Handout 8

[Chapters 2.1-2.3, 3.1, 3.3 and 5.1.1 in Brockwell and Davis.]

In this handout we will be mostly concerned with autoregressive models. In handout 6 concepts of $AR(1)$, $AR(2)$, $MA(1)$ and $MA(2)$ models were introduced. We can have autoregressive models of order higher than 2. Similarly, we can have moving average models of order higher than 2. Yet there is another class of models which have both the autoregressive and moving average parts and they are called $ARMA$ models. As in regression analysis, the coefficients of any model are not known and have to be estimated from the data.

In all the cases here, $\{\varepsilon_t\}$ denotes a white noise sequence (i.e., with mean zero) and variance σ^2. [In the text, white noise sequence is denoted by $\{Z_t\}$].

$AR(p)$ model: A sequence $\{X_t\}$ with mean μ is said to be an autoregressive model of order p (written as $AR(p)$) if it is of the form

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \cdots + \phi_p(X_{t-p} - \mu) + \varepsilon_t, \tag{2}$$

$MA(q)$ model: A sequence $\{X_t\}$ with mean μ is said to be a moving average model of order q if it is of the form

$$X_t - \mu = \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q},$$

$ARMA(1,1)$ model: A sequence $\{X_t\}$ with mean μ is said to have an $ARMA(1,1)$ model (autoregressive order=1 and moving average order=1) if it can be written in the form

$$X_t - \mu = \phi(X_{t-1} - \mu) + \varepsilon_t + \theta\varepsilon_{t-1},$$

$ARMA(2,1)$ model: A sequence $\{X_t\}$ with mean μ is said to have an $ARMA(2,1)$ model (autoregressive order=2 and moving average order=1) if it can be written in the form

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \phi_2(X_{t-2} - \mu) + \varepsilon_t + \theta_1\varepsilon_{t-1},$$

$ARMA(p,q)$ model: A sequence $\{X_t\}$ with mean μ is said to have an $ARMA(p,q)$ model (autoregressive order=p and moving average order=q) if it can be written in the form

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \cdots + \phi_p(X_{t-p} - \mu) + \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q},$$

Remarks

1. Any moving average sequence $MA(q)$ is stationary.

2. An $AR(p)$ sequence is not necessarily stationary. Conditions are needed on the coefficients $\{\phi_1, \phi_2, ...\}$ for the sequence to be stationary. Similarly, an $ARMA(p,q)$ model is not stationary (unless $p = 0$), and conditions are needed on the coefficients $\{\phi_1, \phi_2, ...\}$ of the autoregressive part in order for the sequence $\{X_t\}$ to be stationary. We will discuss these issues later.

3. As in regression analysis, one fits various $ARMA$ models and chooses an appropriate one by using a model selection criterion such as AIC, AICC, BIC etc.

Autoregressive models

One important class of time series models is given by autoregressive models. A mean zero sequence $\{X_t\}$ is said to be an autoregressive model of order p if it is of the form

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t, \tag{1}$$

where $\{\varepsilon_t\}$ is mean zero while noise with variance σ^2. Similarly, a sequence $\{X_t\}$ with mean μ is said to be an autoregressive model of order p if it is of the form

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \cdots + \phi_p(X_{t-p} - \mu) + \varepsilon_t, \tag{2}$$

where, as before, $\{\varepsilon_t\}$ is mean zero while noise with variance σ^2.

Note that the model in (2) can be simply rewritten as

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t,$$

where $\phi_0 = \mu(1 - \phi_1 - \cdots - \phi_p)$. This is simply a linear regression model with the dependent variable X_t and the independent variables $X_{t-1}, ..., X_{t-p}$. Clearly, this model is particularly simple for forecasting.

Backshift notation: A common notation in time series is that of a "backshift operator B". It is defined to be $X_{t-1} = BX_t$. So $X_{t-2} = B^2X_t$, $X_{t-3} = B^3X_t$ etc.

