Chapter 28

MSMYS2 Time Series And Forecasting

(28.1) Analysis Over Time

(28.1.1) Time Series & Trends

Much of statistical theory is concerned with independent and identically distributed random variables. What happens when this is not the case is, essentially, the subject of this chapter. Take for example the price of a stock: one expects the price tomorrow to be similar to that today, and that perhaps from the past predictions about the future can be made.

Perhaps the most prominent feature of time series data is an 'overall trend', closely followed by cyclicity. These become apparent when data is plotted, and plotting data is advisable as it can be very revealing. Some trends are 'fake', that is they arise randomly, perhaps to the variance of the data increasing over time—this would suggest divergence from a central level. Other trends are not, and these can be modeled with a function of the form $Y(t) = \mu(t) + X(t)$ where μ represents the trend and X is a stochastic process.

The lack of independence of the data means that it is somehow dependent on itself. Whereas two random variables may be correlated in the ordinary sense it should not now come as much of a surprise that for a time series the data is correlated with itself.

Definition 1 *For a time series $Y(t)$ define the autocovariance function*

$$\gamma(s,t) = \mathbb{E}\left(Y(s) - \mathbb{E}\,Y(s)\right)\left(Y(t) - \mathbb{E}\,Y(t)\right)$$

and the autocorrelation function

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}$$

Definition 2 *Let $Y(y)$ be a time series. $Y(t)$ is stationary if*

$$Y(t_1 + s) = Y(t_2 + s) = \cdots = Y(t_m + s)$$

for all m. The process is said to be second order stationary if

$$\gamma(s,t) = F(|t-s|) \quad \textit{or equivalently} \quad \gamma(t, t+k) = F(|k|)$$

As stationarity cannot usually be verified the second order stationarity definition is usually used. This definition means that the autocovariance is a function only of 'the gap'.

Returning to the model $Y(t) = \mu(t) + X(t)$ it is clear that X should be a stationary process such as a Normal random variable with mean 0 and variance σ_t^2. When working with dependent random variables it is usually fairly easy to calculate means. However, for variances recall that

$$\operatorname{var} XY = \operatorname{var} X + \operatorname{var} Y - 2 \operatorname{cov}(X, Y)$$

Smoothing Data

In order to observe trends in data it is advantageous to 'smooth' the data, in the hope of removing local variations and revealing global variations. A smooth line through the data may be plotted with the aid of a moving average.

Definition 3 *Let $y_1, y_2, \ldots, y_n$ be observations then*

$$s_t = \sum_{j=-p}^{p} w_j y_{t+j}$$

where t takes values $p+1, p+2, \ldots, n-p$ is said to be a moving average of order $2p+1$. Usually $w_j \geqslant 0 \forall j$ and $\sum_j w_j = 1$.

If a moving average is calculated with order equal to a seasonal period then cyclicity in the data will be smoothed out.

Another method to smooth data is differencing. This is usually represented in terms of the differencing operator D where

$$Dy_t = y_t - y_{t-1} \quad \text{and} \quad D^2 y_t = D(y_t - y_{t-1}) = y_t - 2y_{t-1} + y_{t-2}$$

Plotting differences will make apparent when a stationary process is reached, and at most D^2 tends to suffice.

Processes

As usual linearity is highly desirable. A general linear process may be expressed in the form

$$Y(t) = \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \ldots$$

where $\psi_0 = 1$ and $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$ are independent. Calculating the variance of this,

$$\operatorname{var} Y(t) = \sigma^2 \sum_{i=1}^{\infty} \psi_i$$

so it is necessary for this sum to converge for a linear process. Absolute convergence is also required. A useful choice for the ψs is, say $\psi_j = \phi^j$ for some $|\phi| < 1$, so the distant past has much less effect on the value if $Y(t)$ than the recent past.

Definition 4 *Let $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$ be independent, then the moving average process $Y \sim \mathrm{MA}\left(q\right)$ is defined as*

$$Y(t) = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

which is a stationary process.

Definition 5 *Let $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$ be independent, then the autoregressive process $Y \sim \mathrm{AR}\left(p\right)$ is defined as*

$$Y(t) = \phi_1 y_{t-1} + \phi_2 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

An autoregressive process is not necessarily stationary. It is of interest, therefore, as to under what conditions such a process is stationary. If $\mathbb{E}\,Y_{t-j} = 0$ for all $j \geqslant 1$ then clearly the process is stationary, but of course this is not generally the case. In the case of an AR (1) process

$$\begin{aligned} y_t &= \phi y_{t-1} + \varepsilon_t \\ &= \phi\left(\phi y_{t-2} + \varepsilon_{t-1}\right) + \varepsilon_t \quad \text{by substituting for } y_{t-1} \\ &= \sum_{j=0}^{J-1} \phi^j \varepsilon_{t-j} + \phi^J y_{t-J} \end{aligned}$$

Hence when $|\phi| < 1$ $\mathbb{E}\,Y_t \to 0$ as $J \to \infty$ because all the εs have mean 0. In the general case

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \\ \mathbb{E}\,Y_t &= \phi_1\,\mathbb{E}\,y_{t-1} + \phi_2\,\mathbb{E}\,y_{t-2} + \cdots + \phi_p\,\mathbb{E}\,y_{t-p} \\ \mu_t - \phi_1\mu_{t-1} - \cdots - \phi_p\mu_{t-p} = 0 \end{aligned}$$

where $\mu_i = \mathbb{E}\,Y_i$. This last equation is a difference equation[*] Looking for solutions of the form $\mu_i = Ax^i$ where A is some constant gives

$$\begin{aligned} Ax^t - A\phi_1 x^{t-1} - \cdots - A\phi_p x^{t-p} &= 0 \\ Ax^{t-p}\left(x^p + \phi_1 x^{p-1} - \cdots - \phi_p\right) &= 0 \end{aligned}$$

