# Chapter 3

# Estimation, Model Fitting and Prediction for Time Series

## 3.1 Introduction

In the previous chapter we described in some detail the family of mathematical models that are used to model the processes that produce real–world time series. No mention was made of the data in that chapter, as the models exist in the world of mathematics. Such models have proved themselves to be of good utility, in that they provide reasonably sound approximations to the sort of data that often occur in practice.

Our first aim in this chapter is to marry the right theoretical model to the data in hand. How might we decide which of our models will be a good "fit" for the data we are analysing? The methods, as we shall see, are not entirely prescriptive, and occasionally more than one model may seem acceptable.

Having decided on a suitable model for a time series, it is often the case that we wish to *forecast* some future values. Much interest, particularly in financial markets, rests on accurately predicting what is to happen in the near future.

## 3.2 Estimation in the time domain

We outline here the basic theory of how one might estimate parameters of model from the data, which we denote $x(1), \ldots, x(N)$. We will assume that the series in question is either stationary to begin with, or has been suitably

pre-processed (say by removing of trend and/or seasonal variation) to look stationary.

In essence we are trying to match the data with the best mathematical model. The models we consider can all be thought of as ARMA$(p, q)$ processes, and we would aim to fit a simple, *parsimonious* model where possible. We denote the underlying model – which may be thought of as the *data generating process* – by $X(t)$, and denote the mean, variance and autocovariance at lag $k$ by $\mu$, $\sigma_X^2$ and $\gamma(k)$ respectively.

All the models under consideration have parameters that require *estimation* before they can be fitted. To begin with, we discuss how the acf, acvf, and mean of a model can be estimated using the data.

### 3.2.1 Properties of the sample acvf and acf

We can estimate the acvf $\gamma(k)$ of a time series model from the data by using the sample autocovariance function defined in chapter 1, namely

$$c_k = \frac{1}{N}\sum_{t=1}^{N-k}\left(x(t)-\bar{x}\right)\left(x(t+k)-\bar{x}\right).$$

This is a sensible choice, and actually is a very good estimator of the underlying function $\gamma(k)$ when $N$ is large. However it is not unbiased, as

$$E(c_k) = \gamma(k) + O\left(\frac{1}{N}\right),$$

meaning that the bias in $c_k$ from a sample of size $N$, $B_N$ say, is such that there is a number $M$ such that

$$N|B_N| \leq M$$

for all $N$. However $c_k$ is asymptotically unbiased, since clearly from the above

$$\lim_{N\to\infty} E(c_k) = \gamma(k).$$

One way to reduce the bias is to *jack-knife:* split the data series into halves, and work out the sample acvf at lag $k$ in each, giving $c_{k1}$ and $c_{k2}$ say. Then let

$$\bar{c}_k := 2c_k - \frac{(c_{k1}+c_{k2})}{2},$$

with $c_k$ the acvf from the full sample as before. It can be shown that

$$E\left(\bar{c}_k\right) = \gamma\left(k\right) + O\left(\frac{1}{N^2}\right),$$

and so $\bar{c}_k$ is somewhat better than $c_k$. That said, few if any software packages use this estimator. Neither is another alternative,

$$c'_k := \frac{1}{N-k}\sum_{t=1}^{N-k}\left(x\left(t\right) - \bar{x}\right)\left(x\left(t+k\right) - \bar{x}\right)$$

greatly favoured, though it tends to be of smaller bias than $c_k$ under certain conditions. The estimator $c_k$ as defined wins out typically since it has some useful properties when estimating in the *frequency domain,* as considered later in the course.

In fact, it can be shown that successive values of $c_k$ tend to be highly correlated, something which should be kept in mind when examining a correlogram. Specifically,

$$\mathrm{Cov}\left(c_j, c_k\right) \approx \frac{1}{N}\sum_{n=-\infty}^{\infty}\left(\gamma\left(n\right)\gamma\left(n+k-j\right) + \gamma\left(n+k\right)\gamma\left(n-j\right)\right),$$

which yields $\mathrm{Var}(c_k)$ when $j = k$.

We estimate the underlying acf $\rho\left(k\right)$ by the sample acf

$$r_k = \frac{c_k}{c_0}.$$

Recall from chapter 1 that if the data $x\left(1\right), \ldots, x\left(N\right)$ are from a completely random process (such as a white noise process) then for large $N$ under some weak conditions, for $k \neq 0$,

$$r_k \sim N\left(0, \frac{1}{N}\right).$$

The correlogram is simply a plot of $r_k$ against $k$, and should always be examined in order to help determine which ARMA model might be appropriate. We will summarise how the acf should behave for different models, but remember that (i) for an MA($q$) process the acf should cut off sharply at lag $q$ and (ii) for an AR($p$) process the acf will decay.

**Remark 1** *A common convention in Statistics writing is to denote the estimate of a parameter, $\alpha$ say, by the same symbol with a "hat" on it, $\hat{\alpha}$ in this case. Those who have studied Stat 305 should be well–versed in the topic of statistical estimation. Just think of an estimator $\hat{\alpha}$ as being a "good guess" at the unknown value $\alpha$, based on the data.*

### 3.2.2 Estimating the mean of the process

Whilst an initial problem in classical statistics involving i.i.d. observations is the estimation of the mean of the underlying model, the situation in time series estimation is somewhat less straightforward. It is true that if $M$ series of observations, each of length $N$, were obtained and the sample mean $\bar{x}_j$ found in each, for $j = 1, \ldots, M$, then the mean of these sample means

$$\hat{\mu} = \frac{1}{M} \sum_{j=1}^{M} \bar{x}_j$$

converges to $\mu$ in mean square, meaning that

$$\lim_{M \to \infty} E\left( [\hat{\mu} - \mu]^2 \right) = 0.$$

This is all very well, but in practice $M = 1$, as only one (with luck reasonably long) sequence of observations is available.

So the question arises as to what extent the mean of a sample,

$$\bar{x} = \frac{1}{N} \sum_{t=1}^{N} x\left(t\right)$$

constitutes a useful estimator of $\mu$. However, it can be shown that provided we are sampling from a stationary process for which $\gamma\left(k\right) \to 0$ as $k \to \infty$ we have

1. $\bar{x}$ is unbiased for $\mu$ and
2. $\bar{x}$ is consistent, in that

$$\lim_{N \to \infty} \mathrm{Var}\left(\bar{x}\right) = 0.$$

It is possible to strengthen the statements to incorporate certain processes which are not stationary, as the following theorem shows.

**Theorem 1** *If $X(t)$ is a process for which*

$$\lim_{t\to\infty} E(X(t)) = \mu$$

*and*

$$\lim_{N\to\infty} Cov(\bar{x}, X(N)) = 0$$

*where*

$$\bar{x} = \frac{1}{N}\sum_{t=1}^{N} X(t)$$

*then*

$$\lim_{N\to\infty} E\left([\bar{x}-\mu]^2\right) = 0.$$

Recalling that for an i.i.d. sample of size $N$ we know that

$$\mathrm{Var}(\bar{x}) = \frac{\sigma_X^2}{N}$$

it might be hoped that something similar is true for correlated data. In fact (see exercises)

$$\mathrm{Var}(\bar{x}) = \frac{\sigma_X^2}{N}\left(1 + 2\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)\rho(k)\right)$$

where the data are from a process with acf $\rho(\cdot)$. Now the second term in the brackets above will differ from zero substantially when autocorrelations are large.

**Example 2** *Recall for an AR(1) process that*

$$\rho(k) = \alpha^k$$

*for $k = 1, 2, \ldots,$ and*

$$Var(X(t)) = \sigma_X^2 = \frac{\sigma^2}{(1-\alpha^2)},$$

*with $\sigma^2$ the variance of the white noise process as usual. So*

$$\begin{aligned} Var(\bar{x}) &= \frac{\sigma_X^2}{N}\left(1 + 2\sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)\alpha^k\right) \\ &= \frac{\sigma_X^2}{N}\left(1 + 2\left(\sum_{k=1}^{N-1}\alpha^k - \frac{1}{N}\sum_{k=1}^{N-1}k\alpha^k\right)\right). \end{aligned}$$

*But since $|\alpha| < 1$ we have*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N-1}k\alpha^k = 0$$

*and*

$$\sum_{k=1}^{N-1}\alpha^k = \frac{\alpha - \alpha^N}{1-\alpha} \approx \frac{\alpha}{1-\alpha}.$$

*Hence for large $N$*

$$\begin{aligned} Var(\bar{x}) &\approx \frac{\sigma_X^2}{N}\left(1 + \frac{2\alpha}{1-\alpha}\right) \\ &= \frac{\sigma_X^2}{N}\left(\frac{1+\alpha}{1-\alpha}\right). \end{aligned}$$

*So the scaling factor on the variance from what would be expected from an independent sample is the term*

$$\frac{1+\alpha}{1-\alpha}.$$

*Now this is positive and greater than unity when $\alpha > 0$, so there is higher variance in the sample mean in this case due to the positive autocorrelation in the data – note in this situation one observation being above $\mu$ makes subsequent observations more likely to be above $\mu$, which adversely impacts on $\bar{x}$ as an estimate of $\mu$. However when $\alpha < 0$, the variance of $\bar{x}$ is less than in the i.i.d. case.*

We will describe the most common modern methods for model fitting in time series, starting with probably the most straightforward case.

## 3.3 Fitting an AR model

If we decide that an AR model might be suitable for the data in hand, we must do two things initially: estimate the parameters, and determine the order, $p$. Whilst this might naturally be done in practice in the reverse order, we will first discuss estimating the parameters.

### Estimating the parameters

Consider the general AR($p$) process with mean $\mu$,

$$X(t) - \mu = \alpha_1 (X(t-1) - \mu) + \cdots + \alpha_p (X(t-p) - \mu) + Z(t).$$

We have data $x(1), \ldots, x(N)$ supposedly from this process, with mean $\bar{x}$. The natural procedure for estimating the parameters is one you have already encountered: *least squares estimation.* This is the obvious approach since as we recognise, the model is essentially a linear regression model.

Least squares estimation here entails choosing the estimators $\hat{\mu}, \hat{\alpha}_1, \ldots, \hat{\alpha}_p$ which minimise the sum of squares

$$S := \sum_{t=p+1}^{N} \left(x(t) - \hat{\mu} - \hat{\alpha}_1 (x(t-1) - \hat{\mu}) - \cdots - \hat{\alpha}_p (x(t-p) - \hat{\mu})\right)^2.$$

Note that we are minimising the sum of squares of the differences between the observed values and the estimated values, the summation only starting at $t = p + 1$ since we have no way of estimating the first $p$ terms using the model.

Let us consider the case of fitting an AR(1) model with mean $\mu$,

$$X(t) - \mu = \alpha (X(t-1) - \mu) + Z(t).$$

Akin to simple linear regression, we aim to minimise

$$S(\mu, \alpha) = \sum_{t=2}^{N} \left((x(t) - \mu) - \alpha (x(t-1) - \mu)\right)^2$$

with respect to the arguments $\mu$ and $\alpha$. We differentiate partially with respect

to these variables in turn; firstly with respect to $\mu$, giving

$$\begin{aligned}\frac{\partial S}{\partial \mu} &= 2\sum_{t=2}^{N}\left[\left(x\left(t\right)-\mu\right)-\alpha\left(x\left(t-1\right)-\mu\right)\right]\left(\alpha-1\right)\\ &= 2\left(\alpha-1\right)\sum_{t=2}^{N}\left[\left(x\left(t\right)-\alpha x\left(t-1\right)\right)+\left(\alpha-1\right)\mu\right],\end{aligned}$$

and then with respect to $\alpha$ to yield

$$\begin{aligned}\frac{\partial S}{\partial \alpha} &= -2\sum_{t=2}^{N}\left[\left(x\left(t\right)-\mu\right)-\alpha\left(x\left(t-1\right)-\mu\right)\right]\left(x\left(t-1\right)-\mu\right)\\ &= -2\left(\sum_{t=2}^{N}\left(x\left(t\right)-\mu\right)\left(x\left(t-1\right)-\mu\right)-\alpha\sum_{t=2}^{N}\left(x\left(t-1\right)-\mu\right)^{2}\right).\end{aligned}$$

Now for a minimum we set these equations to zero, giving

$$\begin{aligned}\hat{\mu} &= \frac{\sum_{t=2}^{N}\left(x\left(t\right)-\hat{\alpha}x\left(t-1\right)\right)}{\left(1-\hat{\alpha}\right)\left(N-1\right)}\\ &= \frac{\bar{x}_{(2)}-\hat{\alpha}_1\bar{x}_{(1)}}{1-\hat{\alpha}_1}\end{aligned}$$

where $\bar{x}_{(1)}$ is the mean of the first $N-1$ observations, $\bar{x}_{(2)}$ is the mean of the second $N-1$ observations and

$$\hat{\alpha}_1=\frac{\sum_{t=1}^{N-1}\left(x\left(t\right)-\hat{\mu}\right)\left(x\left(t+1\right)-\hat{\mu}\right)}{\sum_{t=1}^{N-1}\left(x\left(t\right)-\hat{\mu}\right)^{2}}.$$

However, since $\bar{x}_{(1)}\approx\bar{x}_{(2)}\approx\bar{x}$, it is possible to take

$$\begin{aligned}\hat{\mu} &= \bar{x}\\ \hat{\alpha}_1 &= \frac{\sum_{t=1}^{N-1}\left(x\left(t\right)-\bar{x}\right)\left(x\left(t+1\right)-\bar{x}\right)}{\sum_{t=1}^{N-1}\left(x\left(t\right)-\bar{x}\right)^{2}}.\end{aligned}$$

Since the denominator in the estimator of $\alpha$ above is approximately $\sum_{t=1}^{N}\left(x\left(t\right)-\bar{x}\right)^{2}$, it is usual to let

$$\hat{\alpha}_1=\frac{c_1}{c_0}=r_1.$$

This is quite appealing, since we recall that for an AR(1) process $\rho(k) = \alpha_1^{|k|}$, and so $\rho(1) = \alpha_1$.

There is another parameter to estimate, namely $\sigma^2$, and recalling the procedure in simple linear regression (from Stat 200) we might sensibly use the residual mean square for this, where the *residual* at time $t$ is

$$z(t) = (x(t) - \hat{\mu}) - \hat{\alpha}(x(t-1) - \hat{\mu})$$

and the residual mean square is

$$\hat{\sigma}^2 = \frac{\sum_{t=2}^{N} z(t)^2}{N-1}.$$

There is some sense in which the divisor in the estimator above might better be taken as $N-3$, though this makes little difference in practice.

