Time Series Analysis

Stationary ARMA Process

Binzhi Chen

Department of Economics
University of Birmingham

March 4, 2020

Conditional Expectation

If we want to forecast Y_{t+1} based on a set of variables $\mathbf{X_t}$ observed at date t

Let $Y^*_{t+1|t}$ denote a forecast of Y_{t+1} based on $\mathbf{X_t}$, assuming a quadratic loss function, also known as *mean square error* (MSE):

$$MSE(Y^*_{t+1|t}) = E(Y_{t+1} - Y^*_{t+1|t})^2$$

Where $Y^*_{t+1|t}$ is the value of Y_{t+1} conditional on $\mathbf{X_t}$, or any function of $g(\mathbf{X_t})$

$$Y^*_{t+1|t} = E(Y_{t+1}|\mathbf{X_t}) = g(\mathbf{X_t})$$

Conditional Expectation

To verify this rule, we construct the MSE as:

$$\begin{aligned} E[Y_{t+1} - g(\mathbf{X_t})]^2 &= E[Y_{t+1} - E(Y_{t+1}|\mathbf{X_t}) + E(Y_{t+1}|\mathbf{X_t}) - g(\mathbf{X_t})]^2 \\ &= E[Y_{t+1} - E(Y_{t+1}|\mathbf{X_t})]^2 \\ &+ 2E\{[Y_{t+1} - E(Y_{t+1}|\mathbf{X_t})][E(Y_{t+1}|\mathbf{X_t}) - g(\mathbf{X_t})]\} \\ &+ E\{[E(Y_{t+1}|\mathbf{X_t}) - g(\mathbf{X_t})]^2\} \end{aligned} \tag{1}$$

Rewrite the second term on the right side of equation (1) as:

$$2E[\eta_{t+1}]$$

the second term would be zero as the law of iterated expectations:

$$E[\eta_{t+1}] = E_{\mathbf{X_t}}(E[\eta_{t+1}|\mathbf{X_t}])] = 0$$

Conditional Expectation

Now equation (1) becomes:

$$E[Y_{t+1} - g(\mathbf{X_t})]^2 = E[Y_{t+1} - E(Y_{t+1}|\mathbf{X_t})]^2 + E\{[E(Y_{t+1}|\mathbf{X_t}) - g(\mathbf{X_t})]^2\} \tag{2}$$

The second term on equation (2) cannot be smaller than zero, and the first term does not depend on $g(\mathbf{X_t})$, the only way to minimize the equation is to set the second term to zero:

$$E(Y_{t+1}|\mathbf{X_t}) = g(\mathbf{X_t})$$

Thus the MSE optimal forecast is:

$$E[Y_{t+1} - g(\mathbf{X_t})]^2 = E[Y_{t+1} - E(Y_{t+1}|\mathbf{X_t})]^2$$

In this lecture, we will use information set at time t ($\boldsymbol{\Omega}_t$) to represent the past variables (lagged ϵ, lagged Y or both) in the conditional expectation

Forecasting Based on Lagged $\epsilon' s$

Consider a MA(∞) process:

$$Y_t - \mu = \psi(L)\epsilon_t$$

Where ϵ_t is white noise and $\psi(L) = \sum_{j=0}^{\infty} \psi_j L^j$

If we want to forecast Y_{t+s}:

$$\begin{aligned} Y_{t+s} =& \mu + \epsilon_{t+s} + \psi_1 \epsilon_{t+s-1} \\ & + \cdots + \psi_{s-1}\epsilon_{t+1} + \psi_s \epsilon_t \\ & + \cdots + \psi_{s+1}\epsilon_{t-1} + \cdots + \end{aligned}$$

The condintional expectation of Y_{t+s} is:

$$E(Y_{t+s}|\epsilon_t, \epsilon_{t-1}, \dots) = \mu + \psi_s \epsilon_t + \psi_{s+1}\epsilon_{t-1} + \psi_{s+2}\epsilon_{t-2} + \dots \tag{3}$$

Where the unknown future ϵ's are set to the expected value of zero

Forecasting Based on Lagged $\epsilon's$

The error associated with the forecast is:

$$Y_{t+s} - E(Y_{t+s}|\epsilon_t, \epsilon_{t-1}, \dots) = \epsilon_{t+s} + \psi_1 \epsilon_{t+s-1} + \dots + \psi_{s-1}\epsilon_{t+1} \qquad (4)$$

