Mariano (1995) test would need to be amended, such that:

$$t_W = \frac{\hat{\phi} - 0}{\sqrt{\mathsf{var} \left(\sum_{t=R}^{T} \Delta \mathscr{L}_{t+h} \right) + z_t}} \sqrt{P}$$

where z_t is the contribution of parameter estimation error that is due to the variance of the losses. This modification allowed West (1996) to show that if the estimation error decreases when the in-sample period increases, then the effect of parameter estimation decreases when R increases for fixed values of P.

6.2 Nested models

When models are nested then some of the coefficients from the larger model are not available in the restricted model. This would imply that the variance for the sequence of errors in the smaller model would be equal to zero and as such the Diebold and Mariano (1995) test should not be used. This is an important consideration, as the random walk model would in most cases be nested within other models.

Clark and McCracken (2001) suggest that the adjusted encompassing test should be used in such cases and can be expressed as follows:

$$ENCNEW = P \frac{\frac{1}{P} \sum_{t=R}^{T} \left(e_{1,t+h}^2 - e_{1,t+h} e_{2,t+h} \right)}{\frac{1}{P} \sum_{t=R}^{T} e_{2,t+h}^2}$$

where $e_{1,t+h}$ are the forecast errors of the small model and $e_{2,t+h}$ are the forecast errors of the large model. Note that as the denominator is not represented by the variance of $\Delta \mathscr{L}_T$, this statistic will not be subject to the same problems as the Diebold and Mariano (1995) statistic, as the variance of the forecast errors of the large model will not be zero. This statistic does not have a standard distribution and as such the critical values would need to be calculated from a Monte Carlo simulation. These critical values, which are to be used for linear models, are included in the paper of Clark and McCracken (2001).

To generate a test statistic that could be compared to a normal distribution, Clark and West (2007) compare the properties of nested models with the properties of a model that would be subject to a normal distribution. In doing so, they look to adjust the test statistic in a way that would allow it to be compared to a normal distribution. This allows them to compare the forecasts of a large model, where the forecasts are generated by $\mathbb{E}_t \left(y_{t+h}\right) = \hat{\phi}^\top x_t$ along with a small random-walk model that has forecasts $\mathbb{E}_t \left(y_{t+h}\right)$. In the case of both models we could calculate the mean-squared forecast errors (MSFE) as:

$$
\begin{aligned}
MSFE_{large} &= \frac{1}{P} \sum_{t=R}^{T} \left(y_{t+h} - \hat{\phi}^\top x_t\right)^2 \\
MSFE_{small} &= \frac{1}{P} \sum_{t=R}^{T} \left(y_{t+h}\right)^2
\end{aligned}
$$

The essential idea behind this statistic is that we are looking to adjust the mean-squared forecast error of the large model to correct for the effects of parameter uncertainty when calculating the test statistic. If one were to apply the Diebold and Mariano (1995) test to this problem we would proceed as follows:

$$
\begin{aligned}
\Delta \bar{\mathscr{L}}_T &= \frac{1}{P} \sum_{t=R}^{T} \left(y_{t+h} - \hat{\phi}^\top x_t\right)^2 - \frac{1}{P} \sum_{t=R}^{T} \left(y_{t+h}\right)^2 \\
&= \frac{1}{P} \sum_{t=R}^{T} \left(y_{t+h}\right)^2 + \frac{1}{P} \sum_{t=R}^{T} \left(\hat{\phi}^\top x_t\right)^2 - \frac{2}{P} \sum_{t=R}^{T} \left(y_{t+h}\right) \left(\hat{\phi}^\top x_t\right) - \frac{1}{P} \sum_{t=R}^{T} \left(y_{t+h}\right)^2 \\
&= \frac{1}{P} \sum_{t=R}^{T} \left(\hat{\phi}^\top x_t\right)^2 - \frac{2}{P} \sum_{t=R}^{T} \left(y_{t+h}\right) \left(\hat{\phi}^\top x_t\right)
\end{aligned}
$$

Now under the null hypothesis $y_{t+h} = \epsilon_{t+h}$ such that,

$$
\begin{aligned}
H_0 &: \frac{1}{P} \sum_{t=R}^{T} \left(\hat{\phi}^\top x_t\right)^2 - \frac{2}{P} \sum_{t=R}^{T} \left(\epsilon_{t+h}\right) \left(\hat{\phi}^\top x_t\right) \\
&: \frac{1}{P} \sum_{t=R}^{T} \left(\hat{\phi}^\top x_t\right)^2 - \frac{2}{P} \hat{\phi} \sum_{t=R}^{T} \left(\epsilon_{t+h}\right) \left(x_t\right)
\end{aligned}
$$

Note that if we assume that ϵ_{t+h} are independent, then $\frac{2}{P} \hat{\phi} \sum_{t=R}^{T} \epsilon_{t+h} x_t = 0$, and since the value for $\frac{1}{P} \sum_{t=R}^{T} \left(\hat{\phi}^\top x_t\right)^2$ will always be positive, the loss of the large model will always exceed the small model. This statistic is used in Clark and West (2007) to correct the Diebold and Mariano (1995) statistic, where the smaller model has an unfair advantage. Since the distribution for $\frac{1}{P} \sum_{t=R}^{T} \left(\hat{\phi}^\top x_t\right)^2$ would represent an asymptotic normal distribution of something that is also normally distributed, we are left with a test statistic that is asymptotically normal. Therefore, the Clark and West (2007) statistic applies a similar framework to what is used in the Diebold and Mariano (1995) test, but where the measure of $\Delta \bar{\mathscr{L}}_T$ is adjusted for the estimation error.

$$t_{CW} = \frac{\Delta \bar{\mathscr{L}}_T^{adj} - 0}{\sqrt{\mathsf{var}\left(\bar{\mathscr{L}}_T^{adj}\right)}} \quad \underset{H_0}{\longrightarrow} \quad \mathcal{N}(0,1)$$

The ultimate effect of this is that it would be easier for the large model to outperform the small model when using the Clark and West (2007) test (as opposed to using the Diebold and Mariano (1995) test on nested models). Note that this is a one-sided test, where we are only testing the null that the models have equal predictive ability, or that the large model provides more accurate estimates. Hence, this statistic does not test whether or not the small model provides more accurate forecast estimates. In addition, all of these tests focus on the null hypothesis, $H_0 : \phi = 0$, based on $\mathbb{E}\left(\Delta \mathscr{L}_{t+h}^{\phi}\right) = 0$, where ϕ is the true parameter estimate. To make use of a different null hypothesis, $H_0 : \mathbb{E}\left(\Delta \mathscr{L}_{t+h}^{\hat{\phi}}\right) = 0$, where $\hat{\phi}$ is the sample estimate of the population parameter, consult the work of Giacomini and White (2006).