**Remark 2** *An AR(p) can be written*

$$Z(t) = X(t) - \alpha_1 X(t-1) - \cdots - \alpha_p X(t-p).$$

*Multiplying the above equation by $X(t)$ gives*

$$Z(t)X(t) = X(t)^2 - \alpha_1 X(t-1)X(t) - \cdots - \alpha_p X(t-p)X(t).$$

*Now take the expectation operator on both sides, giving (check this!)*

$$\sigma^2 = Var(X(t))(1 - \alpha_1\rho(1) - \cdots - \alpha_p\rho(p)).$$

*So an alternative, "model–based" estimator for $\sigma^2$ in the AR case is therefore*

$$\hat{\sigma}^2 = c_0(1 - \hat{\alpha}_1 r_1 - \cdots - \hat{\alpha}_p r_p).$$

For an AR(2) process, the least squares procedure leads to the following estimates:

$$\begin{aligned} \hat{\mu} &= \bar{x}, \\ \hat{\alpha}_1 &= \frac{r_1(1-r_2)}{(1-r_1^2)}, \\ \hat{\alpha}_2 &= \frac{(r_2 - r_1^2)}{(1-r_1^2)}. \end{aligned}$$

So if we attempted to fit an AR(2) process to what was really an AR(1) (for which $\rho(2) = \alpha_1^2$) then the equations above would reduce to $\hat{\alpha}_1 \approx r_1$, $\hat{\alpha}_2 \approx 0$, which is intuitively appealing.

The value $\hat{\alpha}_2$ is often called the *partial autocorrelation coefficient* (pacf) of order 2, since it measures the "extra" correlation between $X(t)$ and $X(t-2)$ not accounted for by $\hat{\alpha}_1$.

The least squares estimation procedure described briefly above is essentially routine, and can be performed for an AR($p$) using a software package such as R or Minitab. We briefly discuss an alternative approach, maximum likelihood estimation, in a later section.

### 3.3.1 Estimating the order

Deciding what value of $p$ might be suitable for the AR model in question can be quite hard. Looking at the correlogram is not always a great help – in fact the acf will decay away to zero with time (behaving like a "damped" sine curve or exponential) for an AR, but gives little indication as to what order should be adopted.

If we define the residuals of a fitted model to be the differences

$$x(t) - \hat{x}(t),$$

where $\hat{x}(t)$ is the value predicted for $x(t)$ by the model under consideration, then the sum of *squares* of the residuals (the Res SS, for short) is often useful in model fitting. We might fit successive AR($p$) models, for $p = 1, 2, \ldots$, work out the Res SS for each, and choose the value of $p$ above which there appears to be no "significant" reduction in the Res SS.

We have already defined the pacf of order 2, and the definition extends to general order, and provides a useful tool in model fitting. When fitting an AR($k$) process, the last coefficient $\alpha_k$ measures the "extra" correlation (between $X(t)$ and $X(t-k)$) not accounted for by the AR($k-1$) model. The estimate of the coefficient on the term in the model of highest lag, which we could denote $\hat{\alpha}_{kk}$, is called the *partial autocorrelation coefficient (pacf) of order* $k$, and is often plotted against $k$ when successively larger models are being fitted and considered. For a true AR($p$) process, the pacf should "cut–off" at order $p$.

As a rule of thumb, and to clarify the above, values of the pacf outside the range $\pm\frac{2}{\sqrt{N}}$ are deemed "significantly" different from zero. Indeed, it can

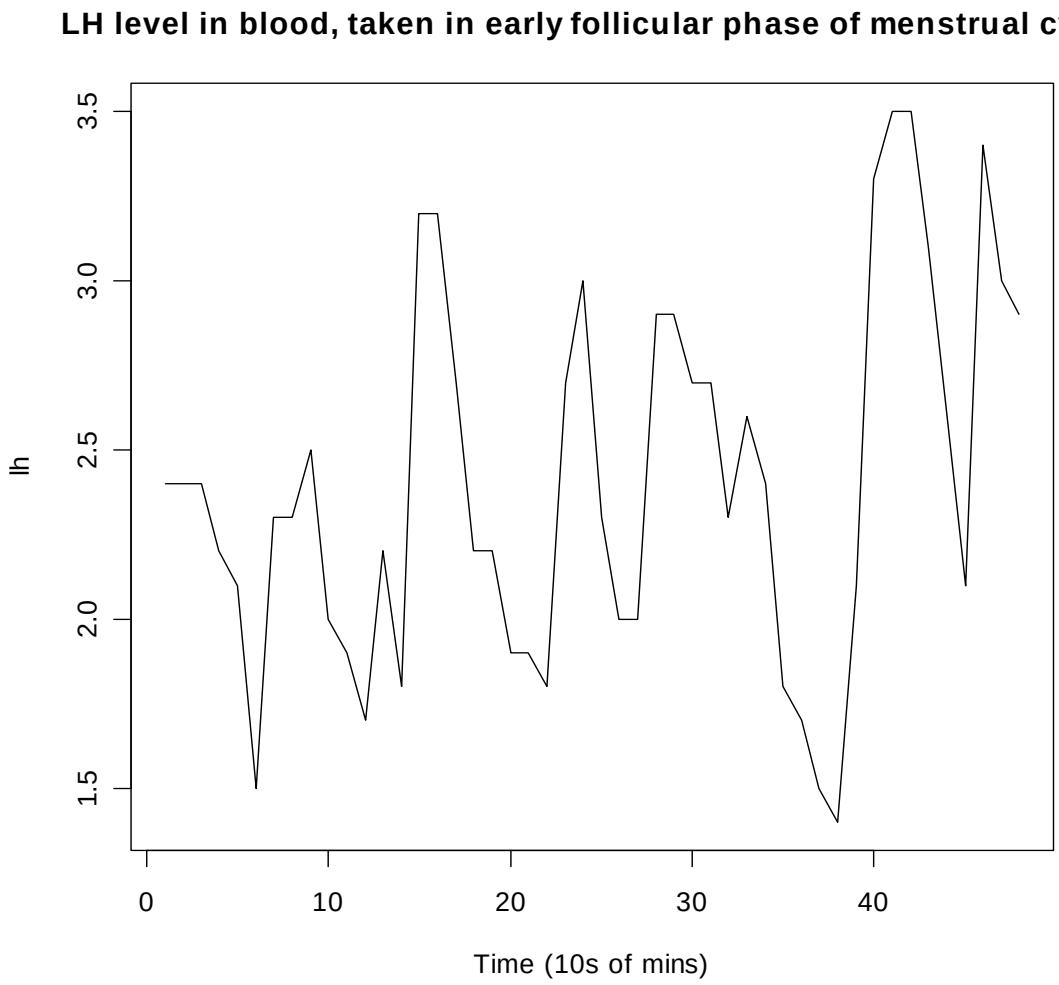
be shown that for an underlying AR($p$) process, approximately

$$\hat{\alpha}_{kk} \sim N\left(0, \frac{1}{N}\right)$$

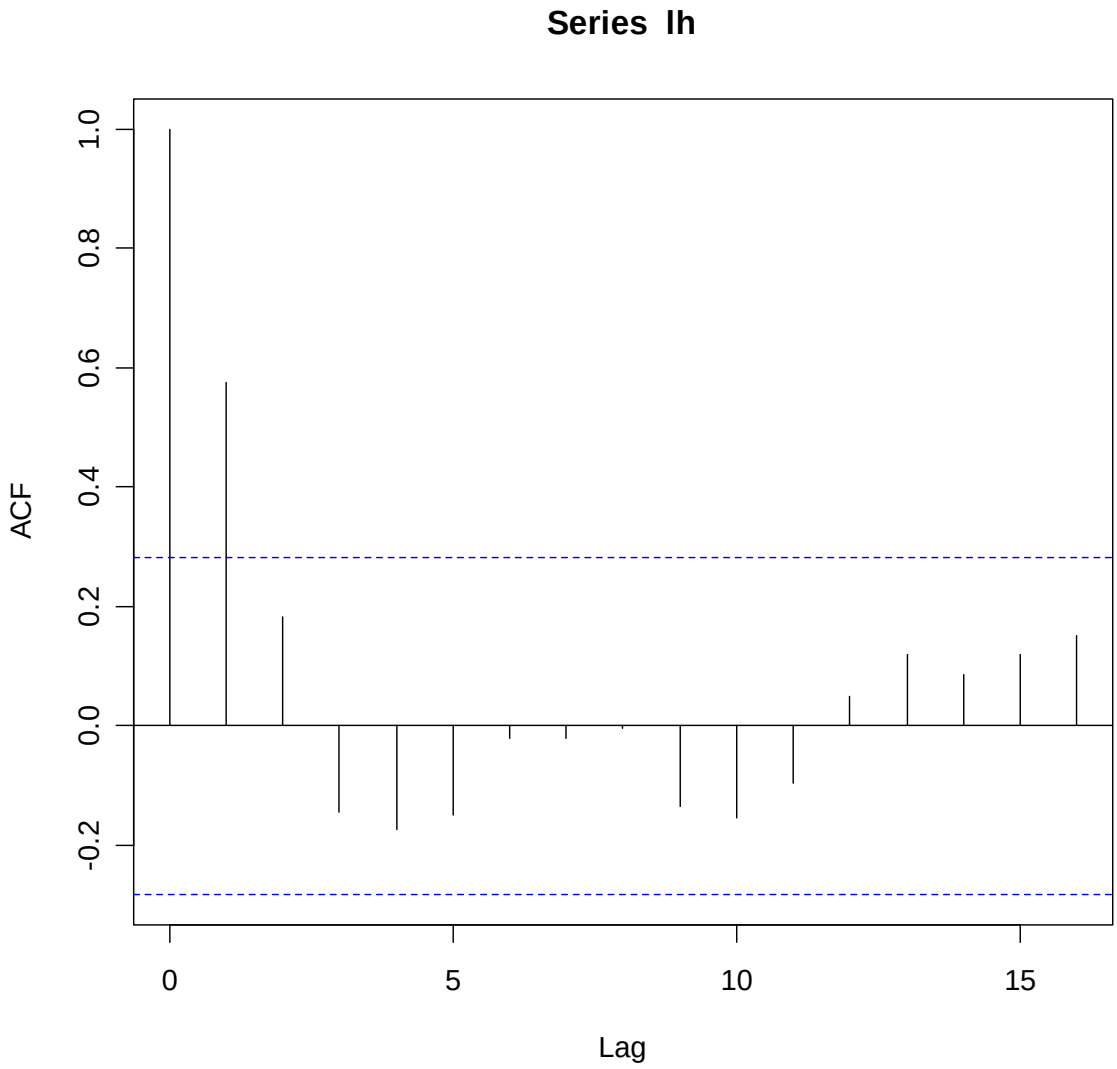
for $k > p$. If in the fitting of a AR($p$) process the value of the pacf of order $p$ is outside the range $\pm\frac{2}{\sqrt{N}}$, it is sensible to consider the next model up, the AR($p+1$).

**Example 3** *Diggle (1990) reports a study of the level of luteinizing hormone (lh, which is important in the reproductive process) in the blood of a healthy woman taken at ten minute intervals over an eight-hour period. Three separate series were collected, and below is the plot of the lh levels taken during*

*the early follicular phase of the subject's menstrual cycle.*

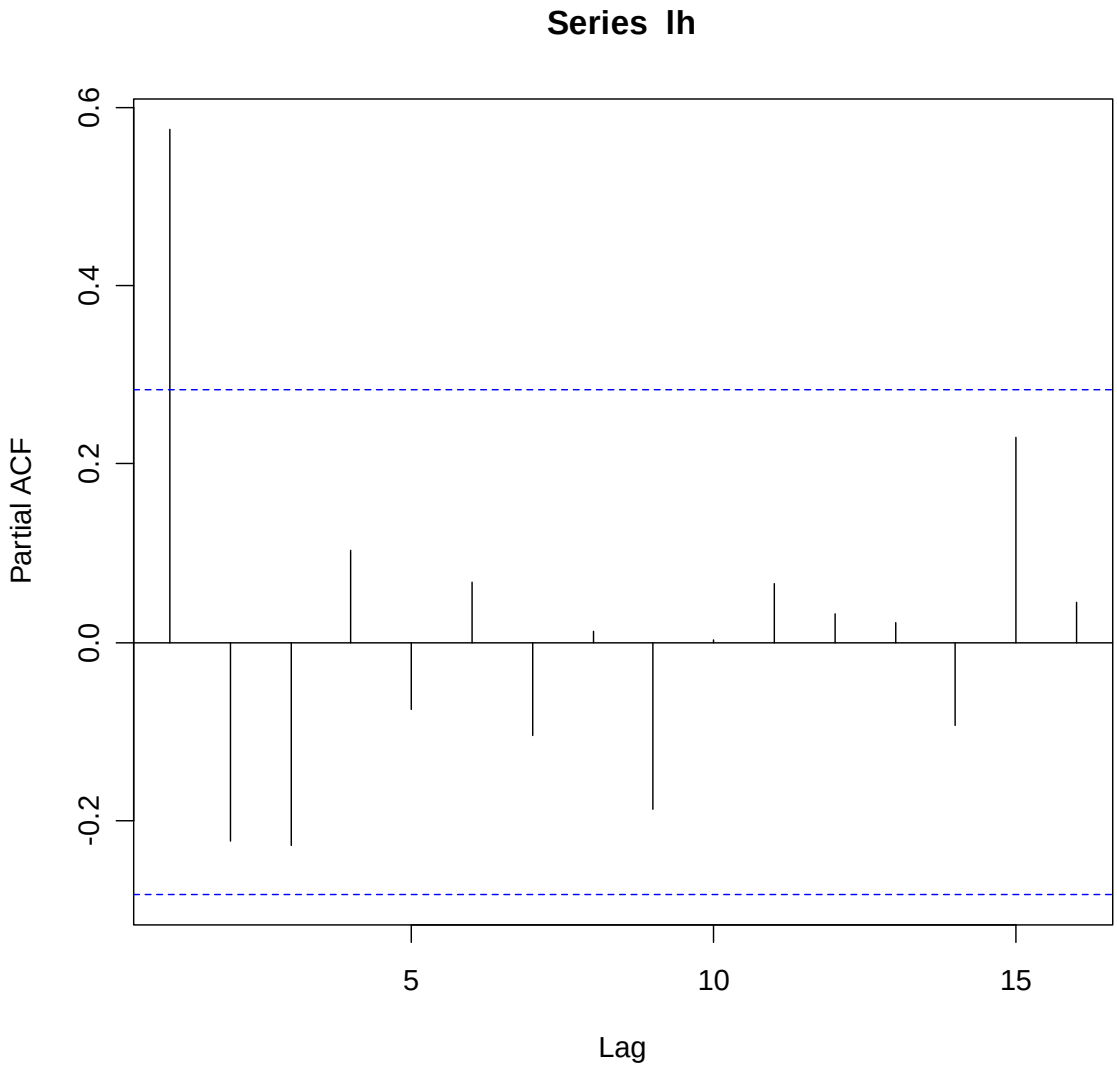


*The sample acf here is plotted below:*




*It appears to show some decaying, as consistent with a stationary process,*

*there being no "significant" values after lag 1. Now the pacf is insightful:*




*There appears a sharp cut–off in the pacf after lag 1, suggesting an* $AR(1)$ *model might be appropriate. Fitting this by least squares (using the* `ar.ols` *function in R) we fit the model with* $\hat{\alpha} = 0.586$ *and* $\hat{\mu} = 2.4$*. The estimate of* $\sigma^2$ *is 0.2016. Note here R chose the order "automatically", though the maximal order can be specified if so desired.*

## 3.4 Fitting an MA model

As with fitting an AR there are two problems, though the relative difficulties are reversed for the MA case.

