Estimating ARMA Parameters

Ruoqi Yu

Week 12

Announcements

- Grades of Homework 4 was posted.
  - The average is 81/100.
  - Regrading request (to Laura) deadline: Sunday, Nov 14.
- Homework 5 was due today (Nov 10), 3 pm.
- Final exam: Dec 10, 7-9 PM, Evans 10.
  - There is no make-up final exam.
  - You have to take the final exam in person. Otherwise, you will get a failing grade.

Section 1

Recap

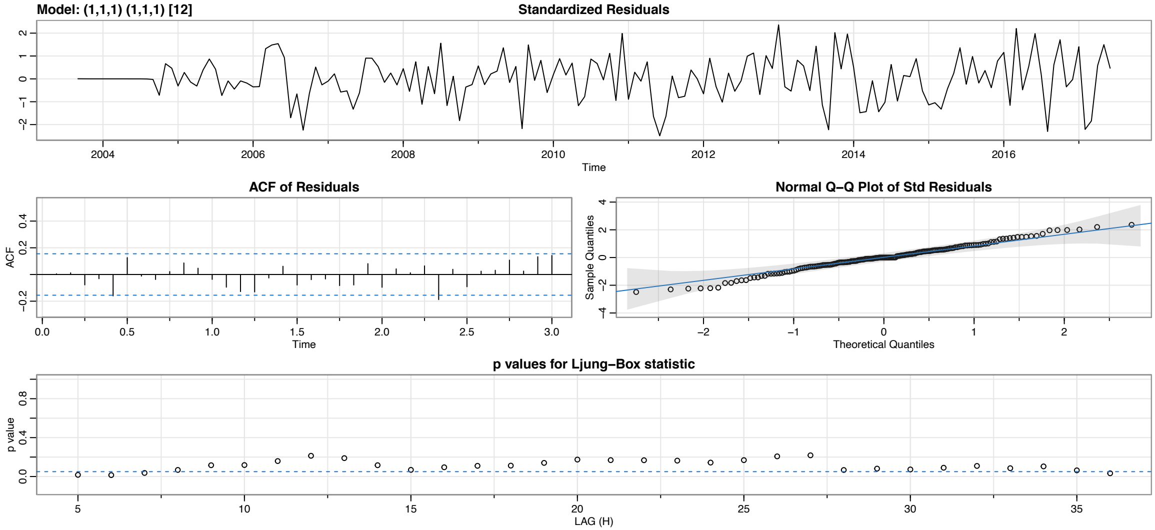
Definition: Ljung-Box-Pierce test

- Fix a maximum lag k (typically $k = 20$).
- Reject the hypothesis that data $x_1, \ldots, x_n$ was generated from a causal and invertible ARMA(p,q) model if

$$\tilde{Q}(x_1, \ldots, x_n) = n(n+2) \sum_{i=1}^{k} \frac{\hat{r}_i^2}{n-i} > q_{1-\alpha},$$

- where $q_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the χ^2 distribution with $k-p-q$ degrees of freedom.
- It is common to use a Ljung-Box test to check that the residuals from a time series model resemble white noise.

Ljung-Box in sarima() diagnostics



Practical Consideration

If the test gives $p > 0.05$ for all k, then obviously the time series (residuals) pass the test. How to interpret the test if $p < 0.05$ for some values of k and not for other values?

Which k should we use in Ljung-Box test?

- The value k is chosen somewhat arbitrarily, typically, $k = 20$. (Shumway and Stoffer, 2016)
- $h = 10$ for non-seasonal data and $h = 2S$ for seasonal data with period S. (Hyndman and Athanasopoulos, 2018)
- $h = \min(10, n/5)$ for non-seasonal data and $h = \min(2S, n/5)$ for seasonal data with period S. (https://robjhyndman.com/hyndsight/ljung-box-test/)

Practical Consideration

If the test gives $p > 0.05$ for all k, then obviously the time series (residuals) pass the test. How to interpret the test if $p < 0.05$ for some values of k and not for other values?

- Alternatively, consider a joint significance test.
- For instance, incorporate multiple testing techniques.

