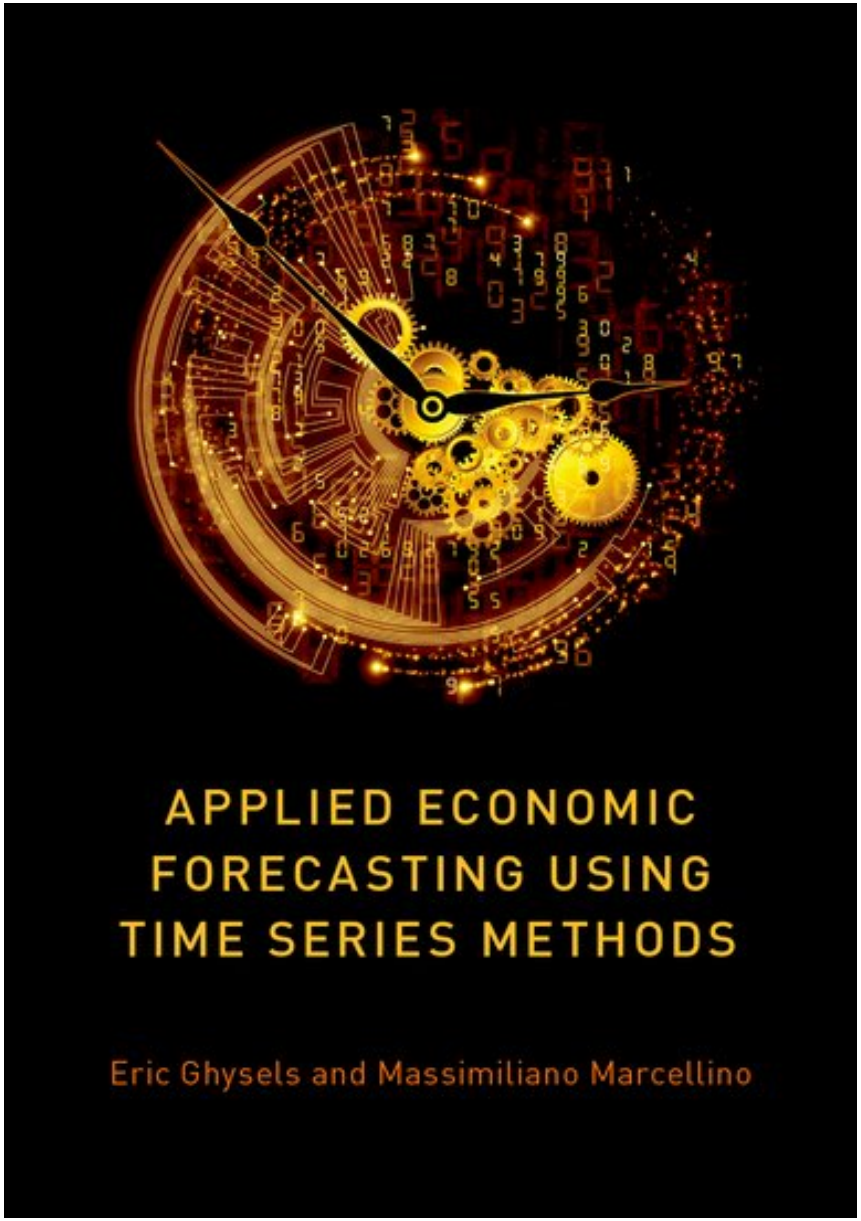
APPLIED ECONOMIC
FORECASTING USING
TIME SERIES METHODS
Eric Ghysels and Massimiliano Marcellino

Applied Economic Forecasting using Time Series Methods

Eric Ghysels and Massimiliano Marcellino

Companion Slides - Chapter 4 Forecast Evaluation and Combination

Forecast Evaluation and Combination

Introduction

Given different competing forecasts, in this chapter we try to answer the following questions:

(i) How "good," in some sense, is a particular set of forecasts?

(ii) Is one set of forecasts better than another one?

(iii) Is it possible to get a better forecast as a combination of various forecasts?

Introduction

To address these questions:

(i) we define some key properties a good forecast should have and discuss how to test them.

(ii) we introduce some basic statistics to assess whether one forecast is equivalent or better than another with respect to a given criterion.

(iii) we discuss how to combine the forecasts and why the resulting pooled forecast can be expected to perform well.

Forecast Evaluation and Combination

Unbiasedness and efficiency

- The optimal forecast under a MSFE loss function is the conditional expectation of the variable, so that it should be $E_t(y_{t+h}) = \hat{y}_{t+h|t}$, where $E_t(\cdot)$ is the conditional expectation given information at time t.
- **Unbiasedness** means that the expected value of the forecast error should be equal to zero, implying that on average the forecast should be correct.

Unbiasedness and efficiency

- **Efficiency** is related to the efficient use of the available information.

  So that the optimal h-steps ahead forecast error should be at most correlated of order $h-1$, and uncorrelated with available information at the time the forecast is made.

Unbiasedness and efficiency

Inefficient forecasts can still be unbiased, and biased forecasts can be efficient.

- If y_t is a random walk, namely,

$$y_t = y_{t-1} + \varepsilon_t,$$

and we use as a forecast

$$\widetilde{y}_{t+h|t} = y_{t-g},$$

then $\widetilde{y}_{t+h|t}$ is unbiased but not efficient.

- If y_t is a random walk with drift, namely,

$$y_t = a + y_{t-1} + \varepsilon_t,$$

and we use as a forecast

$$\widetilde{y}_{t+h|t} = y_t,$$

then $\widetilde{y}_{t+h|t}$ is biased but efficient.