7 Forecast uncertainty

7.1 Mean squared forecast error (MSFE)

The mean squared forecast error is a quadratic loss function that is widely used to evaluate the forecasting accuracy of a particular model. In addition, this statistic may also be used as a measure of the forecast error variance, which would usually need to be calculated when constructing forecast intervals. We may denote the MSFE for the h-step ahead forecast, as σ_{t+h}^{f}, which may be derived as follows

$$\begin{aligned} \mathbb{E}_t\left[\sigma_{t+h}^{f}\right] &= \mathbb{E}_t\left[\left(y_{t+h} - y_{t+h}^{f}\right)^2\right] \\ &= \mathbb{E}_t\left[\left(\sum_{i=0}^{h-1} \phi_1^i \varepsilon_{t+h-i}\right)\left(\sum_{i=0}^{h-1} \phi_1^i \varepsilon_{t+h-i}\right)\right] \end{aligned} \tag{7.1}$$

Note that as the ϕ_1 terms are know *a priori*, they are not governed by the expectations operator and can be moved to the front of the expression. In addition, since $\mathbb{E}_t\left[\varepsilon_{t+h-i}\varepsilon_{t+h-i}\right] = \sigma_{\varepsilon}^2$ for all h. Hence,

$$\sigma_{t+h}^{f} = \sum_{i=0}^{h-1} \phi_1^i \sigma_{\varepsilon}^2 \sum_{i=0}^{h-1} \phi_1^i,$$

which would allow us to derive values for $1, 2, 3, \ldots, h$, with the aid of the following expressions,

$$
\begin{aligned}
\sigma^f_{t+1} &= \sigma^2_\varepsilon \\
\sigma^f_{t+2} &= \sigma^2_\varepsilon + \phi_1^2\sigma^2_\varepsilon = \sigma^f_{t+1} + \phi_1^2\sigma^2_\varepsilon \\
\sigma^f_{t+3} &= \sigma^2_\varepsilon + \phi_1^2\sigma^2_\varepsilon + \phi_1^4\sigma^2_\varepsilon = \sigma^f_{t+2} + \phi_1^4\sigma^2_\varepsilon \\
&\vdots \\
\sigma^f_{t+h} &= \sigma^2_\varepsilon(1 + \phi_1^2 + \phi_1^4 + \ldots + \phi_1^{2(h-1)}) = \sigma^f_{t+h-1} + \phi_1^{h-1}\sigma^2_\varepsilon\phi_1^{h-1}
\end{aligned}
\tag{7.2}
$$

If we let $h \rightarrow \infty$, (7.2) implies that $\sigma^2_\varepsilon(1 + \phi_1^2 + \phi_1^4 + \ldots + \phi_1^\infty) = \frac{\sigma^2_\varepsilon}{1-\phi_1^2}$. Note that this sequence has the unconditional variance of the process that was provided previously. Thus,

$$
\sigma^f_{t+h} \rightarrow \frac{\sigma^2_\varepsilon}{1-\phi_1^2} \quad \text{where} \quad h \rightarrow \infty \tag{7.3}
$$

Note that the *expected* variance of the forecasts converges to the unconditional variance of the process, which should not be surprising as the *expected* forecast errors are derived from the *expected* values for the future errors of the process. Hence, if we assume that the errors in the process are distributed $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ then the *expected* forecast errors could also be described by a similar distribution. In which case, $\mathbb{E}_t\left[\sigma^f_{t+h}\right] = \sigma^2$.

7.2 Uncertainty

The previous section suggests that we can use the conditional expectation to derive predicted values of a variable that may be described by a stable autoregressive model that has Gaussian white noise errors. In this case, the forecast will on average be equal to the true value of the variable that we want to forecast, which would imply that the forecasts are unbiased. However, the forecasts will not necessarily be equal to the true value of the process at all periods of time. Therefore, the forecast errors will have a positive variance, which may be measured by the mean squared forecast error.

In many instances it is desirable to report on both the point forecast and some measure of uncertainty that relates to the forecast. For example, most central banks publish fan charts together with their inflation forecasts. These fan charts communicate the central banks view on possible paths for future inflation. In addition, a number of central banks also publish fan charts when referring to the relative success of their past forecasts.

An example of this type of communication is presented in Figure 4, which was included in the South African Reserve Bank Monetary Policy Review (June, 2014). Note that towards the end of the forecast horizon, uncertainty increases as the bands become wider over time. This graph allows the central bank to suggest that although they largely expect inflation to initially increase to a point that is beyond the target range, it is expected to decline after a year whereupon it will return to a point that is within the range. It also suggests that there is quite a large degree of uncertainty surrounding the expected future values of inflation and as such we could realise a relatively wide array of values for the rate of inflation in the future.