This yields a characteristic polynomial the solutions to which, $m_1, m_2, \ldots, m_p$, determine the solutions for the various values of μ for example $\mu_t = A_1 m_1^t + A_2 m_2^t + \cdots + A_p m_p^t$. In the case of a repeated $m_1 = m_2$ use

$$\mu_t = (A_1 + A_2 t) m_1^t$$

In the case of complex roots $m_1 = \alpha + i\beta$ and $m_2 = \alpha - i\beta$ use

$$\begin{aligned} \mu_t &= A_1(\alpha + i\beta)^t + A_2(\alpha - i\beta)^t \\ &= r^t C \cos\left(\theta t + D\right) \end{aligned}$$

Whatever roots are found it is clear that if all roots lie within the unit circle in the Argand plane then $\mu_t \to 0$ as $t \to \infty$. Hence a condition for stationarity has been determined.

Definition 6 *The lag operator (or backward shift operator) L has the effect $Ly_t = y_{t-1}$ and hence $L^k y_t = y_{t-k}$. The operator L^{-1} is called the lead operator (or forward shift operator) and has the effect $L^{-1}y_t = y_{t+1}$.*

Use of the lag operator can much simplify the expression of processes. While the autoregressive and moving average processes are useful they may be combined to form an ARMA (p, q) process. If $y_t \sim$ ARMA (p, q) then

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

Using the lag operator this may be written as

$$\begin{aligned} \phi(L) y_t &= \theta(L)\varepsilon_t \\ \text{where } \phi(L) &= 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p \\ \text{and } \theta(L) &= 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q \end{aligned}$$

[*] Difference equations behave rather like ordinary differential equations with constant coefficients.

The polynomials ϕ and θ are called the associated polynomials.

Similar to the autoregressive moving average process is the autoregressive integrated moving average process, ARMA (p, d, q). This is defined by

$$\phi(L)D^d y_t = \theta(L)\varepsilon_t$$

where $D = 1 - L$ is the difference operator.

Consider now an AR (p) process, $\phi(L)y_t = \varepsilon_t$. If ϕ could be inverted this could be expressed as an infinite moving average process $y_t = \psi(L)\varepsilon_t$, say. Using this

$$\begin{aligned} y_t &= \psi(L)\varepsilon_t \\ \phi(L)y_t &= \phi(L)\psi(L)\varepsilon_t \end{aligned}$$

from which it must be the case that $\phi(L)\psi(L) = 1$. This method as avoided the issue of the existence of $\phi^{-1}(L)$, though if it does exist is is easy to construct a circular argument.

Invertability

Thus far processes have been presented and their autocorrelation and autocovariance calculated. In reality it is more usual to receive data and then to ask "what process is this". The estimation of parameters is discussed in Section 28.2.

It is important, therefore, that given a sample autocorrelation and an MA (1) process different people should infer the same value for θ_1. Consider the MA (1) processes

$$Y_t = \varepsilon_t + \theta\varepsilon_{t-1} \qquad X_t = \varepsilon_t + \frac{1}{\theta}\varepsilon_{ti1}$$

These processes may now be 'inverted' to give

$$\begin{aligned} \varepsilon_t &= \frac{Y_t}{1+\theta L} = 1 - \theta L Y_t + \theta^2 L^2 Y_t - \dots \\ \varepsilon_t &= \frac{X_t}{1+\frac{1}{\theta}L} = 1 - \frac{1}{\theta}L X_t + \frac{1}{\theta^2}L^2 X_t - \dots \end{aligned}$$

Assuming stationarity and that $\mathbb{E}\, Y_t = 0$ the autocovariance and autocorrelation functions for the process in Y_t can now be found.

$$\begin{aligned} \gamma(0) &= \operatorname{var} Y_t = (1+\theta^2)\sigma^2 \\ \gamma(1) = \operatorname{cov}(Y_t, Y_{t-1}) = \mathbb{E}(Y_t Y_{t-1}) = \mathbb{E}\left(\varepsilon_t\varepsilon_{t-1} + \theta\varepsilon_{t-1}^2 + \theta\varepsilon_t\varepsilon_{t-2} + \theta^2\varepsilon_{t-1}\varepsilon_{t-2}\right) &= \theta\sigma^2 \\ \gamma(k) &= 0 \quad \text{for } k \geqslant 2 \end{aligned}$$

Dividing by $\gamma(0)$ now gives

$$\rho(0) = 1 \qquad \rho(1) = \frac{\theta}{1+\theta^2}$$

with $\rho(k) = 0$ for $k \geqslant 2$. However, if θ is replaced by $\frac{1}{\theta}$ the expression for ρ is unaltered. Now, suppose that $\hat{\rho}(1)$ is a sample autocorrelation. The equation $\rho(1)$ can then be solved for θ, and is a quadratic hence yielding two solutions $\hat{\theta}_1$ and $\hat{\theta}_2$, but which of these is correct? In some special cases the roots may be the reciprocal of eachother in which case it doesn't matter, but this is not generally the case. In order to resolve this take by definition $|\theta| < 1$ so that the sequence expansion for Y_t is convergent (and hence the process is invertible).

