Time Series Analysis – Lecture 7
Model Fitting and Checking

Dr. Kevin Burke

1 Introduction

In the previous lecture we looked at methods for identifying some reasonable models for a given set of data, i.e., the first step in the Box-Jenkins approach. We now consider the next two steps:

2. Model fitting: estimating model parameters based on the observed time series.
3. Model checking: assessing model adequacy by inspecting residuals).

In the context of this lecture, for ease of notation, we will use Y_t to represent a **stationary** series. That is, we assume that this series has already been transformed using a Box-Cox transformation and/or differencing and, hence, that Y_t is the transformed series.

Another important point is that in Lecture 4 we assumed for convenience that $E(Y_t) = 0$. However, in reality we may have a non-zero mean, $E(Y_t) = \mu$. We can incorporate a non-zero mean into an ARMA model by writing:

$$\begin{aligned} Y_t - \mu &= \phi_1(Y_{t-1} - \mu) + \cdots + \phi_p(Y_{t-p} - \mu) \\ &\quad + e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}. \end{aligned}$$

which can be written as

$$\begin{aligned} (1 - \phi_1 B - \cdots - \phi_p B^p)\, Y_t &= (1 - \phi_1 - \cdots - \phi_p)\mu \\ &\quad + (1 - \theta_1 B - \cdots - \theta_q B^q)\, e_t \\ \Rightarrow \phi(B)\, Y_t &= \beta_0 + \theta(B)\, e_t \end{aligned}$$

where $\beta_0 = (1 - \phi_1 - \cdots - \phi_p)\mu$ is called the **intercept**.

Note: For models incorporating differencing (i.e., ARIMA(p,d,q) with $d \neq 0$) it is customary *not* to have an intercept as doing so introduces deterministic trend, e.g., a random walk with drift $Y_t = \beta_0 + Y_{t-1} + e_t$.

2 Estimation Procedures

There are three main methods for estimating parameters:

- **Method of Moments**: equate sample moments to theoretical moments and solve for the parameters.
- **Least Squares**: find the parameters which minimise the squared-error (i.e., observed minus predicted).
- **Maximum Likelihood**: find the "most likely" parameter values for a given set of data - for this method we need to assume some distribution for the errors (typically a normal distribution).

3 Method of Moments

The method of moments is generally easy to implement and typically yields simple estimates. However, for time series models, this **only works well for AR models**. Its performance for MA or ARMA models is generally poor (Cryer & Chan, Chaper 7).

3.1 Approach for Time Series Models

For a given model we have a *theoretical* formula for the autocorrelations, ρ_k, in terms of the model parameters, ϕ and θ, i.e., we may write $\rho_k(\phi, \theta)$ to make this explicitly clear.

Recall from Lecture 2 that, based on the observed data, we can calculate the *sample* autocorrelations using:

$$r_k = \frac{\sum_{t=1}^{n-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}.$$

Thus, we set the theoretical and sample autocorrelations equal to each other and solve:

$$\rho_k(\theta, \phi) = r_k$$

to get estimates of the ARMA parameters. Note that there are $p + q$ parameters in an ARMA(p,q) model and, hence, we solve such $p + q$ equations:

$$\begin{aligned} \rho_1(\phi, \theta) &= r_1 \\ \rho_2(\phi, \theta) &= r_2 \\ &\vdots \\ \rho_{p+q}(\phi, \theta) &= r_{p+q}. \end{aligned}$$

The method of moments estimator for μ the sample mean, $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$, and, hence, the intercept is estimated as:

$$\hat{\beta}_0 = (1 - \hat{\phi}_1 - \cdots - \hat{\phi}_p)\bar{y}$$

where $\hat{\phi}_1 \dots \hat{\phi}_p$ are estimates coming from the previous step.

The last parameter that needs to be estimated is the variance parameter, σ_e^2. Since the variance of the process, $\text{Var}(Y_t) = \gamma_0$, will be a function of σ_e^2 we simply solve:

$$\gamma_0(\sigma_e^2) = \hat{\gamma}_0 = \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2$$

i.e., we set the theoretical variance equal to the sample variance.

Note that for sample variance we usually divide by $n-1$ but, in the case of time series data, this is less important as n is usually quite large.

3.2 Autoregressive

AR(1)

Since $\rho_1 = \phi$, we can estimate ϕ as

$$\hat{\phi} = r_1.$$

From this we can get the intercept term

$$\hat{\beta}_0 = (1 - \hat{\phi})\bar{y}.$$

In order to estimate the variance parameter, σ_e^2, note that $\gamma_0 = \frac{\sigma_e^2}{1-\phi^2}$ for an AR(1) process. Hence, we can solve

$$\frac{\hat{\sigma}_e^2}{1-\hat{\phi}^2} = \hat{\gamma}_0$$

where $\hat{\gamma}_0 = \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2$ is the sample variance estimate.

$$\begin{aligned} \Rightarrow \hat{\sigma}_e^2 &= \hat{\gamma}_0(1 - \hat{\phi}^2) \\ &= \hat{\gamma}_0(1 - r_1^2) \end{aligned}$$

since, from above, $\hat{\phi} = r_1$.