Note that the autocovariances and autocorrelations of the two sequences given in (1) and (2) are the same. For the purpose of discussion we will refer to model (1) for notational simplicity, since we can always subtract the mean μ or estimated mean $\bar{X}$ from the series.

A common notation in time series is that of what is called a backshift operator B. For instance $X_{t-1} = BX_t$, $X_{t-2} = B^2X_t$ etc. For instance $AR(1)$ model can be written as

$$X_t = \phi_1 BX_t + \varepsilon_t, \text{ or } X_t - \phi_1 BX_t = \varepsilon_t, \text{ or } (1 - \phi_1 B)X_t = \varepsilon_t.$$

Similarly an $AR(2)$ model can be expressed as

$$X_t = \phi_1 BX_t + \phi_2 X_{t-2} + \varepsilon_t, \text{ or } X_t - \phi_1 BX_t - \phi_2 B^2 X_t = \varepsilon_t, \text{ or } (1 - \phi_1 B - \phi_2 B^2)X_t = \varepsilon_t, \textit{or } \phi(B)X_t = \varepsilon_t,$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2$. Similarly, an $AR(p)$ model can be written as

$$(1 - \phi_1 B - \cdots - \phi_p B^p)X_t = \varepsilon_t, \text{ or } \phi(B)X_t = \varepsilon_t, \text{ with } \phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p.$$

Forecasting with autoregressive models

We assume that we have observations $X_1, \ldots, X_n$ from an AR(p) series. The goal is to forecast $X_{n+1}, X_{n+2}, \ldots$ when the parameters of the model are known, or when the parameters have been estimated which is usually the case in practice. Since an autoregressive model is like an usual regression model, forecasting is rather straightforward.

Example 1 ($AR(1)$ model for Sunspot series) This series has observations from $1700-2012$, thus length of the series is $n = 313$. Histogram of the original series is quite skewed and, for this reason, we consider the square root of the series which leads to a much less skewed histogram. Call the square rooted transformed series $\{S_t\}$. Mean of the series is 6.3977 and we will work with the mean subtracted series $\{X_t = S_t - 6.3977\}$ (thus $\bar{X} = 0$). Computer output shows,

$$\hat{\phi} = 0.8299, s(\hat{\phi}) = 0.0314, \hat{\sigma}^2 = 2.683, AIC = 1202.34.$$

As in regression, the fitted values are given by $\hat{X}_t = \hat{\phi} X_{t-1}$, $t = 2, ..., n$. Now suppose that we want to predict the sunspot numbers for the years 2013 and 2014, i.e., at times $t = n+1, n+2$. Since $X_n = 1.198302$, the predicted values for X_{n+1} is

$$\hat{X}_{n+1} = \hat{\phi} X_n = (0.8299)(1.198302) = 0.9945$$

Now the predicted value of $\hat{X}_{n+2} = \hat{\phi} X_{n+1}$. Since we do not have the data at time $t = n+1$, we substitute X_{n+1} by its guessed value. Hence the predicted value of X_{n+2} is

$$\hat{X}_{n+2} = \hat{\phi} \hat{X}_{n+1} = (0.8299)(0.9945) = 0.8253.$$

Consequently, the predicted values of S_{n+1} and S_{n+2} are

$$\hat{S}_{n+1} = 6.3977 + \hat{X}_{n+1} = 7.3922, \ \hat{S}_{n+2} = 6.3977 + \hat{X}_{n+2} = 7.2230.$$

Example 2. (AR(2) model for Sunspot series) Suppose that we decide to fit an $AR(2)$ model instead of $AR(1)$, then let us see how the forecasting is done. Of course, a reasonable question is which model is better: $AR(1)$ or $AR(2)$? This can be decided by using a model selection criterion such as AIC, AICC or BIC. We will take up this issue later. For the sunspot data, we have defined $\{X_t\}$ to be the mean subtracted series as in Example 1. The estimated $AR(2)$ parameters are

$$\hat{\phi}_1 = 1.4109, \hat{\phi}_2 = -0.7000, s(\hat{\phi}_1) = 0.0401, s(\hat{\phi}_2) = 0.0401, \hat{\sigma}^2 = 1.361, AIC = 993.28.$$

