Time Series Lectures for ST280/380 (continued) - Fall 2005
VI. Fitting ARIMA Models (continued)

(iv) Model Estimation: Yule-Walker, LSE, Approximate and Exact MLE

Consider estimation of the parameters of the ARMA\((p, q)\) process

\[
\phi(B)X_t = \theta(B)Z_t, \text{ where } Z_t \sim WN(0, \sigma^2)
\]

based on observations \(x_1, \cdots, x_n\).

Assume the process has zero mean. If the mean of the process was non-zero, say \(\mu\), we estimate it by the sample mean \(\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}\), and obtain the mean-subtracted process, which has zero-mean.

Let \(\phi = (\phi_1, \cdots, \phi_p)'\), \(\theta = (\theta_1, \cdots, \theta_q)'\) and \(\sigma^2\) be the unknown parameters of the process.

Yule-Walker Procedure:

This is useful with purely AR models. Recall the Yule-Walker equations which we discussed earlier for the AR(p) model

\[ X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t. \]

where \( Z_t \sim WN(0, \sigma^2) \). Let \( \gamma(h) \) denote the autocovariance of lag \( h \), let

\[
\Gamma_p = \begin{pmatrix}
\gamma(0) & \gamma(1) & \gamma(2) & \cdots & \gamma(p-1) \\
\gamma(1) & \gamma(0) & \gamma(1) & \cdots & \gamma(p-2) \\
& & & \ddots & \\
\gamma(p-1) & \gamma(p-2) & \gamma(p-3) & \cdots & \gamma(0)
\end{pmatrix}
\]

and \( \gamma_p = (\gamma(1), \cdots, \gamma(p))' \).
The Yule-Walker equations can be written as

\[ \Gamma_p \phi = \gamma_p \]
\[ \sigma^2 = \gamma(0) - \phi' \gamma_p \]

Replace \( \gamma(h) \) by \( \hat{\gamma}(h) \) for \( h = 0, 1, \ldots, p \). We get a set of equations whose solution yields Yule-Walker estimates for \( \phi_1, \ldots, \phi_p \) and \( \sigma^2 \) in terms of the sample autocovariances via

\[ \hat{\phi} = \Gamma_p^{-1} \hat{\gamma}_p \]
\[ \hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\phi}' \hat{\gamma}_p \]

In terms of the sample autocorrelations, since \( \hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0) \), the Yule-Walker estimates are

\[ \hat{\phi} = R_p^{-1} \hat{\rho}_p \]
\[ \hat{\sigma}^2 = \hat{\gamma}(0)[1 - \hat{\phi}' \hat{\gamma}_p] \]
\[ = \hat{\gamma}(0)[1 - \hat{\rho}_p R_p^{-1} \hat{\rho}_p] \]

The fitted Yule-Walker AR(p) model is
\[ X_t = \hat{\phi}_1 X_{t-1} + \cdots + \hat{\phi}_p X_{t-p} \]

where

\[ \hat{\phi} = \hat{R}_p^{-1} \hat{\rho}_p, \]

and the estimated WN variance is

\[ \hat{\nu}_p = \hat{\gamma}(0)[1 - \hat{\rho}'_p \hat{R}_p^{-1} \hat{\rho}_p] \]

The Yule-Walker method yields method of moments estimators for the parameters of the AR(p) process.

The large sample distribution of the Yule-Walker estimators: if \( x_1, \cdots, x_n \) is a large sample from an AR(p) process, then

\[ \hat{\phi} = (\hat{\phi}_1, \cdots, \hat{\phi}_p)' \approx N(\phi, \sigma^2 \Gamma_p^{-1}/n). \]

We can replace \( \Gamma_p \) by \( \hat{\Gamma}_p \) and \( \sigma^2 \) by \( \hat{\sigma}^2 \).
Let $\hat{v}_{jj}$ denote the $j$th diagonal element of $\hat{v}_p \hat{\Gamma}_p^{-1}$. Assuming that the true AR model has been fitted, i.e., $p$ is the true order, the large sample $100(1 - \alpha)$% confidence interval for $\phi_j$ is given for $j = 1, \cdots, p$ by

$$\hat{\phi}_j \pm z_{\alpha/2} n^{-1/2} \hat{v}_{jj}^{1/2}.$$  

The Yule-Walker estimators can also be derived for ARMA(p,q) process for $q > 0$ in the same way. However, the equations that result are nonlinear in the unknown parameters $\phi_1, \cdots, \phi_p$, $\theta_1, \cdots, \theta_q$ and $\sigma^2$:

$$\hat{\gamma}(k) - \phi_1 \hat{\gamma}(k - 1) - \cdots - \phi_p \hat{\gamma}(k - p) = \sigma^2 \sum_{j=k}^{q} \theta_j \psi_{j-k}, \quad 0 \leq k \leq p + q$$  

where $\psi_j$ are the weights of the MA representation of the process.
Example 5.1.2 in B&D shows the method of moments estimators for the parameters of an MA(1) process.