### 3.4.1 Estimating the parameters

This is somewhat harder than for the AR case, and there are several suggested approaches. We will only cover one method which is commonly adopted.

Let us begin by considering the MA(1) case. The model with mean $\mu$ takes the form

$$X(t) = \mu + Z(t) + \beta_1 Z(t-1). \tag{3.1}$$

Being the mean of the process, it would seem sensible to take as estimator of $\mu$ the sample mean $\bar{x}$. Bearing in mind the form of the acf for the MA(1) (from chapter 2), it might also seem reasonable to choose $\hat{\beta}_1$ which satisfies

$$r_1 = \frac{\hat{\beta}_1}{1 + \hat{\beta}_1^2}, \tag{3.2}$$

where $r_1$ is the sample acf at lag 1 as usual. Imposing that $|\hat{\beta}_1| < 1$ would add an extra constraint to ensure invertibility.

The approach described above, whilst seemingly intuitive and simple, can be shown to give rise to rather inefficient estimators – that is, estimators which vary a great deal depending on the data values.

**Exercise 3.4.1** *You might like to try solving (3.2) for a given value of $r_1$, and choose the solution $\hat{\beta}_1$ such that $|\hat{\beta}_1| < 1$ (should one exist). Then change the value of $r_1$ a small amount (say 0.05). What happens to the solution $\hat{\beta}_1$ now?*

A better approach is the following: choose suitable starting estimates for $\mu$ and $\beta_1$ – probably those suggested above, $\hat{\mu} = \bar{x}$ and $\hat{\beta}_1$ the solution to (3.2). Then the residual sum of squares for the model fitted can be found by repeatedly using (3.1) in the form

$$Z(t) = X(t) - \mu - \beta_1 Z(t-1)$$

to calculate the residuals for the model. Specifically, start with $z(0) = 0$, then take

$$\begin{aligned} z(1) &= x(1) - \mu \\ z(2) &= x(2) - \mu - \beta_1 z(1) \\ &\vdots \end{aligned}$$

and in this way calculate the Res SS, $\sum_{i=1}^{N} z(i)^2$. Then perform a *grid search* over a range of values for the pair $\left(\hat{\mu}, \hat{\beta}_1\right)$, and choose as the final estimates the pair which minimises the Res SS above. In this example we can think of the Res SS as a function of the two variables $\hat{\mu}$ and $\hat{\beta}_1$, and we seek to find the minimum of this function.

**Remark 3** *The process described above is an example of function* optimisation*, and is an area which has much relevance to modern Statistics. There are various methods that can be used for trying to find the minimum (or maximum) of a multivariate function, including Newton–Raphson, steepest descent, and the simplex method.*

Of course, one would naturally not do the above procedure by hand, with modern computers being well–suited to such calculations. The idea extends naturally to models of higher order: for the MA(2)

$$X(t) = \mu + \beta_1 Z(t-1) + \beta_2 Z(t-2) + Z(t)$$

we could choose starting estimates of $\mu$, $\beta_1$ and $\beta_2$ in some way, calculate the residuals $z(t)$, $t = 1, \ldots, N$, and then find the Res SS, $\sum_{i=1}^{N} z(t)^2$. Then make further choices for the parameters, perhaps moving around a grid of values for the triple $(\mu, \beta_1, \beta_2)$, and choose the set of estimates which minimises Res SS. To re-iterate, this is a procedure best performed using a computer.

To estimate $\sigma^2$ here, the residual mean square is a reasonable candidate, and works quite well in practice. Note also though for an MA($q$) process we have

$$\mathrm{Var}(X(t)) = \sigma^2 \left(1 + \sum_{i=1}^{q} \beta_i^2\right),$$

and so an "obvious" estimator of $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{c_0}{\left(1 + \sum_{i=1}^{q} \hat{\beta}_i^2\right)}.$$

### 3.4.2 Determining the order

In principle this should be quite straightforward, as the acf for an MA($q$) process should "cut–off" to zero at lag $q$. The pacf provides no useful information here, however.

## 3.5 Fitting an ARIMA model

Fitting the most general type of model (for which AR and MA processes are special cases) is not an exact science, as with the simpler models. Series that arise in practice are usually *not* stationary – the plot will usually indicate departures from stationarity, and be aware that the correlogram for a non-stationary series will generally not decay to zero for high lags.

As we know, a common approach used for making a series appear stationary is to *difference* it. Applying the difference operator $\nabla$ once is usually sufficient, more than twice being rarely, if ever, required. So the value of the parameter $d$ in an ARIMA($p, d, q$) model is almost always 0 or 1, occasionally perhaps 2 or 3.

To determine the orders $p$ and $q$ an iterative procedure must be undertaken, as in the fitting of a MA process described in the previous subsection. The idea is the same so the details will not be repeated here, and of course the method is only sensibly implemented on a computer. Modern statistical software packages will often allow ARIMA($p, d, q$) models to be fitted up to specified values of $p$, $d$ and $q$, and may suggest a choice from those models considered, though experience and parsimony should determine the final model adopted.

We have seen that the acf and pacf help shed some light on which model might be most appropriate, and included below is a table re-iterating how these functions should behave for the processes we have discussed, which can be used as a rule of thumb when examining the sample acf and pacf.

| Model | Acf | Pacf |
|---|---|---|
| MA($q$) | Cuts–off at lag $q$ | Tails off, no pattern |
| AR($p$) | Tails off, like a "damped" sine wave or exponential | Cuts–off at lag $p$ |
| ARMA($p, q$) | No pattern up to lag $q$, then tails off as in AR case | Tails off, no pattern |

Having fitted an ARMA$(p, q)$ model, the residual sum of squares

$$\sum_{t=1}^{N} z\left(t\right)^2$$

should provide an estimate of $\sigma^2$, once divided by $N-p-q$, or $N-p-q-1$ if the model has a non-zero mean term $\mu$ included.

The fitting of ARIMA and, to some extent, SARIMA models differs in practice very little from general ARMA modelling, other than a choice has to be made regarding how many times the initial series should be differenced before it appears to arise from a stationary process. Usually first or second order differencing suffices. The choice of $s$ in SARIMA models may be obvious (12 say, in the case of monthly data), but some trial–and–error with the values of $p$, $P$, $q$ and $Q$ may be involved in order to find what appears to be the best model for a data set.

## 3.6 Maximum likelihood estimation

Without doubt the most commonly used method for parameter estimation in modern Statistics, the concept of maximum likelihood estimation (m.l.e.) should be familiar to you from Stat 305. The idea is in essence a simple one: choose the values of the parameters which makes the data observed "most likely", in the sense of maximising the joint density function of the data as a function of the parameter values. In the early studies of time series this procedure was well beyond the computational facilities of the day, but modern software packages such as R allow m.l.e., or close variants of the approach, to be routinely applied for time series models.

Given a time series

$$\mathbf{X} := \left(X\left(1\right), \ldots, X\left(N\right)\right)$$

the general ARMA model is of the form

$$X\left(t\right) - \mu = \sum_{j=1}^{p} \alpha_j \left(X\left(t-j\right) - \mu\right) + Z\left(t\right) + \sum_{j=1}^{q} \beta_j Z\left(t-j\right),$$

and so depends on $\mu$, $\mathrm{Var}(Z\left(t\right)) = \sigma^2$ and the parameter vectors

$$\boldsymbol{\alpha} := \left(\alpha_1, \ldots, \alpha_p\right)$$

and

$$\boldsymbol{\beta} := \left(\beta_1, \ldots, \beta_q\right).$$

Hence there are $p + q + 2$ parameters to estimate.

Now the variance–covariance matrix of $\mathbf{X}$ can be written as

$$\mathrm{Var}\left(\mathbf{X}\right) = \sigma^2 \mathbf{V}\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right),$$

being an $N \times N$ matrix. Under the assumption that $Z\left(t\right)$ is Normally distributed for all $t$, it follows that $\mathbf{X}$ is multivariate Normal. The log-likelihood function is therefore of the form

$$l\left(\mu, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{x}\right) = -\frac{1}{2}\left(N \log\left(\sigma^2\right) + \log\left(\mathbf{V}\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right)\right) + \frac{\left(\mathbf{x}-\mu\mathbf{1}\right)' \mathbf{V}\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right)^{-1} \left(\mathbf{x}-\mu\mathbf{1}\right)}{\sigma^2}\right),$$

in which $\mathbf{x}$ represents the data and $\mathbf{1}$ a column vector of 1's.

Now maximisation of $l\left(\mu, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{x}\right)$ is possible in theory, and involves the inversion of an $N \times N$ matrix at each step. One way to improve the efficiency of the method is to adopt a *conditional* likelihood based approach – we optimise conditional on a given set of observations.

To illustrate this, consider the AR(1) model

$$X\left(t\right) - \mu = \alpha\left(X\left(t-1\right) - \mu\right) + Z\left(t\right).$$

We derive the likelihood conditional on the first observation, $x\left(1\right)$, for it is clear that distribution of $X\left(t\right)$ depends on the past values only through $X\left(t-1\right)$, for all $t = 2, 3, \ldots, N$. So the density function of $X\left(2\right), \ldots, X\left(N\right)$, conditional on the value $x\left(1\right)$, must be of the form

$$f\left(x\left(2\right), \ldots, x\left(N\right) | X\left(1\right) = x\left(1\right)\right) = \prod_{t=2}^{N} g\left(x\left(t\right) | x\left(t-1\right)\right)$$

where $g\left(\cdot\right)$ represents the conditional density function of $X\left(t\right)$ given $x\left(t-1\right)$. Now under the assumption that $Z\left(t\right) \sim N\left(0, \sigma^2\right)$, then conditional on the value of $x\left(t-1\right)$ we have

$$X\left(t\right) \sim N\left(\mu + \alpha\left(x\left(t-1\right) - \mu\right), \sigma^2\right).$$

Hence the conditional density function of $X\left(t\right)$ given $x\left(t-1\right)$ is

$$g\left(x\left(t\right) | x\left(t-1\right)\right) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left(-\frac{\left(x\left(t\right) - \mu - \alpha\left(x\left(t-1\right) - \mu\right)\right)^2}{2\sigma^2}\right).$$

Consequently the conditional likelihood can be written

$$f\left(x\left(2\right),\ldots,x\left(N\right)|x\left(1\right)\right)=\frac{1}{\left(2\pi\sigma^{2}\right)^{\left(N-1\right)/2}}\exp\left(-\sum_{t=2}^{N}\frac{\left(x\left(t\right)-\mu-\alpha\left(x\left(t-1\right)-\mu\right)\right)^{2}}{2\sigma^{2}}\right).$$

Taking logarithms of this gives the conditional log–likelihood as being

$$l_{c}\left(\mu,\sigma^{2},\alpha;\mathbf{x}\right):=-\frac{\left(N-1\right)\log\left(\sigma^{2}\right)}{2}-\sum_{t=2}^{N}\frac{\left(x\left(t\right)-\mu-\alpha\left(x\left(t-1\right)-\mu\right)\right)^{2}}{2\sigma^{2}}$$

ignoring a term not involving the parameters of interest. Now to maximise $l_{c}\left(\mu,\sigma^{2},\alpha;\mathbf{x}\right)$ with respect to $\mu$ and $\alpha$ we must minimise

$$\sum_{t=2}^{N}\left(x\left(t\right)-\mu-\alpha\left(x\left(t-1\right)-\mu\right)\right)^{2}$$

which is exactly the same minimisation problem we addressed when fitting an AR(1) model by least squares. *Hence least squares estimation and conditional m.l.e. coincide here.*

To estimate $\sigma^{2}$, differentiation gives

$$\frac{\partial l_{c}}{\partial\sigma^{2}}=\frac{\left(N-1\right)}{2\sigma^{2}}+\frac{\sum_{t=2}^{N}\left(x\left(t\right)-\mu-\alpha\left(x\left(t-1\right)-\mu\right)\right)^{2}}{2\sigma^{4}}$$

and setting this to zero gives the (conditional) m.l.e. of $\sigma^{2}$ to be

$$\frac{\sum_{t=2}^{N}\left(x\left(t\right)-\mu-\alpha\left(x\left(t-1\right)-\mu\right)\right)^{2}}{N-1}=\frac{\sum_{t=2}^{N}z\left(t\right)^{2}}{N-1}.$$

This is the "natural" estimator suggested earlier.

In principle "full–blown" m.l.e. can be applied to time series model fitting via a computer package, but for long series this can still be computationally prohibitive. Conditional m.l.e., as in the example above, is often adopted, and in many cases the estimates obtained are either identical or very similar to those found via least squares estimation when the white noise process is assumed Normal.

## 3.7 Model diagnostics

Having fitted what might be considered a suitable model from the ARIMA family, the final stage in the process is to check that the model fits the data reasonably well, and that there are no patterns in the data that the model is not detecting. There are no hard–and–fast rules as to how one should do this, but a general approach to model diagnostics in Statistics involves consideration of the *residuals* of the model, the differences between the observed values and the corresponding fitted values. So let

$$\hat{z}(t) := x(t) - \hat{x}(t)$$

(for $t = 1, 2, \ldots, N$) define the residuals of a model, where $\hat{x}(t)$ is the value fitted by the model at time $t$.

**Example 4** *In fitting an AR(1) model to a data set, we must estimate the parameter $\alpha$, by $\hat{\alpha}$, as described above. The residual at time $t$ for the fitted model would therefore be*

$$\hat{z}(t) = x(t) - \hat{\alpha}x(t-1).$$

*If you like, the above is an estimate of the white noise term $z(t)$, which appears in the definition of the model.*

Naturally we would like the residuals to be small, and indeed if the model fits quite well and is not leaving "residual" pattern in the data, the $\hat{z}(t)$ should look "random", like a realisation of a white noise process. They should not, for example, be a series with an acf much outside the range $\pm 2/\sqrt{N}$ at any lag values.

Several tests are suggested based on the residuals, and these can often be used as guidelines. Let the acf at lag $k$ of the residuals be denoted $r_k(\hat{z})$.