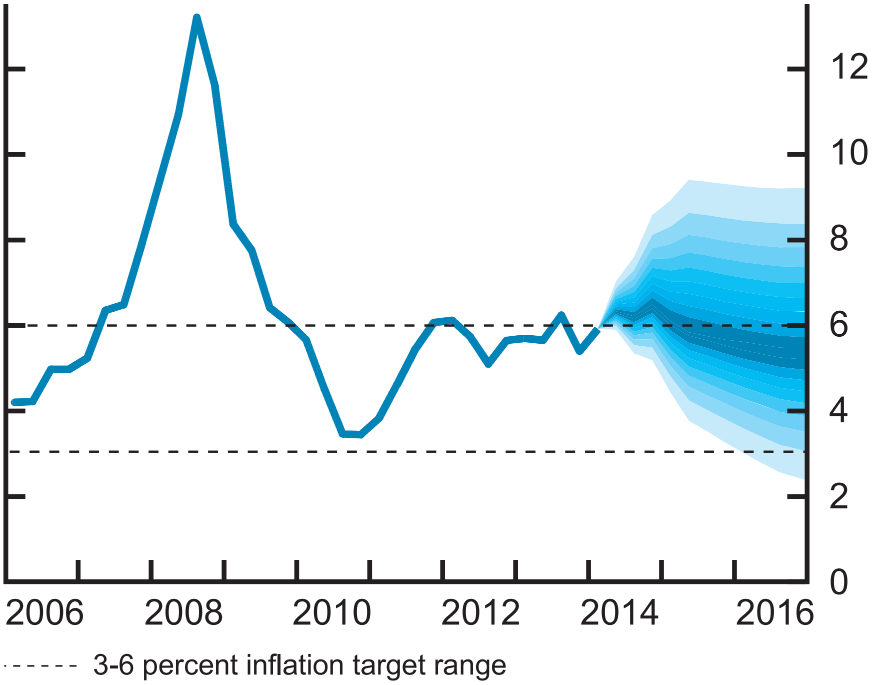


Figure 4: **South African inflation fan chart (SARB June 2014)**

There are a number of different methods that may be used to construct these fan charts. When all the coefficients in the model are used to provide point estimates, we could calculate the MSFE to generate distributions for the distribution of future forecast estimates.[1] Note that if we assume that the residuals in the model are, $\varepsilon_t \sim \mathsf{i.i.d.}\; \mathcal{N}(0, \sigma^2_\varepsilon)$, it implies that the forecast errors should also be normally distributed,

$$y_{t+h} - y^f_{t+h} \sim \; \mathcal{N}(0, \sigma^f_{t+h}) \tag{7.4}$$

Such that,

$$\frac{y_{t+h} - y^f_{t+h}}{\sqrt{\sigma^f_{t+h}}} \sim \; \mathcal{N}(0,1) \tag{7.5}$$

In this case we could make use of a normal distribution, with z_α defining the upper and lower bounds that may be used to derive the forecast interval for the h-period ahead forecast,

$$\left[y^f_{t+h} - z_{\alpha/2}\sqrt{\sigma^f_{t+h}} \quad , \quad y^f_{t+h} + z_{\alpha/2}\sqrt{\sigma^f_{t+h}} \right] \tag{7.6}$$

This allows for the construction of standard confidence intervals for the parameter estimates, where the predictor, y^f_{t+h}, and the MSFE, σ^f_{t+h}, are used to derive the appropriate interval.

7.3 Example: Forecasts with confidence intervals

Let us show how this works with a numerical example for the AR(1) model used in the previous section. That is, we assumed that we have the model,

- AR(1) with $\mu = 0.4$, $\phi_1 = 0.7$, and $y_t = 2$

Let $\varepsilon_t \sim \mathcal{N}(0, 0.1)$, and where we elect to work with a 95% confidence interval, $\alpha = 0.05$. This implies that $z_{\alpha/2} = 1.96$ in large samples. For forecast horizon $h = 1, 5, 10$ we can then use the point forecasts in Table 1 and (7.2) to derive the forecast error variance. Then lastly, with the use of (7.6) we can construct forecast intervals that are provided in Table 3.

	Point Estimate	MSFE	Lower Bound	Upper Bound
y^f_{t+1}	1.80	0.10	1.18	2.42
y^f_{t+5}	1.45	0.19	0.59	2.30
y^f_{t+10}	1.35	0.20	0.48	2.22

Table 3: **Forecast intervals for AR(1) model**

Note that the point forecasts are the same as those that are provided in the upper row in Table 3. The second column shows MSFE, which converges to the unconditional variance of the AR process as the forecast horizon increases. The confidence intervals suggest that if the errors continue to be distributed $\varepsilon_t \sim \mathcal{N}(0, 0.1)$, then there is a 95% probability that the intervals will contain the future value of the random variable, y_{t+h}.

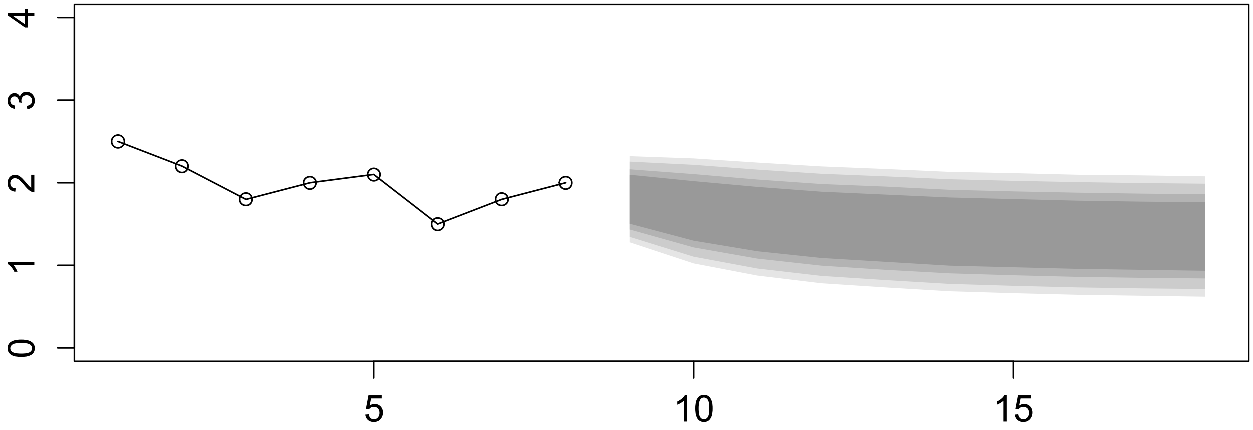


Figure 5: **AR(1) fan chart from simulation**

After putting together many forecast intervals for different significance levels (i.e. for different values of α in (7.6)), a fan chart can be constructed for the 30, 50, 70 and 90 percentiles of the forecast distribution, as in Figure 6. This would provide a visual display of the results, as per those of the central bank in Figure 5.