Unbiasedness and efficiency

In order to test whether a forecast is **unbiased**, let us consider the regression

$$y_{i+h} = \alpha + \beta \hat{y}_{i+h|i} + \varepsilon_{i+h}, \quad i = T, ..., T + H - h, \qquad h < H \tag{1a}$$

where h is the forecast horizon and $T + 1, ..., T + H$ the evaluation sample. The forecasts $\hat{y}_{i+h|i}$ are recursively updated in periods $i = T, \ldots, T + H - h$.

- Sufficient condition:

$$\alpha = 0, \ \beta = 1 \tag{1b}$$

- Necessary condition:

$$\alpha = (1 - \beta) E(\hat{y}_{i+h|i}). \tag{1c}$$

Unbiasedness and efficiency

- The sufficient condition $(\alpha = 0, \beta = 1)$ can be tested by a "robust" $F-$test, where the fact that ε_{i+h} is in general autocorrelated (at least) of order $h-1$ is taken into account in the derivation of the HAC variance.

- The necessary condition (formula 1c) is equivalent to $\tau = 0$ in the regression:

$$e_{i+h} = y_{i+h} - \hat{y}_{i+h|i} = \tau + \varepsilon_{i+h},$$

which can be tested with a robust version of the $t-$test.

Unbiasedness and efficiency

- Note that $(\alpha = 0, \beta = 1)$ also implies that the forecast and forecast errors are uncorrelated. (formula 1a) can be rewritten as

$$e_{i+h} = \alpha + (\beta - 1)\hat{y}_{i+h|i} + \varepsilon_{i+h},$$

so that

$$E(\hat{y}_{i+h|i}e_{i+h}) = \alpha E(\hat{y}_{i+h|i}) + (\beta - 1)E(\hat{y}_{i+h|i}) + E(\underbrace{\hat{y}_{i+h|i}\varepsilon_{i+h}}_{=0}) = 0.$$

- Moreover, when $(\alpha = 0, \beta = 1)$, we have that

$$var(y_{i+h}) = var(\hat{y}_{i+h|i}) + var(e_{i+h}),$$

implying that the volatility of the variable should be larger than that of the (optimal) forecast, the more so the larger the variance of the forecast error.

Unbiasedness and efficiency

- The coefficient of determination (R^2) from formula 1a can also be used as an indicator of the forecast quality, with good forecasts associated with high R^2.

  However, one should also consider that persistent variables are easier to forecast than volatile variables, given that their past is a useful leading indicator.

Unbiasedness and efficiency

- To test for **weak efficiency**: e_{i+h} is correlated, across time, at most of order h - 1, so that no lagged information beyond $h-1$ can explain the forecast errors.
- We can fit a moving average model of order $h-1$, $MA(h-1)$, to the h-steps ahead forecast error and testing that the resulting residuals are white noise.

Unbiasedness and efficiency

- To test for **strong efficiency**: No indicators available when the forecasts were formulated can improve $h-$step forecast and therefore explain the $h-$step ahead forecast error.

- We can test if $\gamma = 0$ in the regression

$$e_{i+h} = \gamma' z_i + \varepsilon_{i+h},$$

where z_i is a vector of potentially relevant variables for explaining the forecast errors. e_{i+h} are realizations of a set of h-steps ahead forecast errors, namely, $e_{i+h} = y_{i+h} - \hat{y}_{i+h|i}$.

Forecast Evaluation and Combination

Evaluation of fixed event forecasts

- So far we have considered forecasts for period $T + h$ made in period T where T progressively increases, namely, $\{\hat{y}_{i+h|i}\}$ for $i = T, \ldots, T + H - h$.

  As an alternative, we can consider $\{\hat{y}_{\tau|\tau-h}\}$, $h = 1, 2, \ldots$ i.e., forecasts for a fixed target value (y_τ) made at different time periods that become closer and closer to τ. The $\{\hat{y}_{\tau|\tau-h}\}$ are known as **fixed event forecasts**. For example, for an AR(1) process, it is

$$\hat{y}_{\tau|\tau-h} = \rho^h y_{\tau-h}.$$

Evaluation of fixed event forecasts

- The properties of fixed events forecasts were studied by, e.g., Clements(1997). Let us decompose the forecast error as:

$$e_{\tau|\tau-h} = y_\tau - \hat{y}_{\tau|\tau-h} = v_{\tau|\tau-h+1} + v_{\tau|\tau-h+2} + ... + v_{\tau|\tau}, \tag{2}$$

where

$$v_{\tau|J} = \hat{y}_{\tau|J} - \hat{y}_{\tau|J-1}, \quad \hat{y}_{\tau|\tau} = y_\tau, \quad J = \tau - h + 1, \ldots, \tau. \tag{3}$$

For the AR(1) example, it is

$$v_{\tau|J} = \rho^{\tau-J}\varepsilon_{J,}$$

and

$$e_{\tau|\tau-h} = \rho^{h-1}\varepsilon_{\tau-h+1} + ... + \varepsilon_\tau.$$

- **Unbiasedness** requires that

$$E(e_{\tau|\tau-h}) = 0, \quad \forall \quad \tau - h.$$

Evaluation of fixed event forecasts

- For **weak efficiency**, the following should hold:

$$E(e_{\tau|\tau-h}|v_{\tau|\tau-h},...,v_{\tau|1}) = 0, \quad \forall \quad \tau - h,$$

This condition is equivalent to

$$E(v_{\tau|\tau-h}|v_{\tau|\tau-h-1},...,v_{\tau|1}) = 0, \quad \forall \quad \tau - h$$

These conditions also imply that

$$\hat{y}_{\tau|J} - \tilde{y}_{\tau|J-1} = \rho^{(\tau-J)}\varepsilon_J, \tag{4}$$

- This property can be assessed by testing if forecast revisions are white noise.