Here $X_{n-1} = 1.065494$, $X_n = 1.198302$. So the predicted value of X_{n+1} is

$$\hat{X}_{n+1} = \hat{\phi}_1 X_n + \hat{\phi}_2 X_{n-1} = 0.9448$$

Now the predicted value of X_{n+2} is $\hat{\phi}_1 X_{n+1} + \hat{\phi}_2 X_n$. The value of X_n is available, but the value of X_{n+1} is not. Hence, if substitute X_{n+1} by $\hat{X}_{n+1}$ the predicted value of X_{n+2} is

$$\hat{X}_{n+2} = \hat{\phi}_1 \hat{X}_{n+1} + \hat{\phi}_2 X_n = 0.4942.$$

Consequently the predicted values of S_{n+1} and S_{n+2} are

$$\hat{S}_{n+1} = 6.3977 + \hat{X}_{n+1} = 7.3425, \hat{S}_{n+2} = 6.3977 + \hat{X}_{n+2} = 6.8919.$$

Note that the predicted values using $AR(1)$ and $AR(2)$ are somewhat different. Note that we should choose $AR(2)$ model over $AR(1)$ since it has a smaller AIC value.

Example 3. (Temperature series). Recall that the temperature series $\{Y_t : t = 1, ..., n = 163\}$ was expressed as $Y_t = m_t + X_t$, where m_t is the trend. We estimated the trend by a loess method (with $span = 41/163 \approx 0.25$) and obtained the estimated X_t values $Y_t - \hat{m}_t$. Recall that $t = 163$ corresponds to the year 2012. Now we would like to predict the temperature for the years 2013 and 2014, which corresponds to times $n + 1$ and $n + 2$. In this case the predicted values of Y_{n+1} and Y_{n+2} are

$$\hat{Y}_{n+1} = \hat{m}_{n+1} + \hat{X}_{n+1} \text{ and } \hat{Y}_{n+2} = \hat{m}_{n+2} + \hat{X}_{n+2}.$$

The values of $\hat{m}_{n+1}$ and $\hat{m}_{n+2}$ are obtained by using extrapolation of $\{\hat{m}_1, ..., \hat{m}_n\}$, and the values of $\hat{X}_{n+1}$ and $\hat{X}_{n+2}$ are obtained by forecasting the series $\{X_t : t = 1, ..., n\}$. At present, we will try only AR models up to order 3. It turns out that the AICC value is the smallest for the $AR(3)$ model. The ACF plot shows that the residuals of this model can be taken to be the white noise. The parameters of the fitted model are

$$\begin{aligned} \hat{\phi}_1 &= 0.2172, \hat{\phi}_2 = -0.0468, \hat{\phi}_3 = -0.2001, \hat{\mu} = 0.0004 \\ s(\hat{\phi}_1) &= 0.0768, s(\hat{\phi}_2) = 0.0788, s(\hat{\phi}_3) = 0.0770 \end{aligned}$$

Predicted values $\hat{X}_{n+1}$ and $\hat{X}_{n+2}$ are 0.1015689 and 0.1039362, respectively.

Now the estimated values of m_{n+1} and m_{n+2} are $0.8184320, 0.8233173$, respectively. [Using the function "approxExtrap" in Hmisc package].

So the predicted temperature in years 2013 and 2014 are

$$\begin{aligned} \hat{Y}_{n+1} &= \hat{m}_{n+1} + \hat{X}_{n+1} = 0.8184320 + 0.1015689 = 0.9200, \\ \hat{Y}_{n+2} &= \hat{m}_{n+2} + \hat{X}_{n+2} = 0.8233173 + 0.1039362 = 0.9273. \end{aligned}$$

How do we select the order of the $AR(p)$ model

We can select an appropriate model by using a criterion such as AIC, AICC or BIC. We will use the sunspot data as the example here for tentative identification of a model. For the mean subtracted sunspot

data, the table here gives the values of AIC, AICC and BIC for orders 0 through 5. [$AR(0)$ means $X_t = \varepsilon_t$, i.e., $\{X_t\}$ is white noise.]