An alternative method of constructing a density forecast would be to simulate n number of forecasts from the normal distribution with mean y^f_{t+h} and variance σ^f_{t+h} across all horizons, h. From the vector of forecasts $(n \times 1)$ for each h, a forecast interval can be derived by sorting the numbers (e.g. from lowest to highest) and then choosing the percentile of interest from this sorted vector. Of course, if we are primarily interested in given forecast intervals, then the method that uses (7.6) would be more efficient. Moreover, when approximating the forecast distribution with n draws, the results would be dependent on the size of n. If n is not big enough, the procedure for simulating a number of forecasts from a normal distribution will not be equal to the direct approach, which made use of (7.6).

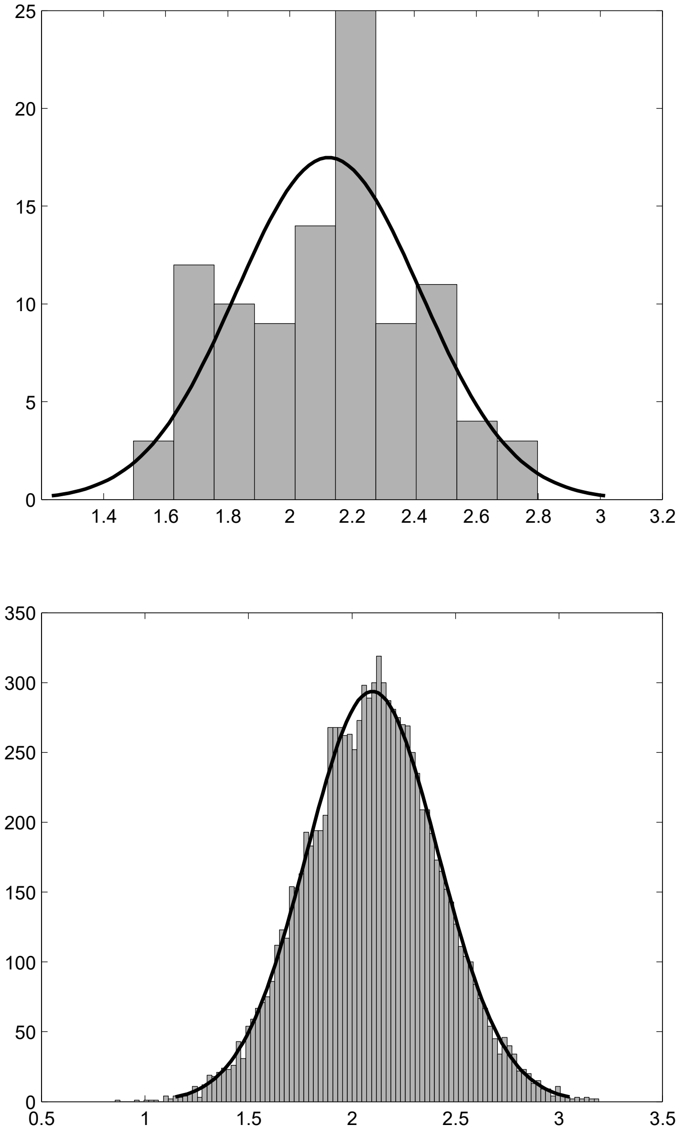


Figure 6: **AR(1) histograms from simulation**

Figure 6 shows two histograms of the forecast distribution for $h = 1$, given the values in Table 3, which are reported above. As we see, when n is small, simulating the forecast distribution provides a poor approximation to the normal distribution; however, when n is large, the simulation works well. If we were to construct a fan chart with the aid of this method we would need to repeat this simulation for each h-step.

8 Displaying the forecasting results

It is usually a good idea to plot each of the forecasts against the observed data to visualise the results of this forecasting experiment. Such an example is provided in Figure 7, where we have made use of an autoregressive model to generate the forecasts. Note that the forecasts converge on the mean and the time that is taken to converge on this mean value is dependent on the values of the coefficients.[2]