(28.1.2) Autocovariance And Autocorrelation For Processes

Moving Average Processes

A moving average process MA (q) is defined by

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

the autocovariance of this can be calculated as follows

$$\begin{aligned}\gamma(k) &= \mathbb{E}\,(Y_t - \mathbb{E}\,Y_t)\,(Y_{t+k} - \mathbb{E}\,Y_{t+k}) \\ &= \mathbb{E}\,Y_t Y_{t+k} \quad \text{but } \operatorname{cov}(X, Z) = \mathbb{E}\,(XZ) - \mathbb{E}\,X\,\mathbb{E}\,Z \\ &= \operatorname{cov}(Y_t, Y_{t+k})\end{aligned}$$

Now,

$$y_{t-k} = \varepsilon_{t-k} + \theta_1 \varepsilon_{t-k-1} + \theta_2 \varepsilon_{t-k-2} + \cdots + \varepsilon_{t-k-q}$$

Since $\mathbb{E}\,\theta_i \theta_j = \sigma^2$ this gives

$$\gamma(k) = \begin{cases} \sigma^2 \sum_{i=1}^{q-k} \theta_i \theta_{i+k} & \text{if } k \leqslant q \\ 0 & \text{if } k > q \end{cases} \tag{7}$$

Hence the autocorrelation, $\rho(k) = \frac{\gamma(k)}{\gamma(0)}$, is zero for $k > q$ which should be noted when viewing plots of sample autocorrelation, and such a feature may suggest the use of a moving average model.

Autoregressive Processes

An autoregressive process AR (p) is defined by

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

Stationarity is assumed and without further loss of generality assume $\mathbb{E}\,Y_i = 0$ for all i. By the independence of ε_i, $\mathbb{E}\,Y_t \varepsilon_t = \mathbb{E}\,\varepsilon_t^2 = \sigma^2$. Hence $\gamma(k) = \mathbb{E}\,Y_t Y_{t-k}$. Now,

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \quad \text{now multiply by } y_{t-k} \\ y_t y_{t-k} &= \phi_1 y_{t-1} y_{t-k} + \phi_2 y_{t-2} y_{t-k} + \cdots + \phi_p y_{t-p} y_{t-k} + \varepsilon_t y_{t-k} \quad \text{now take expected values} \\ \gamma(k) &= \phi_1 \gamma(k-1) + \phi_2 \gamma(k-2) + \cdots + \phi_p \gamma(k-p) \end{aligned}$$

Now, $\operatorname{var} Y_t = \mathbb{E}\,(Y_t^2) - (\mathbb{E}\,Y_t)^2$ and since $\mathbb{E}\,Y_t = 0$ this gives $\operatorname{var} Y_t = \gamma(0)$. Assuming this is not zero the autocovariance function can be divided by it for all values of k to give

$$\begin{aligned} 1 &= \phi_1 \rho_1 + \phi_2 \rho_2 + \cdots + \phi_p \rho_p \\ \rho(1) &= \phi_1 + \phi_2 \rho(1) + \phi_3 \rho(2) + \cdots + \phi_p \rho_{p-1} \\ \rho(2) &= \phi_1 \rho(1) + \phi_2 + \phi_3 \rho(1) + \cdots + \phi_p \rho(p-2) \\ \rho(3) &= \phi_1 \rho_2 + \phi_1 \rho_1 + \phi_3 + \phi_4 \rho_1 + \cdots + \phi_p \rho_{p-3} \\ &\vdots \\ \rho(p) &= \phi_1 \rho(p-1) + \phi_2 \rho(p-2) + \cdots + \phi_p \end{aligned} \tag{8}$$

These are the Yule-Walker equations, with the exception of the first. These p equations in p unknowns (the ϕs) can be used to estimate the values of ϕ from sample autocorrelations $\hat{\rho}_i$. Alternatively the autocorrelations of a known autoregressive process could be found. In matrix form they may be expressed as

$$\mathbf{\ae} = \mathbf{\OE} + \Phi\mathbf{\ae} \quad \text{or} \quad \mathbf{\OE} = (I - \Phi)^{-1}\mathbf{\ae}$$

Besides using linear algebra the kth Yule-Walker equation can also be solved as difference equation allowing any of the equations to be calculated without knowing the others. In

$$\rho(k) = \phi_1\rho(k-1) + \phi_2\rho(k-2) + \cdots + \phi_p\rho(k-p) \tag{9}$$

put $\rho(k) = Ac^{|k|}$ to give a pth order polynomial in c identical to the characteristic equation.

At this point attention is drawn to the difference between the characteristic equation and $\phi(L)$. The way the characteristic equation is constructed means that if

$$\phi L = \phi_1 L + \phi_2 L^2 + \cdots + \phi_p L^p$$

then the characteristic equation is

$$\phi_1 x^p + \phi_2 x^{p-1} + \cdots + \phi_p$$

The roots to these equations are closely related: they are reciprocals.

Solving the characteristic equation gives values for c so that

$$\rho(k) = \sum_{i=1}^{p} A_i c_i^{|k|}$$

Using $k = 0$ gives $\sum_{i=1}^{p} A_i = 1 = \rho(0)$. Now, as the Yule-Walker equations are known equation (9) can be substituted into them and hence the A_is determined. Hence when the unknown constants in equation (9) are found it is possible to find any of the autocorrelations.

Autoregressive Moving Average Processes

Generally an ARMA (p, q) process is of the form $\phi(L)y_t = \theta(L)\varepsilon_t$ and has rather a lot of parameters. It is worth noting that should ϕ and θ have a common linear factor then it can be cancelled to give an ARMA $(p-1, q-1)$ process.

Autocovariance

Given the involvement of polynomials it should come as little surprise that a generating function for autocovariance may be useful. Since an autoregressive process has an infinite moving average process representation it is sufficient to consider such representations and in doing so cover general ARMA (p, q) processes.

Theorem 10 *Where* $g(L) = \sum_{k=-\infty}^{\infty} \gamma(k)L^k$, $g(L) = \psi(L)\psi(L^{-1})\sigma^2$.