Note that AIC, $AICC$ and BIC values are (with some modifications for time series)

$$\begin{aligned} AIC &= n\ln(\hat{\sigma}^2) + 2p, \\ AICC &= n\ln(\hat{\sigma}^2) + 2p + 2p(p+1)/(n-p-1) = AIC + 2p(p+1)/(n-p-1), \\ BIC &= n\ln(\hat{\sigma}^2) + [\ln(n)]p. \end{aligned}$$

It should be pointed out that different packages (or even different programs in the same package) may report different values of AIC, AICC and BIC. For instance, output using the "sarima" function in ASTSA (in R) will produce the values of $AIC/n + 1$, $AICC/n + 1$ and BIC/n, but they are still labeled (by R) as $AIC, AICC$ and BIC, respectively.

p	0	1	2	3	4	5
AIC	3.153	1.993	1.321	1.318	1.324	1.328
$AICC$	3.159	2.000	1.328	1.324	1.331	1.335
BIC	2.153	1.005	.345	.353	.371	.388

According to AIC and AICC criteria, out of these six models, $AR(3)$ seems to be the most appropriate one, whereas BIC selects AR(2). The ACF plot of $AR(3)$ is given here. The plot does indicate that the residuals (i.e., $X_t - \hat{X}_t$) is not a white noise series. And this indicates that further analysis (and perhaps a model different from AR(3)) is needed for the sunspot data.

Estimation of parameters in AR.

Let us begin the discussion with the AR(1) case when the observations $X_1, \ldots, X_n$ are available. The AR(1) series can be expressed as $X_t = \phi_0 + \phi_1 X_{t-1} + \varepsilon_t$ with $\phi_0 = (1 - \phi_1)\mu$. Note that X_t is the dependent variable, and X_{t-1} is the independent variable. The method of least squares minimizes the sum of squares $\sum(X_t - f_0 - f_1 X_{t-1})^2$ with respect to f_0 and f_1 leading to the usual normal equations. Estimates of ϕ_0 and ϕ_1 are obtained by solving the normal equations. Let us now examine the range of 't' in the summation. We cannot have $t = 1$ in the summation since X_0 is not available. Similarly, we cannot use the independent variable X_n (ie, when $t - 1 = n$) since X_{n+1} is not available. Thus in order to follow the least squares method, we need to minimize $\sum_{t=2}^{n}(X_t - f_0 - f_1 X_{t-1})^2$ with rerspect to f_0 and f_1. However, in order to use the entire data, we would like to minimize

$$(X_1 - f_0 - f_1 X_0)^2 + (X_2 - f_0 - f_1 X_1)^2 + \cdots + (X_n - f_0 - f_1 X_{n-1})^2 + (X_{n+1} - f_0 - f_1 X_n)^2$$

with respect to f_0 and f_1 since, in this case, we have used the observed data both as independent and dependent variables at $t = 1$ and ant $t = n + 1$.

The preceding discussion leads to two methods: least squares and modified least squares.

The Least squares.

Minimizing $\sum_{t=2}^{n}(X_t - f_0 - f_1 X_{t-1})^2$ with respect to f_0 and f_1 leads to the normal equations

$$\begin{aligned}(n-1)f_0 + \left(\sum_{t=2}^{n} X_{t-1}\right) f_1 &= \sum_{t=2}^{n} X_t, \\ \left(\sum_{t=2}^{n} X_{t-1}\right) f_0 + \left(\sum_{t=2}^{n} X_{t-1}^2\right) f_1 &= \sum_{t=2}^{n} X_{t-1}X_t.\end{aligned}$$

The solutions $\hat{\phi}_0$ and $\hat{\phi}_1$ of these equations are the least squares estimates of ϕ_0 and ϕ_1. It can be verified that in general $\hat{\phi}_0 \neq (1-\hat{\phi}_1)\bar{X}$ and $\hat{\mu} \neq \bar{X}$. We know that we must have $|\phi_1| < 1$ for an AR(1) series to be stationary. Unfortunately, the least squares estimate $\hat{\phi}_1$ is not guaranteed to be smaller than 1 in magnitude. The modified least squares estimation method remedies this defect.