**Remark 4** *The standard tests that are used here are based on comparing some number, the* test statistic*, with the* Chi-squared distribution*. This probability distribution should be familiar to you, and we only mention here that it is continuous, unimodal and has a density function over positive real numbers. It depends on one parameter, usually called the* degrees of freedom *of the distribution. The Chi-squared distribution on $n$ degrees of freedom is sometimes denoted $\chi^2_n$. Tables exist of the cumulative distribution function for this distribution, for various degrees of freedom.*

**Remark 5** *In the following tests we are assuming an ARMA$(p,q)$ model is under scrutiny, having been fitted to $N$ data points. If actually we had to difference the initial series in order to achieve stationarity, the length of the series after differencing should be taken as our $N$. To clarify, if we have differenced the initial series $d$ times before fitting the ARMA model, then $N$ in what follows is the initial length minus $d$.*

One test that is often performed is called the *portmanteau lack–of–fit test.* This has test statistic that we will denote as $Q$, where

$$Q := N \sum_{k=1}^{M} r_k \left(\hat{z}\right)^2,$$

where $N$ is the number of terms in the series (possibly after differencing) and $M$ is an integer rather less than $N$, usually between 15 and 30. If the ARMA$(p,q)$ model we have fitted to the data is reasonable, we would expect

$$Q \sim \chi^2_{M-p-q},$$

that is to say, the value of $Q$ should be consistent with the $\chi^2_{M-p-q}$ distribution. If the model fitted is a poor one, the value of $Q$ becomes inflated, and would appear to be too large to be consistent with the $\chi^2_{M-p-q}$ distribution. The rule of thumb might be to re-consider the model if the value of $Q$ lies above the 95% point of $\chi^2_{M-p-q}$, a figure available from tables.

A variant of the above test is the *Ljung–Box–Pierce test*, which suggests test statistic

$$N\left(N+2\right) \sum_{k=1}^{M} \frac{r_k \left(\hat{z}\right)^2}{\left(N-k\right)}$$

instead of $Q$. Once again, the above should be consistent with $\chi^2_{M-p-q}$ if the model fitted is adequate.

In truth, whilst there is no loss in considering the outcomes of these and similar tests, they are by nature rather inconclusive. In fact they tend to accept an unacceptable model as being adequate rather too often. That said, some software packages provide them as part of their output, so little time is wasted in considering these indicators.

A wise step is to consider the correlogram for the residuals. As the residuals should be "random", we would not expect there to be significantly large values of the acf, even at small lags. The odd value outside the interval $\pm 2/\sqrt{N}$ can be tolerated, of course, especially if occurring at lags with no obvious physical interpretation.

### 3.7.1 Example 1

A stationary time series of length $N = 200$ gave the following results:

| $k$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $r_k$ | 0.427 | 0.475 | 0.169 | 0.253 | 0.126 |
| $\hat{\alpha}_{kk}$ | 0.427 | 0.358 | –0.160 | 0.106 | 0.035 |

For the series, the sample mean was $\bar{x} = 0.09$ and $c_0 = 1.15$. We are required to determine the most appropriate ARMA model for the data, and then fit the model in question. Now noting that $2/\sqrt{200} = 0.141$ we see that $r_1, \ldots, r_4$ and $\hat{\alpha}_{11}$, $\hat{\alpha}_{22}$ and $\hat{\alpha}_{33}$ are all "significant". Whilst a MA(4) model is feasible, we would prefer a smaller model if possible, as probably the acf is just decaying, which indicates an AR model. We might try an AR(2) or AR(3), or possibly an ARMA(1,1).

One question which should be asked is: should a non-zero mean $\mu$ be included? The sample mean 0.09 suggests not – in fact this figure should be standardised by the estimate of its standard deviation in order to make a sound decision, and it turns out that setting the mean of the model to be fitted to zero is acceptable, whichever model we are considering.

For the AR(2) model, our starting estimates of $\alpha_1$ and $\alpha_2$ would be

$$\begin{aligned} \hat{\alpha}_1 &= \frac{r_1 (1 - r_2)}{1 - r_1^2} = 0.274, \\ \hat{\alpha}_2 &= \frac{r_2 - r_1^2}{1 - r_1^2} = 0.358. \end{aligned}$$

Strictly, we must check that this "initial" model is *admissible* – does the underlying polynomial have roots outside the unit circle? Here we are fine, but be aware that some software packages will blindly fit inadmissible (that is, non-stationary and/or non-invertible) models to data, without providing the user with any warning.

An estimate of the white noise variance is

$$\hat{\sigma}^2 = c_0 (1 - \hat{\alpha}_1 r_1 - \hat{\alpha}_2 r_2) = 0.820,$$

so the "initial" model is

$$X(t) = 0.274X(t-1) + 0.358X(t-2) + Z(t),$$

where $Z(t) \sim N(0, 0.82)$.

The above would provide the starting point for an iterative routine in order to find optimal parameter estimates. For this and the other models under comparison the computations give the following:

| Model | Parameter estimates | $\hat{\sigma}^2$ | $\chi^2$ statistic |
|---|---|---|---|
| AR(2) | $\hat{\alpha}_1 = 0.285$ | 0.805 | 18.43 |
| | $\hat{\alpha}_2 = 0.360$ | | |
| AR(3) | $\hat{\alpha}_1 = 0.348$ | 0.785 | 12.89 |
| | $\hat{\alpha}_2 = 0.405$ | | |
| | $\hat{\alpha}_3 = -0.168$ | | |
| ARMA(1, 1) | $\hat{\alpha}_1 = 0.781$ | 0.861 | 36.59 |
| | $\hat{\beta}_1 = -0.424$ | | |

What might we conclude from the above? Well the Box–Pierce Chi-squared statistic is on (approximately) 20 degrees of freedom for each model. Since the 95-percentile point of that distribution is 31.41, the ARMA model can be discarded. The two AR models are harder to choose between, as both give good fits. The fact that (i) $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are similar for the two models, (ii) the diagnostic test statistic is about 30% smaller for the AR(3) and (iii) $\hat{\alpha}_3$ is quite significantly different from zero in the larger model suggests fitting the AR(3). That is, choose as our model

$$X(t) = 0.348X(t-1) + 0.405X(t-2) - 0.168X(t-3) + Z(t)$$

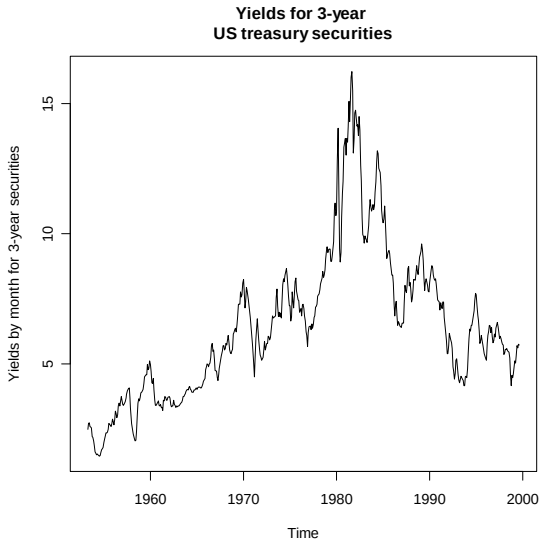
where $Z(t) \sim N(0, 0.785)$.

**Remark 6** *In fact the model used to simulate the data was an AR(2).*

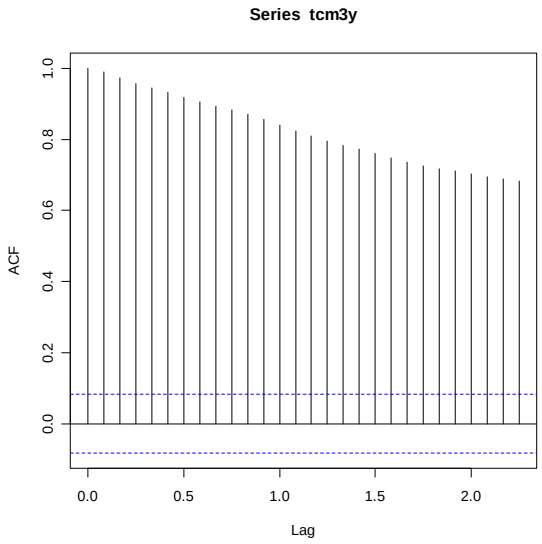
### 3.7.2 Example 2

We consider the data set `tcm3y` from the `tcm` data set in R, using
> `data(tcm)`
The series is of yields on 3–year US treasury securities by month, based on the most–traded securities, and includes 558 observations. A plot of the data

looks like:




The series exhibits some cyclical variations but no obvious long-term trend. It is unclear whether the series is stationary. Plotting the acf we see:

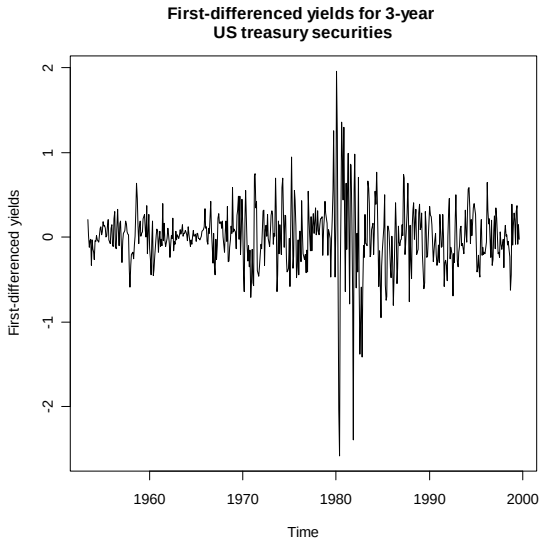



This is obviously not the acf of a stationary series, as the acf is decaying far too slowly.

**Remark 7** *R chooses the horizontal axis scale to be time here, so there are twelve lags per unit time in the acf plots in this example.*

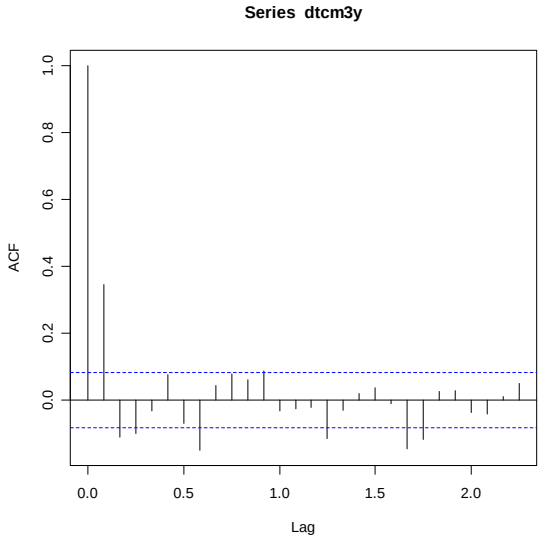
We should difference to see if that helps:

```
> dtcm3y <- diff(dtcm3y)
> plot(dtcm3y)
```

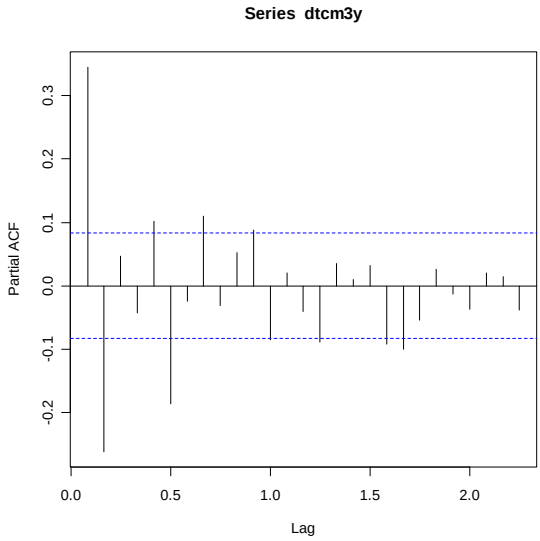



This has clearly removed some of the cyclical variation. Note the series has some major "spikes" in the early 1980's, which will be difficult for a model

to sensibly interpret. The acf of the series looks like:




The above is more consistent with what one might expect from a stationary series, although there are "significant" values at some high lags. There is perhaps some suspicion from this that an MA(2) or MA(3) could be appropriate. The pacf is plotted below:



Worryingly this is not decaying particularly quickly, with some large values (in absolute terms) at lags 5, 6 and 8, for example. Certainly a pure AR model does not look appropriate based on the above plot.

We initially try fitting an ARMA$(1,1)$ model – to see a summary of this model fit enter

```
> summary(dtcm3y.arma <- arma(dtcm3y, order = c(1,1)), include.mean=F)
```

Note no mean has been included, as the series has been differenced initially. Using (conditional) least squares here, R fits the following model

$$Y(t) = -0.164Y(t-1) + Z(t) + 0.634Z(t-1)$$

and estimates $\sigma^2$ to be 0.1205. Trying an MA(2) model via

```
> summary(dtcm3y.arma <- arma(dtcm3y, order = c(0,2)), include.mean=F,
include.intercept=F)
```

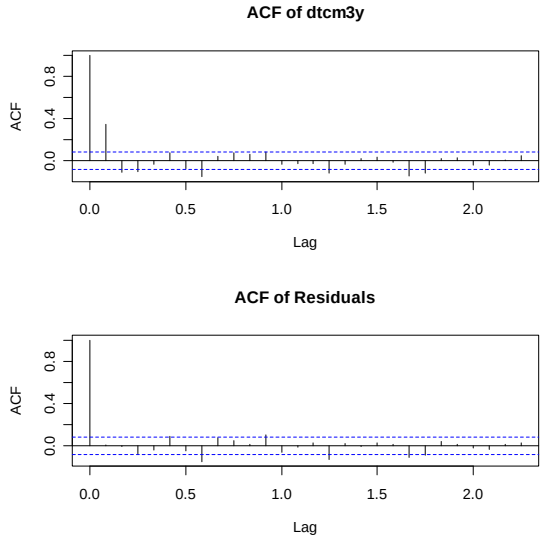
we fit

$$Y(t) = Z(t) + 0.462Z(t-1) - 0.0980Z(t-2).$$

The higher–order MA(3) does not give a substantial improvement, and consideration of the residuals for the MA(2) model indicates a reasonable fit, the largest residuals by far occurring at the times of the "spikes", hardly surprisingly. Pleasingly there is no obvious autocorrelation structure left in the residuals, which is checked via

```
> plot(dtcm3y.arma)
```

This shows an array of plots useful for model checking, a sample pair being given below.




Certainly this only represents an initial analysis of this time series; there may be an argument for removing or smoothing out particular values in the early 1980's as here the behaviour seems rather more volatile than elsewhere. Nonetheless, the two suggested models both appear to give reasonable fits to the data, with the MA(2) being perhaps preferable due to it producing a slightly smaller residual sum of squares

## 3.8 Forecasting

One of the principle aims of studying time series is to enable *predictions* to be made. That is, for a given series we wish to estimate a certain number of future values. Types of forecasting procedures can be categorized in several ways: one such is to distinguish between

1. *subjective* forecasting methods – procedures based on personal knowledge, opinions and "inside" information, and

2. *model–based* approaches, which rely on fitting some kind of model to the data in hand and using the model to predict future values.