Proof. Extending equation (7) to an infinite moving average process;

$$\begin{aligned}\gamma(k) &= \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k} \\ \text{so } g(L) &= \sigma^2 \sum_{-\infty}^{\infty} \sum_{j=0}^{\infty} \psi_j \psi_{j+k} L^k \\ &= \sigma^2 \sum_{j=0}^{\infty} \sum_{k=-j}^{\infty} \psi_j \psi_{j+k} L^k \quad \text{because } \psi_j = 0 \text{ for } j < 0 \\ &= \sigma^2 \sum_{j=0}^{\infty} \sum_{j=0}^{\infty} \sum_{l=0}^{\infty} \psi_j \psi_l L^{l-j} \quad \text{where} l = k+j \\ &= \sigma^2 \left(\sum_{j=0}^{\infty} \psi_j L^{-j} \right) \left(\sum_{l=0}^{\infty} \psi_j L^l \right)\end{aligned}$$

$$\sigma^2 \psi(L) \psi(L^{-1})$$

□

(28.2) Parameter Estimation

(28.2.1) Effect Of Dependence

Estimating Mean And Variance

It is usual to estimate the mean and variance of a distribution with the estimators

$$\hat{\mu} = \overline{y} \quad \text{and} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

However, when the assumption of independence is removed the variance estimate may become very inaccurate because of the covariance structure of dependent random variables.

Suppose that $Y_1, Y_2, \ldots, Y_n$ are not independent, then $\mathbb{E}\,\overline{Y} = \mu$ still holds and so can be estimated by the sample mean. However, for the variance,

$$\begin{aligned}\operatorname{var} \overline{Y} &= \operatorname{var} \left(\frac{1}{n} \sum_{i=1}^{n} Y_i \right) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{cov} \left(Y_i, Y_j \right) \\ &= \frac{1}{n^2} \left(2\sigma^2 + 2\sigma^2 \sum_{k=i}^{n-1} (n-k) \rho_k \right) \quad \text{recalling that } \rho(Y_i, Y_j) = \frac{\operatorname{cov} \left(Y_i, Y_j \right)}{\sqrt{(\operatorname{var} Y_i)(\operatorname{var} Y_j)}}\end{aligned}$$

The summation term here may make the variance of $\overline{Y}$ much larger, or perhaps much smaller, than $\frac{\sigma^2}{n}$, so clearly s^2 cannot be used as an estimator. The size of this change is dependent on the autocorrelation function, ρ_k.

Calculating the variance in practise is just a case of substituting in some numbers. For example the MA (1) process

$$y_t = \varepsilon_t - \frac{1}{2} \varepsilon_{t-1}$$

gives autocorrelation function

$$\gamma(k) = \begin{cases} \frac{5}{4}\sigma^2 & \text{if } k = 0 \\ \frac{-1}{2}\sigma^2 & \text{if } |k| = 1 \end{cases}$$

$$\text{so } \rho_1 = \frac{\gamma(1)}{\sqrt{\gamma(0)\gamma(0)}} = \frac{-2}{5}$$

All other values of the autocorrelation are zero as $\gamma(k) = 0$ for $|k| \geqslant 2$. Hence substituting

$$\operatorname{var} \overline{Y} = \frac{1}{n^2}\left(n\sigma^2 + 2\sigma^2 \frac{-2(n-1)}{5}\right) \approx \frac{3\sigma^2}{5n} \text{ by using } \frac{n-1}{n} \approx 1$$

Similarly for an AR (1) process, $y_t = \phi y_{t-1} + \varepsilon_t$ the autocovariances are given by $\rho_k = \phi^{|k|}$

Sample Autocovariance & Autocorrelation

Besides estimating the mean and variance the autocovariance can be estimated by

$$\gamma^*(k) = \frac{1}{n-k}\sum_{t=0}^{n-k}(y_{k+t} - \mu)(y_t - \mu)$$

Replacing μ by its estimate, $\overline{y}$, produces a small bias which tends to zero as n increases. A different estimator is

$$c(k) = \frac{1}{n}\frac{1}{n}\sum_{t=0}^{n-k}(y_{k+t} - \overline{y})(y_t - \overline{y})$$

This has a smaller mean square error and is easier to calculate due to the simpler factor. Now, as k increases there are less and less data from which the autocovariance estimate can be calculated. For values of k that are near n the estimate is, therefore, very inaccurate. Furthermore, assessing such an estimator by calculating its variance is a rather unwieldy process, involving the expected value of a ratio. While asymptotic approximations may be found this is still, essentially, a job for a computer.

Having observed data the objective is to find a suitable model. Instantly this suggests the use of a hypothesis test. The null hypothesis is that the data are independent, and there is no time series structure. The sample autocorrelation, $r_k = \frac{c(k)}{c(0)}$ can be used for this. The expected value of r_k is approximately zero, and the variance is approximately $\frac{1}{n}$. The interval $\left(\frac{-1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right)$ is an approximate 95% confidence interval, and so values of r_k lying outside this are sought.

Testing For Randomness

In conducting hypothesis tests for independence it is important to remember that while independence implies (theoretical) uncorrelation, the reverse implication is not true. Nothing can be deduced, therefore, if such a hypothesis test does not reject the null hypothesis.

It is usual to test using r_0 as any model is likely to have a high correlation with the first of past observations. Choosing to test at other lags suggests at least part of the model is already known, though: under what justification does one test r_4, say, if the model is unknown?

The Van-Neumann Ratio

The Van-Neumann ration is defined as

$$\mathrm{VNR} = \frac{n}{n-1} \frac{\sum_{k=2}^{n} (y_k - y_{k-1})^2}{\sum_{k=1}^{n} (y_k - \overline{y})^2} \approx 2(1 - r_1) \tag{11}$$

From the approximation observe that

$$\mathrm{VNR} \to 2 \text{ as } r_1 \to 0 \qquad \mathrm{VNR} \to 0^+ \text{ as } r_i \to 1^- \qquad \mathrm{VNR} \to 4^- \text{ as } r_1 \to -1^-$$

From this a number of tests may be devised.