AR(2)

Recall that we developed the "Yule-Walker" equations

$$\begin{aligned} \rho_1 &= \phi_1 + \phi_2\rho_1 \\ \rho_2 &= \phi_1\rho_1 + \phi_2 \end{aligned}$$

Thus, replacing ρ_1 and ρ_2 with r_1 and r_2 we get

$$\begin{aligned} r_1 &= \phi_1 + \phi_2 r_1 \\ r_2 &= \phi_1 r_1 + \phi_2 \end{aligned}$$

which are linear simultaneous equations and, hence, can easily be solved (check this) to yield the estimates

$$\hat{\phi}_1 = \frac{r_1(1-r_2)}{1-r_1^2}$$

$$\hat{\phi}_2 = \frac{r_2 - r_1^2}{1-r_1^2}.$$

Of course then the intercept term is

$$\hat{\beta}_0 = (1 - \hat{\phi}_1 - \hat{\phi}_2)\bar{y}.$$

To estimate the error variance σ_e^2, we use the following relationship derived in Lecture 4:

$$\begin{aligned} \gamma_0 &= \phi_1\gamma_1 + \phi_2\gamma_2 + \sigma_e^2 \\ \Rightarrow \sigma_e^2 &= \gamma_0 - \phi_1\gamma_1 - \phi_2\gamma_2 \\ &= (1 - \phi_1\rho_1 - \phi_2\rho_2)\gamma_0 \\ &\quad \text{(factoring out } \gamma_0 \text{ and } \rho_k = \gamma_k/\gamma_0) \\ \Rightarrow \hat{\sigma}_e^2 &= (1 - \hat{\phi}_1 r_1 - \hat{\phi}_2 r_2)\hat{\gamma}_0 \end{aligned}$$

where $\hat{\gamma}_0$ is the sample variance.

AR(p)

For any AR process, we may derive the Yule-Walker equations:

$$\begin{aligned} \rho_1 &= \phi_1 + \phi_2\,\rho_1 + \cdots + \phi_p\,\rho_{p-1} \\ \rho_2 &= \phi_1\,\rho_1 + \phi_2 + \cdots + \phi_p\,\rho_{p-2} \\ &\vdots \\ \rho_p &= \phi_1\,\rho_{p-1} + \phi_2\,\rho_{p-2} + \cdots + \phi_p \end{aligned}$$

where we can replace ρ_k with r_k (for $k = 1, \ldots, p$) to get

$$\begin{aligned} r_1 &= \phi_1 + \phi_2\,r_1 + \cdots + \phi_p\,r_{p-1} \\ r_2 &= \phi_1\,r_1 + \phi_2 + \cdots + \phi_p\,r_{p-2} \\ &\vdots \\ r_p &= \phi_1\,r_{p-1} + \phi_2\,r_{p-2} + \cdots + \phi_p \end{aligned}$$

which is a system of p linear simultaneous equations. This can easily be solved to yield estimates $\hat{\phi}_1, \ldots, \hat{\phi}_p$.

With this we can estimate the intercept term

$$\hat{\beta}_0 = (1 - \hat{\phi}_1 - \cdots - \hat{\phi}_p)\bar{y}.$$

For an AR process, we showed in Lecture 4 that

$$\gamma_0 = \frac{\sigma_e^2}{1 - \phi_1\rho_1 - \cdots - \phi_p\rho_p}.$$

This relationship can be used to estimate σ_e^2:

$$\hat{\sigma}_e^2 = (1 - \hat{\phi}_1 r_1 - \cdots - \hat{\phi}_p r_p)\gamma_0.$$

3.3 Moving Average

MA(1)

For an MA(1) process, we have that

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

Thus, replacing ρ_1 with r_1 gives

$$r_1 = \frac{-\theta}{1+\theta^2}$$

$$\Rightarrow r_1\theta^2 + \theta + r_1 = 0.$$

Solving this quadratic yields two possible estimates for θ, i.e.,

$$\hat{\theta} = \frac{-1 \pm \sqrt{1-4r_1^2}}{2r_1}.$$

Let $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ represent these two estimates. Recall that, for invertibility of an MA(1) process, we require $|\theta| < 1$. It is very easy to show that

$$\hat{\theta}^{(1)}\hat{\theta}^{(2)} = 1$$

$$\Rightarrow \hat{\theta}^{(1)} = \frac{1}{\hat{\theta}^{(2)}}$$

and, hence, only one of the estimates will satisfy $|\hat{\theta}| < 1$. Thus, we use the invertible solution.

Due to the appearance of $\sqrt{1-4r_1^2}$ in the estimate, it is clear that there will only exist a real solution when $|r_1| < 0.5$.

For an MA(1) process, it is true that in theory $|\rho_1| < 0.5$. However, due to sampling error, in a given sample we can observe $|r_1| > 0.5$ (particularly when n is small and θ is close to 1). This can easily be confirmed by running the following code several times:

```
acf(arima.sim(n=40, model=list(ma=c(0.9)) ), plot=F)
```

Clearly then it is not very satisfactory that the method of moments does not yield an estimate in cases where $|r_1| > 0.5$.

MA(2)

For an MA(2) process we have that

$$\rho_1 = \frac{-\theta_1 + \theta_1\theta_2}{1+\theta_1^2+\theta_2^2}$$

$$\rho_2 = \frac{-\theta_2}{1+\theta_1^2+\theta_2^2}$$

Replacing ρ_1 and ρ_2 with r_1 and r_2 and solving can produce estimates $\hat{\theta}_1$ and $\hat{\theta}_2$. These non-linear equations typically require a Newton-Raphson procedure to find the solution.