Although these estimates are simple to obtain for AR models and have asymptotic efficiency similar to that of MLE’s (see p. 146 in B&D), the estimates are computationally expensive for ARMA models with $q > 0$ and are not as efficient as other estimators. Hence, the Yule-Walker estimates are usually not used with ARMA models with $q > 0$, but are suitable for purely AR($p$) models.

**Innovations Algorithm.**

Let $X_t$ be a time series with zero mean and finite second moment (the time series is not necessarily stationary). Let $E(X_iX_j) = \kappa(i,j)$ be the ACVF, which here is written as a function of the two time points. Let $X_n(1)$ denote the MMSE one-step ahead prediction for $X_{n+1}$,
which is defined as $E(X_{n+1}|X_n, X_{n-1}, \cdots)$. Let the MMSE prediction for $X_1$ be $X_0(1)$ which is equal to the mean, 0. Let

$$v_n = E(X_{n+1} - X_n(1))^2$$

denote the mean squared error of the prediction. Let

$$U_{n+1} = X_{n+1} - X_n(1)$$

denote the innovation or one-step-ahead prediction error.

See Section 2.5.2 in B&D.

Write

$$X_n(1) = \theta_{n1}(X_n - X_{n-1}(1)) + \theta_{n2}(X_{n-1} - X_{n-2}(1)) \cdots + \theta_{nn}(X_1 - X_0(1))$$
where the coefficients $\theta_{n1}, \cdots, \theta_{nn}$ and $v_n$’s are computed recursively from the equations

\[
\begin{align*}
\theta_{n,n-k} &= v_k^{-1} \left\{ \kappa(n + 1, k + 1) \right. \\
&- \left. \sum_{j=0}^{k-1} \theta_{k,k-j} \theta_{n,n-j} v_j \right\}, \ k = 0, 1, \cdots, n - 1
\end{align*}
\]

\[

\begin{align*}
v_0 &= \kappa(1, 1) \\
v_n &= \kappa(n + 1, n + 1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j
\end{align*}
\]

The innovations algorithm provides a recursive method for obtaining forecasts from zero-mean time series with finite second-order moments. The time series need not be stationary.

Note that the Durbin-Levinson algorithm expresses the one-step-ahead prediction in terms of past observations $X_n, \cdots, X_1$. By contrast, the innovations algorithm expresses the one-step-ahead prediction in terms of past innovations $U_n, \cdots, U_1$, which are uncorrelated.
Example: MA(1) Process. Here, $\kappa(i, j) = 0$ for $|i - j| > 1$; $\kappa(i, i) = \sigma^2(1 + \theta^2)$; and $\kappa(i, i + 1) = \theta \sigma^2$. Use the innovations algorithm:

\[
\begin{align*}
v_0 & = \kappa(1, 1) = \sigma^2(1 + \theta^2) \\
\theta_{n,n-k} & = 0, \quad k = 0, 1, \ldots, n-2 \\
\theta_{n,1} & = v_{n-1}^{-1} \kappa(n + 1, n) = v_{n-1}^{-1} \theta \sigma^2 \\
v_n & = \sigma^2(1 + \theta^2) - \theta_{n,1}^2 v_{n-1} \\
& = \sigma^2(1 + \theta^2) - v_{n-1}^{-2} \theta^2 \sigma^2 v_{n-1} \\
& = [1 + \theta^2 - v_{n-1}^{-1} \theta^2 \sigma^2] \sigma^2
\end{align*}
\]

which follows from the definition of $\kappa(i, j)$. See page 74 of B&D for a numerical example.
The innovations algorithm is obtained for ARMA processes. Consider the causal ARMA(p,q) process

\[ \phi(B)X_t = \theta(B)Z_t, \quad Z_t \sim WN(0,\sigma^2) \]