The MSE is:

$$E[Y_{t+s} - E(Y_{t+s}|\epsilon_t, \epsilon_{t-1}, \dots)]^2 = (1 + \psi_1^2 + \psi_2^2 + \dots + \psi_{s-1}^2)\sigma^2 \qquad (5)$$

Forecasting Based on Lagged $\epsilon's$

For an MA(q) process:

$$\psi(L) = 1 + \theta_1 L + \theta_2 L^2 + \theta_3 L^3 + \cdots + \theta_q L^q$$

The optimal linear forecast is:

$$E(Y_{t+s}|\epsilon_t, \epsilon_{t-1}, \ldots)$$

$$=$$

$$\begin{cases} \mu + \theta_s\epsilon_t + \theta_{s+1}\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q+s} & \text{for } s = 1,\ 2, \ldots,\ q \\ \mu & \text{for } s = q+1,\ q+2, \ldots \end{cases}$$

Forecasting Based on Lagged $\epsilon' s$

The MSE is:

$$\begin{cases} \sigma^2 & \text{for } s = 1 \\ (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_{s-1}^2)\sigma^2 & \text{for } s = 2, 3, \ldots, q \\ (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma^2 & \text{for } s = q+1, q+2, \ldots \end{cases}$$

Forecasting Based on Lagged $\epsilon's$

Sometimes we can use polynomial to represent equation (3)

$$\begin{aligned}\frac{\psi(L)}{L^s} =& L^{-s} + \psi_1 L^{1-s} + \psi_2 L^{2-s} + \cdots + \psi_{s-1} L^{-1} + \psi_s L^0 \\ & + \psi_{s+1} L^1 + \psi_{s+2} L^2 + \ldots\end{aligned}$$

The *annihilation operator* (indicated by $[.]_+$) replaces negative powers of L by zero, for example:

$$[\frac{\psi(L)}{L^s}]_+ = \psi_s L^0 + \psi_{s+1} L^1 + \psi_{s+2} L^2 + \ldots$$

So that equation (3) can be represented by:

$$E(Y_{t+s}|\epsilon_t, \epsilon_{t-1}, \ldots) = \mu + [\frac{\psi(L)}{L^s}]_+ \epsilon_t \tag{6}$$

Forecasting Based on Lagged $Y's$

For an AR (1) process:

$$(1-\phi L)(Y_t-\mu)=\epsilon_t$$

The value of ϵ can be constructed for:

$$\epsilon_t=(Y_t-\mu)-\phi(Y_{t-1}-\mu)$$

Where $\mu=1/(1-\phi)$

For an MA (1) process $Y_t-\mu=\psi(L)\epsilon_t$ written in invertible form as AR(∞):

$$(1-\theta L)^{-1}(Y_t-\mu)=\eta(L)(Y_t-\mu)=\epsilon_t$$

Or we could construct ϵ_t as:

$$\epsilon_t=(Y_t-\mu)-\theta(Y_{t-1}-\mu)+\theta^2(Y_{t-2}-\mu)-\theta^3(Y_{t-3}-\mu)+\dots$$

Forecasting Based on Lagged $Y's$

Under these conditions, the function of forecasting that based on lagged $Y's$ is the same as equation (6)

$$E(Y_{t+s}|Y_t, Y_{t-1}, \ldots) = \mu + [\frac{\psi(L)}{L^s}]_+ \eta(L)(Y_t - \mu) \tag{7}$$

Or:

$$E(Y_{t+s}|Y_t, Y_{t-1}, \ldots) = \mu + [\frac{\psi(L)}{L^s}]_+ \frac{1}{\psi(L)}(Y_t - \mu) \tag{8}$$

Equation (8) is known as *Wiener-Kolmogorov prediction Formula*

Forecasting an AR(1) Process

Consider an AR(1) process:

$$(1 - \phi L)(Y_t - \mu) = \epsilon_t$$

$$(Y_t - \mu) = \psi(L)\epsilon_t$$

Where

$$\psi(L) = 1/(1 - \phi L) = 1 + \phi L + \phi^2 L + \dots$$

And

$$[\frac{\psi(L)}{L^s}]_+ = \phi^s + \phi^{s+1} L + \phi^{s+2} L^+ \dots = \frac{\phi^s}{1 - \phi L}$$