In the AR(p) case, the series can be expressed as

$$\begin{aligned}X_t &= \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t \text{ with} \\ \phi_0 &= (1 - \phi_1 - \cdots - \phi_p)\mu.\end{aligned}$$

The method of least squares minimizes $\sum_{t=p+1}^{n}(X_t - f_0 - f_1 X_{t-1} - \cdots - f_p X_{t-p})^2$ with respect to $f_0, f_1, \ldots, f_p$ leading to the normal equations. Solutions of these equations $\hat{\phi}_0, \hat{\phi}_1, \ldots, \hat{\phi}_p$ are estimates of the AR(p) parameters $\phi_0, \phi_1, \ldots, \phi_p$. Once again, in general $\hat{\phi}_0 \neq (1-\hat{\phi}_1)\bar{X}$ and $\hat{\mu} \neq \bar{X}$. Similar to the AR(1) case, the estimated parameters $\hat{\phi}_1, \ldots, \hat{\phi}_p$ may not satisfy the conditions for stationarity.

The Modified least squares (Yule-Walker).

Once again, let us begin with the AR(1) case. If we approximate X_0 and X_{n+1} by the sample mean $\bar{X} = (X_1 + \cdots + X_n)/n$, then we can run the least squares method for the padded data $\{\bar{X}, X_1, \ldots, X_n, \bar{X}\}$, ie, minimize $\sum_{t=1}^{n+1}(\tilde{X}_t - f_0 - f_1 \tilde{X}_{t-1})^2$ with respect to f_0 and f_1, where $\tilde{X}_0 = \bar{X}, \tilde{X}_1 = X_1, \ldots, \tilde{X}_n = X_n, \tilde{X}_{n+1} = \bar{X}$. The normal equations can be simplified to

$$\begin{aligned}f_0 &= (1 - f_1)\bar{X}, \\ \hat{\gamma}(0) f_1 &= \hat{\gamma}(1).\end{aligned}$$

The second equation $\hat{\gamma}(0)f_1 = \hat{\gamma}(1)$ is known as the Yule-Walker equation. Note that the solutions are:

$$\hat{\phi}_1 = \hat{\gamma}(1)/\hat{\gamma}(0) \text{ and } \hat{\phi}_0 = (1 - \hat{\phi}_1)\bar{X},$$

and the estimate of μ is $\bar{X}$.

For the AR(p) case, we can pad the data with with $2p$ $\bar{X}$ values: p values at the beginning and p values at the end, ie, the padded data are now of the form $\{\bar{X}, \ldots, \bar{X}, X_1, \ldots, X_n, \bar{X}, \ldots, \bar{X}\}$. Let

$$\tilde{X}_t = \begin{cases} \bar{X} & t = 1-p, \ldots, 0 \\ X_t & t = 1, \ldots, n \\ \bar{X} & t = n+1, \ldots, n+p. \end{cases}.$$

.If we minimize $\sum_{t=1}^{n+p}(\tilde{X}_t - f_0 - f_1\tilde{X}_{t-1} - \cdots - f_p\tilde{X}_{t-p})^2$ with respect to $f_0, \ldots, f_p$, it can be shown that the last p of the normal equations (there are $p+1$ equations in total) are

$$\begin{aligned}
\hat{\gamma}(0)f_1 + \cdots + \hat{\gamma}(p-1)f_p &= \hat{\gamma}(1) \\
\hat{\gamma}(1)f_1 + \cdots + \hat{\gamma}(p-2)f_p &= \hat{\gamma}(2). \\
&\vdots \quad\quad \vdots \\
\hat{\gamma}(p-1)f_1 + \cdots + \hat{\gamma}(0)f_p &= \hat{\gamma}(p).
\end{aligned}$$