where \( p \geq 1 \) and \( q \geq 1 \). Let \( m = \max(p,q) \). For recursive forecasting of this process, apply the innovations algorithm to a transformed time series \( \{W_t\} \) defined by

\[
W_t = \begin{cases} 
\sigma^{-1}X_t & t = 1, \ldots, m \\
\sigma^{-1}\phi(B)X_t & t > m 
\end{cases}
\]

where \( \theta_0 = 1 \) and \( \theta_j = 0 \) for \( j > q \). Then, the ACVF of \( \{W_t\} \) are (see (3.3.3) in B&D)
\[ \kappa(i, j) = \sigma^{-2} \gamma_X(i - j) \text{ if } 1 \leq i, j \leq m \]
\[ = \sigma^{-2} \left\{ \gamma_X(i - j) - \sum_{r=1}^{p} \phi_r \gamma_X(r - |i - j|) \right\} \]
\[ \text{if } \min(i, j) \leq m < \max(i, j) \leq 2m \]
\[ = \sum_{r=0}^{q} \theta_r \theta_{r + |i - j|} \text{ if } \min(i, j) > m \]
\[ = 0 \text{ otherwise} \]

We may then obtain the coefficients \( \theta_{nj} \) and \( r_n \) from the innovations algorithm and thus the one-step-ahead predeictors \( W_n(1) \) (see (3.3.4) in B&D), and from those obtain the one-step-ahead preeictors \( X_n(1) \) as given in (3.3.7) in B&D. Note that

\[ r_n = E(W_{n+1} - W_n(1))^2 \]
The innovations algorithm is used in preliminary parameter estimation for an MA(q) process as follows. For \( m = 1, 2, \ldots \), suppose we use the innovations algorithm to fit MA(m) models

\[
X_t = Z_t + \hat{\theta}_1 Z_{t-1} + \cdots + \hat{\theta}_m Z_{t-m}
\]

where \( Z_t \sim WN(0, \hat{\nu}_m) \). That is, the coefficients \( \hat{\theta}_1, \ldots, \hat{\theta}_m \) and \( \hat{\nu}_m \) are obtained from the recursive equations under the innovations algorithm, with \( \kappa(i, j) \) replaced by \( \hat{\kappa}(i, j) \). The first \( q \) values denote the preliminary estimates for an MA(q) process. See Example 5.1.5 in B&D.

The innovations algorithm may also be used to obtain preliminary estimates from ARMA(p,q) models. Recall that the \( \psi_j \) weights satisfy (see (3.1.7) in B&D):
\[ \psi_j = \theta_j + \sum_{i=1}^{\min(j,p)} \phi_i \psi_{j-i}, \quad j = 0, 1, \ldots \]

where \( \theta_0 = 1 \) and \( \theta_j = 0 \) for \( j > q \). Use the innovations algorithm and obtain the coefficients \( \hat{\theta}_m, \cdots, \hat{\theta}_{m,p+q} \) and substitute these for \( \psi_1, \cdots, \psi_{p+q} \) in the equation above to obtain for \( j = 1, \cdots, p + q \),

\[ \hat{\theta}_{mj} = \theta_j + \sum_{i=1}^{\min(j,p)} \phi_i \hat{\theta}_{m,j-i}, \quad j = 0, 1, \ldots \]

Solve these equations for \( \phi_1, \cdots, \phi_p \) and \( \theta_1, \cdots, \theta_q \) using these steps:

First, solve for \( \phi_1, \cdots, \phi_p \) from the last \( q \) equations

\[
\begin{pmatrix}
\hat{\theta}_{m,q+1} \\
\hat{\theta}_{m,q+2} \\
\vdots \\
\hat{\theta}_{m,q+p}
\end{pmatrix} = 
\begin{pmatrix}
\hat{\theta}_{m,q} & \hat{\theta}_{m,q-1} & \cdots & \hat{\theta}_{m,q+1-p} \\
\hat{\theta}_{m,q+1} & \hat{\theta}_{m,q} & \cdots & \hat{\theta}_{m,q+2-p} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{\theta}_{m,q+p-1} & \hat{\theta}_{m,q+p-2} & \cdots & \hat{\theta}_{m,q}
\end{pmatrix} 
\begin{pmatrix}
\phi_1 \\
\phi_2 \\
\vdots \\
\phi_p
\end{pmatrix}
\]
to obtain $\hat{\phi}_1, \cdots, \hat{\phi}_p$.