- Testing for positive correlation. Take as the null hypothesis that the data are independent. The alternative is positive correlation, so reject if the statistic is small.
- Testing for negative correlation. Take as the null hypothesis that the data are independent. The alternative is negative correlation, so reject if the statistic is large.

The reasoning for this test is made clear by looking at the graph of the approximate van Neumann ratio. Note that for large values of n the ratio has an approximate

$$\mathcal{N}\left(\frac{2n}{n-1}, \frac{4}{n}\right)$$

distribution.

The Portmanteau Test Statistic

The more general Portmanteau test was proposed by Box and Pierce in 1970 and can be used to test whether sample autocorrelations are non-zero. For an ARMA (p, d, q) model the statistic is

$$Q = (n - d) \sum_{i=1}^{k} r_i^2$$

where r_i is the ith sample autocorrelation. This statistic was thought to have a χ^2_{k-p-q} distribution, becoming large when the model is inappropriate. However, it is not very powerful and a better Portmanteau statistic is

$$Q^* = n(n+2) \sum_{i=1}^{k} \frac{r_i^2}{n-k}$$

which is larger than Q (suggesting the power of the test is greater).

The Portmanteau test uses autocorrelations to test whether the sample is nothing more than noise. If some of the autocorrelations are high then the sample may have some structure and the Portmanteau statistic will be large. It may be useful to consider many Portmanteau statistics, summing to 12, then to 24, etc. However, if the first 12 are tested and a conclusion drawn the test for the first 24 must surely be somehow conditional. Nevertheless, such tests are common.

(28.2.2) Estimating Parameters

Once the form of a model has been decided upon, the next stage is to estimate the parameters of the model.

Moments

The method of moments[†] for parameter estimation simply estimates moments by the sample moments. For an AR (p) process there are the p values for ϕ to estimate, and the variance of the innovation, σ^2. The Yule-Walker equations (equations (8)) provide a method by which suitable estimates may be found. The equations **æ** = **Œ** + Φ**æ** can be solved for **Œ** when **æ** is replaced by **r**, the sample autocorrelations. The quantity

$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2$$

estimates the variance of the process Y_t, but not σ^2.

Recall the problem of invertability, Section 28.1.1, that when faced with an option about the value of a parameter the one with size less than 1 should be chosen. For an MA (1) process it has already been shown that

$$\rho_1=\frac{\theta}{1+\theta^2}$$

However, certain values of the sample autocorrelation may produce complex roots in which case it is not really possible to estimate the parameter. For a MA (q) process the equations to solve are rather unwieldy;

$$\rho_k=\begin{cases}\frac{\sum_{i=0}^{q-k}\theta_i\theta_{i+k}}{\sum_{i=0}^{q}\theta_i^2} & \text{for } 1\leqslant k\leqslant 1\\ 0 & \text{otherwise}\end{cases}$$

These equations are very non-linear, so should they need solving numerical methods would need to be used.

Conditional Least Squares

With least squares estimation parameters are chosen so that the sum of the squares of the differences between the data and the values predicted by the model—the expression for ε_i—is minimised. Hence

$$S^*=\sum_{t=p+1}^{n}\big((y_t-\mu)-\phi_1(y_{t-1}-\mu)-\phi_2(y_{t-2}-\mu)-\cdots-\phi_p(y_{t-p}-\mu)\big)^2$$

is to be minimised for an AR (p) process. This gives

$$\begin{aligned}\frac{\partial S^*}{\partial\mu}&=0=-2(1-\phi_1-\cdots-\phi_p)(n-p-1)\sum_{t=p+1}^{n}\big((y_t-\mu)-\phi_1(y_{t-1}-\mu)-\cdots-\phi_p(y_{t-p}-\mu)\big)\\ 0&=-2\sum_{t=p+1}^{n}\big(y_t-\phi_1y_{t-1}-\cdots-y_{t-p}\phi_{t-p}\big)-2(n-p-1)\mu(1-\phi_1-\cdots-\phi_p)\\ \hat{\mu}&=\frac{1}{(n-p-1)(1-\phi_1-\cdots-\phi_p)}\sum_{t=p+1}^{n}y_t-\phi_1y_{t-1}-\cdots-y_{t-p}\phi_{t-p}\end{aligned}$$

[†] The first moment is the mean and the second is the variance.

Observe that $\hat{\mu}$ is approximately $\overline{y}$. As $\overline{y}$ is much simpler it is preferred as the estimator for the mean of the process. For the other parameters,

$$\begin{aligned}\frac{\partial S^*}{\partial \phi_i} = 0 = -2 \sum_{t=p+1}^{n} (y_i - \overline{y}) \left((y_t - \mu) - \phi_1 (y_{t-1} - \mu) - \phi_2 (y_{t-2} - \mu) - \cdots - \phi_p (y_{t-p} - \mu) \right) \\ 0 = \operatorname{cov}(y_t, y_{t-i}) - \phi_1 \operatorname{cov}(y_{t-1}, y_{t-i}) - \cdots - \phi_{i-1} \operatorname{cov}(y_{t-1}, y_{t-i}) - \phi_i \operatorname{var} y_i - \\ - \phi_{i+1} \operatorname{cov}\left(y_{y-i-1}, y_i\right) - \cdots - \phi_p \operatorname{cov}(y_{t-p}, y_i)\end{aligned}$$

Dividing through by $\operatorname{var} y_i$ gives one of the Yule-Walker equations, and producing such equations for each i produces a set of p linear equations in p unknowns—the ϕs. Note the covariance stated above is the sample covariance, so this produces the same estimates as the method of moments.