Overfit

- The Ljung-Box-Pierce test provides a strategy to evaluate, for given p, q, whether or not an ARMA(p,q) model is appropriate for data $x_1, \ldots, x_n$.
- But how should we choose the parameters p and q in the first place?
- Clearly every ARMA(p,q) model can be arbitrarily-well approximated by an ARMA(p', q') model with $p' > p$ and $q' > q$.
- However, a overfitted model is likely to be useless to predict future values.
- Model selection: we want the number of model parameters to be large enough, so that it can fit the data well. At the same time the number of model parameters should not be too large, which would result in overfitting: fitting the data and not the true underlying process.
- A solution: measure in-sample fit and penalize for model size/complexity.

Information Criterion

- $AIC = -2\log(\text{likelihood}) + 2k$
- $AICc = AIC + \frac{2k(k+1)}{n-k-1}$
- $BIC = -2\log(\text{likelihood}) + k\log n$

Cross Validation

- General idea: we want to know how the model will perform out-of-sample, so let's reserve some of our data to check this out!
- Basic Idea:
    - Divide dataset into "training" and "testing" subsets
    - Fit all candidate models with the training data
    - Evaluate the performance of each model on the testing data
    - Repeat as needed
    - Select the model that performed the best

Notes about this example

- Can use metric other than SSE/MSE
- Probably don't want to start with so little data. For this example, start with at least half of the data in the training set.

Section 2

Estimating Parameters of AR(p)

Estimating AR(p)

Assume our given data $x_1, \dots, x_n$ was generated by a causal AR(p) model with mean μ, that is,

$$(X_t - \mu) - \phi_1(X_{t-1} - \mu) - \cdots - \phi_p(X_{t-p} - \mu) = W_t.$$

with a white noise process $\{W_t\}$ with variance σ_W^2.

We are interested in finding estimates $\hat{\mu}, \hat{\phi}_1, \dots, \hat{\phi}_p, \hat{\sigma}_W^2$ the parameters $\mu, \phi_1, \dots, \phi_p, \sigma_W^2$.

How do you think we could estimate these?

Different Methods

We will look at three different methods:

1. Method of moments (Yule-Walker),
2. Least squares (LS), and
3. Maximum Likelihood (MLE).

Section 3

Yule-Walker Method (Method of Moments)

Method of Moments

The method of moments is using the sample moments to estimate the true/population moments.

The basic idea behind this form of the method is to:

1. Equate the first sample moment about the origin $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ to the first theoretical moment $E(X)$.
2. Equate the second sample moment about the origin $M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ to the second theoretical moment $E(X^2)$.
3. Continue equating sample moments about the origin, M_k, with the corresponding theoretical moments $E(X^k)$, $k = 3, 4, \ldots$ until you have as many equations as you have parameters.
4. Solve for the parameters.

Method of Moments

For example, for sterotypical $X \sim N(\mu, \sigma^2)$:

1. $\hat{\mu} \stackrel{set}{=} \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. $\widehat{\sigma^2} \stackrel{set}{=} s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Yule-Walker Method

For all t we have that $E(X_t) = \mu$. Therefore, the method of moments simply estimates μ by the sample mean:

$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t.$$

For estimating the other parameters $\phi_1, \ldots, \phi_p$ and and σ_W^2, recall the Yule-Walker equations from the ARMA-ACVF Lecture

$$\gamma_X(0) - \phi_1\gamma_X(1) - \cdots - \phi_p\gamma_X(p) = \sigma_W^2, \tag{1}$$
$$\gamma_X(k) - \phi_1\gamma_X(k-1) - \cdots - \phi_p\gamma_X(k-p) = 0 \text{ for } k \geq 1. \tag{2}$$

Yule-Walker Method

- Previously, we considered solving these equations to write $\gamma_X(k)$ in terms of σ_W^2 and $\phi_1, \dots, \phi_p$.
- But these same equations can be used to estimate σ_W^2 and $\phi_1, \dots, \phi_p$ from the data $x_1, \dots, x_n$:
- Definition: The Yule-Walker estimates $\hat{\phi}_1, \dots, \hat{\phi}_p, \hat{\sigma}_W^2$ for the parameters $\phi_1, \dots, \phi_p, \sigma_W^2$ in an AR(p) model are obtained by
  1. estimate the autocovariances $\gamma_X(h)$ by the sample autocovariances $\hat{\gamma}_X(h)$.
  2. solve the above equations for the unknown parameters σ_W^2 and $\phi_1, \dots, \phi_p$.

