Forecasting: General Remarks

Suppose you wish to forecast a random quantity X (say, the temperature at noon two days from now).

You collect some relevant information $\mathcal{I}$. This doesn't determine X precisely; some uncertainty in X remains, which is described by a probability distribution called the conditional distribution of X given $\mathcal{I}$.

What is the best prediction (forecast) of X?

That depends on what you mean by "best".

Suppose you will be punished (or rewarded) depending on the accuracy of your forecast. Then the "best" forecast is the one which minimizes your expected punishment (or maximizes your expected reward).

Let $\widehat{X}$ denote your forecast (guess) for X.

Squared Error Loss:

Suppose that you will be forced to pay $(X - \widehat{X})^2$ dollars if you guess $\widehat{X}$ and the actual value is X. This is called "squared error loss".

Then the best strategy is: **guess the mean**.

More precisely, your best guess is the mean of the conditional distribution of X given $\mathcal{I}$, written as

$$\widehat{X} = E(X \mid \mathcal{I}) .$$

This is the value of $\widehat{X}$ which minimizes the mean squared error given $\mathcal{I}$, written as

$$E[\,(X - \widehat{X})^2 \mid \mathcal{I}\,] .$$

"Squared error loss" is the most mathematically convenient "loss function" and is reasonable in many situations. It is the one we will use. But it is not the only possibility. Another is . . .

Absolute Error Loss

Suppose that you will be forced to pay $|X - \widehat{X}|$ dollars if you guess $\widehat{X}$ and the actual value is X. This is called "absolute error loss".

Then the best strategy is: **guess the median**.

More precisely, the best forecast is the median of the conditional distribution of X given $\mathcal{I}$. This is the value which minimizes the mean absolute error given $\mathcal{I}$:

$$E\left(|X - \widehat{X}| \,\middle|\, \mathcal{I}\right) .$$

And there are still other loss functions.

For example, if you are forced to pay \$100 unless your guess $\widehat{X}$ is within ε of X (where ε is small), then your best guess is the **mode** of the conditional distribution.

For normal distributions **mean = median = mode** so that all three loss functions lead to the same "best" prediction.

But for skewed distributions or bimodal distributions, they give different predictions.

Luckily, for
ARIMA models with normally distributed random shocks, $(*)$ the future values we wish to predict have normal distributions, so all three loss functions give the same best prediction.

For situations like $(*)$, we summarize the forecasting procedure on the next page.

If the conditional distribution of X given the information $\mathcal{I}$ is:

Normal with mean $= E(X \mid \mathcal{I})$ and variance $= \mathrm{Var}(X \mid \mathcal{I})$,

Then

- The best forecast of X is $\widehat{X} = E(X \mid \mathcal{I})$
- The forecast error $X - \widehat{X}$ has variance $= \mathrm{Var}(X \mid \mathcal{I})$.
- A $100(1-\alpha)\%$ confidence interval for X is given by

$$(\widehat{X} - z_{\alpha/2}\,\mathrm{SE}\,,\ \widehat{X} + z_{\alpha/2}\,\mathrm{SE})$$

where $\mathrm{SE} = \sqrt{\mathrm{Var}(X \mid \mathcal{I})}$ is called the "standard error" of the forecast. (As usual, for 95% confidence use $z_{.025} = 1.96$.)

Computing Forecasts

Forecasts are just conditional expected values (or means), which obey rules like those of ordinary expected values. Here are some examples.

Let X, Y, Z be random variables, and a, b, c be constants. Let $\widehat{X} = E(X \,|\, \mathcal{I})$, $\widehat{Y} = E(Y \,|\, \mathcal{I})$, $\widehat{Z} = E(Z \,|\, \mathcal{I})$.

If $X = Y + Z$, then $\widehat{X} = \widehat{Y} + \widehat{Z}$.

If $X = aY + bZ + c$, then $\widehat{X} = a\widehat{Y} + b\widehat{Z} + c$.

And so on for summations involving any number of random variables.

Forecasting Future Values Given Observations Up To Time n

Suppose $\{z_t\}$ is a realization of a **known** ARIMA(p, d, q) process; we know the orders p, d, q and the values of all parameters.