For a moving average process finding least squares estimates is not so easy. For an MA (1) process assume $\varepsilon_0 = 0$. The error for the ith datum is then given by $\varepsilon_i = y_i - \theta \varepsilon_{i-1}$ so that

$$S^* = \sum_{i=1}^{n} \varepsilon_i^2$$

this is a polynomial of order n in θ and must be minimised for $\theta \in (-1, 1)$, which is best done using numerical methods.

Maximum Likelihood

Maximum likelihood works by taking a joint density function evaluated at the observed data and treating it as a function of the parameters.

Take for example an AR (1) process $y_t - \mu = \phi(y_{t-1} - \mu) + \varepsilon_t$ where

$$\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right) \text{ so } f(\varepsilon_t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\varepsilon_t^2}{2\sigma^2}\right)$$

The joint distribution of the errors is then the product of these. Conditioning is now done on y_1, which essentially means that it is necessary to know y_1 or assume some value for it. The joint density function for the data is then given by changing variables from ε to y in the above. Hence

$$f(y_2, y_3, \ldots, y_n \mid y_1) = \left(2\pi\sigma^2\right)^{\frac{-(n-1)}{2}} \exp\left(\frac{-1}{2\sigma^2} \sum_{t=2}^{n} \left((y_t - \mu) - \phi(y_{t-1} - \mu)\right)^2\right)$$

Now, y_1 is Normally distributed, has mean μ, and variance $\frac{\sigma^2}{1-\phi^2}$ as is easily shown. This gives a probability density function for y_1

$$f(y_1) = \frac{\sqrt{1-\phi^2}}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y_1 - \mu)^2(1-\phi^2)}{2\sigma^2}\right)$$

Multiplying this with the conditional density function already found gives the joint density function required: the likelihood.

$$\begin{aligned}L(\mathbf{y}, \phi, \mu) = \frac{\sqrt{1-\phi^2}}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_1 - \mu)^2(1-\phi^2)}{2\sigma^2} - \frac{1}{2\sigma^2} \sum_{t=2}^{n} \left((y_t - \mu) - \phi(y_{t-1} - \mu)\right)^2\right) \\ l = \frac{1}{2} \ln(1-\phi^2) - \frac{n}{2} \ln(2\pi\sigma^2) + \frac{-(y_1 - \mu)^2(1-\phi^2)}{2\sigma^2} - \frac{1}{2\sigma^2} \sum_{t=2}^{n} \left((y_t - \mu) - \phi(y_{t-1} - \mu)\right)^2\end{aligned}$$

Now, the third term is in the form of $t = 1$ in the summation, putting $y_0 = y_1$. This simplifies to the unconditional sum of squares so that

$$= \frac{1}{2}\ln(1-\phi^2) - \frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}S(\mu,\phi)$$

This can now be differentiated to find parameter estimates. As may be expected $\hat{\sigma}^2 = \frac{1}{n}S(\hat{\mu},\hat{\phi})$ though the equations for $\hat{\mu}$ and $\hat{\phi}$ must be solved numerically.

(28.2.3) Quality Of Estimates

Variance Of Estimates

From a maximum likelihood perspective parameters $\mathbf{ff}$ are estimated by solving $\frac{\partial l}{\partial \mathbf{ff}} = \mathbf{0}$ and the central limit theorem ensures that $\sqrt{n}(\hat{\mathbf{ff}} - \mathbf{ff})$ is asymptotic to a multivariate Normal distribution with the information matrix being asymptotically proportional to the variance-covariance matrix.

Residuals

HAving fitted an ARMA (p,q) model it is necessary to test its adequacy which is usually done by comparing the estimated residual $\hat{\varepsilon}_t$ and comparing to ε_t. For an AR (p) process

$$\begin{aligned} y_t - \mu &= \sum_{i=1}^{p}\phi_i(y_{t-i}-\mu) + \varepsilon_t \\ \text{so } \hat{\varepsilon}_t &= (y_t - \mu) - \sum_{i=1}^{p}\hat{\phi}_i(y_{t-i}-\hat{\mu}) \end{aligned}$$

This can be calculated for $p + 1 \leqslant t \leqslant n$.

For an MA (q) process

$$\begin{aligned} y_t &= \mu + \varepsilon_t + \sum_{i=1}^{q}\theta_q\varepsilon_{t-i} \\ \hat{\varepsilon}_1 &= y_1 - \hat{\mu} \\ \hat{\varepsilon}_2 &= y_2 - \hat{\mu} - \hat{\theta}_1\hat{\varepsilon}_1 \\ &\vdots \\ \hat{\varepsilon}_n &= y_n - \hat{\mu} - \sum_{i=1}^{q}\hat{\theta}_i\hat{\varepsilon}_{n-j} \end{aligned}$$

However, the first q terms are discarded as many terms are missing from their estimation. While this has an effect on the last $n - q$ terms these are estimated by the correct number of past terms and so are more reliable.

For an ARMA (p,q) process the result is simply a combination of the above 2 so that

$$\hat{\varepsilon}_t = y_t - \hat{\mu} - \sum_{i=1}^{p}\hat{\phi}_i(y_{t-i}-\hat{\mu}) - \sum_{i=1}^{q}\hat{\theta}_i\hat{\varepsilon}_{t-j}$$

However, the first few terms must be disregarded. The number dropped is the larger of p and q.

Having calculated the residuals it is now possible to perform a Portmanteau test.

(28.2.4) Prediction

Minimum Mean Square Error Estimation

Having estimated parameters it is now desirable to use the model to estimate the future. Consider first of all prediction when the parameters are known.