Evaluation of fixed event forecasts

- **Strong efficiency** can be defined as the lack of explanatory power for $v_{\tau|J}$ of variables z included in the information set for period J.
- This property can be assessed by regressing $v_{\tau|J}$ on z_J and testing for the non-significance of z_J.

Forecast Evaluation and Combination

Tests of predictive accuracy

- Tests of predictive accuracy compare an estimate of the forecast error variance obtained from the past residuals with the actual MSFE of the forecasts.
- Hence, they provide a measure of how well the model performs in the future relative to the past.
- We will focus on **Wald-type tests**.

Tests of predictive accuracy

- if y_t admits the Wold MA(∞) representation:

$$y_t = \psi(L)\varepsilon_t,$$

then the $h-step$ ahead minimum MSFE predictor is

$$\hat{y}_{T+h|i} = \sum_{J=h}^{\infty} \psi_J \varepsilon_{T+h-J},$$

with associated forecast error

$$e_{T+h} = \sum_{J=0}^{h-1} \psi_J \varepsilon_{T+h-J},$$

where $\psi_0 = 1$.

Tests of predictive accuracy

- We can group the errors in forecasting $(y_{T+1}, ..., y_{T+h})$ conditional on period T in

$$e_h = \psi \varepsilon_h, \tag{5}$$

where $e_h = (e_{T+1}, ..., e_{T+h})'$, $\varepsilon_h = (\varepsilon_{T+1}, ..., \varepsilon_{T+h})'$ and

$$\psi = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 & 0 \\ \psi_1 & 1 & \dots & \dots & 0 & 0 \\ \psi_2 & \psi_1 & \dots & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & & \vdots & \vdots \\ \vdots & \vdots & & \ddots & 1 & 0 \\ \psi_{h-1} & \psi_{h-2} & \dots & \dots & \psi_1 & 1 \end{pmatrix}.$$

Tests of predictive accuracy

- If we define

$$\Phi_h = E(e_h e_h') = \psi E(\varepsilon_h \varepsilon_h')\psi' = \sigma_\varepsilon^2 \psi\psi',$$

and if the appropriate model over $[1, ..., T]$ remains valid over the forecast horizon and $\varepsilon \sim N$, then

$$Q = e_h' \Phi_h^{-1} e_h \sim \chi^2(h),$$

where $\Phi_h^{-1} = \sigma_\varepsilon^{-2}(\psi^{-1})'\psi^{-1}$.

Tests of predictive accuracy

- Writing the autoregressive (AR) approximation of the Wold representation (see again the next chapter for details) as

$$\varphi(L)y_t = \varepsilon_t, \quad \varphi(L) = \psi(L)^{-1},$$

we also have

$$\varepsilon_h = \varphi e_h, \quad \varphi = \psi^{-1},$$

$$\Phi_h^{-1} = \sigma_\varepsilon^{-2}\varphi'\varphi.$$

We can therefore rewrite the statistic Q as

$$Q = \frac{e_h'\varphi'\varphi e_h}{\sigma_\varepsilon^2} = \frac{\varepsilon_h'\varepsilon_h}{\sigma_\varepsilon^2} = \frac{1}{\sigma_\varepsilon^2}\sum_{J=1}^{h}\varepsilon_{T+J}^2.$$

Tests of predictive accuracy

- An operational version of the test is therefore:

$$\hat{Q} = \frac{1}{\hat{\sigma}_{\varepsilon}^2} \sum_{J=1}^{h} e_{T+J|T+J-1}^2 \quad \sim \quad F(h, T-p),$$

where p is the number of parameters used in the model and $e_{T+J|T+J-1}$ indicates for clarity the one-step ahead forecast error.

Forecast Evaluation and Combination

Forecast comparison tests

- The most common approach to compare alternative predictions is to rank them according to the associated loss function, typically the MSFE or MAFE. However, these comparisons are deterministic.
- We will now consider two tests for the hypothesis that two forecasts are equivalent, in the sense that the associated loss difference is not statistically different from zero:

(i) Morgan - Granger - Newbold test

(ii) Diebold - Mariano test

Morgan - Granger - Newbold test

- The test requires the forecast errors to be zero mean, normally distributed, and uncorrelated.
- If we indicate by e_1 and e_2 the forecast errors from the competing models, the test is based on the auxiliary variables:

$$u_{1,T+J} = e_{1,T+J} - e_{2,T+J}, \quad u_{2,T+J} = e_{1,T+J} + e_{2,T+J}. \tag{6}$$

It is

$$E(u_1 u_2) = MSFE_1 - MSFE_2,$$

so that the hypothesis of interest is whether u_1 and u_2 are correlated or not.

Morgan - Granger - Newbold test

- The proposed **Morgan - Granger - Newbold test statistic** is

$$\frac{r}{\sqrt{(H-1)^{-1}(1-r^2)}} \quad \sim \quad t_{H-1},$$

where (1) t_{H-1} a Student t distribution with $H-1$ degrees of freedom, (2) H is the length of the evaluation sample and

$$r = \frac{\sum_1^H u_{1,T+i}\, u_{2,T+i}}{\sqrt{\sum_1^H u_{1,T+i}^2 \sum_1^H u_{2,T+i}^2}}.$$

Diebold - Mariano test

- It relaxes the requirements on the forecast errors and can deal with the comparison of general loss functions.
- Let us define the **Diebold-Mariano test statistic** as

$$DM = H^{1/2} \frac{\sum_{j=1}^{H} d_j / H}{\sigma_d} = H^{1/2} \frac{\overline{d}}{\sigma_d}, \tag{7}$$

where

$$d_j = g(e_{1j}) - g(e_{2j}),$$

g is the loss function of interest, e.g., the quadratic loss $g(e) = e^2$ or the absolute loss $g(e) = |e|$, e_1 and e_2 are the errors from the two competing forecasts, and σ_d^2 is the variance of $\overline{d}$.