There is no sense in which these are mutually exclusive, noting as we should that our model–fitting methods described above do rely on subjective judgement in choosing what is the most appropriate model, and moreover the two approaches can be combined in some situations, with perhaps a weight being given in the forecasting process to input from both "statistical" and "non-statistical" analysts. That said, we are not well–placed to discuss judgemental forecasting in any detail, so will naturally focus on methods based on a statistical footing.

A further way to categorise methods is into

1. "automatic" and

2. "non-automatic"

approaches. By "automatic" here, we infer that once a forecasting method has been chosen, no further input is required from the user – typically a computer package will apply the approach, using some criteria for eliciting the best model and forecasting accordingly. Now the distinction is not clearly defined, since subjectivity will play a part in any *choice* of so–called automatic methods, and even methods which would usually be termed non-automatic can be automated by computer in some way or other.

An over–arching concern relating to time series forecasting is in the identification of the *purpose* of the forecast. That is, to what end will the forecast be used? In some cases it may be more important to predict the chance that a future value of a series will fall outside some range than actually forecast the value itself, and not all methods will sensibly facilitate this. For example, in forecasting tidal behaviour, it is typically *extremal* values which are of interest, since particularly high tides might exceed sea walls and cause extensive floodings. In this scenario there would be little point in investing great effort in predicting exact future tidal levels.

Let us suppose we have observed a time series $x(1), \ldots, x(N)$, up to time $N$, and wish to predict the value $x(N+l)$ which is $l$ time steps in the future. The integer value $l$ is known as the *lead time,* and we will denote our forecasted value by $\hat{x}(N, l)$. There are several approaches that can be used for statistical forecasting, but note that the ones we describe here are *univariate* in nature. That is to say, we consider a single time series in isolation. This may in practice often be sub-optimal, since many time series are intrinsically linked to at least one other. For example, time series for national inflation

rates and unemployment tend to be quite highly *cross–correlated*. However, we have not yet discussed how one might model *bivariate* (or, more generally, *multivariate*) time series, an advanced topic met briefly near the end of this course.

### 3.8.1 Extrapolation of trend curves

We might simply extrapolate a fitted trend curve, which is a crude but easy approach. As we have discussed in chapter 1, one of a variety of curves may be fitted to a time series, and once fitted for the data in hand, such a model can be extended to predict future values. The fitting of polynomials by least squares is quite routinely performed by computer.

This approach has simplicity in its favour. However, the method is undynamic, assuming the model fitted not only is acceptable for the values observed but also will be sensible for future values, and the choice between possible curves may be difficult. Often two competing curves may fit the data almost equally well, but would give very different predictions when extrapolated to future times. Worse still, undue influence regarding the choice of curve to be fitted is placed on observations near the start of the series, which for obvious reasons should have least impact on future behaviour.

All in all, unless a series is too short for anything more sophisticated, or it is clear that a simple parametric curve captures the key features in the data and is likely to hold for future values, this approach is best avoided.

### 3.8.2 Exponential smoothing

This is a more sophisticated approach, and is useful only for *stationary* series. Therefore, if our original series appears to have a trend and seasonality, these effects must be removed before we can commence with this procedure.

Given a series $x(1), \ldots, x(N)$ it seems plausible to take as our estimate of the value $x(N+1)$ a weighted sum of the past observations, i.e., something of the form

$$\hat{x}(N,1) = w_0 x(N) + w_1 x(N-1) + \cdots$$

where the $\{w_i\}$ are a sequence of weights. As it seems sensible to choose these weights to be decreasing (so as most weight is given to the recent past), a possible choice is

$$w_i = \alpha(1-\alpha)^i$$

for $i = 0, 1, \ldots,$ and $0 < \alpha < 1$. These weights are actually the probabilities in a *Geometric* distribution, $w_i$ being the probability of waiting until the $(i+1)$th toss before observing a head in a sequence of coin tosses with $P(H) = \alpha$. Therefore the sum of the weights is unity, and we could call this process *geometric smoothing.* With these weights, we have

$$\hat{x}(N,1) = \alpha x(N) + \alpha(1-\alpha)x(N-1) + \alpha(1-\alpha)^2 x(N-2) + \cdots .$$

Strictly, the above equation implies an infinite number of past observations, so we usually write it as

$$\begin{aligned} \hat{x}(N,1) &= \alpha x(N) + (1-\alpha)\left(\alpha x(N-1) + \alpha(1-\alpha)x(N-2) + \cdots\right) \\ &= \alpha x(N) + (1-\alpha)\hat{x}(N-1,1). \end{aligned} \tag{3.3}$$

Setting $\hat{x}(1,1) = x(1)$ we can use (3.3) recursively to compute our forecasts. In addition, (3.3) reduces the amount of arithmetic involved, since forecasts are obtained from the latest observation and the previous forecast.

We sometimes write (3.3) as

$$\begin{aligned} \hat{x}(N,1) &= \alpha\left(x(N) - \hat{x}(N-1,1)\right) + \hat{x}(N-1,1) \\ &= \alpha e(N) + \hat{x}(N-1,1) \end{aligned}$$

where

$$e(t) := x(t) - \hat{x}(t-1,1)$$

is the "prediction error" at time $t$, $t = 2, \ldots, N$.

So far we have said nothing about the choice of the parameter $\alpha$. Note that small values of $\alpha$ indicate a dependence on a large number of past observations, whereas a choice of $\alpha$ close to unity suggests that dependence is greatly influenced by recent observations. Similar to the estimation of the parameters in a MA process, the choosing of $\alpha$ is not an exact science, and we might choose a value close to one which we know has worked well for similar data in the past. Rather better is to use the sum of the squared errors,

$$\sum_{t=1}^{N} e(t)^2 = \sum_{t=1}^{N} \left(x(t) - \hat{x}(t-1,1)\right)^2 \tag{3.4}$$

to guide our choice, or some related quantity. Naturally the value of (3.4) is determined by $\alpha$. Broadly, there are two strategies:

1. Perform a *grid search* over possible values of $\alpha$, say starting at 0.1 and going up to 0.9 in steps of 0.05. Choose the value of $\alpha$ on this grid for which (3.4) is least.

2. Perform an *iterative* search, by which we mean a process which attempts to take "intelligent" steps over the set of values for $\alpha$ from a given starting point. The steps are designed in such a way as to make final convergence to the minimum of (3.4) in a relatively small number of steps highly likely. It is computationally more intensive than a grid search, but more likely to find the optimal value of $\alpha$.

In many cases, however, the sum of squares surface given by (3.4) is relatively flat, so the exact choice of $\alpha$ is not critical. Either method can be adopted using a suitable software package.

**Example 5** *The de-seasonalised values for the death rate data are given below. Applying exponential smoothing to this (supposedly stationary) series,* Minitab *gives the estimate of $\alpha$ to be 0.297. The fit for this model is, as the figure below indicates, quite good. The next predicted value in the series is therefore found to be 21.55. Adding in the seasonal effect found earlier, this gives $\hat{x}(17) = 23.95$. Note there is some suggestion that a trend should be removed from the data also.*

22.3000
22.5875
22.3625
22.1500
22.3000
22.6875
21.6625
22.3500
22.8000
22.3875
22.3625
22.5500
21.3000
21.4875
20.9625
21.7500

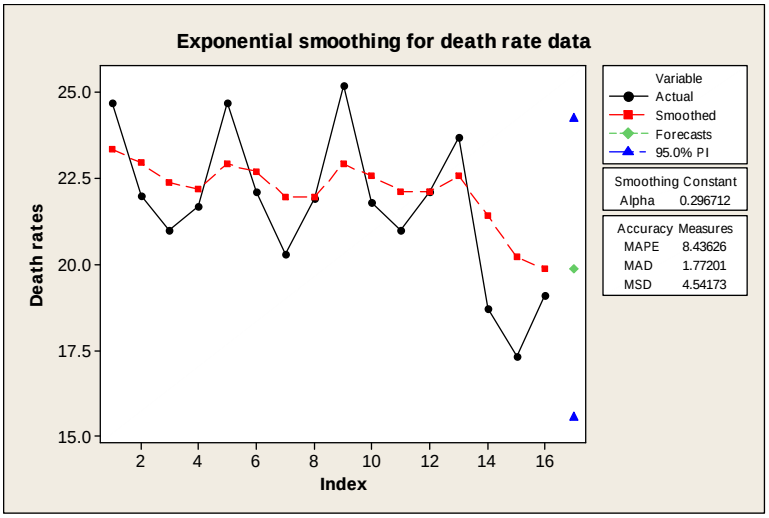



Figure 3.1: Exponential smoothing for death rates

The above method can in fact be extended to cases where trend and seasonal effects are present, and this is known as *Holt–Winters* forecasting, an extension we discuss shortly.

Exponential smoothing is a technique for forecasting which has been shown to work quite well in practice. Indeed, for certain types of time series, in particular ARIMA$(0, 1, 1)$ processes, it is optimal in some respect as will be demonstrated later. The only assumption is that the series in some sense has a memory, with the future depending on the past. The main problem is that it is impossible to indicate the distribution of the prediction errors, $e\,(t)$, and hence confidence intervals for predicted values, for example, cannot be given. The method produces the same forecasts for each lead time $l = 1, 2, \ldots$

### 3.8.3 Holt and Holt–Winters forecasting

Applied to a stationary series as is appropriate, exponential smoothing can be thought of as providing a forecast for the expected "level" of the series at the next time point. A natural extension to this approach to incorporate series with a trend was suggested by Holt in a paper in 1957, and this in turn was expanded to seasonal series by Winter in 1960. The blended method, often termed Holt–Winters forecasting, has found much favour in practice,

being both fairly straightforward to implement and comparatively accurate in its forecasts if adopted with due care.

We begin with a description of Holt's method, which is sometimes referred to as *double exponential smoothing.* Let the "level" of the series at time $t$ be denoted $L(t)$. Now rather than writing $\hat{x}(t,1)$ as $\alpha x(t) + (1-\alpha)\hat{x}(t-1,1)$ as in exponential smoothing, write

$$L(t) = \alpha x(t) + (1-\alpha) L(t-1)$$

for some $\alpha$. Suppose now that the series has a trend, with the expected change per unit time at time $t$ being $T(t)$. This is likely to depend on the level at that time. Holt suggested the following relationships: firstly

$$L(t) = \alpha x(t) + (1-\alpha)(L(t-1) + T(t-1)),$$

in which the $L(t-1)$ term is augmented by adding a change in trend at that time, and secondly

$$T(t) = \beta(L(t) - L(t-1)) + (1-\beta) T(t-1),$$

in which the change due to the trend at time $t$ is a weighted average the change in level and the trend at time $t-1$. Based on this model, the forecast at time $t$ to time $t+l$ will be of the form

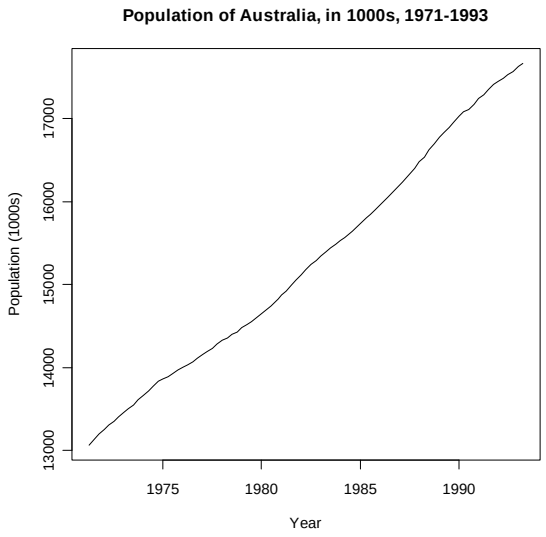
$$\hat{x}(t,l) = L(t) + lT(t)$$

for $l = 1, 2, \ldots$ In effect this predicts the value at time $t+l$ as having the same level as at time $t$, with $l$ trend values added on.

There are two parameters to estimate now, $\alpha$ and $\beta$, with $\beta = 0$ defaulting to simple exponential smoothing. Typically these parameters are restricted to the set $(0,1)$.

**Example 6** *The plot below shows the population (in thousands) of Australia,*

*measured quarterly between March 1971 and June 1993.*




*Applying Holt's method to this data set via the* `HoltWinters` *command in R, the values of* $\alpha$ *and* $\beta$ *are chosen respectively as 1 and 0.4062. The next four quarterly populations (in thousands) would be predicted as being*

$$17704.75, \quad 17747.99, \quad 17791.24, \quad 17834.49.$$

*In fact the population of Australia in June 1994 was approximately 17,855000.*

The above approach extends to include a seasonal component, which is commonly referred to as Holt–Winters smoothing. Let $I(t)$ denote the seasonal effect at time $t$, both additive and multiplicative effects being possible, so that de-seasonalising may be brought about by considering the series $x(t) - I(t)$ or alternatively $x(t)/I(t)$. There are now three quantities to up-date dynamically with $t$, namely $L$, $T$ and $I$.