Given $y_1, y_2, \ldots, y_n$ predict l steps ahead 'optimally' by the value

$$\tilde{y}_{T+l|T} = \mathbb{E}\left(y_{T+l|T}\right)$$

where "$y_{T+l|T}$" means that values of y up to y_T are known. The error in prediction for some general predictor $\hat{y}$ is $y_{T+l|T} - \hat{y}_{T+l|T}$ so

$$\begin{aligned} y_{T+l|T} - \hat{y}_{T+l|T} &= y_{T+l|T} - \mathbb{E}\, y_{T+l|T} - \left(\hat{y}_{T+l|T} - \mathbb{E}\, y_{T+l|T}\right) \\ \text{MSE}\left(\hat{y}_{T+l|T}\right) &= \operatorname{var} y_{T+l|T} + \left(\hat{y}_{T+l|T} - \mathbb{E}\, y_{T+l|T}\right)^2 \end{aligned}$$

So $\hat{y} = \tilde{y}$ is optimal in the sense that the mean square error is minimised.

Actually calculating an estimate in practise is quite easy as it is just a case of calculating expectation. Either ε_i is known in which case $\mathbb{E}\,\varepsilon_i = \varepsilon_i$ or it is not known and so $\mathbb{E}\,\varepsilon_i = 0$. Similarly if y_i is known then $\mathbb{E}\, y_i = y_i$ whereas if it is not known then y_i must be substituted for using the autoregressive process given so that the index can be reduced to give values of y that are known.

To predict many steps ahead it is necessary to predict all previous steps ahead, which may be expressed in the relationship $y_{T+l} = f(\mathbf{y}_T + {}''_T) + \varepsilon_{T+1}$.

Having predicted values the prediction error is of interest. Finding the infinite moving average representation of the general ARMA (p, q) process and the prediction for it,

$$\begin{aligned} y_{T+l} &= \sum_{i=0}^{\infty} \phi_i y_{T+l-i} \\ &= \sum_{j=l}^{1} \psi_j \varepsilon_{T+l-j} + \sum_{j=l}^{\infty} \psi_j \varepsilon_{T+l-j} \\ &= \sum_{j=0}^{l} \psi_{l-j} \varepsilon_{T+j} + \sum_{j=0}^{\infty} \psi_{l+j} \varepsilon_{T-j} \quad \text{putting } j \mapsto l-j \text{ and } j \mapsto j+l \end{aligned}$$

Now taking conditional expectation (conditional on T) the second summation is completely determined and gives the (minimum) mean square error of y_{T+l}, which is also the prediction of y_{T+l}. The other summation is therefore the error in prediction. The mean square error of this prediction error is

$$\mathbb{E}\left(y_{T+l} - \tilde{y}_{T+l|T}\right)^2 = \sum_{j=1}^{l} \psi_{l-j}^2 \sigma^2 = \sigma^2\left(1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{l-1}^2\right)$$

If the ε_is are Normally distributed then

$$\tilde{y}_{T+l|T} \pm 1.96\sigma\sqrt{1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{l-1}^2}$$

is a 95% prediction interval for y_{T+l}.

Using Estimated Parameters To Make Predictions

Without knowledge of what the model parameters are, or even what the model is, it would be expected that the predictions will be less precise. The same estimation process is used, but the estimate is now called $\tilde{y}^*_{T+l|T}$ and is calculated using the estimates ˆ and **Œ̂** for ` and **Œ**.

Consider the simple AR (1) case where $\hat{y}_{T+l|T} = \phi^l y_T$ and $\tilde{y}^*_{T+l|T} = \hat{\phi}^l y_T$. The prediction error is

$$\begin{aligned} y_{T+l} - \tilde{y}^*_{T+l|T} &= \left(y_{T+l} - \tilde{y}_{T+l|T}\right) + \left(\tilde{y}_{T+l|T} - \tilde{y}^*_{T+l|T}\right) \\ &= \left(y_{T+l} - \tilde{y}_{T+l|T}\right) + y_T \left(\hat{\phi}^l - \phi^l\right) \\ \mathrm{MSE}\left(\tilde{y}^*_{T+l|T}\right) &= \mathrm{MSE}\left(\tilde{y}_{T+l|T}\right) + y_T^2\, \mathbb{E}\left(\hat{\phi}^l - \phi^l\right)^2 \end{aligned}$$

Predicting many steps ahead requires many parameter estimates. Consider predicting only 1 step ahead, then $\mathbb{E}\left(\hat{\phi}^l - \phi^l\right)^2 = \operatorname{var}\hat{\phi}$.

Hypothesis Testing

Likelihood ratio tests can be used to test the relative worth of various models. As usual the models must be nested, but further restrictions apply as an ARMA (1, 1) cannot be tested against an ARMA (2, 2) because of difficulties estimating the extra parameters from the null hypothesis. Note also that the models ARMA (1, 2) and ARMA (2, 1) are not nested.

The likelihood ratio test statistic

$$\lambda = \frac{L(\hat{\alpha}_0)}{L(\hat{\alpha}}$$

is calculated where $\hat{\alpha}_0$ is the estimate of α (or **ff**) constrained to the conditions of the null hypothesis. $\hat{\alpha}$ is the maximum likelihood estimate. $-s \ln \lambda$ has an approximate χ^2_m distribution where there are m constraints in the null hypothesis.

Sums Of ARMA Processes

The sum of two uncorrelated ARMA processes is an ARMA process known as the reduced form. Say X_t is an ARMA (p_1, q_1) process and that Y_t is an ARMA (p_2, q_2) process. Then $Z_t = X_t + Y_t$ is an ARMA (p, q) process where $p \leqslant p_1 + p_2$ and $q \leqslant \max\{p_1 + q_2, p_2 + q_1\}$.