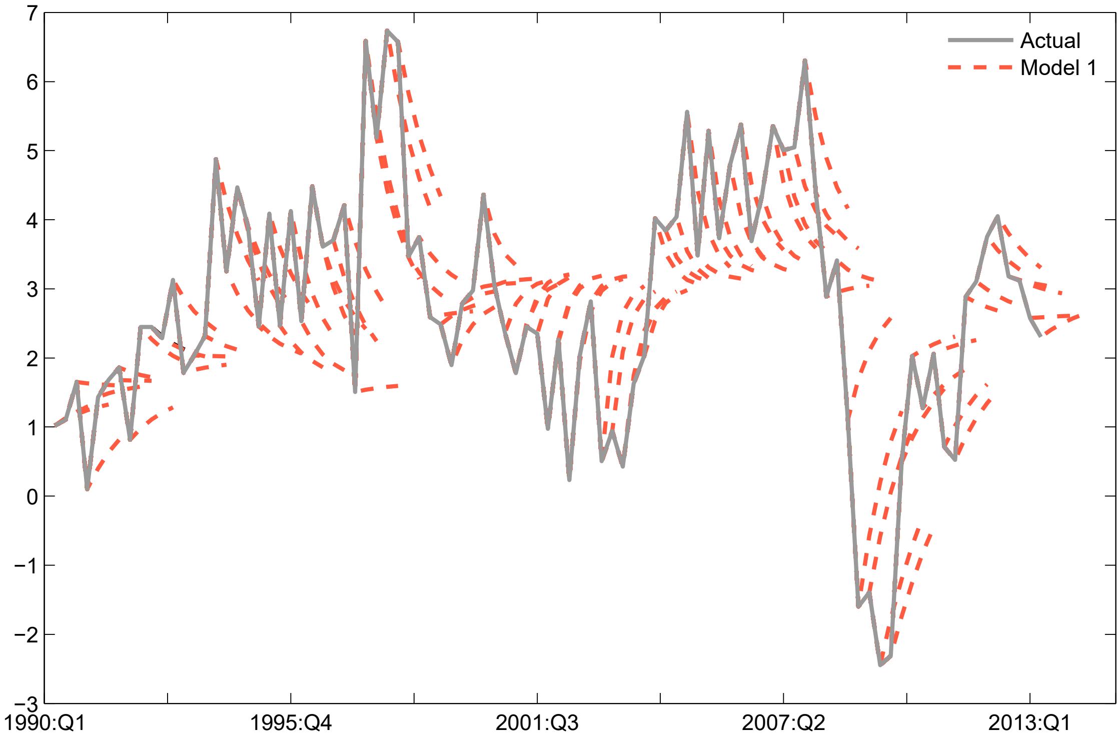
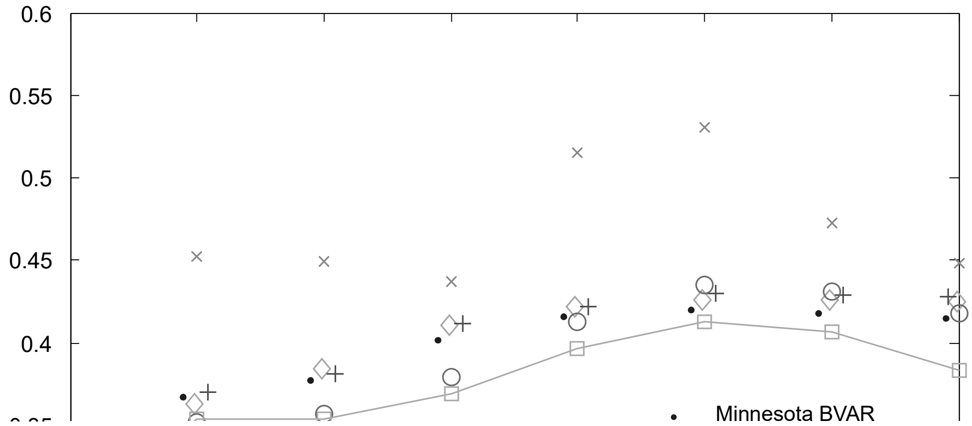


Figure 7: **Hairy plot of realised and forecasted values**

It is also usually a good idea to plot the results of the different RMSE errors for a range of different models. An example of such a plot is displayed in Figure 8, which are taken from Balcilar, Gupta, and Kotzé (2015), where the different horizons are placed on the horizontal axis and the value of the RMSE is displayed on the vertical axis.



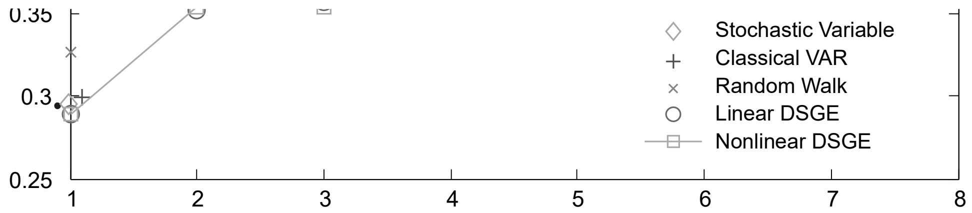


Figure 8: **Comparing RMSE for one- to eight-step ahead forecasts**

Although we have focused our attention on the evaluation of models for different forecasting horizons, as is common in the literature, we could also consider how the RMSE statistics may vary over time for a combination of the different h-step ahead horizons. For example, rather than calculate the RMSE for each column of the forecasting error matrix (2.1), we could also calculate the RMSE for each row of this matrix. This could be used to show how the forecasting error may have changed over time, where such a figure, which is also taken from Balcilar, Gupta, and Kotzé (2015), is displayed in Figure 9. Note that in this case, time is displayed on the horizontal axis and the grey lines display the forecasts for the competing models.

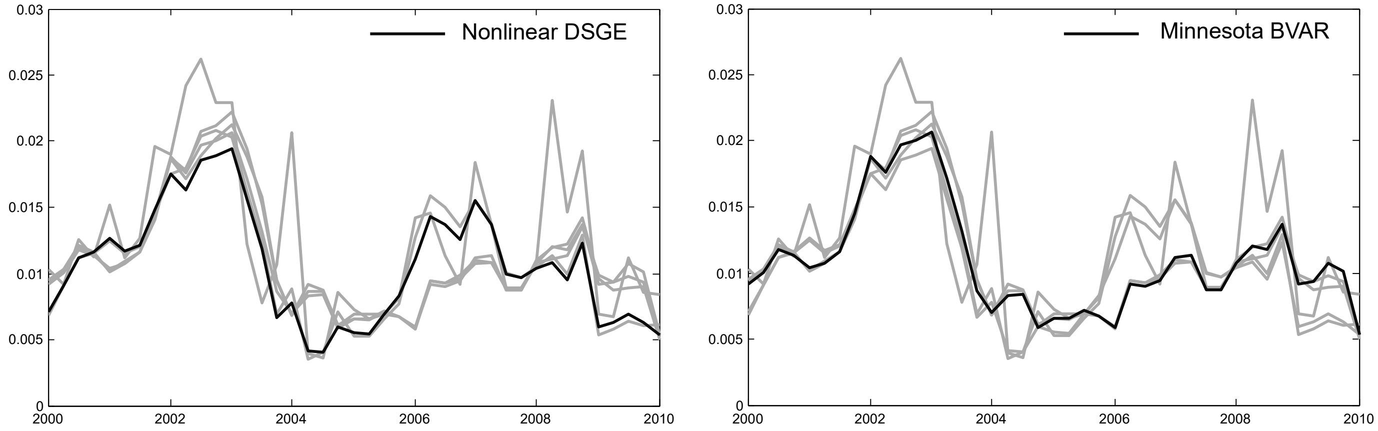


Figure 9: **Average RMSE over time for different models**

9 Model combination

In the previous section we described how forecasts for the same variable could be compared, when these forecasts are generated by different models. Implicitly, we wanted to find the model that would provide the superior forecast, when comparing two alternative model specifications. An alternative approach would be to combine these forecasts from the set of all of the models under consideration. Model combination has a long history in economics, dating back to Bates and Granger (1969), and possibly even further.