Next, solve for $\hat{\theta}_1, \cdots, \hat{\theta}_q$ from

$$\hat{\theta}_j = \hat{\theta}_{mj} - \sum_{i=1}^{\min(j,p)} \hat{\phi}_i \hat{\theta}_{m,j-i}$$

Finally, compute the one-step-ahead predictions $X_t(1)$ for $t = 0, 1, \cdots, n$ with $X_0(1) = 0$. The WN variance is then estimated by

$$\hat{\sigma}^2 = n^{-1} \sum_{t=1}^{n} (X_t - X_t(1))^2 / r_{t-1}$$

The ITSM program described in B&D may be used to fit the innovations algorithm and provide preliminary estimates for the ARMA model parameters.
Final Estimation

- Conditional Least Squares
- Unconditional Least Squares
- Maximum Likelihood

Let \( \{X_t\} \) be a Gaussian time series with mean zero, i.e., \( EX_t = 0 \) and AVCF \( \kappa(i, j) \). Let \( X_n = (X_1, \cdots, X_n)' \) and let \( \Gamma_n = \{\kappa_{i,j}\} \) be the variance-covariance matrix of \( X_n \). The likelihood function based on the data and the unknown parameters \( \Lambda \) is

\[
L(\Lambda; X_n) = (2\pi)^{-n/2} |\Gamma_n|^{-1/2} \exp(-\frac{1}{2}X_n'\Gamma_n^{-1}X_n)
\]
Note that if \( \{X_t\} \) is a Gaussian ARMA\((p,q)\) process, then the vector of parameters is \( \Lambda = (\phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q, \sigma^2) \). The MLE of \( \Lambda \) is obtained as the values that maximize \( L(\Lambda; X_n) \). The direct evaluation of \( |\Gamma_n| \) and \( \Gamma_n^{-1} \) is cumbersome for large \( n \).

Innovations Form of the Likelihood. Let \( X_n = (X_1, \cdots, X_n)' \) be the time series data, let \( \hat{X}_n = (\hat{X}_1, \cdots, \hat{X}_n)' \) denote the vector of one-step-ahead predictions, and let \( U_n = (U_1, \cdots, U_n)' \) denote the vector of innovations (with uncorrelated components). Let \( \theta_{ij}, \ j = 1, \cdots, i; \ i = 1, 2, \cdots \) and the \( v_n \) denote the coefficients in the recursive innovations algorithm. Let

\[
C_n = \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 \\
\theta_{11} & 1 & 0 & \cdots & 0 \\
\theta_{22} & \theta_{21} & 1 & \cdots & 0 \\
\vdots & & \ddots & \ddots & \ddots \\
\theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \cdots & 1
\end{pmatrix}
\]
Then,
\[
\hat{X}_n = X_n - U_n = C_n U_n - U_n, \text{ or } X_n = C_n [X_n - \hat{X}_n] = C_n U_n
\]

Let \( D_n = \text{diag}(v_0, v_1, \cdots, v_{n-1}) = \text{Cov}(U_n) \).

Using properties of Gaussian distribution,
\[
\Gamma_n = \text{Cov}(X_n) = C_n \text{Cov}(U_n) C_n' = C_n D_n C_n'
\]

so that
\[
\Gamma_n^{-1} = (C_n^{-1})' D_n^{-1} C_n^{-1}
\]

and
\[
X_n' \Gamma_n^{-1} X_n = (C_n U_n)' \Gamma_n^{-1} C_n U_n = U_n C_n' \Gamma_n^{-1} C_n U_n = U_n C_n' C_n^{-1})' D_n^{-1} C_n^{-1} C_n U_n = U_n D_n^{-1} U_n = (X_n - \hat{X}_n)' D_n^{-1} (X_n - \hat{X}_n) = \sum_{j=1}^{n} \frac{U_j^2}{v_{j-1}} = \sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{v_{j-1}}
\]
Also,

\[
\left| \Gamma_n \right| = \left| C_n D_n C'_n \right| = |C_n|^2 |C_n| = v_0 v_1 \cdots v_{n-1}
\]

The likelihood function can then be written as

\[
L(\Lambda; X_n) = (2\pi)^{-n/2} (v_0 v_1 \cdots v_{n-1})^{-1/2} \exp \left( -\frac{1}{2} \sum_{j=1}^{n} (X_j - \hat{X}_j)^2 / v_{j-1} \right)
\]