MA(q)

The above can be extended to an MA(q) process to estimate the θ parameters. However, the equations quickly get complicated. Furthermore, the performance of the method of moments is generally poor for MA models.

We have not mentioned above how to estimate β_0 or σ_e^2 for an MA process. We will now see how this is done.

From Section 3.1, we have that $\hat{\beta}_0 = (1-\hat{\phi}_1 - \cdots - \hat{\phi}_p)\bar{y}$ for an ARMA process. Therefore, since an MA process does not have any ϕ parameters,

$$\hat{\beta}_0 = \bar{y}.$$

It is easy to show that for an MA(q) process,

$$\gamma_0 = (1+\theta_1^2+\cdots+\theta_q^2)\sigma_e^2$$

from which we can estimate the error variance:

$$\hat{\sigma}_e^2 = \frac{\hat{\gamma}_0}{1+\hat{\theta}_1^2+\cdots+\hat{\theta}_q^2}$$

where, as before, $\hat{\gamma}_0 = \frac{1}{n}\sum_{t=1}^{n}(y_t-\bar{y})^2$ is the sample variance and $\hat{\theta}_1, \ldots, \hat{\theta}_q$ come from the first step.

3.4 Autoregressive Moving Average

Since the method of moments does not perform well for models with MA terms, it is not satisfactory for general ARMA models. Nonetheless, we will briefly discuss the ARMA(1,1) case.

ARMA(1,1)

For an ARMA(1,1) process we have that

$$\rho_1 = \frac{(\phi-\theta)(1-\phi\theta)}{1-2\phi\theta+\theta^2}$$

$$\rho_2 = \phi\frac{(\phi-\theta)(1-\phi\theta)}{1-2\phi\theta+\theta^2}.$$

Hence, as before, replace ρ_k with r_k and solve for ϕ and θ. It is worth noting that $\frac{\rho_2}{\rho_1} = \phi$ which yields a straightforward estimate for ϕ:

$$\hat{\phi} = \frac{r_2}{r_1}$$

and, hence, we can then obtain $\hat{\theta}$ by solving

$$r_1 = \frac{(\hat{\phi}-\theta)(1-\hat{\phi}\theta)}{1-2\hat{\phi}\theta+\theta^2}$$

which produces two possible estimates. Only one of these corresponds to an invertible process.

Since

$$\gamma_0 = \sigma_e^2 \frac{1 - 2\phi\theta + \theta^2}{1 - \phi^2},$$

the error variance can be estimated as follows:

$$\hat{\sigma}_e^2 = \hat{\gamma}_0 \frac{1 - \hat{\phi}^2}{1 - 2\hat{\phi}\hat{\theta} + \hat{\theta}^2}.$$

Finally, the intercept is given by

$$\hat{\beta}_0 = (1 - \hat{\phi})\bar{y}.$$

4 Least Squares

Since the method of moments is not satisfactory for models with MA terms, we need to look towards other approaches.

The idea behind this is that we wish to minimise the *sum of squared errors*:

$$S(\beta_0, \phi, \theta) = \sum e_t^2 = \sum (y_t - \hat{y}_t)$$

where y_t is the observed value and $\hat{y}_t$ is the predicted value from the ARMA model. Thus, the aim of least squares is to find the values of β_0, ϕ and θ that minimise this sum; this can be achieved through differentiation.

As is usual in least squares settings, the error variance can be estimated as

$$\begin{aligned} \hat{\sigma}_e^2 &= \frac{1}{n-1} \sum \hat{e}_t^2 \\ &\approx \frac{1}{n} \sum \hat{e}_t^2 \qquad \text{(when } n \text{ is large)} \end{aligned}$$

4.1 Autoregressive

Consider an AR(1) model with intercept

$$Y_t = \beta_0 + \phi Y_{t-1} + e_t$$

$$\Rightarrow e_t = Y_t - \beta_0 - \phi Y_{t-1}$$

$$\Rightarrow S(\beta_0, \phi) = \sum e_t^2 = \sum_{t=2}^{n} (y_t - \beta_0 - \phi y_{t-1})^2$$

where the sum starts at $t = 2$ since at $t = 1$ we would have $e_1 = Y_1 - \beta_0 - \phi Y_0$. The issue here is that Y_0 is not a value we have observed. Thus, we start the sum at $t = 2$ to avoid this.