These are called the Yule-Walker equations. The solutions $\hat{\phi}_1, \ldots, \hat{\phi}_p$ are the estimates of $\phi_1, \ldots, \phi_p$. The first equation in the normal equations turns out to be $f_0 = (1 - f_1 - \cdots - f_p)\bar{X}$. Estimates of ϕ_0 and μ are $\hat{\phi}_0 = (1 - \hat{\phi}_1 - \cdots - \hat{\phi}_p)\bar{X}$ and $\bar{X}$, respectively.

When $p = 2$, the Yule-Walker equations are

$$\begin{aligned}
\hat{\gamma}(0)f_1 + \hat{\gamma}(1)f_2 &= \hat{\gamma}(1), \\
\hat{\gamma}(1)f_1 + \hat{\gamma}(0)f_2 &= \hat{\gamma}(2).
\end{aligned}$$

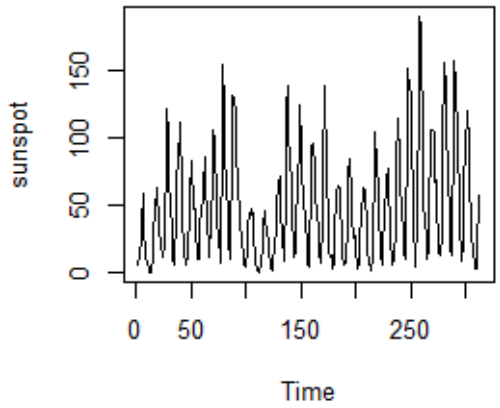
Mathematical results show that the AR(p) model obtained via Yule-Walker estimates is stationary. Using matrix notations, the Yule-Waler equations can be written as $\hat{\mathbf{\Gamma}}_p \boldsymbol{f} = \hat{\boldsymbol{\gamma}}_p$, where $\hat{\mathbf{\Gamma}}_p$ is a $p \times p$ matrix whose element (i, j) is $\hat{\gamma}(i-j)$ (note that $\hat{\gamma}(-h) = \hat{\gamma}(h)$ for any integer h), $\boldsymbol{f}$ is the column vector of $f_1, \ldots, f_p$, and $\hat{\boldsymbol{\gamma}}_p$ is the column vector of $\hat{\gamma}(1), \ldots, \hat{\gamma}(p)$. The estimates are (denoting the $p \times 1$ vector with elements $\hat{\phi}_1, \ldots, \hat{\phi}_p$ by $\hat{\boldsymbol{\phi}}$)

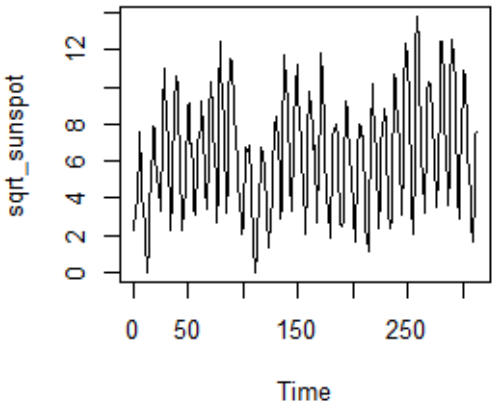
$$\hat{\boldsymbol{\phi}} = \hat{\mathbf{\Gamma}}_p^{-1}\hat{\boldsymbol{\gamma}}_p \text{ and } \hat{\sigma}^2 = \hat{\gamma}(0)(1 - \hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{\gamma}}_p) = \hat{\gamma}(0)(1 - \hat{\boldsymbol{\gamma}}_p^T\hat{\mathbf{\Gamma}}_p\hat{\boldsymbol{\gamma}}_p).$$