This is maximized wrt the model parameters to yield the MLE’s of the parameters.
Another form of the Likelihood for an ARMA Process

The zero-mean ARMA model

\[ X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \]

can be written as

\[ Z_t = X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} - \theta_1 Z_{t-1} - \cdots - \theta_q Z_{t-q} \]

where we assume that \( Z_t \sim WN(0, \sigma^2) \). Let \( \phi = (\phi_1, \cdots, \phi_p)' \) and \( \theta = (\theta_1, \cdots, \theta_q)' \). The joint density of \( Z = (Z_1, \cdots, Z_n)' \) is

\[ p(Z; \phi, \theta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp[-\sum_{t=1}^{n} Z_t^2 / 2\sigma^2] \]

The likelihood function

\[ L(\phi, \theta, \sigma^2; X_n) = g_1(\phi, \theta, \sigma^2) \exp[-S(\phi, \theta)/2\sigma^2] \]
where $g_1(.)$ is a function of the model parameters. The sum of squares function is

$$S(\phi, \theta) = \sum_{t=1-p-q}^{n} [E(u_t|X_n, \phi, \theta, \sigma^2)]^2$$

where

$$u_t = \begin{cases} Z_t & t = 1, 2, \cdots, n \\ g_2(Z_*, X_*) & t \leq 0 \end{cases}$$

and $g_2(.)$ is a linear function of the initial unobservable vectors

$$Z_* = (Z_{1-q}, \cdots, Z_{-1}, Z_0)' \text{ and } X_* = (X_{1-p}, \cdots, X_{-1}, X_0)'.$$

The specific forms of the functions $g_1(.)$ and $g_2(.)$ depend on the particular ARMA(p,q) model, and the exact likelihood function follows.
For example, for the AR(1) model,

\[ L(\phi, \sigma^2; X_n) = (2\pi\sigma^2)^{-n/2}(1 - \phi^2)^{1/2} \times \exp\left\{-\frac{1}{2\sigma^2}\left[(1 - \phi^2)X_1^2 + \sum_{t=2}^{n}(X_t - \phi X_{t-1})^2\right]\right\} \]

The maximization of the likelihood function involves solution of a cubic equation and is most easily achieved via numerical iterative methods.

For example, for the MA(1) model,

\[ L(\theta, \sigma^2; X_n) = (2\pi\sigma^2)^{-n/2}\left[\frac{1 - \theta^2}{1 - \theta^2(n+1)}\right]^{1/2} \times \exp\left\{-\frac{1}{2\sigma^2} \sum_{t=0}^{n} [E(Z_t|X_n)]^2\right\} \]

Given \( \theta \) and \( X_n \), we obtain

\[ E(Z_t|X_n) = X_t - \theta E(Z_{t-1}|X_n) \]
provided we can compute $E(Z_0|X_n)$. The resulting likelihood function is a complicated non-linear function of $\theta$ and must be numerically maximized.

In general, the exact likelihood function is complicated because of the function $g_1(\phi, \theta, \sigma^2)$, and because of the need to compute the conditional expectations $E(Z_t|X_n)$ for $t = 1-q, \cdots, 0$ and $E(X_t|X_n)$ for $t = 1-p, \cdots, 0$. Numerical algorithms (for example, Ansley, 1979 Biometrika) have been used to compute the MLE’s. Further, some approximations to the likelihood function have been discussed in the time series literature. If the model parameters do not lie very close to the boundary of the invertibility region, there is negligible effect if we ignore $g_1(.)$ and minimize $S(\phi, \theta)$, to obtain least squares estimates of the parameters. Even in the computation of $S(\phi, \theta)$, we still need to compute the conditional expectations corresponding to the initial values of the process.
and white noise given the data. We discuss Unconditional Least Squares Estimation and Conditional Least Squares Estimation.

Under Unconditional Least Squares, the method of backcasting is used to compute the required conditional expectations of the initial unknown process and white noise variables. Consider the backward model

\[ et = X_t - \phi_1 X_{t+1} - \cdots - \phi_p X_{t+p} - \theta_1 e_{t+1} - \cdots - \theta_q e_{t+q} \]

where \( e_t \sim WN(0, \sigma^2) \).