The model adopted in the case where we have data with a multiplicative seasonal effect of period $p$ has

$$\begin{aligned} L(t) &= \alpha\left(\frac{x(t)}{I(t-p)}\right) + (1-\alpha)\left(L(t-1) + T(t-1)\right), \\ T(t) &= \beta\left(L(t) - L(t-1)\right) + (1-\beta)\,T(t-1), \\ I(t) &= \gamma\left(\frac{x(t)}{L(t)}\right) + (1-\gamma)\,I(t-p). \end{aligned}$$

The time $t$ forecast for the value at time $t+l$ is then

$$\hat{x}(t,l) = (L(t) + lT(t))\, I(t-p+l)$$

for $l = 1, 2, \ldots, p$. The equivalent formulae when the seasonal effect is additive are

$$\begin{aligned} L(t) &= \alpha(x(t) - I(t-p)) + (1-\alpha)(L(t-1) + T(t-1)), \\ T(t) &= \beta(L(t) - L(t-1)) + (1-\beta)T(t-1), \\ I(t) &= \gamma(x(t) - L(t)) + (1-\gamma)I(t-p), \end{aligned}$$

with the forecast value at time $t+l$ being

$$\hat{x}(t,l) = L(t) + lT(t) + I(t-p+l).$$

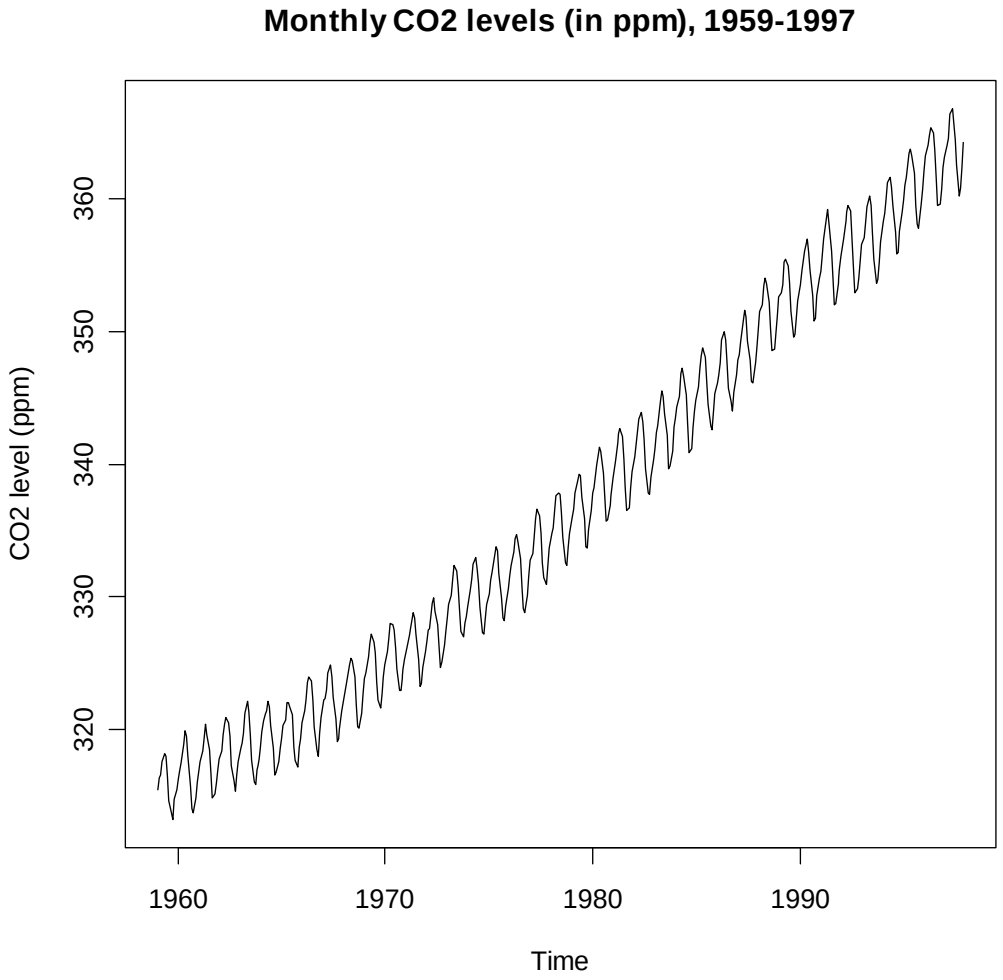
In order to apply successfully the Holt–Winters approach, one must:

1. Determine whether a seasonal affect is present, and if so whether it is additive or multiplicative.
2. Consider what starting values for $L, T$ and $I(1), \ldots, I(p)$ might sensibly be assigned.
3. Decide on the values of $\alpha$, $\beta$ and $\gamma$ – typically this is done by minimising $\sum_t e(t)^2$ over either the whole data set or a reasonable subset.

The above can be performed, if so desired, in a more–or–less fully automated fashion using R.

**Example 7** *The plot below shows the monthly levels of $CO_2$ gas, in parts*

*per million (ppm), taken between 1959 and 1997 in Hawaii.*




*Clearly the above series has both a trend and a seasonal effect, the latter being most likely additive. The* `HoltWinters` *command in R performs Holt–Winters smoothing here, taking* $\alpha$ *as 0.5126,* $\beta$ *to be 0.009497 and* $\gamma$ *as 0.4729. R gives the following predictions for the next four values, for January to April 1998:*

$$365.108,\ 365.966,\ 366.734,\ 368.136.$$

*Actually, the next four values in the series were*

$$365.32,\ 366.15,\ 367.31,\ 368.61$$

*(Source: www.john-daly.com/co2diff.htm).*

### 3.8.4 Box–Jenkins forecasting

Quite a revolution in time series forecasting was brought about by the work of Box and Jenkins in the 1960's, which subsequently appeared in the first edition of their book on the topic in 1970. Whilst ARMA models had been studied for decades before, the key contributions of Box and Jenkins were

1. illustrating that various types of non-stationary effects can be removed by differencing and

2. formulating a coherent strategy for forecasting using ARIMA models.

As the approach gained great popularity in the forecasting community, ARIMA models became known as "Box–Jenkins models", and the methods are used today, greatly enhanced by modern computing power.

There are some variants on the forecasting procedure proposed by Box and Jenkins. In practice, provided the model fitted is reasonably good and the series fairly long, it may make little difference as to which "flavour" of Box–Jenkins prediction is undertaken.

To begin with, the following five–step procedure for model fitting should be performed:

1. If necessary, reduce the observed series to stationarity (usually by differencing and/or trend/seasonal effect removal, but whatever).

2. Having examined facets of the data (such as the acf and pacf), select an appropriate ARMA model for the time series.

3. Estimate the parameters for the fitted model.

4. Perform diagnostic measures to assess the goodness–of–fit of the model fitted.

5. If necessary, examine alternative models for the data from the ARMA family.

The final step in the forecasting process above can be performed in one of three different ways. It is useful to appreciate all three.

### (a) Using the model equation

Probably the most natural approach to finding point estimates of future values is using the equation which defines the model fitted. Provided a satisfactory model can be fitted following the process described above, the Box–Jenkins forecasting procedure can be applied by using

1. previous, observed values of $X$ and $Z$,
2. zero for future values of $Z$ which have not been observed and
3. the expectation of future values of $X$ for the prediction.

**Remark 8** *The final step above describes a* conditional expectation, *a notion which should be familiar to you. Essentially we are estimating* $x(N+l)$ *by the expectation* $E(X(N+l)|X(N), X(N-1), \ldots)$, *where the vertical line indicates that the expectation is* given *(or* conditional on*) the values given to its right.*

**Example 8** *For the AR(1) process defined by*

$$X(t) = \alpha X(t-1) + Z(t), \tag{3.5}$$

*obviously the value at time* $t+1$ *is*

$$X(t+1) = \alpha X(t) + Z(t+1).$$

*Now since* $E(Z(t)) = 0$ *for all* $t$, *the estimate of* $x(t+1)$ *is*

$$\hat{x}(t,1) = \alpha x(t),$$

*assuming we have observed a realisation of the process up to time* $t$. *For larger lead times we can use the forecast values to substitute for future values of* $X(t)$, *i.e., since*

$$X(t+2) = \alpha X(t+1) + Z(t+2)$$

*we let*

$$\begin{aligned} \hat{x}(t,2) &= \alpha \hat{x}(t,1) \\ &= \alpha^2 x(t). \end{aligned}$$

*This is equivalent to back–substituting using (3.5). A suitable estimate of* $\alpha$ *is required, of course.*

Models with an MA component require that we have "observed" values of the noise process $Z(t)$ for values of $t = 2, \ldots, N$, in order to predict future values of $X(t)$. The usual approach is to take the residual at time $t$ as the estimate of $z(t)$.

**Example 9** *Janacek and Swift (1993, p. 151–152) examine a time series which is a record of daily temperature readings. The data are initially deseasonalised, and then differenced to attain stationarity, to give a series* $x(2), \ldots, x(365)$ *(*$x(1)$ *is not defined, due to the initial series having been differenced). They fit an MA model to the data, an MA(4) in fact, of the form*

$$X(t) = Z(t) + 0.0722Z(t-1) - 0.3085Z(t-2) - 0.1312Z(t-3) - 0.2022Z(t-4).$$

*A sample of the data, fitted values and residuals are given below:*

| $t$ | $x(t)$ | $\hat{x}(t)$ | $\hat{z}(t)$ |
|---|---|---|---|
| *2* | *–0.13* | *0.1233* | *–0.2533* |
| *3* | *6.82* | *0.9238* | *5.8962* |
| *4* | *1.03* | *0.8811* | *0.1489* |
| *5* | *–0.02* | *–1.3516* | *1.3316* |
| ⋮ | ⋮ | ⋮ | ⋮ |
| *357* | *4.94* | *0.6258* | *4.3141* |
| *358* | *0.85* | *0.4079* | *0.4420* |
| *359* | *–0.64* | *–1.4111* | *0.77119* |
| *360* | *–4.62* | *–0.5794* | *–4.0406* |

*In order to estimate the following value in the series,* $x(361)$*, we estimate* $z(361)$ *by zero and earlier values of* $z(t)$ *by the residuals provided. This gives*

$$\begin{aligned} \hat{x}(360, 1) &= 0.0722(-4.0406) - 0.3085(0.7711) - 0.1312(0.4420) - 0.2022(4.3141) \\ &= -1.46. \end{aligned}$$

### (b) Using the MA version of the model

We should recall that an ARMA model can be written as an MA of possibly infinite order, specifically of the form

$$X(t) = \sum_{i=0}^{\infty} \psi_i Z(t-i),$$

so that

$$X(N+l) = Z(N+l) + \psi_1 Z(N+l-1) + \psi_2 Z(N+l-2) + \cdots .$$

Now since future values of $Z$ are not known at time $N$, and the best estimate of $Z(t)$ for $t > N$ is zero, we can replace $t$ by $N+l$ and write

$$\hat{x}(N,l) = \sum_{j=l}^{\infty} \psi_j z(N+l-j).$$

The forecast error in this is therefore

$$X(N+l) - \hat{x}(N,l) = Z(N+l) + \psi_1 Z(N+l-1) + \cdots + \psi_{l-1} Z(N+1).$$

This approach is most often used for finding confidence intervals for a predicted value as we will see. Obviously determination of the $\{\psi_i\}$ values is part of the procedure.

### (c) Using the AR version of the model

Recalling that for any ARMA model $X(t)$ there is some polynomial $\pi(B) = 1 - \sum_{i=1}^{\infty} \pi_i B^i$ such that

$$\pi(B) X(t) = Z(t).$$

At time $t = N+l$ we have

$$X(N+l) = \pi_1 X(N+l-1) + \pi_2 X(N+l-2) + \cdots + \pi_l X(N) + \cdots + Z(N+l).$$

Therefore, replacing values at times earlier than $N+l$ by either their observed or predicted value the forecast is

$$\hat{x}(N,l) = \pi_1 \hat{x}(N,l-1) + \cdots + \pi_l x(N) + \pi_{l+1} x(N-1) + \cdots$$

This allows for recursive updating as new observations are obtained.

**Example 10** *Recall from Chapter 2 that the ARMA*$(1,1)$ *process*

$$X(t) = 0.8X(t-1) + Z(t) - 0.2Z(t-1)$$

*may be written*

$$\pi(B) X(t) = Z(t)$$

*where*

$$\pi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$$

*and for* $j = 1, 2, \ldots,$

$$\pi_j = -0.6 \times 0.2^{j-1}.$$

*Hence*

$$X(N+l) = -0.6 \sum_{j=1}^{\infty} 0.2^{j-1} X(N+l-j) + Z(N+l).$$

*So, for example, given observations* $x(1), \ldots, x(N)$, *the forecast for the value at time* $N+1$ *would be*

$$\begin{aligned} \hat{x}(N,1) &= \sum_{j=1}^{\infty} \pi_j x(N+1-j) \\ &= -0.6 \sum_{j=1}^{N} 0.2^{j-1} x(N+1-j). \end{aligned}$$

*More generally the forecast for the value at time* $N+l$ *would be*

$$\begin{aligned} \hat{x}(N,l) &= \sum_{j=1}^{l-1} \pi_j \hat{x}(N,l-j) + \sum_{j=l}^{N+l-1} \pi_j x(N+l-j) \\ &= -0.6 \left( \sum_{j=1}^{l-1} 0.2^{j-1} \hat{x}(N,l-j) + \sum_{j=l}^{N+l-1} 0.2^{j-1} x(N+l-j) \right). \end{aligned}$$

**Justification for the Box–Jenkins approach**

Why should the approaches suggested above be in any sense optimal? Here we explain the way in which Box–Jenkins forecasting methods are "best" according to a key criterion.

Recall that our data–generating process $X(t)$ can be written in the form

$$X(t) = \sum_{i=0}^{\infty} \psi_i Z(t-i) \tag{3.6}$$

for some set of weights $\{\psi_i\}$. This is sometimes known as a *linear filter.* The "forecast generating process" $\hat{X}(N,l)$, of the value $X(N+l)$ based on observations up to time $N$, might reasonably be taken as a linear function of existing observations, so can be written

$$\hat{X}(N,l) = \sum_{j=0}^{N-1} \alpha_j X(N-j)$$

for some weights $\{\alpha_i\}$ – note this statement was the basis for exponential smoothing. Combining the two equations above, it must be that

$$\hat{X}(N,l) = \sum_{i=0}^{\infty} \omega_i Z(N-i) \tag{3.7}$$

for a set of weights $\{\omega_i\}$.

One way to assess the performance of a forecast–generating process is via its *mean squared error* (m.s.e.), defined by

$$E\left[\left(X(N+l) - \hat{X}(N,l)\right)^2\right].$$

Obviously for m.s.e., smaller is better. Now by (3.6) and (3.7) we see that the m.s.e. can be written

$$E\left[\left(\sum_{i=0}^{\infty} \psi_i Z(N+l-i) - \sum_{i=0}^{\infty} \omega_i Z(N-i)\right)^2\right].$$

This in turn can be written

$$E\left[\left(\sum_{i=0}^{l-1} \psi_i Z(N+l-i) + \sum_{i=l}^{\infty} (\psi_i - \omega_{i-l}) Z(N+l-i)\right)^2\right] = \sigma^2\left(\sum_{i=0}^{l-1} \psi_i^2 + \sum_{i=l}^{\infty} (\psi_i - \omega_{i-l})^2\right)$$

recalling that $\mathrm{Cov}(Z(t), Z(t+j)) = 0$ for $j \neq 0$. The choice of $\{w_i\}$ which minimises the m.s.e. above is clearly

$$\omega_i = \psi_{i+l}$$

for $i = 0, 1, 2, \ldots$. Adopting this choice, the forecasts are generated via

$$\begin{aligned}\hat{X}(N, l) &= \sum_{i=0}^{\infty} \omega_i Z(N-i) \\ &= \sum_{i=0}^{\infty} \psi_{i+l} Z(N-i) \\ &= \sum_{j=l}^{\infty} \psi_j Z(N+l-j)\end{aligned}$$

putting $j = i + l$. This is exactly the approach adopted in Box–Jenkins forecasting: a best (in terms of m.s.e.) forecast is found from the linear filter form for $X(N+l)$ by setting future, unobserved values of $Z(t)$ to zero. The forecast errors are then

$$X(N+l) - \hat{X}(N, l) = \sum_{i=0}^{l-1} \psi_i Z(N+l-i)$$

since

$$X(N+l) = \sum_{i=0}^{\infty} \psi_i Z(N+l-i)$$

by (3.6).

