Forecasting

Let $\{y_t\}$ be a covariance stationary are ergodic process, e.g. an ARMA(p,q) process with Wold representation

$$\begin{aligned} y_t &= \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},\ \varepsilon_t \sim WN(0,\sigma^2) \\ &= \mu + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \cdots \end{aligned}$$

Let $I_t = \{y_t, y_{t-1}, \ldots\}$ denote the information set available at time t. Recall,

$$\begin{aligned} E[y_t] &= \mu \\ var(y_t) &= \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 \end{aligned}$$

Goal: Using I_t produce optimal forecasts of y_{t+h} for $h = 1, 2, \ldots, s$

Note:

$$\begin{aligned} y_{t+h} &= \mu + \varepsilon_{t+h} + \psi_1 \varepsilon_{t+h-1} + \cdots \\ &\quad + \psi_{h-1}\varepsilon_{t+1} + \psi_h \varepsilon_t + \psi_{h+1}\varepsilon_{t-1} + \cdots \end{aligned}$$

Define $y_{t+h|t}$ as the forecast of y_{t+h} based on I_t known parameters. The forecast error is

$$\varepsilon_{t+h|t} = y_{t+h} - y_{t+h|t}$$

and the mean squared error of the forecast is

$$\begin{aligned} MSE(\varepsilon_{t+h|t}) &= E[\varepsilon_{t+h|t}^2] \\ &= E[(y_{t+h} - y_{t+h|t})^2] \end{aligned}$$

Theorem: The minimum MSE forecast (best forecast) of y_{t+h} based on I_t is

$$y_{t+h|t} = E[y_{t+h}|I_t]$$

Proof: See Hamilton pages 72-73.

Remarks

1. The computation of $E[y_{t+h}|I_t]$ depends on the distribution of $\{\varepsilon_t\}$ and may be a very complicated nonlinear function of the history of $\{\varepsilon_t\}$. Even if $\{\varepsilon_t\}$ is an uncorrelated process (e.g. white noise) it may be the case that

$$E[\varepsilon_{t+1}|I_t] \neq 0$$

2. If $\{\varepsilon_t\}$ is independent white noise, then $E[\varepsilon_{t+1}|I_t] = 0$ and $E[y_{t+h}|I_t]$ will be a simple linear function of $\{\varepsilon_t\}$

$$y_{t+h|t} = \mu + \psi_h \varepsilon_t + \psi_{h+1}\varepsilon_{t-1} + \cdots$$

Linear Predictors

A linear predictor of $y_{t+h|t}$ is a linear function of the variables in I_t.

Theorem: The minimum MSE linear forecast (best linear predictor) of y_{t+h} based on I_t is

$$y_{t+h|t} = \mu + \psi_h \varepsilon_t + \psi_{h+1}\varepsilon_{t-1} + \cdots$$

Proof. See Hamilton page 74.

The forecast error of the best linear predictor is

$$\begin{aligned}\varepsilon_{t+h|t} &= y_{t+h} - y_{t+h|t} \\ &= \mu + \varepsilon_{t+h} + \psi_1 \varepsilon_{t+h-1} + \cdots \\ &\quad + \psi_{h-1}\varepsilon_{t+1} + \psi_h \varepsilon_t + \cdots \\ &\quad - (\mu + \psi_h \varepsilon_t + \psi_{h+1}\varepsilon_{t-1} + \cdots) \\ &= \varepsilon_{t+h} + \psi_1 \varepsilon_{t+h-1} + \cdots + \psi_{h-1}\varepsilon_{t+1}\end{aligned}$$

and the MSE of the forecast error is

$$\text{MSE}(\varepsilon_{t+h|t}) = \sigma^2(1 + \psi_1^2 + \cdots + \psi_{h-1}^2)$$

Remarks

1. $E[\varepsilon_{t+h|t}] = 0$

2. $\varepsilon_{t+h|t}$ is uncorrelated with any element in I_t

3. The form of $y_{t+h|t}$ is closely related to the IRF

4. $MSE(\varepsilon_{t+h|t}) = var(\varepsilon_{t+h|t}) \leq var(y_t)$

5. $\lim_{h\to\infty} y_{t+h|t} = \mu$

6. $\lim_{h\to\infty} MSE(\varepsilon_{t+h|t}) = var(y_t)$

Example: BLP for MA(1) process

$$y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1},\ \varepsilon_t \sim \mathsf{WN}(0, \sigma^2)$$

Here

$$\psi_1 = \theta,\ \psi_h = 0 \text{ for } h > 1$$

Therefore,

$$\begin{aligned} y_{t+1|t} &= \mu + \theta\varepsilon_t \\ y_{t+2|t} &= \mu \\ y_{t+h|t} &= \mu \text{ for } h > 1 \end{aligned}$$