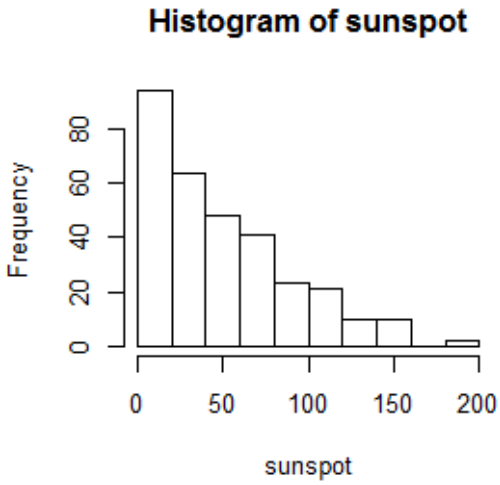
The following results are approximately true

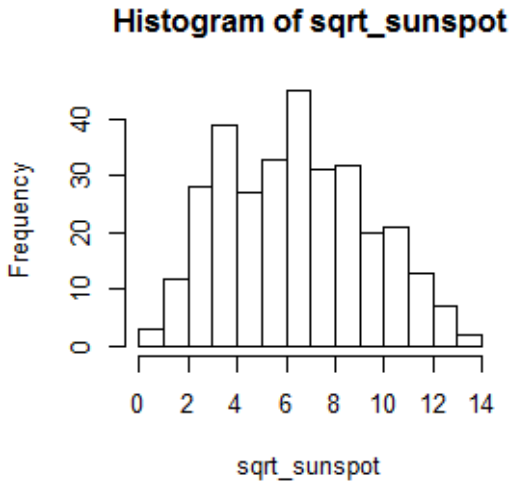
$$E(\hat{\boldsymbol{\phi}}) \approx \boldsymbol{\phi}, Var(\hat{\boldsymbol{\phi}}) \approx (1/n)\sigma^2\mathbf{\Gamma}_p^{-1}, s^2(\hat{\boldsymbol{\phi}}) \approx (1/n)\hat{\sigma}^2\hat{\mathbf{\Gamma}}_p^{-1}.$$

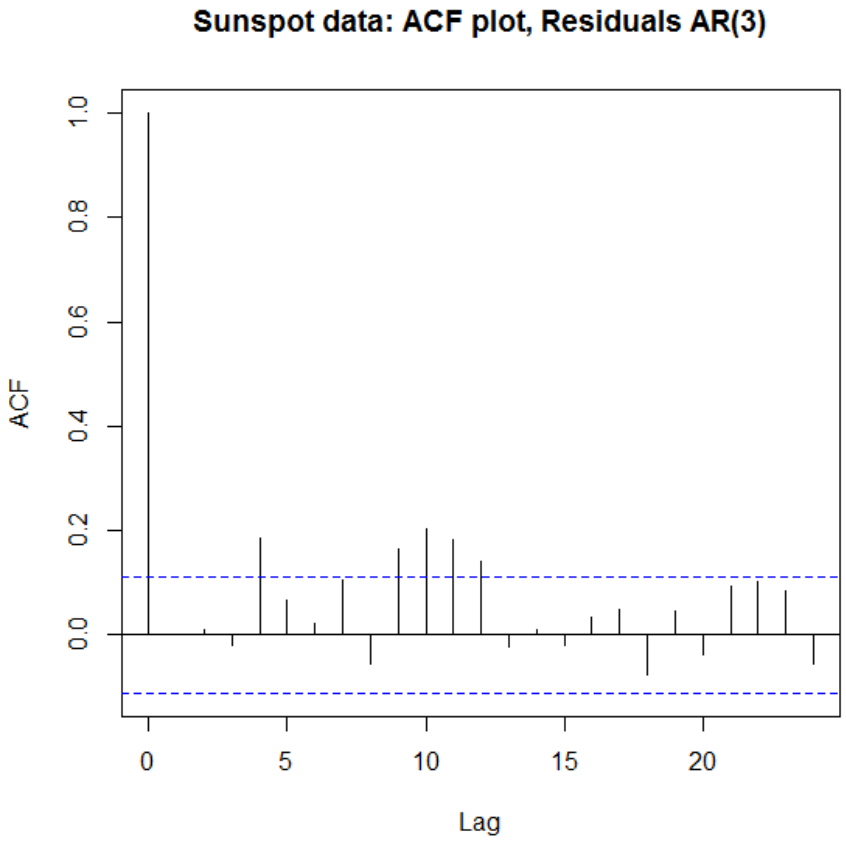
The last result can be used to test (for any $1 \le j \le p$) $H_0 : \phi_j = 0$ against $H_1 : \phi_j \neq 0$. The test statistic is $z^* = \hat{\phi}_j / s(\hat{\phi}_j)$, where $s(\hat{\phi}_j) = \hat{\sigma}\sqrt{\hat{\gamma}^{jj}}/\sqrt{n}$, where $\hat{\gamma}^{jj}$ is the j^{th} diagonal element of $\hat{\mathbf{\Gamma}}_p^{-1}$. If the level of significance $\alpha = 0.05$, then reject the null hypothesis if $|z^*| > 1.96$.