**Example 11** *Suppose that $X(t)$ follows an $ARIMA(0,1,1)$, meaning that $\nabla X(t)$ is an $MA(1)$. This states that*

$$X(t) = X(t-1) + Z(t) + \beta Z(t-1)$$

*for some $\beta$. Now*

$$X(N+1) = X(N) + Z(N+1) + \beta Z(N)$$

*and so setting $Z(N+1) = 0$, and the fact that*

$$Z(N) = X(N) - \hat{X}(N-1, 1)$$

*gives*

$$\begin{aligned}\hat{X}(N, 1) &= X(N) + \beta\left(X(N) - \hat{X}(N-1, 1)\right) \\ &= (1+\beta) X(N) - \beta \hat{X}(N-1, 1).\end{aligned}$$

*If $\beta < 0$ then putting $\alpha = 1 + \beta$ gives*

$$\hat{X}(N,1) = \alpha X(N) + (1-\alpha)\hat{X}(N-1,1),$$

*which is (3.3). So the Box–Jenkins approach provides a model–based underpinning for exponential smoothing.*

**Confidence intervals**

The MA representation (b), via

$$X(t) = \sum_{i=0}^{\infty} \psi_i Z(t-i)$$

gives confidence intervals for forecasts quite readily. These are sometimes termed *prediction intervals*, to distinguish from the estimation of fixed, unknown parameters.

Let the forecast error of the value $X(N+l)$ based on observations up to time $N$ be denoted

$$e(N,l) := X(N+l) - \hat{X}(N,l) = \sum_{i=0}^{l-1} \psi_i Z(N+l-i).$$

In practice this would be estimated by $x(N+l) - \hat{x}(N,l)$. Taking expectations

$$\begin{aligned} E(e(N,l)) &= \sum_{i=0}^{l-1} \psi_i E(Z(N+l-i)) \\ &= 0. \end{aligned}$$

Applying the variance operator we find

$$\begin{aligned} \mathrm{Var}(e(N,l)) &= \sum_{i=0}^{l-1} \psi_i^2 \mathrm{Var}(Z(N+l-i)) \\ &= \sigma^2 \sum_{i=0}^{l-1} \psi_i^2. \end{aligned}$$

The above value can be estimated by entering estimators $\hat{\sigma}^2$ and $\left\{\hat{\psi}_1, \ldots, \hat{\psi}_{l-1}\right\}$ obtained from the data and the model fitted. Moreover, assuming approximate Normality for $e(N,l)$, an approximate $(1-\alpha)\,100\%$ prediction interval is given by

$$\hat{x}(N,l) \pm z_{\alpha/2}\hat{\sigma}\sqrt{\sum_{i=0}^{l-1}\hat{\psi}_i^2},$$

where $z_{\alpha/2}$ is the $\frac{\alpha}{2}$ percentage point of the $N(0,1)$ distribution.

Note that

1. The larger $l$, the wider the prediction interval becomes.
2. The forecast errors are mutually correlated.
3. If $Z(t)$ is Normally distributed, the forecast errors are exactly Normally distributed.
4. For $X(t)$ to be stationary, it is required that $\mathrm{Var}(X(t)) < \infty$ for all $t$. Then

$$\begin{aligned}\lim_{l\to\infty}\mathrm{Var}(e(N,l)) &= \lim_{l\to\infty}\left(\sigma^2\left(\psi_0^2+\psi_1^2+\cdots+\psi_{l-1}^2\right)\right)\\ &= \mathrm{Var}(X(t))\end{aligned}$$

   by (3.6), the MA representation of $X(t)$. For example, if $X(t)$ is an AR(1), then

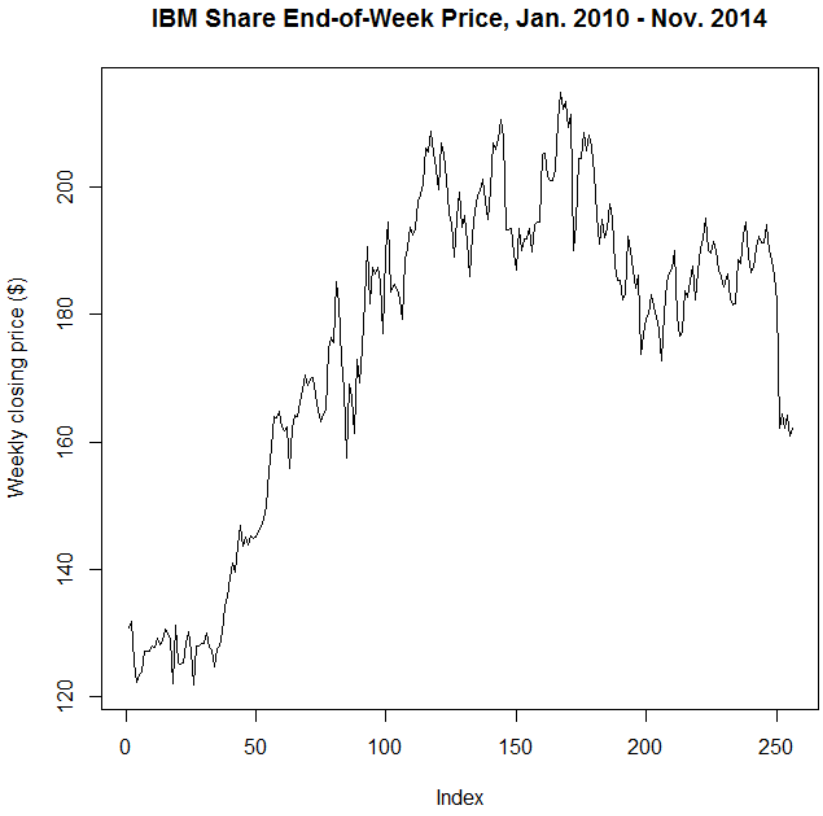
$$\mathrm{Var}(e(N,l)) \to \frac{\sigma^2}{(1-\alpha^2)}$$

   as $l \to \infty$.
5. Anecdotal evidence suggests that prediction intervals tend to be too narrow in practice. This may well be due in many cases to the underlying model changing from the time of the observed values to the time of the forecasts.

### 3.8.5 Examples

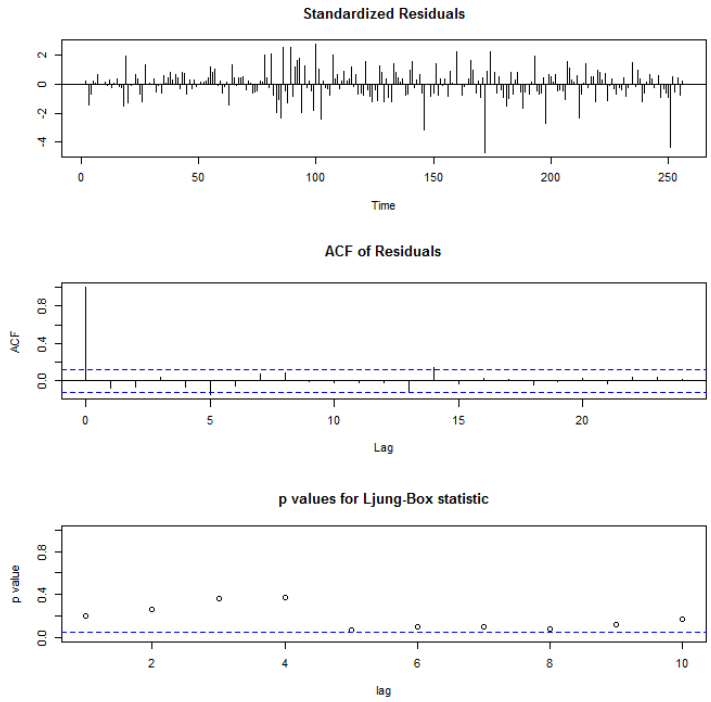
We outline a couple of case studies, using data sets which are available for analysis in R. The reader is urged to re-create the results presented, and also investigate and verify claims made about the data.

**Example 12** *The plot below is of the end–of–week unit closing prices for IBM, January 2010 to November 2014 inclusive. As is typical with such series, it is hard to discern whether there is an underlying trend, or whether perhaps the underlying process is simply noise with non-constant variance.*




*Attempting to fit an ARMA model to the raw data is unsuccessful: the acf decays very slowly, residuals are large in magnitude, and P-values for the Ljung–Box test are small. Differencing once is promising, with R fitting a white noise process to the resulting series. The summary of the model fit is*

*shown below:*




*Only a few of the 255 standardised residuals appear to be outside the range* $\pm 3$*, the acf of the residuals has one or two marginally significant values (but nothing at lags that would have obvious interpretation), and the P-values for the Ljung–Box test are broadly acceptable if somewhat low for certain lags. Although the fit is far from ideal, the model looks plausible with regards to the diagnostics at least. Hence we entertain a model of the form*

$$\nabla X(t) = X(t) - X(t-1) = Z(t),$$

*where* $Z(t)$ *is white noise with mean* $\mu$ *and variance* $\sigma^2$ *unknown. R gives the point estimate of* $\mu$ *to be 0.1228 (with a small estimated variance), and the estimate of the variance of* $Z(t)$ *is 20.9209. Since IBM shares at the end of November 2014 were priced at \$162.17, the one–step forecast would have 95% prediction interval*

$$162.17 + 0.1228 \pm 1.96 \times 4.5741$$

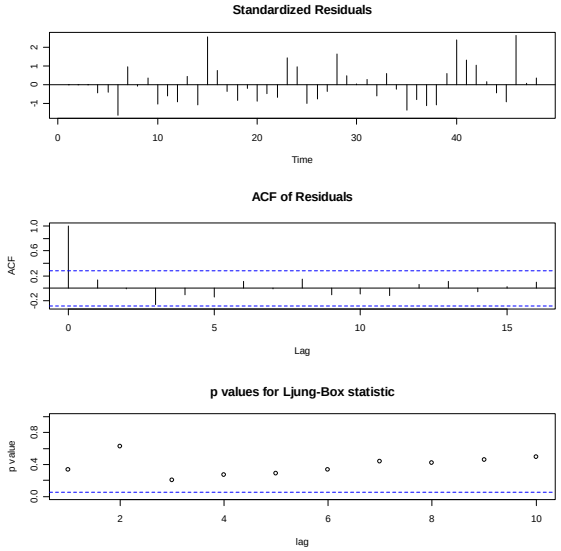
*in dollars. This is not entirely satisfactory as a prediction scheme: we are simply using the mean difference over the observed series to predict future*

*jumps, regardless of the most recent prices. Other models are equally plausible, though for short–term forecasting at least the random walk model would likely give satisfactory interval estimates.*

**Example 13** *Recall the lh data set discussed earlier, which contains 48 observations of the level of luteinizing hormone in the blood of a woman taken at ten minute intervals. This series is relatively short, so high hopes should not be harboured that either the model fitted or resulting forecasts are very reliable. The earlier analysis suggested an* $AR(1)$ *model was reasonable, though cases could be made for both* $MA(1)$ *and* $ARMA(1,1)$ *models being appropriate. Fitting an* $AR(1)$ *by m.l.e., the model is*

$$X(t) - 2.413 = 0.574\,(X(t-1) - 2.413) + Z(t)\,,$$

*where* $\hat{\sigma}^2 = 0.197$. *Note these estimates differ slightly from those found by least squares earlier, as might be expected. A summary of diagnostics for the model are plotted below:*




*Importantly there is no pattern apparent in the residuals, their acf having no significant lags. The plot of the p-values for a sequence of goodness–of–fit tests also indicates nothing untoward about the fit of the model. Using this model and the* `predict` *command in R we can find forecasts for the next four values as being*

$$2.6926,\ 2.5736,\ 2.5053,\ 2.4661.$$

*The estimated standard errors for these are respectively 0.4444, 0.5124, 0.5329 and 0.5395. To verify these computations, note firstly that* $x(48) = 2.9$ *is the final observation, so*

$$\begin{aligned}\hat{x}(48,1) &= \hat{\alpha}(x(48) - \hat{\mu}) + \hat{\mu} \\ &= 0.574(2.9 - 2.413) + 2.413 \\ &= 2.692.\end{aligned}$$

*Other forecasts are found similarly. To find the estimated standard errors, which for* $\hat{x}(48,l)$ *is*

$$\hat{\sigma}\sqrt{\sum_{i=0}^{l-1}\hat{\psi}_i^2},$$

*note that*

$$\begin{aligned}(1-\hat{\alpha})^{-1} &= (1-0.574)^{-1} \\ &= 1 + 0.574 + 0.574^2 + \cdots \\ &= \hat{\psi}_0 + \hat{\psi}_1 + \hat{\psi}_2 + \cdots .\end{aligned}$$

*Hence*

$$\hat{\sigma} = \sqrt{0.197} = 0.444$$

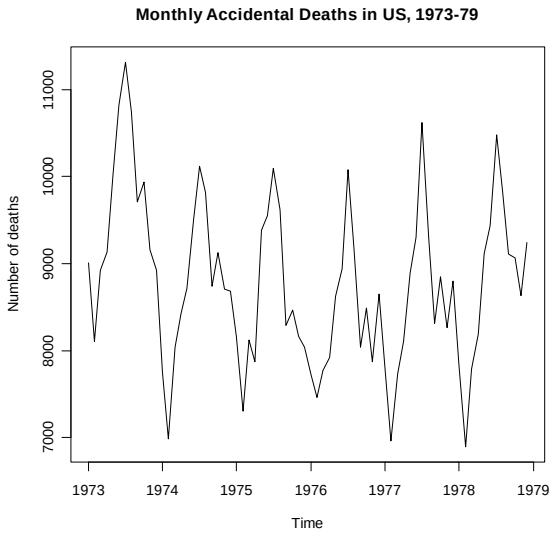
*gives the e.s.e. of* $\hat{x}(48,1)$*. In the same manner we find*

$$\hat{\sigma}\sqrt{1+\hat{\psi}_1^2} = 0.444\sqrt{1+0.574^2} = 0.512$$

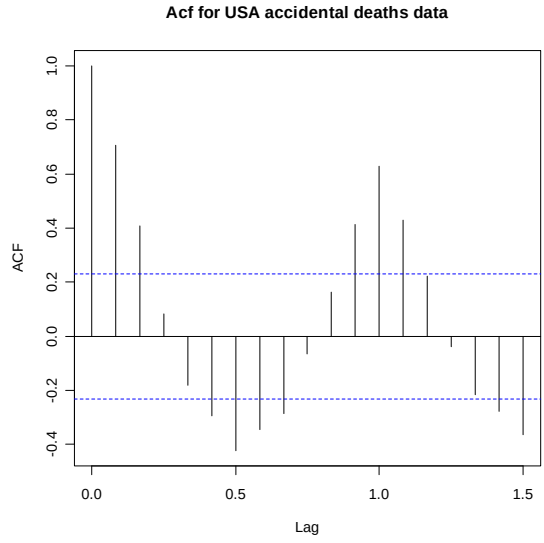
*is the estimate of the standard error of* $\hat{x}(48,2)$*, and so on for the other estimates.*