Differentiating with respect to β_0 gives

$$\frac{\partial S}{\partial \beta_0} = \sum_{t=2}^{n} 2(y_t - \beta_0 - \phi y_{t-1})(-1)$$

which we set equal to zero and then solve:

$$\begin{aligned} -2\sum_{t=2}^{n} (y_t - \hat{\beta}_0 - \phi y_{t-1}) &= 0 \\ \sum_{t=2}^{n} (y_t - \hat{\beta}_0 - \phi y_{t-1}) &= 0 \\ \sum_{t=2}^{n} y_t - (n-1)\hat{\beta}_0 - \phi \sum_{t=2}^{n} y_{t-1} &= 0 \\ &\text{(since there are } n-1 \text{ terms between 2 and } n) \\ \Rightarrow \hat{\beta}_0 = \frac{1}{n-1}\sum_{t=2}^{n} y_t \; - \phi \frac{1}{n-1} \sum_{t=2}^{n} y_{t-1}. \end{aligned}$$

When n is reasonably large we have that

$$\hat{\beta}_0 \approx \bar{y} - \phi\bar{y} = (1 - \phi)\bar{y}.$$

Thus, at $\beta_0 = \hat{\beta}_0$,

$$\begin{aligned} S &\approx \sum_{t=2}^{n} (y_t - \bar{y} + \phi\bar{y} - \phi y_{t-1})^2 \\ &= \sum_{t=2}^{n} [(y_t - \bar{y}) - \phi(y_{t-1} - \bar{y})]^2. \end{aligned}$$

Differentiating this with respect to ϕ gives

$$\frac{\partial S}{\partial \phi} \approx \sum_{t=2}^{n} -2[(y_t - \bar{y}) - \phi(y_{t-1} - \bar{y})](y_{t-1} - \bar{y}).$$

By setting this equal to zero and solving for ϕ, it is easy to obtain

$$\begin{aligned} \hat{\phi} &\approx \frac{\sum_{t=2}^{n} (y_t - \bar{y})(y_{t-1} - \bar{y})}{\sum_{t=2}^{n} (y_{t-1} - \bar{y})^2} \\ &= \frac{\sum_{t=1}^{n-1} (y_{t+1} - \bar{y})(y_t - \bar{y})}{\sum_{t=1}^{n-1} (y_t - \bar{y})^2} \end{aligned}$$

where, in the second line, we have changed the variables in the sums to match notation used in for the sample autocorrelation. Specifically, recall from Section 3.1 (and Lecture 2) that

$$r_1 = \frac{\sum_{t=1}^{n-1} (y_t - \bar{y})(y_{t+1} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2}.$$

Thus, it is clear that $\hat{\phi} \approx r_1$ apart from one missing term from the denominator, i.e., $(y_1 - \bar{y})^2$. However, this is negligible when n is large.

Clearly then the least squares estimates of β_0 and ϕ are numerically close to the method of moments estimates.

The error variance can be estimated as

$$\begin{aligned}\hat{\sigma}_e^2 &= \frac{1}{n}\sum_{t=2}^{n}\hat{e}_t^2\\ &= \frac{1}{n}\sum_{t=2}^{n}(y_t-\hat{\beta}_0-\hat{\phi}y_{t-1})^2.\end{aligned}$$

It is not difficult to show that this is also approximately equal to the method of moments estimator $\hat{\sigma}_e^2=\hat{\gamma}_0(1-r_1^2)$.

Although we will omit the details, it can be shown more generally that for AR(p) processes least squares and method of moments estimates are approximately equal.

4.2 Moving Average

Consider an MA(1) model with intercept

$$Y_t=\beta_0+e_t-\theta e_{t-1}$$

$$\Rightarrow e_t= Y_t-\beta_0+\theta e_{t-1}$$

Since this expression for the error term, e_t, itself depends on an error term, e_{t-1}, it is common to assume an initial value $e_0=0$ (i.e., set e_0 to its expected value) and then define the errors recursively as follows:

$$\begin{aligned}e_1 &= Y_1-\beta_0+\theta e_0= Y_1-\beta_0\\ e_2 &= Y_2-\beta_0+\theta e_1= Y_2-\beta_0+\theta(Y_1-\beta_0)\\ e_3 &= Y_3-\beta_0+\theta e_2= Y_3-\beta_0+\theta(Y_2-\beta_0)+\theta^2(Y_1-\beta_0)\\ &\vdots\end{aligned}$$

$$e_t=Y_t-\beta_0+\theta e_{t-1}=\sum_{i=1}^{t}\theta^{t-i}(Y_i-\beta_0).$$

Hence, the sum of squared errors is given by

$$S(\beta_0,\theta) \;=\; \sum_{t=1}^{n}e_t^2 \;=\; \sum_{t=1}^{n}\left[\sum_{i=1}^{t}\theta^{t-i}(y_i-\beta_0)\right]^2$$

which is a non-linear function in terms of the parameters θ and β_0. Hence, we require numerical methods such as Newton-Raphson to minimize this $S(\beta_0,\theta)$.

For more general MA(q) processes, the sum of squared errors can be defined in the same way as in the above MA(1) case; however, we now have q initial values, $e_0,e_{-1},\ldots,e_{-q+1}$, which are set equal to zero.

4.3 Autoregressive Moving Average

Consider an ARMA(1,1) model with intercept

$$Y_t=\beta_0+\phi Y_{t-1}+e_t-\theta e_{t-1}$$

$$\Rightarrow e_t=Y_t-\beta_0-\phi Y_{t-1}+\theta e_{t-1}.$$

from which we can define the errors recursively. To avoid the presence of Y_0, we start at $t=2$. In this case we $e_1=0$ as an initial value:

$$\begin{aligned}e_2 &= Y_2-\beta_0-\phi Y_1+\theta e_1 = Y_2-\beta_0-\phi Y_1\\ e_3 &= Y_3-\beta_0-\phi Y_2+\theta e_2 = Y_3-\beta_0-\phi Y_2\\ &\qquad\qquad +\theta(Y_2-\beta_0-\phi Y_1)\\ e_4 &= Y_4-\beta_0-\phi Y_3+\theta e_3 = Y_4-\beta_0-\phi Y_3\\ &\qquad\qquad +\theta(Y_3-\beta_0-\phi Y_2)\\ &\qquad\qquad +\theta^2(Y_2-\beta_0-\phi Y_1)\\ &\vdots\end{aligned}$$

