Forecasting: an introduction

Given data $X_0, \ldots, X_{T-1}$.

Goal: guess, or forecast, X_T or X_{T+r}.

There are a variety of *ad hoc* methods as well as a variety of statistically derived methods.

Illustration of *ad hoc* methods: exponentially weighted moving average (EWMA):

$$\hat{X}_T = \frac{X_{T-1} + aX_{T-2} + a^2 X_{T-3} + \cdots + a^{T-1} X_0}{c(a,T)}$$

where $c(a,T)$ makes it a weighted average:

$$c(a,T) = (1 - a^T)/(1-a).$$

For a near 1 almost using sample mean.

For a near 0 virtually using X_{T-1}.

Choose a to trade off desire to use lots of data against possibility that structure of series has changed over time.

Statistically based methods: use some measure of the size of $X_T - \hat{X}_T$

Mean Squared Prediction Error (MSPE): $\mathrm{E}([X_T - \hat{X}_T]^2)$ is the most common.

In general $\hat{X}_T$ is some function $f(X_0, \ldots, X_{T-1})$.

MSPE is minimized by

$$\hat{X}_T = \mathrm{E}(X_T | X_0, \ldots, X_{T-1})$$

Hard to compute for most X distributions.

For Gaussian processes the solution is the usual linear regression of X_T on the data, namely

$$\hat{X}_T = \mu_T + a_1(X_{T-1} - \mu_{T-1}) + \cdots a_T(X_0 - \mu_0)$$

where the coefficient vector a is given by

$$a = \mathrm{Cov}(X_T, (X_{T-1}, \ldots, X_0)^T) \\ \times \mathrm{Var}(X_{T-1}, \ldots, X_0)^{-1}$$

For large T computation difficult but there are some shortcuts.

Forecasting AR(p) processes

When the process is an AR the computation of the conditional expectation is easier:

$$\begin{aligned}
\hat{X}_T &= \mathsf{E}(X_T | X_0, \ldots, X_{T-1}) \\
&= E(\epsilon_T + \sum_{i=1}^{p} a_i X_{T-i} | X_0, \ldots, X_{T-1}) \\
&= \sum_{i=1}^{p} a_i X_{T-i}
\end{aligned}$$

For $r > 0$ we have the recursion

$$\begin{aligned}
&\mathsf{E}(X_{T+r} | X_0, \ldots, X_{T-1}) \\
&\qquad = E(\epsilon_{T+r} + \sum_{i=1}^{p} a_i X_{T+r-i} | X_0, \ldots, X_{T-1}) \\
&\qquad = \sum_{i=1}^{p} a_i \hat{X}_{T+r-i}
\end{aligned}$$

Note forecast into future uses current values where these are available and forecasts already calculated for other X's.

Forecasting ARMA(p, q) processes

An ARMA(p, q) can be inverted to be an infinite order AR process.

Then use method just given for AR.

But: now formula mentions values of X_t for $t < 0$.

In practice: truncate series, and ignore missing terms in forecast, assuming that the coefficients of these omitted terms are very small.

Remember each term is built up out of a geometric series for $(I - \alpha B)^{-1}$ with $|\alpha| < 1$.

More direct method:

$$\begin{aligned}\hat{X}_{T+r} &= \mathsf{E}(\epsilon_{T+r}|X) + \sum_{i=1}^{p} a_i \hat{X}_{T+r-i} \\ &\quad + \sum_{i=1}^{q} b_i \mathsf{E}(\epsilon_{T+r-i}|X)\end{aligned}$$

where conditioning "$|X$" means given data observed.

For $T+r-i \ge T$ conditional expectation is 0.

For $T+r-i < T$ need to guess value of ϵ_{T+r-i}.

The same recursion can be re-arranged to help compute $\text{E}(\epsilon_t|X)$ for $0 \le t \le T-1$, at least approximately:

$$\begin{aligned}\text{E}(\epsilon_t|X) &= X_t - \sum a_i X_{t-i} \\ &\quad + \sum b_i \text{E}(\epsilon_{t-i}|X)\end{aligned}$$

Recursion works backward; generally start recursion by putting

$$\hat{\epsilon}_t = 0$$

for negative t and then using the recursion.

Coefficients b are such that the effect of getting these values of ϵ wrong is damped out at a geometric rate as we increase t.

So: if we have enough data and the smallest root of the characteristic polynomial for the MA part is not too close to 1 then we will have accurate values for $\hat{\epsilon}_t$ for t near T.

Computed estimates of the epsilons can be improved by backcasting the values of ϵ_t for negative t and then forecasting and backcasting, etc.

Forecasting ARIMA(p, d, q) series

Suppose $Z = (I - B)^d X$ for X ARIMA(p, d, q).

Compute Z, forecast Z and reconstruct X by undoing the differencing.

For $d = 1$ for example we just have

$$\hat{X}_t = \hat{Z}_t + \hat{X}_{t-1}\,.$$

Forecast standard errors

Note: computations of conditional expectations used fact that a's and b's are constants – the true parameter values.

In practice: replace parameter values with estimates.

Quality of forecasts summarized by forecast standard error:

$$\sqrt{\mathsf{E}[(X_t - \hat{X}_t)^2]}\,.$$

We will compute this ignoring the estimation of the parameters and then discuss how much that might have cost us.

If $\hat{X}_t = \mathsf{E}(X_t|X)$ then $\mathsf{E}(\hat{X}_t) + \mathsf{E}(X_t)$ so that our forecast standard error is just the variance of $X_t - \hat{X}_t$.