Diebold - Mariano test

- In order to take into account the serial correlation of the forecast errors, σ_d^2 can be estimated as

$$\widehat{\sigma}_d^2 = \left(\gamma_0 + 2 \sum_{i=1}^{h-1} \gamma_i \right) \quad \text{with } \gamma_k = H^{-1} \sum_{t=k+1}^{H} (d_t - \bar{d})(d_{t-k} - \bar{d}),$$

  where h is the forecast horizon, so that for $h = 1$ there is no correlation and the standard formula for variance estimation can be used.
- Under the null hypothesis that $E(d) = 0$, the statistic DM has an asymptotic standard normal distribution.

Diebold - Mariano test

- Harvey, Leybourne and Newbold (1998) suggested a modified version of the DM statistic,

$$HLN = \left(\frac{H + 1 - 2h + H^{-1}h(h-1)}{HH} \right)^{1/2} DM,$$

to be compared with critical values from the Student t distribution with $H - 1$ degrees of freedom, in order to improve the finite sample properties of the DM test.

Diebold - Mariano test

- When the models underlying the forecasts under comparison are nested, for example an AR(1) and an AR(2), then the asymptotic distribution of the DM test becomes non-standard and a functional of Brownian motions.
- A simple solution in this case is the use of rolling rather than recursive estimation.

Forecast Evaluation and Combination

The combination of forecasts

- Let us assume that two forecasts $\hat{y}_1$ and $\hat{y}_2$ are available for the same target y, with associated forecast errors e_1 and e_2. We want to construct the combined (linear) forecast

$$\hat{y}_c = \alpha\hat{y}_1 + (1-\alpha)\hat{y}_2, \tag{8}$$

where the weights can be chosen in order to minimize the MSFE of $\hat{y}_c$, From formula 8 we have

$$e_c = y - \hat{y}_c = \alpha e_1 + (1-\alpha)e_2 \tag{9}$$

so that

$$\begin{aligned} MSFE_c = \alpha^2 MSFE_1 + (1-\alpha)^2 MSFE_2 \\ + 2\alpha(1-\alpha)\varphi(MSFE_1\, MSFE_2)^{1/2} \end{aligned} \tag{10}$$

where φ is the correlation coefficient between e_1 and e_2.

The combination of forecasts

- The optimal pooling weights, the minimizers of formula10,

$$\alpha^* = \frac{MSFE_2 - \varphi(MSFE_1 MSFE_2)^{1/2}}{MSFE_1 + MSFE_2 - 2\varphi(MSFE_1 MSFE_2)^{1/2}},$$

which yields

$$MSFE_c^* = \frac{MSFE_1\, MSFE_2\, (1-\varphi^2)}{MSFE_1 + MSFE_2 - 2\varphi(MSFE_1 MSFE_2)^{1/2}}$$

and

$$MSFE_c^* \leq min(MSFE_1, MSFE_2),$$

where equality holds if either $\varphi^2 = MSFE_1/MSFE_2$ (i.e., $e_2 = e_1 + u$) or $\varphi^2 = MSFE_2/MSFE_1$ (i.e., $e_1 = e_2 + u$), which implies that $\hat{y}_1$ or $\hat{y}_2$ is the optimal forecasts.

The combination of forecasts

- If the forecast errors are uncorrelated ($\varphi = 0$), α^* only depends on the relative size of $MSFE_1$, and $MSFE_2$, which are commonly used weights in empirical applications even with correlated errors.
- In practice α is not known and must be estimated. An easier way to obtain an estimate of α is to run, over the evaluation sample, the regression

$$y = \alpha\hat{y}_1 + (1-\alpha)\hat{y}_2 + e, \tag{11}$$

or

$$e_2 = \alpha(\hat{y}_1 - \hat{y}_2) + e. \tag{12}$$

The combination of forecasts

- In general, a lower $MSFE_c$ can be obtained by running the unrestricted regression

$$y = \alpha_0 + \alpha_1 \hat{y}_1 + \alpha_2 \hat{y}_2 + u, \tag{13}$$

with combined forecast

$$\tilde{y}_c = \alpha_0 + \hat{\alpha}_1 \hat{y}_1 + \hat{\alpha}_2 \hat{y}_2.$$

But the residuals from formula 13, i.e., $\widehat{u} = y - \tilde{y}_c$, will be in general serially correlated when the restriction $\hat{\alpha}_1 + \hat{\alpha}_2 = 1$ is not imposed. Indeed,

$$\widehat{u} = -\hat{\alpha}_0 + (1 - \sum_{i=1}^{2} \hat{\alpha}_i) y + \sum_{i=1}^{2} \hat{\alpha}_i e_i.$$

Hence, a proper estimation method (such as GLS) should be adopted.

The combination of forecasts

- In the presence of a rather large set of alternative forecasts, from a practical point view a combined forecast obtained by simply averaging all the alternative available forecasts tends to work well.

Forecast Evaluation and Combination

Forecast encompassing

- One model encompasses another with respect to a certain property if from the first model it is possible to deduce the property of interest in the second model.
- **Forecast encompassing** concerns whether the one-step forecast of one model can explain the forecast errors made by another.

Forecast encompassing

- From an operational point of view, we can use the regression $e_2 = \alpha(\hat{y}_1 - \hat{y}_2) + e$ and test for $\alpha = 0$. If $\alpha \neq 0$ the difference between $\hat{y}_1$ and $\hat{y}_2$ can partly explain e_2, and therefore the second model cannot forecast encompass the first one.