**Example 14** *The series comprising monthly accidental deaths in the USA from 1973 to 1978 is much studied, with the first analysis apparently appear-*

*ing in Brockwell and Davis (1991). The series is plotted below:*

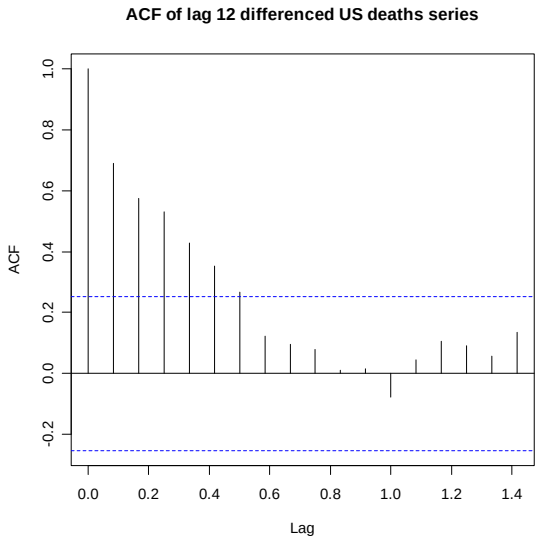



*As might be expected with monthly data, there are clear periodic components. The acf reveals this facet:*




*Note there are significant lags at* $k = 1, 6$ *and* $12$ *months. This suggests applying* $\nabla_{12}$ *to the series; the resulting acf, shown below, still does not resemble*

*that of a stationary series however.*

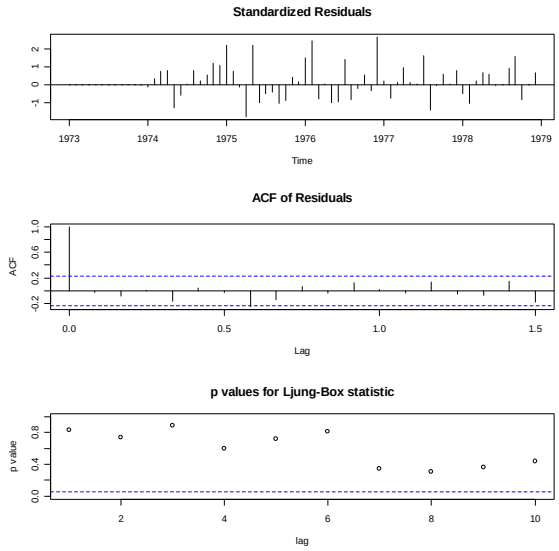



*Differencing once more gives a series with an acf with significant lags only at* $k = 1$ *and* $k = 12$*. This would suggest fitting a model of the SARIMA type of order* $(0,1,1) \times (0,1,1)_{12}$ *to the raw data, as there may well be MA components for both between–month and between–year changes. Fitting this model via m.l.e. using R we obtain*

$$\begin{aligned} W(t) &= \nabla\nabla_{12}X(t) \\ &= 28.31 + (1 - 0.430B)\left(1 - 0.533B^{12}\right)Z(t) \end{aligned}$$

*where the variance of* $Z(t)$ *is estimated to be 99346.86. Some model diag-*

*nostics are summarised as follows:*




*This appears to be a reasonable fit, though this should not discourage the investigation of competing models (Venables and Ripley (1994) suggest fitting an* $MA(13)$ *to* $W(t)$ *but with some of the terms removed* a priori. *It is not clear however that such an approach is possible in R.)*
*Using this model for forecasting the subsequent three months' data, we obtain 8336.06, 7531.83, and 8314.64. In fact the values for the first three months of 1979 were 7798, 7406, and 8363.*

## 3.9 Summary

Neither model fitting nor forecasting can be described as "automatic" procedures, where the user simply follows a set of rules in order to arrive at inferences. Rather, a little knowledge of the data, a grasp of the type of models which are candidates to fit to the data along with at least some experience of analysing time series can lead to fruitful modelling useful for prediction.

Prediction is a dangerous business (particularly predicting the future, as former US Vice President Dan Quayle once famously remarked), and one should not put great faith in *any* predictions, even those based on a seemingly good model which fits well to a lengthy series. Naturally, the further ahead

one expects to predict, the less reliable such predictions are likely to be. It is also beyond the scope of a model to predict outside events which would substantially influence future data. Should there, for example, be a major earthquake in Tokyo today, the effect on the Dow Jones Index, for instance, would be marked, and all predictions made prior to the disaster would "go out of the window". In some sense the world around us is "chaotic", and often signals that are present are swamped by noise.

## 3.10 Model fitting and forecasting in R

The R language has many built-in commands for estimation, model fitting, diagnostics and forecasting. Entering

> `library(tseries)`

will add the `tseries` and `zoo` packages, and entering

> `help.search(''tseries'')`

gives a list of useful commands which can in turn be investigated using the `help` facility. For ARIMA model fitting to a series `x` the two most used commands are `arma` and `arima`, the first of these having syntax

> `arma(x, order=c(p,q), lag=NULL, coef=NULL, ...).`

In the above `p` and `q` give the AR and MA orders respectively, and `coef` can be allocated a vector containing initial estimates of the parameters for the model. This command fits the model by (conditional) least squares. A somewhat more powerful alternative is offered by

> `arima(x, order=c(p,d,q), seasonal=list(order=c(P,D,Q), period=NA), include.mean=T, method=c(''CSS-ML'', ''ML'', ''CSS''), ...).`

Now the above will fit the full range of models to the series in `x`, notation following that given in the notes earlier. The `include.mean` command includes a non-zero $\mu$ by default unless the series has been differenced, in which case this switches to `False` by default. The `method` list indicates the order in which the fitting procedures are attempted, with `CSS-ML` using conditional least squares to obtain parameter estimates to initiate the maximum likelihood estimation.

For a more "automatic" model fitting process, AR models can be fitted via the `ar` and `ar.ols` commands, both of which will determine the order to be fitted using something called the *Akaike Information Criterion* (AIC) – this essentially chooses the model with highest (log) likelihood, but includes a "penalty" factor to reduce the chances of "over–fitting" a model with more

parameters than are necessary.

Model diagnostics can be studied for a fitted model via the command

```
> tsdiag(object, gof.lag, ...)
```

in which `object` is a fitted model and `gof.lag` stipulates the maximum lag in the portmanteau lack–of–fit test. The output is a collection of plots, notably including the acf of the residuals.

The `HoltWinters` command enables both the Holt and Holt–Winters smoothing and forecasting methods to be applied via a single command. The syntax is

```
> HoltWinters(x, alpha = NULL, beta = NULL, gamma = NULL, seasonal
= c("additive", "multiplicative"), start.periods = 3, l.start = NULL,
b.start = NULL, s.start = NULL, optim.start = c(alpha = 0.3, beta
= 0.1, gamma = 0.1), optim.control = list())
```

in which `x` is the time series, `alpha, beta` and `gamma` represent the parameters in the up-dating equations, `seasonal` specifies the type of seasonal effect (additive being the default) and `optim.start` indicates the suggested starting values for $\alpha$, $\beta$ and $\gamma$. Naturally setting `gamma` to zero (or `FALSE`) gives Holt's method, and setting both `gamma` and `beta` to `FALSE` gives simple exponential smoothing. If values for $\alpha$, $\beta$ and $\gamma$ are not specified, R attempts to estimate them by minimising the sum of the squares of the one–step ahead forecasts throughout the data. Setting the smoothed series to be an object, `object$coefficients` contains the estimated level, trend, and seasonal effects.

The `HoltWinters` command in fact defines a class of objects, and certain functions exist which operate on this class. One such is

```
> predict(object, n.ahead=1,...)
```

in which `object` is a `HoltWinters` object and `n.ahead` indicates the number of future values to be predicted. The same command can be used to forecast via an ARIMA model object, the argument `se.fit=TRUE` being added if the estimated standard errors of the predictions are to be computed.

To find the MA representation of an ARMA model, the command

```
> ARMAtoMA(ar = numeric(0), ma = numeric(0), lag.max)
```

in which `ar` is assigned a vector of numerical values, being the coefficients $(\alpha_1, \ldots, \alpha_p)$ in the AR portion of the ARMA model. Similarly `ma` is assigned the vector $\left(\beta_1, \ldots, \beta_q\right)$ in the MA part of the model. The `lag.max` argument gives the highest $k$ for which $\psi_k$ is calculated.

**Example 15** *For the* $ARMA(1,1)$

$$X(t) = 0.8X(t-1) + Z(t) - 0.2Z(t-1)$$

*met in chapter 2, to find the MA representation enter:*

```
>ARMAtoMA(0.8, -0.2, 5)
```

*The output from this is*

```
0.60000 0.48000 0.38400 0.30720 0.24576
```

*verifying the formula given in chapter 2.*

Finally the comman `ARMAacf` can be used to calculate numerical values for the theoretical acf and pacf for an ARMA model. Similar to `ARMAtoMA`, the syntax is

```
> ARMAacf(ar = numeric(0), ma = numeric(0), lag.max = r, pacf =
FALSE)
```

Setting `pacf = TRUE` gives pacf rather than acf values.

## 3.11 Exercises 3

1. Show that if $\bar{x}$ is the sample mean of a realisation of $N$ consecutive observations from a process $X(t)$ with variance $\sigma^2$ and autocorrelation function $\rho(k)$ then

$$\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{N}\left(1 + 2\sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)\rho(k)\right).$$

2. Using R or otherwise, fit the best ARIMA model to the following data, and comment briefly on the goodness of the fit:

   1.34, 1.38, 1.24, 1.45, 1.78, 1.23, 1.47, 1.64, 1.34, 1.55, 1.41, 1.75,
   1.33, 1.56, 1.44, 1.76, 1.87, 1.54, 1.46, 1.23, 1.43, 1.67, 1.66, 1.70.

3. The first 200 terms of a time series gave the following results:

| $k$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| acf $r_k$ | –0.80 | 0.67 | –0.52 | 0.39 | –0.31 |
| pacf $\hat{\alpha}_{kk}$ | –0.80 | 0.085 | 0.112 | –0.046 | –0.061 |

The mean of the observed series was $\bar{x} = 0.03$, and $c_0 = 3.34$. Fit an appropriate model to the series, justifying your choice.

4. Using the Yule–Walker equations, verify the formulae for the initial estimates of the parameters in an AR(2) model. Hence estimate the parameters of the AR(2) model when we observe $r_1 = 0.81$ and $r_2 = 0.43$.

5. The R command `arima.sim` can be used to simulate a series from a specified ARIMA model. Investigate this command by using it to simulate series of length 100 from (i) AR(1), (ii) MA(1) and (iii) ARMA(1,1) processes. For each series simulated, examine the sample acf and pacf.

6. Express an ARMA(1,1) model as a moving average process. Hence show that the series can only be stationary when $|\alpha| < 1$.

7. Show that for the MA(1) process

$$X(t) = Z(t) + \theta Z(t-1)$$

we have

$$\hat{x}(N, 1) = \theta z(N),$$

where in practice a suitable estimate would replace $\theta$. Deduce that forecasts at greater lead times would be zero.

8. For the MA(2) process

$$X(t) = Z(t) - \beta_1 Z(t-1) - \beta_2 Z(t-2)$$

find a formula for $\hat{x}(N, l)$.

9. Forecast the values at times $t = 362$ and $t = 363$ in the Janacek and Swift example.

10. The following model

$$X(t) = 0.5X(t-1) + Z(t) - 0.8Z(t-1) + 0.4Z(t-2)$$

has been fitted to a series, for which $x(N) = 3.24$, $z(N) = 0.64$ and $z(N-1) = 0.95$. Estimate the values of the process at times $N+1$, $N+2$ and $N+3$. If $x(N+1) = 1.6$, find $\hat{x}(N+1, 1)$ and $\hat{x}(N+1, 2)$.

11. For data from the AR(1) model of the form

$$X(t) = \alpha X(t-1) + Z(t),$$

obtain an expression for an approximate 95% confidence interval for $\hat{X}(N, l)$, in terms of $\hat{\sigma}$, $\hat{\alpha}$ and $l$.

12. Obtain an expression for the correlation between successive forecast errors, $e(N, l)$ and $e(N, l+1)$, in Box–Jenkins ARMA–model forecasting. Consider the AR(2) model

$$X(t) = 0.9X(t-1) - 0.4X(t-2) + Z(t)$$

for which we have observed $x(59) = 1.2$ and $x(60) = 0.8$. Forecast the values of the process at times $t = 61, 62$ and 63. The variance of the process $Z(t)$ is estimated to be 0.64. Provide approximate 95% confidence intervals for your forecasted values. Estimate the correlation between successive errors for your forecasts.

13. An approximation to exponential smoothing forecasts the future value of the series $x(1), \ldots, x(N)$ by

$$\hat{x}(N, 1) = \sum_{t=1}^{N} \alpha (1-\alpha)^{t-1} x(t)$$

for some $\alpha$. This can be implemented in R by defining a new function

```
> es <- function(x,alpha)
> {sum(rev(x)*alpha*(1-alpha)^(0:(length(x)-1)))}
```

**(a)** Determine what the R function `rev` does.

**(b)** Remove the final observation from the `lynx` data set, by creating

```
>newlynx <- lynx[1:113]
```

**(c)** Use the function `es` on the `newlynx` series, starting with $\alpha = 0.1$ and working through some other values of $\alpha$. Which value of $\alpha$ appears the best one at predicting the final value of `lynx`? Why is this forecasting approach not very sensible here?

**(d)** To remove the apparent cyclical effect in the data `newlynx`, try differencing the series at lag 10. Plot the new series, called `difflynx`, say.

**(e)** Use exponential smoothing to predict the next value of `difflynx`. Use this estimate and the data to predict the final value of `lynx`.

14. The data set `LakeHuron` in R gives the annual measurements of the level, in feet, of Lake Huron (in the Great Lakes region), 1875–1972.

    **(a)** Examine the acf and pacf for this series. Which ARMA model appears to fit the data best?

    **(b)** Examine appropriate diagnostics for your fitted model. Comment on the fit.

    **(c)** Use the `predict` command in R to forecast the Lake Huron level for the years 1973, 1974 and 1975. Construct 95% prediction intervals for each forecast.

    **(d)** Estimate the values of $\psi_1$ and $\psi_2$ in the MA representation of your chosen model. Use these figures to validate the prediction intervals found in (c).