Yule-Walker Method

- Note that in the definition we have an infinite set of equations in Equation 2 but we only need to estimate $p+1$ parameters.
- So we will only use Equation 1 and the first p of the equations from Equation 2.
- This gives us $p+1$ equations to solve for the $p+1$ unknowns $\phi_1, \ldots, \phi_p$ and σ_W^2.
- Essentially, one is trying to find an AR(p) model whose autocovariance function equals the observed sample autocovariance function at lags $0, 1, \ldots, p$. This is why this method is called the method of moments.

Example

- AR(1)
- AR(2)

Example: AR(1)

For $p = 1$ i.e., the AR(1) case, we just have the two equations:

$$\hat{\gamma}_X(0) - \phi\hat{\gamma}_X(1) = \sigma_W^2 \quad \text{and} \quad \hat{\gamma}_X(1) = \phi\hat{\gamma}_X(0).$$

This of course gives

$$\hat{\phi} = \frac{\hat{\gamma}_X(1)}{\hat{\gamma}_X(0)} = r_1 \quad \text{and} \quad \widehat{\sigma_W^2} := \hat{\gamma}_X(0)\left(1 - r_1^2\right).$$

Example: AR(2)

When $p = 2$ i.e., AR(2), we get the three equations:

$$\begin{aligned}
\hat{\gamma}_X(0) - \phi_1\hat{\gamma}_X(1) - \phi_2\hat{\gamma}_X(2) &= \sigma_W^2 \\
\hat{\gamma}_X(1) - \phi_1\hat{\gamma}_X(0) - \phi_2\hat{\gamma}_X(1) &= 0 \\
\hat{\gamma}_X(2) - \phi_1\hat{\gamma}_X(1) - \phi_2\hat{\gamma}_X(0) &= 0
\end{aligned}$$

The last two equations can used to solve for ϕ_1 and ϕ_2 to yield:

$$\hat{\phi}_1 = \frac{r_1(1 - r_2)}{1 - r_1^2} \quad \text{and} \quad \hat{\phi}_2 = \frac{r_2 - r_1^2}{1 - r_1^2}.$$

Plugging these values for ϕ_1 and ϕ_2 into the top equation gives an estimate for σ_W^2.

Section 4

Least Squares (ols)

Definition: Least Squares

- The (conditional) least squares estimates for the parameters $\mu, \phi_1, \dots, \phi_p$ in an AR(p) model are obtained by minimizing

$$S_c(\phi, \mu) = \sum_{i=p+1}^{n} \left(x_i - \mu - \phi_1(x_{i-1} - \mu) - \dots - \phi_p(x_{i-p} - \mu)\right)^2 .$$

The variance σ_W^2 is then estimated as

$$\hat{\sigma}_W^2 = \frac{1}{n-p} S_c(\hat{\phi}, \hat{\mu}).$$

- "Conditional" as we condition on the first p values of x_i. Unconditional version shown later.

Example: AR(1)

- To minimize the LS equation, let $\beta_0 = \mu(1-\phi)$ and $\beta_1 = \phi$ and rewrite it as

$$\sum_{i=2}^{n} (x_i - \beta_0 - \beta_1 x_{i-1})^2 .$$

- Minimizing this now is exactly linear regression and the answers are given by

$$\hat{\beta}_1 = \frac{\sum_{i=2}^{n} (x_i - \bar{x}_{(2)})(x_{i-1} - \bar{x}_{(1)})}{\sum_{i=2}^{n} (x_{i-1} - \bar{x}_{(1)})^2}$$

where

$$\bar{x}_{(1)} := \frac{x_1 + \cdots + x_{n-1}}{n-1} \quad \text{and} \quad \bar{x}_{(2)} := \frac{x_2 + \cdots + x_n}{n-1}$$

and $\hat{\beta}_0 := \bar{x}_{(2)} - \hat{\beta}_1 \bar{x}_{(1)}$.

Example: AR(1)

This will give

$$\hat{\phi} = \frac{\sum_{i=2}^{n}(x_i - \bar{x}_{(2)})(x_{i-1} - \bar{x}_{(1)})}{\sum_{i=2}^{n}(x_{i-1} - \bar{x}_{(1)})^2} \quad \text{and} \quad \hat{\mu} := \frac{\bar{x}_{(2)} - \hat{\phi}\bar{x}_{(1)}}{1 - \hat{\phi}}.$$

The parameter σ_W^2 is estimated by

$$\hat{\sigma_W^2} := \frac{\sum_{i=2}^{n}\left(x_i - \hat{\mu} - \hat{\phi}(x_{i-1} - \hat{\mu})\right)^2}{n-1}.$$

It is easily seen that these estimates are very close to those obtained by the Yule-Walker method.