- Similarly, if $\beta \neq 0$ in the regression

$$e_1 = \beta(\hat{y}_2 - \hat{y}_1) + v, \tag{14}$$

the first model cannot forecast encompass the second one.

Forecast encompassing

- A second test can be based on the regression

$$e_1 = \delta \hat{y}_2 + \varphi, \tag{15}$$

  and it requires $\delta = 0$ for the second model not to forecast encompass the first one.
- A third alternative is a test for $\alpha_1 = 1$, $\alpha_2 = 0$,

$$y = \alpha_0 + \alpha_1 \hat{y}_1 + \alpha_2 \hat{y}_2 + u.$$

Forecast encompassing

- Even if the procedures for forecast combination and encompassing are similar, the suggestions from the two methods are different.
- The former simply indicates to combine the competing forecasts, the latter to respecify the models that produced the forecasts, because both of them are somewhat misspecified.

Forecast Evaluation and Combination

Evaluation and combination of density forecasts

- Let us indicate the **density forecast** by $f_{T+h|T}$, given information up to T (and X_{T+h}) with horizon h, and its cumulative distribution function (CDF) by $F_{T+h|T}$. Similarly, we indicate the true density of the target variable by $g_{T+h|T}$ and its CDF by $G_{T+h|T}$.
- In the case of the linear regression model considered in Chapter 1, under the assumption of normal errors, we have seen that the (optimal) density forecast ($f_{T+h|T}$) is

$$y_{T+h} \sim N(\hat{y}_{T+h}, V(e_{T+h})),$$

where $\hat{y}_{T+h} = X_{T+h}\hat{\beta}_T$ and $V(e_{T+h})$ denotes the variance of the forecast error. The true density ($g_{T+h|T}$) is instead

$$y_{T+h} \sim N(X_{T+h}\beta, \sigma^2_\varepsilon).$$

Evaluation of density forecasts

- It is convenient to introduce the Probability Integral Transformation (PIT), defined as

$$PIT_t(x) \equiv F_{t+h|t}(x), \tag{16}$$

  for any forecast x.
- It can be shown that if $F_{t+h|t} = G_{t+h|t}$ for all t, then the PIT_ts are independent $U[0,1]$ variables, where U denotes the uniform distribution.
- Therefore, to assess the quality of density forecasts we can check whether their associated $PITs$ are independent and uniformly distributed.

Evaluation of density forecasts

- Uniformity (typically defined as probabilistic calibration) can be evaluated qualitatively, by plotting the histogram of the PIT_ts for the available evaluation sample.
- For a more formal assessment of probabilistic calibration, let us consider the inverse normal transformation:

$$z_t = \Phi^{-1}(PIT_t), \tag{17}$$

where Φ is the CDF of a standard normal variable. If PIT_t is $\overset{iid}{\sim} U(0,1)$ then z_t is $\overset{iid}{\sim} N(0,1)$.

Evaluation of density forecasts

- It is more convenient to assess probabilistic calibration using z_ts rather than PIT_ts.
- Mitchell and Walls (2011) provide a list of tests for uniformity and normality.
- If the z_ts are indeed normally distributed and therefore independence and lack of correlation are equivalent, we can use any of the tests for no correlation in the errors described in Chapter 2.

Comparison of density forecasts

- It is convenient to introduce the logarithmic score, defined as

$$log\, S_j(x) \;=\; log f_{j,t+h|t}(x), \tag{18}$$

where j indicates the alternative densities.

- If one of the densities under comparison coincides with $g_{t+h|t}$ (the true density), then the expected value of the differences in the logarithmic scores coincides with the Kullback-Leibler Information Criterion (KLIC):

$$KLIC_{j,t} = E_g[log\, g_{t+h|t}(x) \;-\; log f_{j,t+h|t}(x)] = E[d_{j,t}(x)].$$

We can interpret $d_{j,t}$ as a density forecast error, so that the $KLIC$ is a kind of "mean density error".

Comparison of density forecasts

- To compare two densities, f_j and f_k, we can then use:

$$\triangle L_t = logS_j(x_t) - log\, S_k\,(x_t). \tag{19}$$

- To assess whether statistically the two densities are different (basically, to construct the counterpart of the Diebold-Mariano test in a density context), we can use:

$$\sqrt{T}(\frac{\sum \triangle L_t}{T}/std.dev.) \rightarrow N(0,1),$$

Combination of density forecasts

- Following, e.g.,Wallis(2005), starting from n forecast densities f_j, $j = 1, ..., n$, the combined density forecast is

$$f_c = \sum_{j=1}^{n} w_j f_j,$$

  where $w_j \geq 0,\ j = 1, ..., n$, and $\sum_{j=1}^{n} w_j = 1$. The combined density f_c is therefore a finite mixture distribution.
- Defining values for the weights $w_j,\ j = 1, ..., n$, is not easy. A simple solution that often works well in practice , is to set $w_j = 1/n, j = 1, \ldots, n$. Alternatively, the weights could be chosen optimally, to maximize a certain objective function or minimize the KLIC with respect to the true unknown density.

Forecast Evaluation and Combination

Examples using simulated data

- In the previous three chapters we have seen forecasts produced by linear regressions, the same models augmented with dummy variables and also models with lagged variables,

$$
\begin{aligned}
y_t = &\quad \alpha_1 + \alpha_2 x_t + \varepsilon_t \\
y_t = &\quad \alpha_1 + \beta_1 D_t + \alpha_2 x_t + \beta_2 D_t x_t + \varepsilon_t \\
y_t = &\quad \alpha_1 + \beta_1 D_t + \alpha_2 x_t + \beta_2 D_t x_t + \gamma_1^y y_{t-1} + \gamma_1^x x_{t-1} + \varepsilon_t
\end{aligned}
$$

  Naturally, model (mis-)specification affects the forecast accuracy. In this chapter we will compare the performance of all three specifications.