The forecast errors and MSEs are

$$\begin{aligned} \varepsilon_{t+1|t} &= \varepsilon_{t+1},\ \mathsf{MSE}(\varepsilon_{t+1|t}) = \sigma^2 \\ \varepsilon_{t+2|t} &= \varepsilon_{t+2} + \theta\varepsilon_{t+1},\ \mathsf{MSE}(\varepsilon_{t+2|t}) = \sigma^2(1 + \theta^2) \end{aligned}$$

Prediction Confidence Intervals

If $\{\varepsilon_t\}$ is Gaussian then

$$y_{t+h}|I_t \sim N(y_{t+h|t}, \sigma^2(1+\psi_1^2+\cdots+\psi_{h-1}^2))$$

A 95% confidence interval for the $h-$step prediction has the form

$$y_{t+h|t} \pm 1.96 \cdot \sqrt{\sigma^2(1+\psi_1^2+\cdots+\psi_{h-1}^2)}$$

Predictions with Estimated Parameters

Let $\hat{y}_{t+h|t}$ denote the BLP with estimated parameters:

$$\hat{y}_{t+h|t} = \hat{\mu} + \hat{\psi}_h \hat{\varepsilon}_t + \hat{\psi}_{h+1}\hat{\varepsilon}_{t-1} + \cdots$$

where $\hat{\varepsilon}_t$ is the estimated residual from the fitted model. The forecast error with estimated parameters is

$$\begin{aligned}
\hat{\varepsilon}_{t+h|t} &= y_{t+h} - \hat{y}_{t+h|t} \\
&= (\mu - \hat{\mu}) + \varepsilon_{t+h} + \psi_1 \varepsilon_{t+h-1} + \cdots + \psi_{h-1}\varepsilon_{t+1} \\
&\quad + \left(\psi_h \varepsilon_t - \hat{\psi}_h \hat{\varepsilon}_t\right) + \left(\psi_{h+1}\varepsilon_{t-1} - \hat{\psi}_{h+1}\hat{\varepsilon}_{t-1}\right) \\
&\quad + \cdots
\end{aligned}$$

Obviously,

$$\mathrm{MSE}(\hat{\varepsilon}_{t+h|t}) \neq \mathrm{MSE}(\varepsilon_{t+h|t}) = \sigma^2(1+\psi_1^2+\cdots+\psi_{h-1}^2)$$

Note: Most software computes

$$\widehat{\mathrm{MSE}}(\varepsilon_{t+h|t}) = \hat{\sigma}^2(1+\hat{\psi}_1^2+\cdots+\hat{\psi}_{h-1}^2)$$

Computing the Best Linear Predictor

The BLP $y_{t+h|t}$ may be computed in many different but equivalent ways. The algorithm for computing $y_{t+h|t}$ from an AR(1) model is simple and the methodology allows for the computation of forecasts for general ARMA models as well as multivariate models.

Example: AR(1) Model

$$\begin{aligned} y_t - \mu &= \phi(y_{t-1} - \mu) + \varepsilon_t \\ & \varepsilon_t \sim WN(0, \sigma^2) \\ & \mu, \phi, \sigma^2 \text{ are known} \end{aligned}$$

In the Wold representation $\psi_j = \phi^j$. Starting at t and iterating forward h periods gives

$$\begin{aligned} y_{t+h} &= \mu + \phi^h(y_t - \mu) + \varepsilon_{t+h} + \phi\varepsilon_{t+h-1} + \cdots \\ & \quad + \phi^{h-1}\varepsilon_{t+1} \\ &= \mu + \phi^h(y_t - \mu) + \varepsilon_{t+h} + \psi_1\varepsilon_{t+h-1} + \cdots \\ & \quad + \psi_{h-1}\varepsilon_{t+1} \end{aligned}$$

The best linear forecasts of $y_{t+1}, y_{t+2}, \ldots, y_{t+h}$ are computed using the *chain-rule of forecasting* (law of iterated projections)

$$\begin{aligned} y_{t+1|t} &= \mu + \phi(y_t - \mu) \\ y_{t+2|t} &= \mu + \phi(y_{t+1|t} - \mu) = \mu + \phi(\phi(y_t - \mu)) \\ &= \mu + \phi^2(y_t - \mu) \\ & \vdots \\ y_{t+h|t} &= \mu + \phi(y_{t+h-1|t} - \mu) = \mu + \phi^h(y_t - \mu) \end{aligned}$$