Suppose we observe **all** the values z_t and a_t (the random shocks) up to time n. Call this set of information $\mathcal{I}_n$:

$$\mathcal{I}_n = \{z_n, z_{n-1}, z_{n-2}, \ldots, a_n, a_{n-1}, a_{n-2}, \ldots\}$$

(Note: The random shocks $a_n, a_{n-1}, a_{n-2}, \ldots$ are not directly observed, but can be calculated from $z_n, z_{n-1}, z_{n-2}, \ldots$ if the ARIMA process is invertible.)

Given $\mathcal{I}_n$, the forecast of z_{n+k} is $\widehat{z}_{n+k} = E(z_{n+k} \,|\, \mathcal{I}_n)$.

How de we compute these forecasts?

The forecasts $\widehat{z}_{n+1}$, $\widehat{z}_{n+2}$, $\widehat{z}_{n+3}$, ... may be computed sequentially using the facts:

- $\widehat{z}_t = E(z_t \,|\, \mathcal{I}_n) = z_t$ for $t \leq n$ (since $z_t \in \mathcal{I}_n$ for $t \leq n$).
- $\widehat{a}_t = E(a_t \,|\, \mathcal{I}_n) = a_t$ for $t \leq n$ (since $a_t \in \mathcal{I}_n$ for $t \leq n$).
- $\widehat{a}_t = E(a_t \,|\, \mathcal{I}_n) = 0$ for $t > n$. The future shocks (a_t for $t > n$) have mean zero and are **independent** of $\mathcal{I}_n$ (the past). Each new shock is independent of what has gone before. Therefore, knowing $\mathcal{I}_n$ does not help in predicting a_t for $t > n$, and the best we can do is "guess the mean" which is zero.

Example: Forecasting an AR(1) process

$$\begin{aligned} z_{n+1} &= C + \phi_1 z_n + a_{n+1} \\ \Rightarrow\ \widehat{z}_{n+1} &= C + \phi_1 \widehat{z}_n + \widehat{a}_{n+1} = C + \phi_1 z_n + 0 = C + \phi_1 z_n \end{aligned}$$

$$\begin{aligned} z_{n+2} &= C + \phi_1 z_{n+1} + a_{n+2} \\ \Rightarrow\ \widehat{z}_{n+2} &= C + \phi_1 \widehat{z}_{n+1} + \widehat{a}_{n+2} = C + \phi_1(C + \phi_1 z_n) + 0 \\ &= C + \phi_1 C + \phi_1^2 z_n \end{aligned}$$

$$\begin{aligned} z_{n+3} &= C + \phi_1 z_{n+2} + a_{n+3} \\ \Rightarrow\ \widehat{z}_{n+3} &= C + \phi_1 \widehat{z}_{n+2} + \widehat{a}_{n+3} = C + \phi_1(C + \phi_1 C + \phi_1^2 z_n) \\ &= C + \phi_1 C + \phi_1^2 C + \phi_1^3 z_n \end{aligned}$$

and so on.

Note: The process is a little simpler if we "mean center" (i.e., $\tilde{z}_t = z_t - \mu_z$) to get rid of the C's.

One-Step-Ahead Prediction Errors

The one-step-ahead prection error is

$$z_{n+1} - \widehat{z}_{n+1} = a_{n+1}\,,$$

which has variance

$$\mathrm{Var}(z_{n+1} - \widehat{z}_{n+1}) = \mathrm{Var}(a_{n+1}) = \sigma_a^2$$

(which is the same as $\mathrm{Var}(z_{n+1} \,|\, \mathcal{I}_n)$).
The 95% confidence interval for z_{n+1} is

$$\widehat{z}_{n+1} \pm 1.96\,\sigma_a\,.$$

We have shown these statements for AR(1) processes, but they are also true for any ARIMA process.

Long Range Forecasts

For an AR(1) process:

$$\widehat{z}_{n+k} = C + \phi_1 C + \phi_1^2 C + \cdots + \phi_1^{k-1} C + \phi_1^k z_n$$

For a stationary process (i.e., $|\phi_1| < 1$), as $k \to \infty$ we have

$$\begin{aligned} &\phi_1^k z_n \to 0 \\ &C + \phi_1 C + \phi_1^2 C + \cdots + \phi_1^{k-1} C \to \frac{C}{1 - \phi_1} = \mu_z \end{aligned}$$

since the sum above is a geometric series.