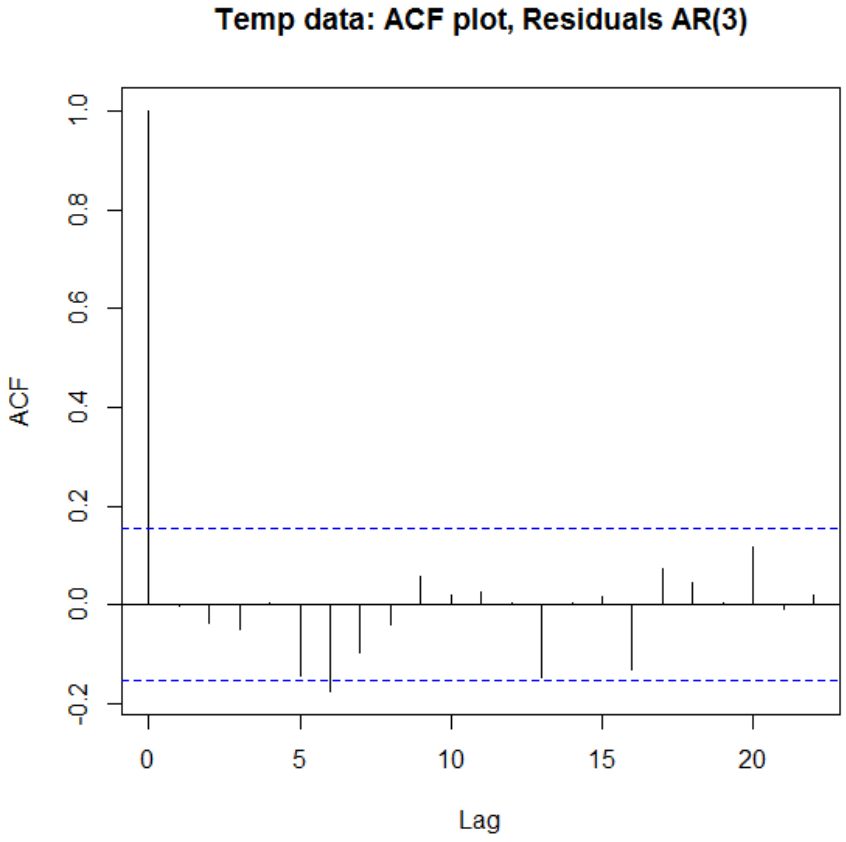

sunspot
0
50
100
150
0
50
150
250
Time


sqrt_sunspot
0
2
4
6
8
12
0
50
150
250
Time


Histogram of sunspot
Frequency
0
20
40
60
80
0
50
100
150
200
sunspot


Histogram of sqrt_sunspot
Frequency
0
10
20
30
40
0
2
4
6
8
10
12
14
sqrt_sunspot


Sunspot data: ACF plot, Residuals AR(3)
ACF
Lag
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20


Temp data: ACF plot, Residuals AR(3)
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
20
Lag