The corresponding forecast errors are

$$\begin{aligned} \varepsilon_{t+1|t} &= y_{t+1} - y_{t+1|t} = \varepsilon_{t+1} \\ \varepsilon_{t+2|t} &= y_{t+2} - y_{t+2|t} = \varepsilon_{t+2} + \phi\varepsilon_{t+1} \\ &= \varepsilon_{t+2} + \psi_1\varepsilon_{t+1} \\ & \vdots \\ \varepsilon_{t+h|t} &= y_{t+h} - y_{t+h|t} = \varepsilon_{t+h} + \phi\varepsilon_{t+h-1} + \cdots \\ & \quad + \phi^{h-1}\varepsilon_{t+1} \\ &= \varepsilon_{t+h} + \psi_1\varepsilon_{t+h-1} + \cdots + \psi_{h-1}\varepsilon_{t+1} \end{aligned}$$

The forecast error variances are

$$
\begin{aligned}
var(\varepsilon_{t+1|t}) &= \sigma^2 \\
var(\varepsilon_{t+2|t}) &= \sigma^2(1+\phi^2) = \sigma^2(1+\psi_1^2) \\
&\vdots \\
var(\varepsilon_{t+h|t}) &= \sigma^2(1+\phi^2+\cdots+\phi^{2(h-1)}) = \sigma^2\frac{1-\phi^{2h}}{1-\phi^2} \\
&= \sigma^2(1+\psi_1^2+\cdots+\psi_{h-1}^2)
\end{aligned}
$$

Clearly,

$$
\begin{aligned}
\lim_{h\to\infty} y_{t+h|t} &= \mu = E[y_t] \\
\lim_{h\to\infty} var(\varepsilon_{t+h|t}) &= \frac{\sigma^2}{1-\phi^2} \\
&= \sigma^2\sum_{h=0}^{\infty}\psi_h^2 = var(y_t)
\end{aligned}
$$

AR(p) Models

Consider the AR(p) model

$$
\begin{aligned}
\phi(L)(y_t-\mu) &= \varepsilon_t,\ \varepsilon_t \sim WN(0,\sigma^2) \\
\phi(L) &= 1-\phi_1 L - \cdots \phi_p L^p
\end{aligned}
$$

The forecasting algorithm for the AR(p) models is essentially the same as that for AR(1) models once we put the AR(p) model in state space form. Let $X_t = y_t - \mu$. The AR(p) in state space form is

$$
\begin{pmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-p+1} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_p \\ 1 & 0 & \cdots & 0 \\ & \ddots & & \vdots \\ 0 & & 1 & 0 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ X_{t-2} \\ \vdots \\ X_{t-p} \end{pmatrix} + \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}
$$

or

$$
\begin{aligned}
\boldsymbol{\xi}_t &= \mathbf{F}\boldsymbol{\xi}_{t-1}+\mathbf{w}_t \\
var(\mathbf{w}_t) &= \boldsymbol{\Sigma}_w
\end{aligned}
$$

Starting at t and iterating forward h periods gives

$$\boldsymbol{\xi}_{t+h} = \mathbf{F}^h \boldsymbol{\xi}_t + \mathbf{w}_{t+h} + \mathbf{F}\mathbf{w}_{t+h-1} + \cdots + \mathbf{F}^{h-1}\mathbf{w}_{t+1}$$

Then the best linear forecasts of $y_{t+1}, y_{t+2}, \ldots, y_{t+h}$ are computed using the *chain-rule of forecasting* are

$$\begin{aligned}
\boldsymbol{\xi}_{t+1|t} &= \mathbf{F}\boldsymbol{\xi}_t \\
\boldsymbol{\xi}_{t+2|t} &= \mathbf{F}\boldsymbol{\xi}_{t+1|t} = \mathbf{F}^2\boldsymbol{\xi}_t \\
&\vdots \\
\boldsymbol{\xi}_{t+h|t} &= \mathbf{F}\boldsymbol{\xi}_{t+h-1|t} = \mathbf{F}^h\boldsymbol{\xi}_t
\end{aligned}$$

The forecast for y_{t+h} is given by μ plus the first row of $\boldsymbol{\xi}_{t+h|t} = \mathbf{F}^h\boldsymbol{\xi}_t$:

$$\boldsymbol{\xi}_{t+h|t} = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_p \\ 1 & 0 & \cdots & 0 \\ & \ddots & & \vdots \\ 0 & & 1 & 0 \end{pmatrix}^h \begin{pmatrix} y_t - \mu \\ y_{t-1} - \mu \\ \vdots \\ y_{t-p+1} - \mu \end{pmatrix}$$