The ACF of the backward model is identical to the ACF of the usual (forward) ARMA(p,q) model.

MMSE forecasting with the backward ARMA model enables the computation of \( E(Z_t|X_n) \) for \( t = 1 - q, \cdots, 0 \) and \( E(X_t|X_n) \) for \( t = 1 - p, \cdots, 0 \). This is called backcasting. Then
$S(\phi, \theta)$ can be evaluated and minimized to yield the Unconditional Least Squares Estimates of the parameters.

Example: MA(1) Process: $Z_t = X_t - \theta Z_{t-1}$.
Backward form of the MA(1) model: $e_t = X_t - \theta e_{t+1}$.
Since the process is going backwards,

$$E(e_t|X_n) = 0 \text{ for } t \leq 0$$

i.e., past random shocks of the backward process are independent of $X_1, \cdots, X_n$.
Also, $E(Z_t|X_n) = 0$ for $t \leq -1$ because this is a 1-correlated process.
Note that for $t = 1, \cdots, n$

$$E(X_t|X_n) = X_t$$

To find $E(X_0|X_n)$, use the backward form of the MA(1) model at $t = 0$, i.e.,
\[ e_0 = X_0 - \theta e_1 \]

Take conditional expectation on both sides

\[ E(e_0|X_n) = E(X_0|X_n) - \theta E(e_1|X_n) \]

Since \( E(e_0|X_n) = 0 \), it follows that

\[ E(X_0|X_n) = \theta E(e_1|X_n) \]

We must get \( E(e_1|X_n) \) as a function of \( E(e_2|X_n), \ldots, E(e_n|X_n) \).

Assume that \( E(e_n|X_n) = 0 \). Then

\[ E(e_{n-1}|X_n) = E(X_{n-1}|X_n) - \theta E(e_n|X_n) = X_{n-1}; \]

\[ E(e_{n-2}|X_n) = E(X_{n-2}|X_n) - \theta E(e_{n-1}|X_n) = X_{n-2} - \theta E(e_{n-1}|X_n); \]

etc., until we evaluate
\[ E(e_1 | X_n) = E(X_1 | X_n) - \theta E(e_2 | X_n) = X_1 - \theta E(e_2 | X_n). \]

Then, \( E(X_0 | X_n) = \theta E(e_1 | X_n). \)

Next, use the forward form of the MA(1) model to compute \( E(Z_0 | X_n), \ldots, E(Z_n | X_n), \) and thus compute \( S(\theta) = \sum_{t=0}^{n} [E(Z_t | X_n)]^2. \)

So,

\[
\begin{align*}
E(Z_0 | X_n) &= E(X_0 | X_n) - \theta E(Z_{-1} | X_n) = E(X_0 | X_n) \\
E(Z_1 | X_n) &= E(X_1 | X_n) - \theta E(Z_0 | X_n) = X_1 - \theta E(Z_0 | X_n) \\
E(Z_n | X_n) &= E(X_n | X_n) - \theta E(Z_{n-1} | X_n) = X_n - \theta E(Z_{n-1} | X_n).
\end{align*}
\]

The value of \( \theta \) that minimizes \( S(\theta) \) is the unconditional least squares estimate. This involves nonlinear least squares since the conditional expectations involved in computing the sum of squares is a nonlinear function of \( \phi \) and \( \theta. \)

This approach may be used with any ARMA(p,q) model.
The least squares estimate of $\sigma^2$ is

$$\hat{\sigma}^2 = S(\hat{\phi}, \hat{\theta})/n$$

The estimates $\hat{\phi} = (\hat{\phi}, \hat{\theta})$ are asymptotically unbiased for $\phi$ and $\theta$ and the variance covariance matrix of the vector of estimates are denoted by $\hat{V}(\hat{\beta})$, which is a by-product of the nonlinear least squares procedure. We may use the resulting standard errors to construct C.I.’s for the model parameters.
Conditional Least Squares

The conditional sum of squares is

$$S_c(\phi, \theta) = \sum_{t=p+1}^{n} Z_t^2 = \sum_{t=p+1}^{n} (X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p})^2$$

i.e., all initial values are set to zero (i.e., the initial process values are set to the unconditional process mean $\mu$, which is here assumed to be zero, while the initial WN values are set to its unconditional mean of 0). The values that minimize $S_c(\phi, \theta)$ are the conditional least squares estimates. For the AR(p) process, the estimates are obtained by linear least squares, while for MA and ARMA processes, nonlinear least squares is used.