Examples using simulated data

- Table 1 represents simple forecast evaluation statistics for all the models considered. The **RMSFE** and **MAFE** are the smallest for the dynamic regression model.

Forecasting Model	RMSFE	RMSFE recursive	MAFE	MAFE recursive
Linear Regression	9.536	9.490	7.995	7.950
Dummy variable model	10.130	9.642	8.314	7.938
Dynamic model	1.108	0.965	0.865	0.760

Table 1: *Forecasts evaluation: RMSFE and MAFE*

- The next step is to compare individual forecasts. In section 5 we discussed two main tests: the Diebold-Mariano (DM) and Morgan-Granger-Newbold (MGN) tests. We start with the DM test.

Diebold - Mariano(DM) test

- The null hypothesis of the **DM test** is that the average loss differential between the forecasts of compared models is equal to zero. The DM test results for all three models are described in Table 2.

Comparisons	$M1\ vs\ M2$	$M1\ vs\ M3$	$M2\ vs\ M3$
DM test statistics	-1.113	8.582	11.445
P-value	0.132	0.000	0.000

Table 2: *Diebold-Mariano test for equality of MSFE*

- This result indicates that at 5% significance level, there is no difference between forecasts of Model and 2, but the dynamic models provides statistically significant better forecasts.

Examples using simulated data

- The null hypothesis of the **MGN test** is that the MSFEs associated with two forecasts are equal. The test results appear in Table 3 and are in line with the DM test findings.

Comparisons	$M1$ *vs* $M2$	$M1$ *vs* $M3$	$M2$ *vs* $M3$
MGN test statistics	-1.100	42.932	65.137
P-value	0.273	0.000	0.000

Table 3: *Morgan-Granger-Newbold test*

- We also compare **recursive forecasts** using DM and MGN tests. The results are omitted as they convey the same message.

Unbiasedness tests

- Let us reconsider the formula 1a :

$$y_{i+h} = \alpha + \beta \hat{y}_{i+h|i} + \varepsilon_{i+h}, \quad i = T, ..., T + H - h, \qquad h < H$$

- **Unbiasedness** test can be implemented with a $t-$test (we do not need a robust one with $h = 1$) in the following regression:

$$e_{t+1} = y_{t+1} - \hat{y}_{t+1|t} = \tau + \varepsilon_{t+1}.$$

Unbiasedness tests

- The results for Models 1 and 2 appear in respectively Tables 4 and 5.

	Coefficient	Std. Error	t-Statistic	Prob.
τ	-1.744	0.953	-1.830	0.069
R-squared	0.000	Mean dep var		-1.744
Adjusted R-squared	0.000	S.D.dependent var		9.609
S.E.of regression	9.609	Akaike IC		7.368
Sum squared resid	18375.640	Schwarz IC		7.385
Log likelihood	-735.834	Hannan-Quinn		7.375
DW stat	0.790			

Table 4: *Unbiasedness and weak efficiency (via DW) tests for Model 1*

Unbiasedness tests

	Coefficient	Std. Error	t-Statistic	Prob.
τ	-3.223	0.810	-3.979	0.000
R-squared	0.000	Mean dep var		-3.223
Adjusted R-squared	0.000	S.D.dependent var		9.625
S.E.of regression	9.625	Akaike IC		7.372
Sum squared resid	18436.060	Schwarz IC		7.388
Log likelihood	-736.162	Hannan-Quinn		7.378
DW stat	1.369			

Table 5: *Unbiasedness and weak efficiency (via DW) tests for Model 2*

- For M1 we accept the null $\tau = 0$ at the 5 % level, while for M2 we clearly reject the null. This means that forecasts for M1 appear to be unbiased, while the opposite is true for M2.

Weak efficiency tests

- For **weak efficiency** test, we can look at the same regression output as the DW statistic tells us whether there is order-one autocorrelation in the ε_t, which amounts to testing the autocorrelation of e_t.
- There is evidence for serial correlation in the forecast errors for both models.
- In general, with $h > 1$ we need to fit a moving average model of order $h-1$, $MA(h-1)$, to the h-steps ahead forecast error and test that the resulting residuals are white noise.

Forecast Evaluation and Combination

Forecasting Euro area GDP growth

- Using again the Euro area data series, we consider the Models 1 through 4, estimated over the period 1996Q1 to 2006Q4; and using 2007Q1 to 2013Q2 as the forecast evaluation period. Recall that Models 1 through 4 and the ARDL Model 2 contain different elements in their information sets:

$$y_t = \alpha + X_t\beta + \varepsilon_t$$

  - Model 1: $X_t = (ipr_t,\ su_t,\ pr_t,\ sr_t)$
  - Model 2: $X_t = (ipr_t,\ su_t,\ sr_t)$
  - Model 3: $X_t = (ipr_t,\ su_t)$
  - Model 4: $X_t = (ipr_t,\ pr_t,\ sr_t)$
  - ARDL Model 2: : $X_t = (y_{t-1},\ ipr_t,\ su_{t-1},\ sr_{t-1})$