$$e_t=Y_t-\beta_0-\phi Y_{t-1}+\theta e_{t-1}=\sum_{i=2}^{t}\theta^{t-i}(Y_i-\beta_0-\phi Y_{i-1})$$

and, hence, the sum of squared errors is

$$S(\beta_0,\phi,\theta) \;=\; \sum_{t=2}^{n}e_t^2 \;=\; \sum_{t=2}^{n}\left[\sum_{i=2}^{t}\theta^{t-i}(Y_i-\beta_0-\phi Y_{i-1})\right]^2.$$

For more general ARMA(p,q) models, least squares can be applied in a similar manner.

5 Maximum Likelihood

Although least squares often works well, the step whereby we arbitrarily set “initial” error values is somewhat unsatisfactory; in some cases this can impact the values of the estimates. This arbitrary choice of initial values is avoided by using maximum likelihood.

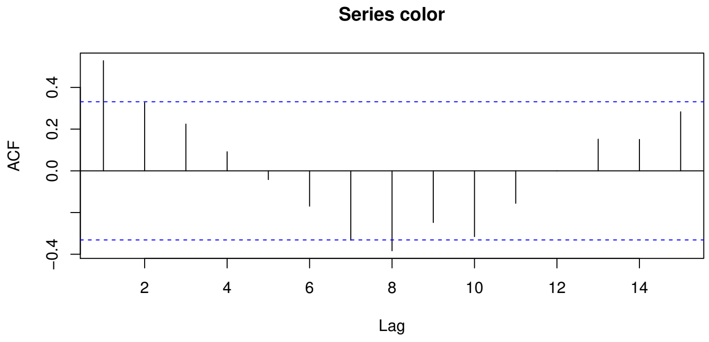
Maximum likelihood aims to find the parameter values for which the observed data is most likely. In order to proceed with this method, a joint probability distribution for the error terms must be assumed - typically a joint normal distribution. Although this is a slight disadvantage relative to least squares, where we did not need to assume a probability distribution, the assumption can be checked in practice, i.e., check that the residuals are approximately normally distributed.

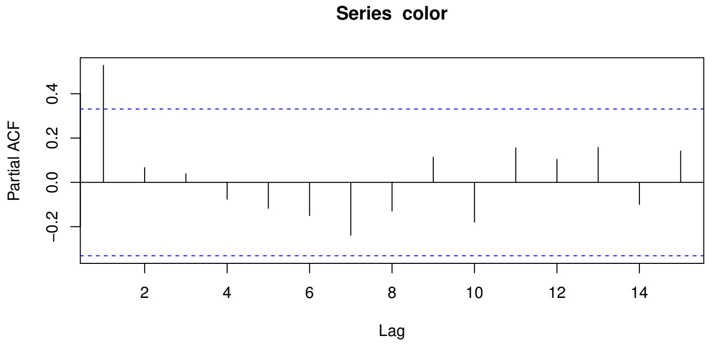
Note: Deriving likelihood functions for ARMA models is beyond the scope of this course.

Example 5.1. *Colour Property – Model Fitting*

We now consider the colour property dataset and fit a ARIMA model using the `arima` function in `R`.

The ACF and PACF were produced using `acf(color)` and `pacf(color)` respectively:





The ACF decays and the PACF cuts off after lag-1 suggesting an AR(1) model. This model is also suggested by the EACF:

```
AR/MA
  0 1 2 3 4 5 6 7
0 x o o o o o o o
1 o o o o o o o o
2 o o o o o o o o
3 x o o o o o o o
4 o o o o o o o o
5 x o o o o o o o
6 x o o o o o o o
7 x o o o o o o o
```

We fit an AR(1) model, i.e., ARIMA(1,0,0), using `arima(color, order=c(1,0,0))`.

```
Coefficients:
         ar1  intercept
      0.5705    74.3293
s.e.  0.1435     1.9151

sigma^2 estimated as 24.83:
log likelihood = -106.07,  aic = 216.15
```

Important: the `intercept` in `R`'s output is in fact an estimate of the process mean μ – not of β_0.

From the above we have $\hat{\phi} = 0.57$, $\hat{\mu} = 74.33$ and $\hat{\sigma}_e^2 = 24.83$. Hence, the estimated model is

$$\begin{aligned}
Y_t - \hat{\mu} &= \hat{\phi}\,(Y_{t-1} - \hat{\mu}) + e_t \\
Y_t - 74.33 &= 0.57\,(Y_{t-1} - 74.33) + e_t \\
Y_t &= (1 - 0.57)74.33 + 0.57\,Y_{t-1} + e_t \\
Y_t &= 31.96 + 0.57\,Y_{t-1} + e_t
\end{aligned}$$

where $\hat{\beta}_0 = 31.96$.