The forecast errors are given by

$$\begin{aligned}
\mathbf{w}_{t+1|t} &= \boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_{t+1|t} = \mathbf{w}_{t+1} \\
\mathbf{w}_{t+2|t} &= \boldsymbol{\xi}_{t+2} - \boldsymbol{\xi}_{t+2|t} = \mathbf{w}_{t+2} + \mathbf{F}\mathbf{w}_{t+1} \\
&\vdots \\
\mathbf{w}_{t+h|t} &= \boldsymbol{\xi}_{t+h} - \boldsymbol{\xi}_{t+h|t} = \mathbf{w}_{t+h} + \mathbf{F}\mathbf{w}_{t+h-1} + \cdots \\
&\quad + \mathbf{F}^{h-1}\mathbf{w}_{t+1}
\end{aligned}$$

and the corresponding forecast MSE matrices are

$$\begin{aligned}
var(\mathbf{w}_{t+1|t}) &= var(\mathbf{w}_t) = \mathbf{\Sigma}_w \\
var(\mathbf{w}_{t+2|t}) &= var(\mathbf{w}_{t+2}) + \mathbf{F}var(\mathbf{w}_{t+1})\mathbf{F}' \\
&= \mathbf{\Sigma}_w + \mathbf{F}\mathbf{\Sigma}_w\mathbf{F}' \\
&\vdots \\
var(\mathbf{w}_{t+h|t}) &= \sum_{j=0}^{h-1} \mathbf{F}^j \mathbf{\Sigma}_w \mathbf{F}^{j'}
\end{aligned}$$

Notice that

$$var(\mathbf{w}_{t+h|t}) = \mathbf{\Sigma}_w + \mathbf{F}var(\mathbf{w}_{t+h-1|t})\mathbf{F}'$$

Forecast Evaluation

Diebold-Mariano Test for Equal Predictive Accuracy

Let $\{y_t\}$ denote the series to be forecast and let $y^1_{t+h|t}$ and $y^2_{t+h|t}$ denote two competing forecasts of y_{t+h} based on I_t. For example, $y^1_{t+h|t}$ could be computed from an AR(p) model and $y^2_{t+h|t}$ could be computed from an ARMA(p, q) model. The forecast errors from the two models are

$$
\begin{aligned}
\varepsilon^1_{t+h|t} &= y_{t+h} - y^1_{t+h|t} \\
\varepsilon^2_{t+h|t} &= y_{t+h} - y^2_{t+h|t}
\end{aligned}
$$

The $h-$step forecasts are assumed to be computed for $t = t_0, \ldots, T$ for a total of T_0 forecasts giving

$$\{\varepsilon^1_{t+h|t}\}^T_{t_0},\ \{\varepsilon^2_{t+h|t}\}^T_{t_0}$$

Because the h-step forecasts use overlapping data the forecast errors in $\{\varepsilon^1_{t+h|t}\}^T_{t_0}$ and $\{\varepsilon^2_{t+h|t}\}^T_{t_0}$ will be serially correlated.

The accuracy of each forecast is measured by a particular loss function

$$L(y_{t+h}, y^i_{t+h|t}) = L(\varepsilon^i_{t+h|t}),\ i = 1, 2$$

Some popular loss functions are:

$$\begin{aligned} L(\varepsilon^i_{t+h|t}) &= \left(\varepsilon^i_{t+h|t}\right)^2 : \text{ squared error loss} \\ L(\varepsilon^i_{t+h|t}) &= \left|\varepsilon^i_{t+h|t}\right| : \text{ absolute value loss} \end{aligned}$$

To determine if one model predicts better than another we may test null hypotheses

$$H_0 : E[L(\varepsilon^1_{t+h|t})] = E[L(\varepsilon^2_{t+h|t})]$$

against the alternative

$$H_1 : E[L(\varepsilon^1_{t+h|t})] \neq E[L(\varepsilon^2_{t+h|t})]$$

The Diebold-Mariano test is based on the loss differential

$$d_t = L(\varepsilon^1_{t+h|t}) - L(\varepsilon^2_{t+h|t})$$

The null of equal predictive accuracy is then

$$H_0 : E[d_t] = 0$$

The Diebold-Mariano test statistic is

$$S = \frac{\bar{d}}{\left(\widehat{avar}(\bar{d})\right)^{1/2}} = \frac{\bar{d}}{\left(\widehat{LRV}_{\bar{d}}/T\right)^{1/2}}$$

where

$$\begin{aligned} \bar{d} &= \frac{1}{T_0}\sum_{t=t_0}^{T} d_t \\ LRV_{\bar{d}} &= \gamma_0 + 2\sum_{j=1}^{\infty}\gamma_j,\ \gamma_j = cov(d_t, d_{t-j}) \end{aligned}$$

Note: The long-run variance is used in the statistic because the sample of loss differentials $\{d_t\}_{t_0}^T$ are serially correlated for $h > 1$.

Diebold and Mariano (1995) show that under the null of equal predictive accuracy

$$S \overset{A}{\sim} N(0,1)$$

So we reject the null of equal predictive accuracy at the 5% level if

$$|S| \; > 1.96$$

One sided tests may also be computed.