Section 5

Maximum Likelihood

Maximum Likelihood

- To write a likelihood, we need a distribution assumption on $\{W_t\}$. Most common assumption is that $\{W_t\}$ are i.i.d normal with mean 0 and variance σ_W^2.
- Then $(x_1, \ldots, x_n)$ are distributed according to the multivariate normal distribution with mean $(\mu, \ldots, \mu)$ and covariance matrix $\Gamma_n := \gamma_X(i-j)$, which has the likelihood function

$$f_{\mu,\Gamma_n}(x_1, \ldots, x_n) = (2\pi)^{-n/2}|\Gamma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T\Gamma^{-1}(x-\mu)\right).$$

Definition

Under Gaussian noise assumption, the maximum likelihood estimator for the parameters $\mu, \phi_1, \dots, \phi_p$ in an AR(p) model are obtained by

- Writing down covariance matrix $\Gamma_n := \gamma_X(i-j)$ as a function of $\phi_1, \dots, \phi_p, \sigma_W^2$,

$$\Gamma_n = \Gamma_n(\phi_1, \dots, \phi_p, \sigma_W^2)$$

- Estimate $\mu, \phi_1, \dots, \phi_p$ by maximizing $f_{\mu, \Gamma_n(\phi_1, \dots, \phi_p, \sigma_W^2)}(x_1, \dots, x_n)$

Example: AR(1)

- In the AR(1) case, it is easy to simplify this likelihood. Decompose the joint density as:

$$f_{\mu,\phi,\sigma^2}(x_1, \dots, x_n) := f(x_1)f(x_2|x_1)f(x_3|x_1, x_2) \dots f(x_n|x_1, \dots, x_{n-1}).$$

- Because of the Gaussian assumption on $\{W_t\}$, it is easy to see that for $i \geq 2$, the conditional distribution of x_i given $x_1, x_2, \dots, x_{i-1}$ is normal with mean $\mu + \phi(x_{i-1} - \mu)$ and variance σ_W^2.

Example: AR(1)

- Moreover x_1 is distributed as a normal with mean μ and variance $\gamma(0) = \sigma_W^2/(1-\phi^2)$. We thus get the following likelihood:

$$f_{\mu,\phi,\sigma_W^2}(x_1, \ldots, x_n) := (2\pi\sigma_W^2)^{-n/2}(1-\phi^2)^{1/2} \exp\left(-\frac{S(\mu,\phi)}{2\sigma_W^2}\right),$$

where

$$S(\mu,\phi) := (1-\phi^2)(x_1-\mu)^2 + \sum_{i=2}^{n} (x_i - \mu - \phi(x_{i-1}-\mu))^2 .$$

- This above sum of squares is called unconditional least squares.

Example: AR(1)

- Maximizing the likelihood or its logarithm results in a non-linear optimization problem. R solves it when you choose the method *mle* in the ar() function.
- A compromise between maximum likelihood and the least squares technique (previous section) is to minimize the unconditional least squares $S(\mu, \phi)$. This also results in a non-linear optimization problem.

Summary

We have studied three different methods to estimate the parameters in an AR(p) model. Assuming that the order p is known, all three methods can be carried out in R by invoking the function *ar*().

1. **Yule Walker or Method of Moments**: Finds the AR(p) model whose acvf equals the sample autocorrelation function at lags $0, 1, \dots, p$. Use *yw* for method in R.
2. **Least Squares**: Minimizes the sum of squares: $\sum_{i=p+1}^{n}(x_i - \mu - \phi_1(x_{i-1} - \mu) - \cdots - \phi_p(x_{i-p} - \mu))^2$ over μ and $\phi_1, \dots, \phi_p$. Use *ols* for method in R. Note the default is $x_t - \bar{x} = intercept + \phi(x_{t-1} - \bar{x}) + \epsilon$.
3. **Maximum Likelihood**: Here one maximizes the likelihood function. Use *mle* for method in R.

It is usually the case that all these three methods yield similar answers. The default method in R is Yule-Walker.

Section 6

Asymptotic Distribution of Estimates

Asymptotic Distribution of Estimates

- Recall that an estimator $\hat{\phi}$ of a parameter ϕ is a function of the data $X_1, \ldots, X_n$, that is $\hat{\phi} = \hat{\phi}(X_1, \ldots, X_n)$.
- Thus, the estimator $\hat{\phi}$ is a random variable which depends on the sample size n. The following theorem gives the approximate distribution of the estimators discussed above when n is large.