Note that the `R` output provides the standard error $\text{se}(\hat{\phi}) = 0.1435$ using standard properties of maximum likelihood. This can be used to produce a 95% confidence interval for $\hat{\phi}$:

$$0.5705 \pm 1.96\,(0.1435)$$

$$\Rightarrow [0.28924,\ 0.85176]$$

which does not include $\phi = 0$ and, hence, the AR parameter is statistically significant.

To show how the `arima` function works, we now show some other fitted other models.

- AR(2) model:

```
arima(x = color, order = c(2, 0, 0))

Coefficients:
         ar1     ar2  intercept
      0.5173  0.1005    74.1551
s.e.  0.1717  0.1815     2.1463

sigma^2 estimated as 24.6:
log likelihood = -105.92,  aic = 217.84
```

Note: the ϕ_2 parameter is not significantly different to zero.

The fitted model is:

$$\begin{aligned}
Y_t &= (1 - \hat{\phi}_1 - \hat{\phi}_2)\,\hat{\mu} + \hat{\phi}_1\,Y_{t-1} + \hat{\phi}_2\,Y_{t-2} + e_t \\
Y_t &= 28.34 + 0.5173\,Y_{t-1} + 0.1005\,Y_{t-2} + e_t
\end{aligned}$$

- MA(1) model:

```
arima(x = color, order = c(0, 0, 1))

Coefficients:
         ma1  intercept
      0.4443    74.7712
s.e.  0.1315     1.2752

sigma^2 estimated as 27.76:
log likelihood = -107.94,  aic = 219.88
```

The fitted model is:

$$\begin{aligned}
Y_t &= \hat{\mu} + e_t - \hat{\theta}\,e_{t-1} \\
Y_t &= 74.77 + e_t + 0.4443\,e_{t-1}
\end{aligned}$$

Note: `R` does not define MA models with minus signs before the coefficients as we have.

Note that of the three models - AR(1), AR(2) and MA(1) - the AR(1) model has the lowest AIC. Recall from Lecture7 (Section 3.5) that we aim to minimize the AIC.

- ARIMA(0,1,1) model, i.e., IMA(1,1):

```
arima(x = color, order = c(0, 1, 1))

Coefficients:
          ma1
      -0.3588
s.e.   0.1799

sigma^2 estimated as 28.15:
log likelihood = -105.05,  aic = 212.1
```

The fitted model is:

$$\begin{aligned}\nabla Y_t &= e_t - \hat{\theta}\, e_{t-1}\\ Y_t - Y_{t-1} &= e_t - 0.3588\, e_{t-1}\\ Y_t &= Y_{t-1} + e_t - 0.3588\, e_{t-1}\end{aligned}$$

Note that when we fit a model which incorporates differencing (i.e., ARIMA with $d \neq 0$), there is **no intercept**. This is standard practice for such models to avoid introducing deterministic trend as mentioned in Section 1.

Important: AIC values arising from models applied to differenced data are not comparable to those from undifferenced data. More generally, we can only compare AIC values from models applied to the same dataset, i.e., if we transform the data in any way (through differencing or power transformation), we can think of this as a new dataset.

6 Model Checking

6.1 Residuals

Recall that we are only interested in models which are invertible, i.e., those which can be written in the form:

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \pi_3 Y_{t-3} + \cdots + e_t$$

where the π weights are functions of the ϕ and θ parameters. This can be written more compactly as

$$\begin{aligned}e_t &= \pi(B)\, Y_t\\ &= (1 - \pi_1 B - \pi_2 B^2 - \pi_3 B^3 - \cdots)\, Y_t.\end{aligned}$$

Once we have fitted the ARIMA model to an observed time series, we can estimate the π weights (since these are functions of $\hat{\phi}$ and $\hat{\theta}$) and, hence, we can estimate the white noise process:

$$\hat{e}_t = Y_t - \hat{\pi}_1 Y_{t-1} - \hat{\pi}_2 Y_{t-2} - \hat{\pi}_3 Y_{t-3} - \cdots$$

$$\hat{e}_t = Y_t - \hat{Y}_t$$

$\Rightarrow$ residual = observed − fitted

Since e_t is a white noise process, **the residuals, $\hat{e}_t$, should behave as white noise if we have fitted an appropriate model**, i.e., there should be no significant time trends or autocorrelations. Aside from this, we also need to assess normality of $\hat{e}_t$ (as the maximum likelihood estimation procedure makes this assumption), check for constant variance and identify outliers.

We typically deal with the standardised residuals

$$\frac{\hat{e}_t}{\hat{\sigma}_e}$$

where $\hat{\sigma}_e$ comes from our estimation procedure (see previous sections). Assuming normality, we expect most standardised residuals to have values in the range ±3 and, hence, this can help us discover outliers.

Note: We have previously considered residual analysis in Lecture 3.

6.2 Normality

Normality can be checked using a histogram, Q-Q plot and the Shapiro-Wilk test.

Note that the histogram will also highlight outliers, i.e., standardised residuals outside of ±3

6.3 Constant Variance

Constant variance can be checked by plotting the fitted values, $\hat{Y}_t$, against the residuals, $\hat{e}_t$. From this plot we would expect a random scatter of points with no obvious pattern.

A common departure from non-constant variance is increasing variance with the level of the series (discussed in Lecture 5, Section 4.1). This would be apparent in the fitted-versus-residual plot where the range of $\hat{e}_t$ values would increase as $\hat{Y}_t$ increases.