Suppose that X_t is the process $\phi_x(L)x_t = \theta_x(L)\varepsilon_t$ that that Y_t is the process $\phi_y(L)y_t = \theta_y \eta_t$. Each of these expressions is multiplied through by the autoregressive part of the other. Hence

$$\begin{aligned} \phi_x(L)x_t = \theta_x(L)\varepsilon_t &\mapsto \phi_y(L)\phi_x(L)x_t = \phi_y(L)\theta_x(L)\varepsilon_t \\ \phi_y(L)y_t = \theta_y(L)\eta_t &\mapsto \phi_x(L)\phi_y(L)y_t = \phi_x(L)\theta_y(L)\eta_t \end{aligned}$$

Now adding,

$$\phi_x(L)\phi_y(L)(x_t + y_t) = \phi_y(L)\theta_x(L)\varepsilon_t + \phi_x(L)\theta_y(L)\eta_t$$

The left hand side is at most an AR $(p_1 + p_2)$ process, and from the right hand side it is evident that $q \leqslant \max\{p_1 + q_2, p_2 + q_1\}$.

The inequalities are used because if $\phi_x(L)$ and $\phi_y(L)$ have common factors then only one copy of the factor is used in the definition of $\phi_z(L) = \phi_x(L)\phi_y(L)$. Therefore if $\phi_x(L) = \phi_y(L)$ then $\phi_z(L) = \phi_x(L) = \phi_y(L)$.

(28.3) Analysis By Frequency

(28.3.1) Modelling Time Series Using Functions

An alternative way to model time series it to use some kind of function of time. The functions may be chosen to represent trends and cyclicity in the data, as is discussed in Section 28.3.2.

Suppose that a time series Y_t has a periodic component of known frequency, then it may be modelled by

$$\begin{aligned} Y_t &= R\cos(\omega t + \theta) + \varepsilon_t \\ &= \alpha\cos(\omega t) + \beta\sin(\omega t) + \varepsilon_t \end{aligned}$$

where ω is the frequency (in radians), R is the amplitude, and θ is the phase. An alternative parameterisation takes $f = \frac{\omega}{2\pi}$ which is the number of cycles per unit time, giving $\frac{1}{f}$ as the wavelength. Linear regression can be used to find α and β which are assumed to be independent random variables with expected value 0, hence giving a stationary process. Combining different cyclic components gives

$$Y_t = \sum_{j=1}^{k} R_j\cos(\omega_j t + \theta_j) + \varepsilon_t$$

which is stationary whenever the R_j are uncorrelated random variables with mean 0, or when the θ_j are uniform on the interval $(0, 2\pi)$.

Spectra

The power spectral distribution function $F(w)$ is defined for a stochastic process with autocovariance function $\gamma(k)$ such that

$$\gamma(k) = \int_0^{\pi} \cos(wk)\,\mathrm{d}F(w)$$

If F is differentiable then where $f(w) = \frac{\mathrm{d}F}{\mathrm{d}w}$ is the spectral density function this may be rewritten as

$$\gamma(k) = \int_0^{\pi} \cos(wk) f(w)\,\mathrm{d}w$$

It is possible to relate $f(w)$ to $\gamma(k)$ by the functions

$$\begin{aligned} f(w) &= \frac{1}{2\pi}\left(\gamma(0) + 2\sum_{k=1}^{\infty}\gamma(k)\cos(wk)\right) \\ &= \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\gamma(k)e^{-iwk} \end{aligned}$$

These expressions are called the spectra, and the second can be calculated by replacing L with e^{ikw} in the autocovariance generating function.

For an MA (1) process $y_t = \varepsilon_t + \theta\varepsilon_{t-1}$ so

$$\begin{aligned} f(w) &= \frac{1}{2\pi}\left((1+\theta^2)\sigma^2 + 2\theta\sigma^2\cos w\right) \\ &= \frac{\sigma^2}{2\pi}\left(1 + \theta^2 + 2\theta\cos w\right) \end{aligned}$$

For an AR (1) process

$$\begin{aligned} f(w) &= \frac{1}{2\pi}\left(\frac{\sigma^2}{1-\phi^2} + 2\sum_{k=1}^{\infty}\frac{\phi^{|k|}}{1-\phi^2}\cos wk\right) \\ &= \frac{\sigma^2}{2\pi(1-\phi^2)}\left(1 + 2\sum_{k=1}^{\infty}\phi^{|k|}\cos wk\right) \end{aligned}$$

(28.3.2) Time Series Decomposition

The decomposition of time series was developed in the 1920s, and as such can be done without the aid of a computer (unlike most of the preceding material). A time series is modelled as a composition of pattern and error, and the pattern may consist of numerous components such as overall trend, seasonality, etc. Such models may be either additive or linear, so

$$\begin{aligned} X_t &= I_t + T_t + C_t + E_t \\ X_t &= I_t T_t C_t E_t \end{aligned}$$

where I is a seasonal index, T is overall trend, C is cyclicity, and E is error. The various components are gradually deduced then removed from the data to allow further analysis. The following discussion is for the multiplicative model, and by replacing division by subtraction is readily modified for the additive model.

First of all use linear regression to fit an overall trend. In exceptional circumstances a quadratic trend may be used. This will give a linear model for the data which is likely to fit poorly. Divide each datum by the corresponding fitted value, hence 'detrending' the data.

The seasonal indices are obtained by dividing the detrended data by a moving average of length equal to the seasonal period. For each season the centre of the distribution for that season is estimated—Minitab uses the median. This value is taken as the seasonal index and each datum may be divided by the corresponding seasonal index in order to obtain the seasonally adjusted data. It is usual to adjust the seasonal indices so that their average is 1. (0 for the additive model.)

The choice of multiplicative or additive model is often down to personal preference. However, note that a multiplicative model is suggested when the size of a seasonal pattern is proportional to the trend.