Substituting into *Wiener-Kolmogorov prediction Formula* we get the optimal linear s-step-ahead forecast value:

$$\begin{aligned} E(Y_{t+s}|Y_t, Y_{t-1}, \dots) &= \mu + \frac{\phi^s}{1 - \phi L}(1 - \phi L)(Y_t - \mu) \\ &= \mu + \phi^s (Y_t - \mu) \end{aligned} \tag{9}$$

Forecasting an MA(1) Process

Consider an invertible MA(1) process:

$$(Y_t - \mu) = (1 + \theta L)\epsilon_t$$

$$(Y_t - \mu) = \psi(L)\epsilon_t$$

Substituting into *Wiener-Kolmogorov prediction Formula* gives:

$$E(Y_{t+s}|Y_t, Y_{t-1}, \dots) = \mu + [\frac{1+\theta L}{L^s}]_+ \frac{1}{1+\theta L}(Y_t - \mu) \tag{10}$$

To forecast an MA (1) process one-step-ahead we get:

$$\begin{aligned} E(Y_{t+1}|Y_t, Y_{t-1}, \dots) &= \mu + \frac{\theta}{1+\theta L}(Y_t - \mu) \\ &= \mu + \theta(Y_t - \mu) - \theta^2(Y_{t-1} - \mu) + \theta^3(Y_{t-2} - \mu) - \dots \end{aligned} \tag{11}$$

To forecast an MA (1) process for $s = 2, 3, \dots$ we get:

$$E(Y_{t+s}|Y_t, Y_{t-1}, \dots) = \mu \tag{12}$$

Forecasting an ARMA(1,1) Process

Consider an ARMA(1,1) process:

$$(1-\phi L)(Y_t-\mu)=(1+\theta L)\epsilon_t$$

Substituting into *Wiener-Kolmogorov prediction Formula* gives:

$$E(Y_{t+s}|Y_t,Y_{t-1},\dots)=\mu+\left[\frac{1+\theta L}{(1-\phi L)L^s}\right]_+\frac{1-\phi L}{1+\theta L}(Y_t-\mu) \tag{13}$$

For the *annihilation operator* term:

$$\begin{aligned}
&\left[\frac{1+\theta L}{(1-\phi L)L^s}\right]_+\\
&=\left[\frac{(1+\phi L+\phi^2L^2+\dots)}{L^s}+\theta L\frac{(1+\phi L+\phi^2L^2+\dots)}{L^s}\right]_+\\
&=(\phi^s+\phi^{s+1}L+\phi^{s+2}L^2+\dots)+\theta(\phi^{s-1}+\phi^sL+\phi^{s+1}L^2+\dots)\\
&=(\phi^s+\theta\phi^{s-1})(1+\phi L+\phi^2L^2+\dots)\\
&=\frac{\phi^s+\theta\phi^{s-1}}{1-\phi L}
\end{aligned}$$

Forecasting an ARMA(1,1) Process

Substituting the annihilation operator term into equation (13) we get:

$$E(Y_{t+s}|Y_t, Y_{t-1}, \dots) = \mu + [\frac{\phi^s + \theta\phi^{s-1}}{1-\phi L}]\frac{1-\phi L}{1+\theta L}(Y_t - \mu)$$
$$= \mu + [\frac{\phi^s + \theta\phi^{s-1}}{1+\theta L}](Y_t - \mu)$$

So one-step-ahead forecast is given by:

$$E(Y_{t+1}|Y_t, Y_{t-1}, \dots) = \mu + \frac{\phi + \theta}{1+\theta L}(Y_t - \mu)$$

So for $s = 2, 3, \dots$ forecast is given by the recursive form:

$$[E(Y_{t+s}|Y_t, Y_{t-1}, \dots) - \mu] = \phi[E(Y_{t+s-1}|Y_t, Y_{t-1}, \dots) - \mu]$$

The Box-Jenkins Modeling Philosophy

Many forecasters are persuaded of the benefits of parsimony, or using as few parameters as possible. Box and Jenkins (1976) have been influential advocates of this view.

The reasons of forming a parsimonious model:

1. Variance of estimators is inversely proportional to the number of degrees of freedom.
2. Models which are profligate might be inclined to fit to data specific features

The approach to forecasting advocated by Box and Jenkins can be broken down into three steps:

The Box-Jenkins Modeling Philosophy

1.Identification:

(a) Involves determining the order of the model.