Thus, as $k \to \infty$,

$$\widehat{z}_{n+k} \to \mu_z$$

and it can also be shown that

$$\mathrm{Var}(z_{n+k} \,|\, \mathcal{I}_n) = \mathrm{Var}(z_{n+k} - \widehat{z}_{n+k}) \to \sigma_z^2$$

so that the 95% confidence interval for z_{n+k} is approximately

$$\mu_z \pm 1.96\, \sigma_z$$

for sufficiently large k.

These properties hold not just for stationary AR(1) processes, but for any stationary ARMA process.

Example: Forecasting an MA(2) process (given $\mathcal{I}_n$)

$$
\begin{aligned}
z_{n+1} &= C + a_{n+1} - \theta_1 a_n - \theta_2 a_{n-1} \\
\Rightarrow \widehat{z}_{n+1} &= C + \widehat{a}_{n+1} - \theta_1 \widehat{a}_n - \theta_2 \widehat{a}_{n-1} = C + 0 - \theta_1 a_n - \theta_2 a_{n-1}
\end{aligned}
$$

$$
\begin{aligned}
z_{n+2} &= C + a_{n+2} - \theta_1 a_{n+1} - \theta_2 a_n \\
\Rightarrow \widehat{z}_{n+2} &= C + \widehat{a}_{n+2} - \theta_1 \widehat{a}_{n+1} - \theta_2 \widehat{a}_n = C + 0 - 0 - \theta_2 a_n
\end{aligned}
$$

$$
\begin{aligned}
z_{n+3} &= C + a_{n+3} - \theta_1 a_{n+2} - \theta_2 a_{n+1} \\
\Rightarrow \widehat{z}_{n+3} &= C + \widehat{a}_{n+3} - \theta_1 \widehat{a}_{n+2} - \theta_2 \widehat{a}_{n+1} = C + 0 - 0 - 0 \\
&= C = \mu_z
\end{aligned}
$$

$\widehat{z}_{n+k} = \mu_z$ for all $k \geq 3$ (by the same argument).

Similarly, for an MA(q) process, $\widehat{z}_{n+k} = \mu_z$ for all $k \geq q + 1$.

Psi-weights and Confidence Interval Widths

Every ARIMA(p,d,q) or ARIMA$(p, d, q)(P, D, Q)_s$ process can be expressed in MA form using the psi-weights ψ_k.

For a **stationary** ARMA process, this has the form

$$z_t = \mu_z + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \psi_3 a_{t-3} + \cdots = \mu_z + \sum_{k=0}^{\infty} \psi_k a_{t-k}$$

where we define $\psi_0 = 1$. For the MA form of a general ARIMA process with $C = 0$, just delete μ_z above. (For non-stationary processes with $C \neq 0$, the formula is a little more complicated; μ_z gets replaced by a deterministic trend.)

If $p > 0$ or $P > 0$, then infinitely many MA terms are required.

• If the process is stationary ($d = 0$ and $D = 0$ and the AR-weights satisfy the appropriate stationarity conditions), then the psi-weights decay to zero, that is, $\psi_k \to 0$ as $k \to \infty$.

• If the process is non-stationary ($d > 0$ or $D > 0$ or the AR-weights violate the stationarity conditions), then the psi-weights do **not** decay to zero. For any lag m, no matter how large, there will be lags $k > m$ with "sizeable" values of ψ_k.

Let $e_n(k)$ denote the k-step-ahead prediction error:

$$e_n(k) = z_{n+k} - \widehat{z}_{n+k}$$

Define $\sigma[e_n(k)] = \sqrt{\mathrm{Var}(e_n(k))}$ (the "standard error"). The 95% confidence interval for z_{n+k} is $\widehat{z}_{n+k} \pm 1.96\,\sigma[e_n(k)]$.

Confidence Interval Widths are governed by the standard error

$$\sigma[e_n(k)] = \sigma_a \sqrt{1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{k-1}^2}\,.$$

For $k = 1$ this becomes

$$\sigma[e_n(1)] = \sigma_a\,.$$

Consequences:

For a stationary ARMA process, the confidence interval widths for long run forecasts converge to a limiting value.

For non-stationary ARIMA processes, the confidence interval widths continue to gradually increase and will (if you forecast far enough into the future) reach arbitrarily large values.

In other words, for stationary processes z_t, the standard error $\sigma[e_n(k)] \to$ a constant value (which equals σ_z) as $k \to \infty$. For nonstationary processes, $\sigma[e_n(k)] \to \infty$ as $k \to \infty$.