Thoerem

- Assume a causal AR(p) process $\{X_t\}$ with acvf $\gamma_X(h)$ and define the $p \times p$ matrix Γ with entries $\Gamma_{ij} = \gamma_X(i - j)$.
- Let $\hat{\phi}$ be from any of the three estimators we've discussed (Yule-Walker, least squares, or MLE).
- Then, under some general conditions on the white noise process $\{W_t\}$, with $var(W_t) = \sigma_W^2$, for n large enough, $\hat{\phi}$ is approximately multivariate normal distributed with mean $\phi = (\phi_1, \ldots, \phi_p)^\top$ and covariance matrix $n^{-1}\sigma_W^2\Gamma^{-1}$, that is

$$\sqrt{n}(\hat{\phi} - \phi) \to N(0, \sigma_W^2\Gamma^{-1}) \quad \text{as } n \to \infty.$$

- Proof is Theorem B.4 in Appendix B of TSA4e

Example

- AR(1)
- AR(2)

Example: AR(1)

In the AR(1) case:

$$\Gamma_p = \Gamma_1 = \gamma_X(0) = \sigma_W^2/(1-\phi^2).$$

Thus $\hat{\phi}$ is approximately normal with mean ϕ and variance $(1-\phi^2)/n$.

Example: AR(2)

For AR(2), using

$$\gamma_X(0) = \frac{1-\phi_2}{1+\phi_2}\frac{\sigma_W^2}{(1-\phi_2)^2-\phi_1^2} \qquad \text{and} \qquad \rho_X(1) = \frac{\phi_1}{1-\phi_2},$$

we can show that $(\hat{\phi}_1, \hat{\phi}_2)$ is approximately normal with mean (ϕ_1, ϕ_2) and covariance matrix is $1/n$ times

$$\begin{pmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1-\phi_2^2 \end{pmatrix}$$

Note that the approximate variances of both $\hat{\phi}_1$ and $\hat{\phi}_2$ are the same. Observe that if we fit AR(2) model to a dataset that comes from AR(1), then the estimate of $\hat{\phi}_1$ might not change much but the standard error will be higher if $|\phi_2| < |\phi|$. We lose precision.

Section 7

Parameter Estimation in ARMA

Method of Moments or Yule-Walker Method

The process, in principle, of solving some subset of equations for the unknown parameters $\theta_1, \ldots, \theta_q, \phi_1, \ldots, \phi_p$ and σ_W^2 (and μ is estimated by the sample mean), by plugging in the sample acvf $\hat{\gamma}(k)$ as an estimate for the true acvf $\gamma(k)$, such as

$$\hat{\gamma}(k) - \phi_1\hat{\gamma}(k-1) - \cdots - \phi_p\hat{\gamma}(k-p) = (\psi_0\theta_k + \psi_1\theta_{k+1} + \cdots + \psi_{q-k}\theta_q)\sigma_W^2$$

for $0 \leq k \leq q$ and

$$\hat{\gamma}(k) - \phi_1\hat{\gamma}(k-1) - \cdots - \phi_p\hat{\gamma}(k-p) = 0 \text{ for } k > q$$

can in principle be applied for ARMA(p,q) models, as well. Note that ψ_j above are functions of $\theta_1, \ldots, \theta_q$ and $\phi_1, \ldots, \phi_p$.

Example: MA(1)

- For an invertible MA(1) model $X_t = W_t + \theta W_{t-1}$,

$$\gamma_X(0) = \sigma_W^2(1+\theta^2) \quad \text{and } \gamma_X(1) = \sigma_W^2\theta.$$

- Thus, with the method of moments one would estimate θ by solving

$$r_1 = \hat{\rho}(1) = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} = \frac{\hat{\theta}}{(1+\hat{\theta}^2)}$$

- Two solutions exist if $|r(1)| \leq 1/2$, so we would pick the invertible one).
- The problem with this estimator is, that the above equation only has a solution when $|r_1| \leq 1/2$. Although $|\rho(1)| \leq 1/2$, because $\hat{\rho}(1)$ is just an estimate, it does not always hold true that $|r_1| \leq 1/2$.