The rationale for combining models as opposed to choosing one model among many candidate models is easily motivated from a decision-making perspective. As noted in Timmermann (2006), unless one can identify a particular forecasting model that generates smaller forecasting errors prior to observing any future data, forecast combinations offer potential diversification gains that may reduce the possibility of making large forecasting errors at particular points in time.[3] This would certainly be the case where there are large outliers in the mis-specification errors of individual models, which may be reduced with the aid of a combined forecasting strategy.

Many different strategies to combine model forecasts exist, both for point and density forecasts. Covering all of these is outside the scope of this course, so we will only focus on two simple strategies for combining point forecasts using a linear aggregation function for the application of equal weighting and MSFE weighting. To ensure that this discussion is relatively intuitive, and without loss of generality, we will assume that we only combine the forecasts of two models. In a more realistic setting, the number of models will typically be much larger.

9.1 Linear opinion pool and weights

In terms of notation, let y^{cf}_{t+h} denote the combined forecast for h-steps ahead. The respective forecasts for each of the two models are denoted $y^{f}_{j,t+h}$, for $j = \{1, 2\}$. The most simple way of aggregating $y^{f}_{j,t+h}$ into one combined forecast y^{cf}_{t+h} is to use a linear function,

$$\begin{aligned} y^{cf}_{t+h} &= w_{h,1} y^{f}_{1,t+h} + w_{h,2} y^{f}_{2,t+h} \\ &= w_h y^{f}_{1,t+h} + \left(1 - w_h\right) y^{f}_{2,t+h} \end{aligned} \tag{9.1}$$

where $w_{h,j}$ is the weight attached to model j. Combining individual forecasts into one combined forecast in this manner is often called a linear opinion pool. Typically we normalise the weights so that they sum to unity, as reflected in the second line of the expression in (9.1).

Of course, it would be simple to compute the result in (9.1), if we knew the weight to attach to each model. Two of the simplest weighting schemes include equal weighting and weighting based on the MSFE. Equal weights are particularly simple to compute, where

$$\text{Equal weights:} \qquad w_{h,j} = \frac{1}{2} \quad \text{for} \;\; j = 1, 2 \tag{9.2}$$

Similarly, one could apply the inverse of the estimates from the MSFE to penalise those models that are associated with greater uncertainty. Thus:

$$\text{MSFE weights:} \qquad \left(1 - w_{h,j}\right) = \frac{\text{MSFE}_{h,j}}{\sum_{j=1}^{2} \text{MSFE}_{h,j}} \quad \text{for} \quad j = 1, 2 \tag{9.3}$$

where the terms for $\text{MSFE}_{h,j}$ can be derived according the specification that has been provided earlier. Both of these weighting schemes are frequently employed in the forecasting literature as they can be computed without any difficulty. In addition, they also have some desirable theoretical properties that are similar to those that may be derived from diversification. For a further discussion of this topic see, Timmermann (2006), Clemen (1989), and Stock and Watson (2004). For more recent work in this area, see Diebold and Shin (2019).

10 Alternative forecasting strategies

Throughout our discussion on forecasting we have focused on the use of autoregressive models as they provide an intuitive appeal when explaining the procedures that are involved in the generation and evaluation of various forecasts. In most cases, the discussion would also apply to other model specifications. However, in certain instances (and for some data) a slightly different approach may be more suitable.

10.1 Direct forecasting

Direct forecasting models are different to iterative forecasting models in that they are specified to forecast at a particular horizon. A simple example of a direct forecasting model could be specified as follows,

$$y_t = \mu + \phi_1 x_{t-h} + \varepsilon_t \tag{10.1}$$

If $h = 1$ and $x = y$, this will just be an AR(1) model as described earlier. However, if $h > 1$ or $x \neq y$, describes a direct forecasting model. For example, if we assume that $h = 4$, then the equation for the model that is estimated would be $y_t = \mu + \phi_1 x_{t-4} + \varepsilon_t$, and a direct 4-period forecast would be

$$y_{t+4} = \mu + \phi_1 x_t + \varepsilon_t \tag{10.2}$$

An example of this forecasting strategy may be found in Stock and Watson (2004). In addition, Marcellino (2006) has noted that in theory, the use of iterated forecasts are more efficient when they are correctly specified; however, in the presence of model mis-specification direct forecasts may be more robust.

10.2 Autoregressive distributed lag model

Yet another time series model that is often applied in empirical work is the autoregressive distributed lag (ADL) model. The ADL model differs to the simple autoregressive models in that they include other variables (i.e. other than the lag of the dependent variable). In a general form, the ADL model can be written as,

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{k=1}^{K} \sum_{j=1}^{J_k} \beta_{j,k} x_{t-j,k} + \varepsilon_t \tag{10.3}$$

where the second term on the right-hand side of is the usual autoregressive term, while the third term summarizes the distributed lag component of the ADL model, with respect to x_t. Therefore, in the ADL model we allow for $k = \{1, \ldots, K\}$ additional regressors, which have J respective lags. In this way, we allow for $x_{t-j,k}$ to be included in the model. Each of the x_k regressors is then allowed to take on a different number of lags. For example,if $K = 2$, where $J_{k=1} = 2$ and $J_{k=2} = 1$ we have

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \beta_{1,1} x_{t-1,1} + \beta_{2,1} x_{t-2,1} + \beta_{1,2} x_{t-1,2} + \varepsilon_t$$

The rationale for including additional regressors is that there might be additional information in these variables that is not captured by lags of the dependent variable. However, the inclusion of additional regressors may result in a number of complications when employing an ADL model for forecasting more than one period ahead. For example, when $h = 3$ and $p = 1$, we have the model

$$y_{t+3} = \mu + \phi_1 y_{t+2} + \beta_{1,1} x_{t+2,1} + \beta_{2,1} x_{t+1,1} + \beta_{1,2} x_{t+2,2} + \varepsilon_{t+3}$$

In this case, the y_{t+2} term can be derived in the usual manner, by plugging in the value for y in the previous period. However, the x terms are not known, as at time t we do not necessarily know what the values of $x_{t+2,1}$, $x_{t+1,1}$ and $x_{t+2,2}$ will be. As such, these values would usually need to be predicted outside of the model. To circumvent this problem, multivariate models are usually employed in such forecasting exercises.