For data from an AR(p) process, define the matrix
\[
B = \begin{pmatrix}
X_p & X_{p-1} & \cdots & X_1 \\
X_{p+1} & X_p & \cdots & X_2 \\
\vdots & & & \\
X_{n-1} & X_{n-2} & \cdots & X_{n-p}
\end{pmatrix}
\]

and the vector \( b = (X_{p+1}, X_{p+2}, \cdots, X_n)' \). Minimizing \( S_c(\phi) \) under linear least squares for an AR(p) model leads to solution of the normal equations

\[B'B\tilde{\phi} = B'b\]

so that the conditional least squares estimates of \( \phi_1, \cdots, \phi_p \) are given by

\[
\tilde{\phi} = (B'B)^{-1}B'b
\]

The least squares estimate of \( \sigma^2 \) is

\[
\tilde{\sigma}^2 = S_c(\tilde{\phi})/n
\]
The estimate $\tilde{\phi}$ is asymptotically unbiased for $\phi$ and the variance covariance matrix of the vector of estimates are

$$\hat{V}(\tilde{\phi}) = \tilde{\sigma}^2 (B'B)^{-1}$$

We may use the resulting standard errors to construct C.I's for the model parameters.
- (v) Model Adequacy and Model Selection

Let us denote the residuals from the $ARMA(p, q)$ process as

$$\hat{Z}_t = \hat{\theta}^{-1}(B)\hat{\phi}(B)(X_t - \hat{\mu}).$$

The following are useful to assess adequacy of the fitted model.

(i) The plot of residuals versus time shows no dependence pattern, and has the form of a WN series.

(ii) The normal probability plot of the residuals indicates a straight line.

(iii) The turning-point test and difference-sign test do not reject the hypothesis of randomness of the residuals.
(iv) The sample ACF plot and the sample PACF plot show WN behavior. That is, none of the sample autocorrelations or sample partial autocorrelations exceed $1.96/\sqrt{n}$ in absolute value.

(v) The portmanteau test statistics (Box-Pierce, Ljung-Box and McLeod-Li) do not reject the null hypothesis that first $K$ autocorrelations of the residuals are zero.

**Model Selection**
This involves the selection of the AR order $p$ and MA order $q$ that yields the “best” model. Several criteria exist, and may not always lead to the same model. The values of $p$ and $q$ that yield the minimum value of each criterion are the preferred choices.
Final Prediction Error (FPE)
The idea behind the FPE criterion is to choose the $AR(p)$ model for the time series $\{X_t\}$ so as to minimize the one-step-ahead mean squared error of prediction when the model fitted to $\{X_t\}$ is used to predict another independent realization $\{Y_t\}$ from the same process. Let $\hat{\sigma}^2$ denote the estimate of $\sigma^2$ based on an $AR(p)$ model fit; then FPE has the form

$$FPE_p = \hat{\sigma}^2 \frac{n + p}{n - p}$$

Choose the value of $p$ which minimizes $FPE_p$. 
Akaike Information Criterion (AIC)

Given data $X_1, \cdots, X_n$ from a Gaussian $ARMA(p, q)$ time series, suppose the MLE’s of parameters are $\hat{\phi}, \hat{\theta},$ and $\hat{\sigma}^2$. Let $Y_1, \cdots, Y_n$ denote a realization from the same $ARMA(p, q)$ process, which is independent of $X_1, \cdots, X_n$. Let $m = p + q$. The AIC criterion is defined as

$$AIC_m = -2 \ln L(\hat{\phi}, \hat{\theta}, \hat{\sigma}^2) + 2(p + q + 1)$$

In general $m$ denotes the number of freely estimated parameters in the model.
**Akaike Information Criterion, Corrected (AICC)**

The bias-corrected AIC is called AICC and has the form

\[
AIC_{Cm} = -2 \ln L(\hat{\phi}, \hat{\theta}, \hat{\sigma}^2) + \frac{2(p + q + 1)n}{(n - p - q - 2)}
\]

The Bayesian Information Criterion (BIC) is defined in (5.5.5) in B & D.
Schwarz Bayesian Information Criterion, (SBC)

\[ SBC_m = -2 \ln L(\hat{\phi}, \hat{\theta}, \hat{\sigma}^2) + \ln(n)(p + q + 1) \]