(b) Use of graphical procedures.

(c) A better procedure is now available.

Implication:

(a) Transform the data, if necessary, so that the assumption of stationarity is a reasonable one.

(b) Make an initial guess of small values for p and q for an ARMA(p, q) model that might describe the transformed series.

The Box-Jenkins Modeling Philosophy

2.Estimation:

(a) Estimation of the parameters.

(b) Can be done using least squares or maximum likelihood depending on the model.

Implication:

Estimate the parameters $\phi(L)$ and $\theta(L)$.

The Box-Jenkins Modeling Philosophy

3.Model Checking:

(a) Deliberate overfitting

(b) Residual diagnostics

Implication:

Perform diagnostic analysis to confirm that the model is indeed consistent with the observed features of the data.

The Box-Jenkins Modeling Philosophy

We will focus on identification in this chapter.

In general, we will use information criteria instead of autocorrelation functions in this step.

The reasons of forming a parsimonious model give motivation for using information criteria:
a) A term which is a function of the RSS
b) Some penalty for adding extra parameters

The object is to choose the number of parameters which minimises the information criterion.

The information criteria vary according to how stiff the penalty term is.

The Box-Jenkins Modeling Philosophy

The three most popular criteria are Akaike's (1974) information criterion (AIC), Schwarz's (1978) Bayesian information criterion (SBIC), and the Hannan-Quinn criterion (HQIC).

$$AIC = ln(\hat{\sigma})^2 + \frac{2k}{T}$$

$$SBIC = ln(\hat{\sigma})^2 + \frac{k}{T} ln(T)$$

$$HQIC = ln(\hat{\sigma})^2 + \frac{2k}{T} ln(ln(T))$$

Where $k = p + q + 1$, $T =$ sample size

The Box-Jenkins Modeling Philosophy

Which IC should be preferred if they suggest different model orders?

a) SBIC is strongly consistent but (inefficient).

b) AIC is not consistent, and will typically pick "bigger" models.

Forecasting Evaluation

Some of the most popular criteria for assessing the accuracy of time series forecasting techniques are:

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(y_{t+s} - \hat{y}_{t+s|t})^2$$

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\left|y_{t+s} - \hat{y}_{t+s|t}\right|$$

$$MAPE = \frac{100}{N}\sum_{t=1}^{N}\frac{\left|y_{t+s} - \hat{y}_{t+s|t}\right|}{y_{t+s}}$$

Forecasting Evaluation

A measure more closely correlated with profit:

$$\%\,correct\ sign\ predictions = \frac{1}{N}\sum_{t=1}^{N} z_{t+s}$$

Where:

$$\begin{aligned} z_{t+s} &= 1 \quad if(y_{t+s} \cdot \hat{y}_{t+s|t}) > 0 \\ z_{t+s} &= 0 \quad Otherwise \end{aligned}$$

Forecasting Evaluation

Mincer-Zarnowitz regressions (MZ) are used to test for the optimality of the forecast and are implemented with a standard regression. If a forecast is correct, it should be the case that a regression of the realized value on its forecast and a constant should produce coefficients of 1 and 0 respectively:

$$y_{t+s} = \beta_0 + \beta_1 \hat{y}_{t+s|t} + \epsilon_t$$

If the forecast is optimal, the coefficients in the MZ regression should be consistent with $\beta_0 = 0$ and $\beta_1 = 1$

Forecasting Evaluation

A Diebold-Mariano test, in contrast to an MZ regression, examines the relative performance of two forecasts. Assume A and B are different types of forecast methods, the loss function can be defined as $L_A = (y_{t+s} - \hat{y}^A_{t+s|t})^2$ and $L_B = (y_{t+s} - \hat{y}^B_{t+s|t})^2$ If the models were equally good (or bad), one would expect that $\bar{L}_A \approx \bar{L}_B$

Define $D_t = L_A - L_B$The Diebold-Mariano test is a test of equal predictive accuracy and is constructed as:

$$DM = \frac{\bar{D}_t}{\sqrt{Var(\bar{D})}}$$

From asymptotic theory, $DM \xrightarrow{L} N(0,1)$ and $Var(\bar{D})$ is the long-run variance of D
If the models are equally accurate, one would expect that $E[D_T] = 0$ which forms the null of the DM test