11 Conclusion

Many aspects of life rely on predictions that may be formulated with the of a forecasting model, where it would be appropriate to evaluate the predictions of such a model with the aid of an out-of-sample forecasting exercise. It is relatively straightforward to employ this methodology and it allows us to evaluate the empirical performance of different forecasting models. The bias, RMSE and MAPE are commonly used as evaluation criteria for such forecasts. To determine whether the forecasts from different models are significantly different from one another, we could employ the Diebold and Mariano (1995) test. However, where such models are nested, then we should rather make use of the tests that have been developed by Clark and West (2007). We may also choose to combined forecasts from many individual models to provide a result that may benefit from diversification gains. In addition, there are also a number of other strategies that could be employed to generate more appropriate forecasts.

12 Appendix: Density forecasting and evaluation

In the most recent forecasting literature, there has been increased attention towards forecasting the whole distribution of future outcomes.[4] If we assume that we know the distribution of the forecast errors, as we did in the previous section, density forecasts can easily be generated as described in that section. If we do not want to assume that we know the forecast distribution, other simulation based methods have to be employed to construct the density forecasts. We will not cover these methods in this course, however, under the maintained normality assumptions, evaluation of density forecasts is not overly complicated.

12.1 Evaluating density forecast

Whilst it is relatively simple to generate density forecast, the evaluation of these densities is more involved. This is partly due to the fact that the true density is not observed, not even *ex-post*. In what follows we refer to two of the most widely used evaluation methods that are applied to this problem, which include the probability integral transform (PIT) and the log-score.[5]

For a particular forecast density, the log score is simply the logarithm of the forecast density evaluated at the outcome of the random variable we forecast. The theoretical justification for using the log-score is somewhat complicated, but the implementation is rather easy, especially if we construct density forecasts according to the methods that was described above. That is, the estimates of y^f_{t+h} and σ^f_{t+h} can be used to compare the forecast density to a normal distribution around the realised outcome. This is easily done with the aid of most software packages.

Therefore, if we were to conduct an out-of-sample forecasting experiment for all forecasts that cover the evaluation period we could then compute the average log-score from the densities. In the density forecasting literature this score is seen upon as an intuitive measure of density fit, or as the density counterpart to the RMSE. However, in contrast to the RMSE, a higher average log score is considered better than a lower value, simply because it reflects a higher likelihood.

The PIT is used to test whether the predictive densities are correctly specified. For example, the 95% forecast interval is an interval that should contain the future value of the series in 95% of repeated applications. Likewise, a 30% forecast interval should contain the future value of the series in 30% of the repeated applications. Considering the whole forecast distribution, if the density forecast is correctly specified we would expect to see realised outcomes in all parts of this distribution, but with more outcomes centred around the peak of the distribution than in the tails. Therefore, if we run a forecasting experiment to construct a density, as described above, this could provide the essence of a repeated application. We could then evaluate whether actual outcomes fall within the forecast densities, which is what the PIT measures.

More formally, it can be shown that data that is modelled as a random variable, from any given distribution can be converted to a random variable that has a uniform distribution using the PIT. Under the maintained normality assumption, it can then be shown that a PIT can be computed by evaluating the normal cumulative probability function for y^f_{t+h} and σ^f_{t+h} at the realised outcome. If the density forecasts are correctly specified, and we do this for all forecasts covering the evaluation sample, P, then the PITs should be more or less uniformly distributed (i.e. a straight line in a histogram with percentiles on the x-axis). Once again, this is easily done in most statistical computer software packages. Moreover, different test-statistics exist and can be applied to determine whether the PIT are indeed uniformly distributed.