First one step ahead forecasting for AR(1):

$$X_T - \hat{X}_T = \epsilon_T \,.$$

The variance of this forecast is σ_ϵ^2 so that the forecast standard error is just σ_ϵ.

For forecasts further ahead in time we have

$$\hat{X}_{T+r} = a\hat{X}_{T+r-1}$$

and

$$X_{T+r} = aX_{T+r-1} + \epsilon_{T+r}$$

Subtracting we see that

$$\begin{aligned} &\text{Var}(X_{T+r} - \hat{X}_{T+r}) \\ &\qquad = \sigma_\epsilon^2 + a^2\text{Var}(X_{T+r-1} - \hat{X}_{T+r-1}) \end{aligned}$$

so may calculate forecast standard errors recursively.

As $r \to \infty$ forecast variance converges to

$$\sigma_\epsilon^2/(1 - a^2)$$

which is simply the variance of individual Xs.

When forecasting a *stationary* series far into future, forecast standard error is just standard deviation of series.

General ARMA(p, q).

Rewrite process as infinite order AR

$$X_t = \sum_{s>0} c_s X_{t-s} + \epsilon_t$$

Ignore truncation of infinite sum in forecast:

$$X_T - \hat{X}_T = \epsilon_T$$

so one step ahead forecast standard error is σ_ϵ.

Parallel to the AR(1) argument:

$$X_{T+r} - \hat{X}_{T+r} = \sum_{j=0}^{r-1} c_{r-j}(X_{T+j} - \hat{X}_{T+j}) + \epsilon_{T+r} \, .$$

Errors on right hand side not independent of one another.

So: computation of variance requires either computation of covariances or recognition of fact that right hand side is a linear combination of $\epsilon_T, \ldots, \epsilon_{T+r}$.

Simpler approach: write process as infinite order MA:

$$X_t = \epsilon_t + \sum_{s>0} d_s \epsilon_{t-s}$$

for suitable coefficients d_s.

Treat conditioning on data as being effectively equivalent to conditioning on all X_t for $t < T$.

Effectively conditioning on ϵ_t for all $t < T$.

This means that

$$\begin{aligned} \mathsf{E}(X_{T+r} | X_{T-1}, X_{T-2}, \ldots) \\ &= \mathsf{E}(X_{T+r} | \epsilon_{T-1}, \epsilon_{T-2}, \ldots) \\ &= \sum_{s>r} d_s \epsilon_{T+r-s} \end{aligned}$$

and the forecast error is just

$$X_{T+r} - \widehat{X}_{T+r} = \epsilon_{T+r} + \sum_{s=1}^{r} d_s \epsilon_{T+r-s}$$

so that the forecast standard error is

$$\sigma_\epsilon \sqrt{1 + \sum_{s=1}^{r} d_s^2}\,.$$

Again as $r \to \infty$ this converges to σ_X.

ARIMA(p, d, q) process: $(I - B)^d X = W$ where W is ARMA(p, q).

Forecast errors in X can be written as a linear combination of forecast errors for W.

So forecast error in X can be written as a linear combination of underlying errors ϵ_t.

Example: ARIMA(0,1,0): $X_t = \epsilon_t + X_{t-1}$.

The forecast of ϵ_{T+r} is 0.

So forecast of X_{T+r} is

$$\hat{X}_{T+r} = \hat{X}_{T+r-1} = \cdots = X_{T-1}\,.$$

The forecast error is

$$\epsilon_{T+r} + \cdots + \epsilon_T$$

whose standard deviation is $\sigma\sqrt{r+1}$.

Notice that the forecast standard error grows to infinity as $r \to \infty$.

For a general ARIMA($p,1,q$) we have

$$\hat{X}_{T+r} = \hat{X}_{T+r-1} + \hat{W}_{T+r}$$

and

$$\begin{aligned} X_{T+r} &- \hat{X}_{T+r} \\ &= (W_{T+r} - \hat{W}_{T+r}) + \cdots + (W_T - \hat{W}_T) \end{aligned}$$

which can be combined with the expression above for the forecast error for an ARMA(p,q) to compute standard errors.

Software

S-Plus function `arima.forecast` can do forecasting. Use `predict.Arima` in **R**.

Comments

Effects of parameter estimation ignored.

In ordinary least squares when we predict the Y corresponding to a new x we get a forecast standard error of

$$\sqrt{Var(Y - x\hat{\beta})} = \sqrt{Var(\epsilon + x(\beta - \hat{\beta}))}$$

which is

$$\sigma\sqrt{1 + x(X^T X)^{-1} x^T}\,.$$

The procedure used here corresponds to ignoring the term $x(X^T X)^{-1} x^T$ which is the variance of the fitted value.

Typically this value is rather smaller than the 1 to which it is added.

In a 1 sample problem for instance it is simply $1/n$.

Generally the major component of forecast error is the standard error of the noise and the effect of parameter estimation is unimportant.

Prediction Intervals

In regression sometimes compute prediction intervals

$$\hat{Y} \pm c\hat{\sigma}_{\hat{Y}}$$

Multiplier c adjusted to make coverage probability P($\frac{|Y-\hat{Y}|}{c\hat{\sigma}} \le 1$) close to desired coverage probability such as 0.95.

If the errors are normal then we can get c by taking $t_{0.025,n-p}\sqrt{1 + x(X^T X)^{-1}x^T}$.

When the errors are not normal, however, the error in $Y - \hat{Y}$ is dominated by ϵ which is not normal so that the coverage probability can be radically different from the nominal.

Moreover, there is no particular theoretical justification for the use of t critical points.

However, even for non-normal errors the prediction standard error is a useful summary of the accuracy of a prediction.