6.4 Trends and Outliers

A timeplot of the residuals should contain no trend (e.g., linear, quadratic, seasonal) if we have adequately modelled the series. If any trends are observed, then we should go back to our model selection step.

The timeplot will also reveal standardised residuals outside of ±3.

6.5 Sample Autocorrelation

There should be no significant autocorrelations in the ACF for the residuals if these are white noise.

Note however that we would expect to see some significant autocorrelations just by chance, i.e., we expect one in twenty outside of the 95% confidence bands $\pm 1.96\frac{1}{\sqrt{n}}$.

6.6 Ljung-Box Test

In addition to looking at individual sample autocorrelations and comparing them to the 95% confidence bands, it is useful to consider the magnitude of these autocorrelations as a group. For example, it may be that no one autocorrelation is significant but taken together as a group, they seem excessive.

The **Ljung-Box test** considers the first K autocorrelations (for the residuals) as a group. The null and alternative hypotheses are as follows:

$$\begin{aligned} H_0 &: \text{No significant autocorrelation} \\ H_1 &: \text{Significant autocorrelation.} \end{aligned}$$

The test statistics is:

$$Q = n\,(n+2)\left(\frac{\tilde{r}_1^2}{n-1} + \frac{\tilde{r}_2^2}{n-2} + \cdots + \frac{\tilde{r}_K^2}{n-K}\right)$$

where $\tilde{r}_k$ is the lag-k autocorrelation for the residual series, $\hat{e}_t$. It can be shown that

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

i.e., the test statistic has a chi-squared distribution with $K - p - q$ degrees of freedom where p and q are the orders of the AR and MA components of the fitted ARIMA(p, d, q) model. Hence, we reject the null hypothesis at the α level if $Q > \chi^2_{K-p-q,\ \alpha}$.

6.7 Overfitting

Another technique used to check the fitted model is called overfitting. Note this does not involve residuals.

After specifying and fitting what we believe to be an adequate model, we fit a slightly more general model. For example, if we thought an AR(1) model was reasonable, we could also fit an AR(2) model or an ARMA(1,1) model. If the original model is reasonable then we should expect that the additional parameters in these more general models will be non-significant and the AIC values will be higher.

Example 6.1. *Colour Property – Residual Analysis*

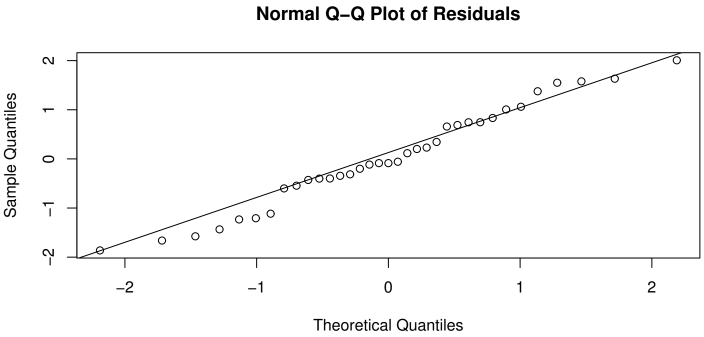
In Example 5.1, we decided that an AR(1) model appeared reasonable. We now analyse the residuals from this fitted model.

Residuals can be extracted as follows:

```
colormodel <- arima(color, order=c(1,0,0))
resid <- rstandard(colormodel)
```

To investigate normality of residuals:

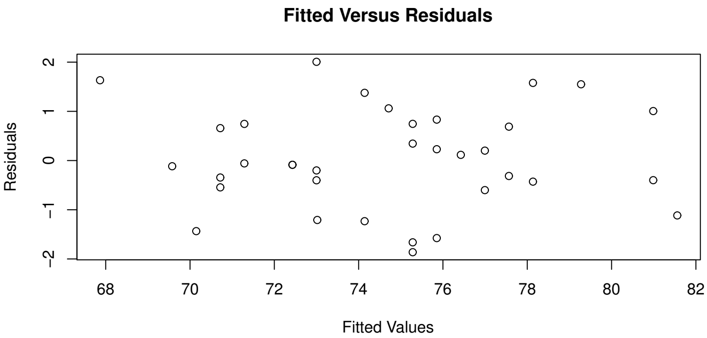
```
qqnorm(resid); qqline(resid);
shapiro.test(resid)
```



Based on the above Q-Q plot, the residuals appear reasonably normal and this is supported by the Shapiro-Wilk test where p-value = 0.6057.

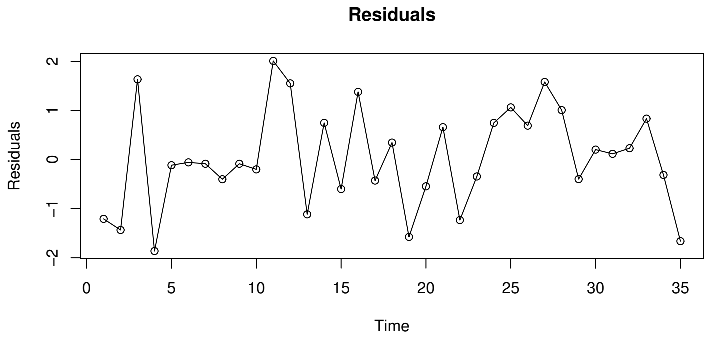
The fitted values were plotted against the residuals as follows:

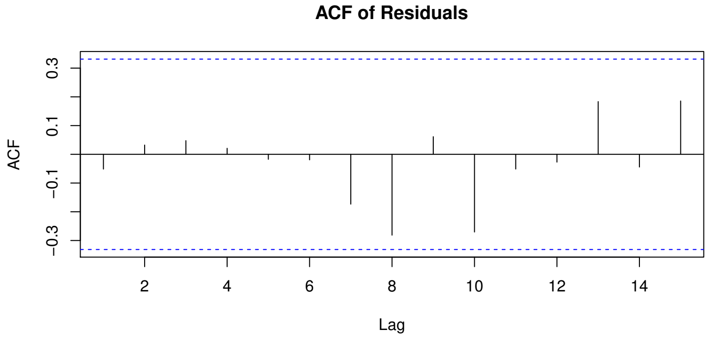
```
fit <- fitted(colormodel)
plot(as.vector(fit), as.vector(resid))
```



The points seem relatively randomly scattered with no evidence of patterns or non-constant variance.

The time series plot of the residuals below, i.e., `plot(resid)`, shows no evidence of trend. Furthermore, there does not appear to be any unusual residuals (i.e., outliers).





The ACF of the residuals (`acf(resid)`) displays no significant values, i.e., the residuals appear to be white noise.

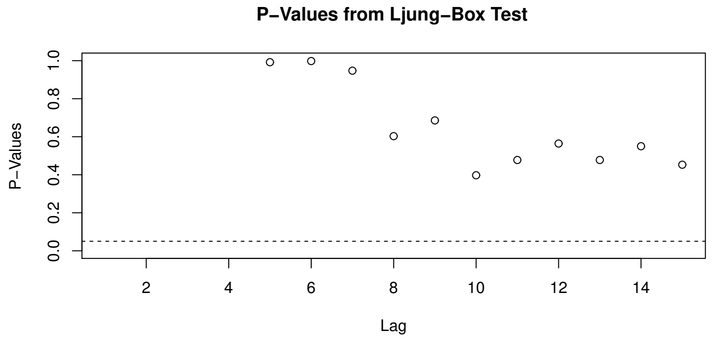
The Ljung-Box test was applied to the first $K = 5$ autocorrelations as a group as well as for $K = 10$ and $K = 15$.

```
LB.test(colormodel, lag=5)
LB.test(colormodel, lag=10)
LB.test(colormodel, lag=15)
```

The resulting p-values are 0.992, 0.3973 and 0.4527 respectively which all support the hypothesis that the residuals are white noise. Note that the runs test of independence (`runs(resid)`) leads to a p-value of 0.76 which also supports the hypothesis that the residuals are white noise.

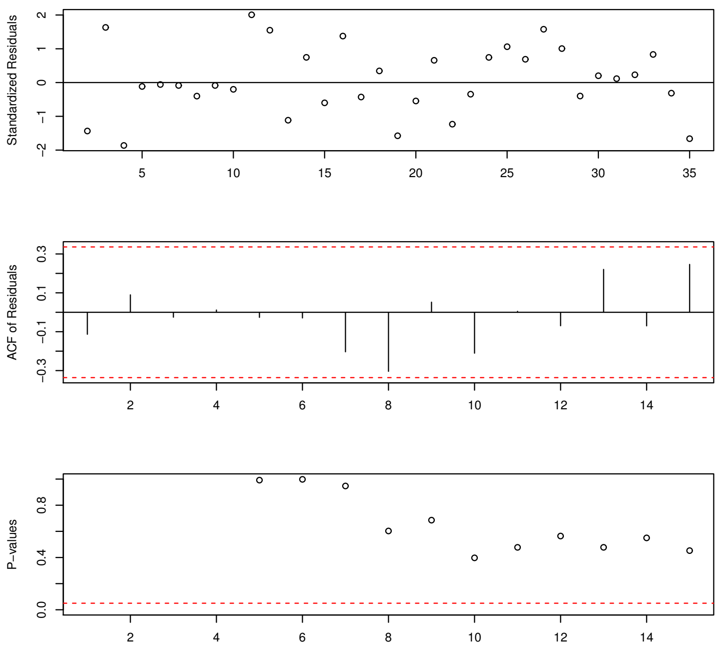
We have considered the Ljung-Box test for three different K values above. However, we may wish to consider a range of different K values all at once. It is common to plot a series of Ljung-Box p-values against K along with a horizontal reference line at 0.05 to help judge the size of the p-values. Such a plot can be produced as follows:

```
LBpvals <- rep(NA, 15)

for(i in 5:15){
  LBpvals[i] <- LB.test(colormodel, lag=i)$p.value
}

plot(LBpvals, ylim=c(0,1)); abline(h=0.05, lty=2)
```



Clearly all of the p-values are well above the 0.05 reference line which supports the hypothesis that the residuals are white noise.

It is worth noting that the last three diagnostic plots (time plot, ACF and Ljung-Box p-values) can be produced simultaneously using `tsdiag(colormodel)`:



All of the above points to the AR(1) model being a good fit for this time series.

Finally, note that in Example 5.1 we also considered an AR(2) model. However, the ϕ_2 parameter is not significant in this model and the AIC was higher than the AR(1) model. Again, this points to the AR(1) model.