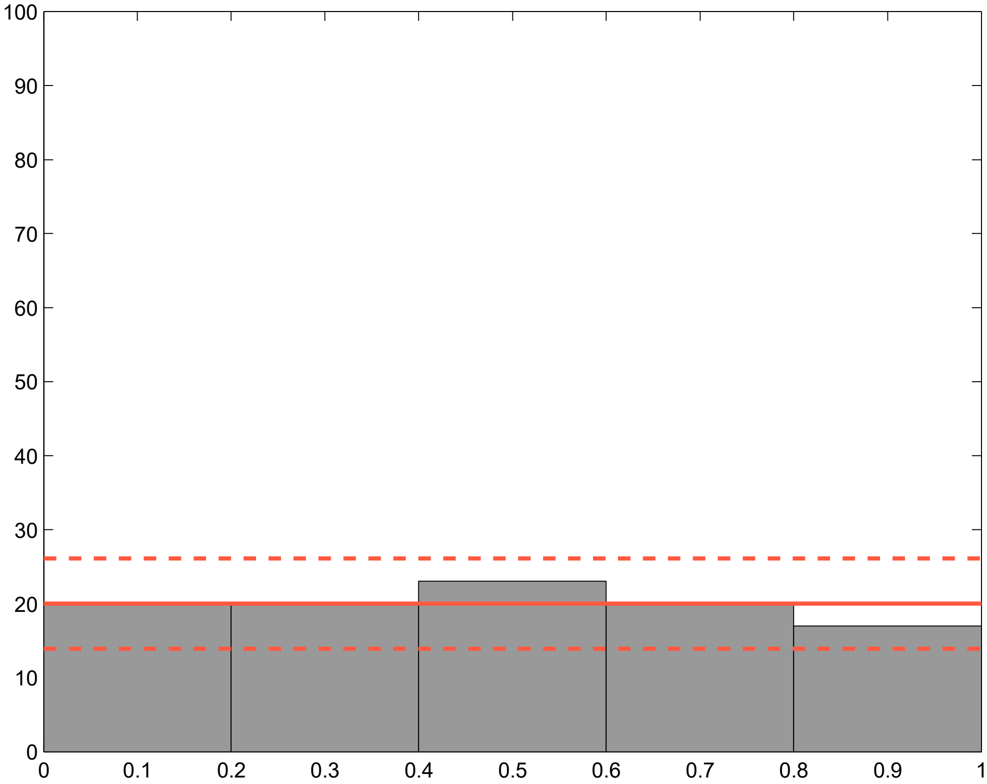


Figure 10: **PIT for $h = 1$**

By way of example, we could compare the probability density of a particular h-step ahead forecast to the distribution of the data from the in-sample period. We would usually start with the evaluation of $h = 1$ as it gets slightly more complicated when $h > 1$, as a result of the potential serial correlation in the forecast errors. To complete this exercise we usually make use histogram to depict the empirical distributions of the PITs. In this example, the solid line represents the number of draws that are expected to be in each bin under a $U(0, 1)$ distribution. The dashed lines represent the 95% confidence interval constructed under the normal approximation of a binomial distribution.

The results from such an exercise are depicted in Figure 10, which would appear to be relatively favourable as there is no part of the transformed $U(0, 1)$ distribution that is beyond the confidence intervals.

13 References

Balcilar, Mehmet, Rangan Gupta, and Kevin Kotzé. 2015. "Forecasting Macroeconomic Data for an Emerging Market with a Nonlinear DSGE Model." *Economic Modelling* 44: 215–28.

Bates, J., and Clive W. J. Granger. 1969. "The Combination of Forecasts." *Operations Research Quarterly* 20(4): 451–68.

Clark, Todd E., and Kenneth D. West. 2007. "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models." *Journal of Econometrics* 138: 291–311.

Clark, Todd, and Michael McCracken. 2001. "Tests of Equal Forecast Accuracy and Encompassing for Nested Models." *Journal of Econometrics* 105 (1): 85–110.

Clemen, Robert T. 1989. "Combining Forecasts: A Review and Annotated Bibliography." *International Journal of Forecasting* 5 (4): 559–83.

Corradi, V., and N. R. Swanson. 2006. "Handbook of Economic Forecasting." In, edited by G. Elliott, C. W. J. Granger, and A. Timmermann, 1:197–284. Elsevier.

Croushore, Dean, and Tom Stark. 2001. "A Real-Time Data Set for Macroeconomists." *Journal of Econometrics* 105 (3): 605–17.

Diebold, Francis X., and Roberto S. Mariano. 1995. "Predictive Accuracy." *Journal of Business and Economic Statistics* 13 (3): 253–63.

Diebold, Francis X., and Minchul Shin. 2019. "Machine Learning for Regularized Survey Forecast Combination: Partially-Egalitarian LASSO and Its Derivatives." *International Journal of Forecasting* 35 (4): 1679–91.

Giacomini, Raffaella, and Halbert White. 2006. "Tests of Conditional Predictive Ability." *Econometrica* 74 (6): 1545–78.

Hall, Stephen G., and James Mitchell. 2007. "Combining Density Forecasts." *International Journal of Forecasting* 23 (1): 1–13.

Jore, Anne Sofie, James Mitchell, and Shaun Vahey. 2008. “Combining Forecast Densities from Vars with Uncertain Instabilities.” Reserve Bank of New Zealand Discussion Paper Series DP2008/18. Reserve Bank of New Zealand.

Kascha, Christian, and Francesco Ravazzolo. 2008. “Combining Inflation Density Forecasts.” Working Paper 2008/22. Norges Bank.

Marcellino, M. 2006. “Handbook of Economic Forecasting.” In, edited by G. Elliott, C. W. J. Granger, and A. Timmermann, 1:879–960. Elsevier.

Meese, Richard A., and Kenneth Rogoff. 1983. “Empirical Exchange Rate Models of the Seventies : Do They Fit Out of Sample?” *Journal of International Economics* 14 (1-2): 3–24.

Newey, W., and K. West. 1987. “A Simple Positive Definite, Heterscedastic and Autocorrelation Consistent Covariance Matrix.” *Econometrica* 55: 703–5.

Stock, J. H., and M. W. Watson. 2004. “Combining Forecasts of Output Growth in Seven-Country Data Set.” *Journal of Forecasting* 23: 405–30.

Timmermann, Allan. 2006. “Handbook of Economic Forecasting.” In, edited by G. Elliott, C. W. J. Granger, and A. Timmermann, 1:136–96. Elsevier.

Wallis, K. F. 2005. “Combining Density and Interval Forecasts: A Modest Proposal.” *Oxford Bulletin of Economics and Statistics* 67(1): 983–94.

West, Mike. 1996. “Asymtotic Inference and Predictive Ability.” *Econometrica* 64(5): 1067–84.

———. 2006. “Handbook of Economic Forecasting.” In, edited by G. Elliott, C. W. J. Granger, and A. Timmermann, 1:99–134. Elsevier.

1. For those models that generate random variables for the parameter estimates (i.e. where Bayesian parameter estimation methods are used) we could use the coefficient estimates to generate distributions for the forecast estimates.↩

2. This is generally the case with autoregressive models.↩

3. The individual forecasts need not come from a formal statistical model and may include a subjective opinions about the future values of variables.↩

4. See, Wallis (2005), Hall and Mitchell (2007), Kascha and Ravazzolo (2008), and Jore, Mitchell, and Vahey (2008), among others.↩

5. A more detailed examination of different density evaluation procedures is given by Corradi and Swanson (2006).↩