Forecasting Euro area GDP growth

- To proceed we need to re-estimate the ARDL Model 2, presented in Chapter 3, over the period 1996Q1 to 2006Q4 using this estimation sample, with the results appearing in Table 6.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.333	0.083	4.037	0.000
Y(-1)	0.197	0.139	1.413	0.166
IPR	0.201	0.069	2.907	0.006
SU(-1)	-0.002	0.018	-0.140	0.889
SR(-1)	0.009	0.006	1.435	0.159
R-squared	0.513	Mean dep var		0.564
Adjusted R-squared	0.463	S.D. dep var		0.355
S.E. of regression	0.260	Akaike IC		0.251
Sum squared resid	2.638	Schwarz IC		0.454
Log likelihood	-0.518	Hannan-Quinn		0.326
F-statistic	10.256	DW stat		2.220
Prob(F-statistic)	0.000			

Table 6: *Estimation output: ARDL Model 2*

Forecasting Euro area GDP growth

- ARDL Model 2 has a better in sample than Models 1 - 3 in terms of AIC, and in most cases, R^2 and $\overline{R}^2$. By the same measures, it seems to be not as good as Model 4.
- However, that a good in-sample fit does not necessarily imply a good out-of-sample performance. To compare the forecasts produced by these models, we need to look at forecast sample evaluation statistics and also to perform forecast comparison tests.

Forecasting Euro area GDP growth

- We first produce one-step ahead static forecasts for the period 2007Q1 to 2013Q2 and compute the RMSFE and the MAFE of the forecasts produced by each of the models, as reported in Table 7.

	M1	M2	M3	M4	ARDL M2
RMSFE	0.383	0.378	0.394	0.415	0.353
MAFE	0.327	0.331	0.344	0.364	0.310

Table 7: *Forecast evaluation measures*

- One-step ahead forecasts produced by the ARDL Model 2 outperforms the other forecasts as indicated by both the values of RMSFE and MAFE.

Forecasting Euro area GDP growth

- **Unbiasedness** can be tested by regressing each forecast error on an intercept and then test its significance. It turns out that it is significant for all the models, indicating the presence of a bias.
- **Weak efficiency** is not rejected for any of the models, however, as the one-step ahead forecasts errors are all serially uncorrelated.

Forecasting Euro area GDP growth

- Assuming the loss function is quadratic, we first compare the **one-step ahead static forecasts** using the use of the **Diebold-Mariano test**. The test statistics are reported in Table 8.

	vs M1	vs M2	vs M3	vs M4
Test Stat.	0.721	0.737	1.331	1.535
p-value	0.235	0.230	0.091	0.062

Table 8: *DM tests ARDL Model 2 against Models 1 through 4*

- The null hypothesis is that the average loss differential, $\overline{d}$, is equal to zero, i.e., $H_0 : \overline{d} = 0$. The test statistic has an asymptotic standard normal distribution, i.e., $DM \overset{a}{\sim} N\left(0,\ 1\right)$.

Forecasting Euro area GDP growth

- If we consider a significance level of 5%, the null hypothesis that the loss differential between the forecasts from ARDL Model 2 and those of the other models being zero cannot be rejected.
- However, at 10% significance level, the test results suggest the loss differential between the forecasts from ARDL Model 2 and those of Model 3 and Model 4.
- Moreover, the reported test statistics here are all positive. This is an indication that the loss associated with Model 1, 2, 3, and 4 is larger than that of ARDL Model 2.

Forecasting Euro area GDP growth

- The computed test statistics and the corresponding p-values of **Morgan-Granger-Newbold (MGN test)** are reported in Table 9.

	vs M1	vs M2	vs M3	vs M4
Test Stat.	1.756	0.735	0.900	1.651
p-value	0.090	0.468	0.375	0.110

Table 9: *MGN tests ARDL Model 2 against Models 1 through 4*

- The null hypothesis is that the mean of the loss differential is zero, meaning the variances of the two forecast errors are equal.The test statistic follows a Student t distribution with $H-1$ degree of freedom (in this case $H = 26$).

Forecasting Euro area GDP growth

- At 5% significance level, the null hypothesis cannot be rejected for all four cases.
- Note however, that the null hypothesis can be rejected for the first case at the 10% significance level.

Forecasting Euro area GDP growth

- We now look into whether there is any change in the results if we consider **recursive forecasts**. Tables 10 and 11 present the test results.

	vs M1	vs M2	vs M3	vs M4
Test stat.	1.716	0.737	1.331	1.535
p-value	0.043	0.230	0.091	0.062

Table 10: *DM tests recursive forecasts ARDL Model 2 vs Models 1 - 4*

	vs M1	vs M2	vs M3	vs M4
Test stat.	1.898	1.281	1.208	1.693
p-value	0.068	0.211	0.237	0.102

Table 11: *MGN tests recursive forecasts ARDL Model 2 vs Models 1 - 4*

Forecasting Euro area GDP growth

- Looking at the Diebold-Mariano test statistics, the null hypothesis that the loss differential between the one-step ahead recursive forecasts from ARDL Model 2 and those from Models 2 through 4 being zero cannot be rejected at the 5% significance level.
- Whereas the null hypothesis that the one-step ahead recursive forecasts from ARDL Model 2 are no difference from those of Model 1 can be rejected at 5% significance level.
- The Morgan-Granger-Newbold test results confirm this finding at the 10 % level.
- At the 10 % level we can also reject the DM test null for the ARDL Model 2 against Models 3 and 4.

Forecasting Euro area GDP growth

- We can construct pooled forecasts by combining the 5 available predictors, using equal weights for simplicity and as the sample size is rather short.

  It turns out that the MSFE of the combined forecasts is lower than that of the forecasts associated with Models 1 through 4, but larger than that of the ARDL Model 2.