Problem

In general, the method of moments for ARMA(p,q) models, has two major problems:

1. It is cumbersome (unless we are in the pure AR case): Solutions might not always exist to these equations (like in this last example). The parameters are estimated in an arbitrary fashion when these equations do not have a solution.
2. The estimators obtained are *inefficient*. Other techniques give much better estimates (smaller standard errors).

- Because of these problems, no one uses method of moments for estimating the parameters of a general ARMA model. R does not even have a function for doing this. Note, however, that both of these problems disappear for the case of the pure AR model.

Conditional least squares

We'll start by looking at two examples.

Example: MA(1)'s CLS

- We want to fit an MA(1) satisfying $X_t - \mu = W_t + \theta W_{t-1}$ to data $x_1, \ldots, x_n$.
- If the data were indeed generated from this model, then

$$
\begin{aligned}
W_1 &= x_1 - \mu - \theta W_0 \\
W_2 &= x_2 - \mu - \theta W_1 \\
&\vdots \\
W_n &= x_n - \mu - \theta W_{n-1}
\end{aligned}
$$

- If we set $W_0 = E(W_0) = 0$, then we can
  - recursively calculate $W_1, \ldots, W_n$ as a function of μ and θ
  - compute the sum of squares $\sum_{i=1}^{n} W_i^2$
  - choose μ and θ such that they minimize this sum of squares
- This is called conditional least squares because this minimization is conditioning on $W_0 = 0$.

Example: ARMA(1,1)'s CLS

- Here the model is $X_t - \mu - \phi(X_{t-1} - \mu) = W_t + \theta W_{t-1}$.
- Here it is convenient to set W_1 to be zero. Then we can write

$$
\begin{aligned}
W_2 &= x_2 - \mu - \phi(x_1 - \mu) \\
W_3 &= x_3 - \mu - \phi(x_2 - \mu) - \theta W_2 \\
&\vdots \\
W_n &= x_n - \mu - \phi(x_{n-1} - \mu) - \theta W_{n-1}
\end{aligned}
$$

- Then the sum of squares $\sum_{i=2}^{n} W_i^2$ is a function of θ, ϕ, and μ, which can be minimized.

Definition: Conditional least squares for ARMA(p,q)

Given some data $x_1, \ldots, x_n$ and $p, q \in N$, define a function $S_c(\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ as follows:

1. Set $W_t = 0$ for all $t \le p$.
2. For $t = p+1, \ldots, n$, recursively calculate:

$$W_t = X_t - \mu - \phi_1(X_{t-1} - \mu) - \cdots - \phi_p(X_{t-p} - \mu) - \theta_1 W_{t-1} - \cdots - \theta_q W_{t-q}$$

3. Let $S_c(\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q) = \sum_{t=p+1}^{n} W_t^2$.

Then the conditional last squares estimator $\hat{\mu}, \hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q$ is defined by minimizing the conditional sum of squares

$$S_c(\hat{\mu}, \hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q) = \min_{\mu,\phi,\theta} S_c(\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$$

Comments on Definition

- This is equivalent to writing the likelihood conditioning on $X_1, \ldots, X_p$ and $W_t = 0$ for $t \leq p$.
- If $q = 0$ (AR models), minimizing the sum of squares is equivalent to linear regression and no iterative technique is needed.
- If $q > 0$, the problem becomes nonlinear regression and numerical optimization routines need to be used.
- In R, this method is performed by calling the function arima() with the method argument set to *CSS* (CSS stands for conditional sum of squares).
- As before, we can estimate the noise variance via

$$\hat{\sigma}_W^2 = \frac{S_c(\hat{\mu}, \hat{\phi}, \hat{\theta})}{n - p}.$$

Maximum Likelihood

- Assume that errors $\{W_t\}$ are Gaussian.
- Write down the likelihood of the observed data $x_1, x_2, \ldots, x_n$ in terms of the unknown parameter values $\mu, \theta_1, \ldots, \theta_q$, $\phi_1, \ldots, \phi_p$ and σ_W^2.
- Maximize over these unknown parameter values.
- R: use the function arima() with the method argument set to *ML*
- ML stands or Maximum Likelihood. R uses an optimization routine to maximize the likelihood. This routine is iterative and needs suitable initial values of the parameters to start.
- You can also set method equal to *CSS-ML*, where R selects the starting values by CSS.

Asymptotic Distribution of Estimators

- See Property 3.10 in TSA4e
- Yields the SE's on coefficients from arima()