Forecasting US GDP growth

- Using again the US data series employed in the previous chapters, we consider Models 1 through 4 and ARDL Models , estimated over the period 1985Q1 to 2006Q4; and using 2007Q1 to 2013Q4 as the forecast evaluation period.
  - Model 1: $X_t = (ipr_t,\ su_t,\ pr_t,\ sr_t)$
  - Model 2: $X_t = (ipr_t,\ su_t,\ sr_t)$
  - Model 3: $X_t = (ipr_t,\ su_t)$
  - Model 4: $X_t = (ipr_t,\ pr_t,\ sr_t)$
  - ARDL Model 2: : $X_t = (y_{t-1},\ ipr_t,\ su_{t-1},\ sr_{t-1})$

Forecasting US GDP growth

- We compute the RMSFE and the MAFE of the forecasts produced by these models, as reported in Table 12.

	M1	M2	M3	M4	ARDL M2
RMSFE	0.537	0.531	0.539	0.537	0.567
MAFE	0.402	0.399	0.403	0.418	0.438

Table 12: *Simple forecast evaluation statistics*

- It is clear that one-step ahead forecasts produced by the ARDL Model 2 is outperformed by the other forecasts

Forecasting US GDP growth

- We first compare the **one-step ahead static forecasts** from ARDL Model 2 with those of Model 1, 2, 3 and 4, using Diebold - Mariano (DM) test and Morgan - Granger - Newbold (MGN) test.
- Table 13 reports the test statistics and the corresponding p-values of **DM test**.

	vs M1	vs M2	vs M3	vs M4
Test stat.	-1.516	-1.748	-1.375	-2.279
p-value	0.064	0.040	0.084	0.011

Table 13: *DM test on one-step ahead forecasts from ARDL model 2 against one-step ahead forecasts from Model 1 through 4.*

Forecasting US GDP growth

- At 5% significance level, the null hypothesis that the loss differential between the forecasts from ARDL Model 2 and those of the other models being zero can be rejected for Models 2 and 4.
- Moreover, at 10% significance level, the test results suggest the loss differential between the forecasts from ARDL Model 2 and all models is non-zero.
- Note, however, that the reported test statistics here are all negative. This is an indication that the loss associated with Model 1, 2, 3 and 4 is smaller than that of ARDL Model 2.

Forecasting US GDP growth

- The computed test statistics and the corresponding p-values of **MGN test** are reported in Table 14.

	vs M1	vs M2	vs M3	vs M4
Test stat.	-1.109	-1.308	-0.946	-2.007
p-value	0.276	0.201	0.352	0.054

Table 14: *MGN test on one-step ahead forecasts from ARDL model 2 against one-step ahead forecasts from Model 1 through 4*

- The null hypothesis of the MGN test is that the mean of the loss differential is zero, meaning the variances of the two forecast errors are equal. The null hypothesis cannot be rejected for first 3 cases, and can be rejected in the last case at the 10% significance level.

Forecasting US GDP growth

- We now look into whether there is any change in the results if we consider **recursive forecasts**. Tables 15 and 16 present the test results.

	vs M1	vs M2	vs M3	vs M4
Test stat.	-1.473	-1.748	-1.375	-2.279
p-value	0.070	0.040	0.084	0.011

Table 15: *DM test on one-step ahead recursive forecasts from ARDL Model 2 against those from Model 1 through 4*

	vs M1	vs M2	vs M3	vs M4
Test stat.	-1.258	-1.243	-0.735	-1.641
p-value	0.219	0.224	0.468	0.112

Table 16: *MGN test on one-step ahead recursive forecasts from ARDL Model 2 against those from Model 1 through 4*

Forecasting US GDP growth

- Looking at the Diebold-Mariano test statistics, the null hypothesis that the loss differential between the one-step ahead recursive forecasts from ARDL Model 2 and those from Model 2, 3, and 4 being zero can be rejected at 10% significance level.
- However, the Morgan-Granger-Newbold test results show that the null hypothesis of the mean of the loss differential being zero cannot be rejected at 10% significance level for all 4 cases.

Default risk

- Let us now revisit the empirical default risk models.
  - ARDL Model 1: $OAS_t = \alpha + \beta_1\ OAS_{t-1} + \beta_2\ VIX_{t-1} + \varepsilon_t$
  - ARDL Model 2: $OAS_t = \alpha + \beta_1\ OAS_{t-1} + \beta_2\ SENT_{t-1} + \varepsilon_t$
  - ARDL Model 3: $OAS_t = \alpha + \beta_1\ OAS_{t-1} + \beta_2\ PMI_{t-1} + \varepsilon_t$
  - ARDL Model 4: $OAS_t = \alpha + \beta_1\ OAS_{t-1} + \beta_2\ SP500_{t-1} + \varepsilon_t$

- We found that the in-sample Model 4 featured the best fit, whereas Model 3 had the best out-of-sample RMSFE.

Default risk

- We use the **Morgan-Granger-Newbold test** for pairwise comparisons of the **one-step ahead forecasts**.

Model 1	Model 2	t-stat	p-val	Superior
VIX	SENT	0.505	0.617	2
VIX	PMI	0.575	0.569	2
VIX	SP500	-1.042	0.305	1
SENT	PMI	0.662	0.513	2
SENT	SP500	-1.410	0.168	1
PMI	SP500	-1.463	0.153	1

Table 17: *MGN tests default risk models*

- The results appear in Table 17. The results indicate that the out-of-sample differences across the different models is not statistically significant. The DM statistics also confirm the same finding.

Forecast Evaluation and Combination

Concluding remarks

- The main focus in this book is forecasting using econometric models. It is important to stress, however, that the methods reviewed in this chapter developed for the purpose of evaluating forecasts apply far beyond the realm of econometric models. Indeed, the forecasts could simply be, say, analyst forecasts at least not explicitly related to any specific model. Hence, the reach of the methods discussed in this chapter is wide.