Ch. 17 Maximum Likelihood Estimation

(July 1, 2022)

1 Introduction

The identification process having led to a tentative formulation for the model, we then need to obtain efficient estimates of the parameters. After the parameters have been estimated, the fitted model will be subjected to diagnostic checks. This chapter contains a general account of likelihood method for estimation of the parameters in the stochastic model.

Consider an $ARMA$ (from model identification) model of the form

$$\begin{aligned} Y_t &= c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} \\ &\quad + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}, \end{aligned}$$

with ε_t white noise:

$$\begin{aligned} E(\varepsilon_t) &= 0 \\ E(\varepsilon_t \varepsilon_\tau) &= \begin{cases} \sigma^2 & for\ t = \tau \\ 0 & otherwise \end{cases} . \end{aligned}$$

This chapter explores how to estimate the value of $(c, \phi_1, ..., \phi_p, \theta_1, ..., \theta_q, \sigma^2)$ on the basis of observations on Y. The primary principle on which estimation will be based is *maximum likelihood estimation.*

Let $\boldsymbol{\theta} = (c, \phi_1, ..., \phi_p, \theta_1, ..., \theta_q, \sigma^2)'$ denote the vector of population parameters. Suppose we have observed a sample of size T , i.e. $\{y_1, y_2, ..., y_T\}$. The approach will be to calculate the joint probability density

$$f_{Y_T, Y_{T-1}, ..., Y_1}(y_T, y_{T-1}, ..., y_1; \boldsymbol{\theta}), \tag{17-1}$$

which might loosely be viewed as the probability of having observed this particular sample. The maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ is the value for which this sample is most likely to have been observed; that is, it is the value of $\boldsymbol{\theta}$ that maximizes (17-1). This approach requires specifying a particular distribution for the white noise process ε_t. Typically we will assume that ε_t is Gaussian white noise: $\varepsilon_t \sim i.i.d.\ N(0, \sigma^2)$.

2 MLE of a Gaussian $AR(1)$ Process

The most important step to study the MLE is to evaluate the sample joint distribution which are also called the likelihood function. In the case of identical and independent sample, the likelihood function is just the product of marginal density of individual sample. However, in the study of time series analysis, the dependence structure of observation is specified and it is not correct to use the product of marginal density to evaluate the likelihood function.[1] To evaluate the sample likelihood, the use of conditional density is needed as seen in the following.

2.1 Evaluating the Likelihood Function Using (Scalar) Conditional Density

A stationary Gaussian $AR(1)$ process takes the form

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t, \qquad (17\text{-}2)$$

with $\varepsilon_t \sim i.i.d.\ N(0, \sigma^2)$ and $|\phi| < 1$ (How do you know at this stage ?). For this case, $\boldsymbol{\theta} = (c, \phi, \sigma^2)'$.

Consider the $p.d.f$ of Y_1, the first observations in the sample. This is a random variable with mean and variance

$$\begin{aligned} E(Y_1) &= \mu = \frac{c}{1-\phi}, \quad and \\ Var(Y_1) &= \frac{\sigma^2}{1-\phi^2}. \end{aligned}$$

Since $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ is Gaussian, Y_1 is also Gaussian. That is, $Y_1 \sim N(c/(1-\phi), \sigma^2/(1-\phi^2))$. Hence,

$$\begin{aligned} f_{Y_1}(y_1; \boldsymbol{\theta}) &= f_{Y_1}(y_1; c, \phi, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2/(1-\phi^2)}} \exp\left[-\frac{1}{2} \cdot \frac{\{y_1 - [c/(1-\phi)]\}^2}{\sigma^2/(1-\phi^2)} \right]. \end{aligned}$$

Next consider the distribution of the second observation Y_2 conditional on the observing $Y_1 = y_1$. From (17-2),

$$Y_2 = c + \phi Y_1 + \varepsilon_2. \qquad (17\text{-}3)$$

[1] It is to be noticed that while ε_t is independently and identically distributed, Y_t is not independent, however.

Conditional on $Y_1 = y_1$ means treating the random variable Y_1 as if it were the deterministic constant y_1. For this case, (17-3) gives Y_2 as the constant $(c + \phi y_1)$ plus the $N(0, \sigma^2)$ variable ε_2. Hence,

$$(Y_2|Y_1 = y_1) \sim N((c + \phi y_1), \sigma^2),$$

meaning that

$$f_{Y_2|Y_1}(y_2|y_1; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \cdot \frac{(y_2 - c - \phi y_1)^2}{\sigma^2}\right].$$

The joint density of observations 1 and 2 is then just

$$f_{Y_2,Y_1}(y_2, y_1; \boldsymbol{\theta}) = f_{Y_2|Y_1}(y_2|y_1; \boldsymbol{\theta}) f_{Y_1}(y_1; \boldsymbol{\theta}).$$

Similarly, the distribution of the third observation conditional on the **first two** is

$$f_{Y_3|Y_2,Y_1}(y_3|y_2, y_1; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \cdot \frac{(y_3 - c - \phi y_2)^2}{\sigma^2}\right]$$

form which

$$\begin{aligned} f_{Y_3,Y_2,Y_1}(y_3, y_2, y_1; \boldsymbol{\theta}) &= f_{Y_3|Y_2,Y_1}(y_3|y_2, y_1; \boldsymbol{\theta}) f_{Y_2,Y_1}(y_2, y_1; \boldsymbol{\theta}) \\ &= f_{Y_3|Y_2,Y_1}(y_3|y_2, y_1; \boldsymbol{\theta}) f_{Y_2|Y_1}(y_2|y_1; \boldsymbol{\theta}) f_{Y_1}(y_1; \boldsymbol{\theta}). \end{aligned}$$

In general, the value of $Y_1, Y_2, ..., Y_{t-1}$ matter for Y_t only through the value Y_{t-1}, and the density of observation t conditional on the preceding $t-1$ observations is given by

$$\begin{aligned} & f_{Y_t|Y_{t-1},Y_{t-2},...,Y_1}(y_t|y_{t-1}, y_{t-2}, ..., y_1; \boldsymbol{\theta}) \\ &= f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \boldsymbol{\theta}) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \cdot \frac{(y_t - c - \phi y_{t-1})^2}{\sigma^2}\right]. \end{aligned}$$

The likelihood of the complete sample can thus be calculated as

$$f_{Y_T,Y_{T-1},Y_{T-2},...,Y_1}(y_T, y_{T-1}, y_{T-2}, ..., y_1; \boldsymbol{\theta}) = f_{Y_1}(y_1; \boldsymbol{\theta}) \cdot \prod_{t=2}^{T} f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \boldsymbol{\theta}). \tag{17-4}$$

The log likelihood function (denoted $\mathcal{L}(\boldsymbol{\theta})$) is therefore

$$\mathcal{L}(\boldsymbol{\theta}) = \log f_{Y_1}(y_1; \boldsymbol{\theta}) + \sum_{t=2}^{T} \log f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \boldsymbol{\theta}). \tag{17-5}$$

The log likelihood for a sample of size T from a Gaussian $AR(1)$ process is seen to be

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log[\sigma^2/(1-\phi^2)] - \frac{\{y_1 - [c/(1-\phi)]\}^2}{2\sigma^2/(1-\phi^2)} \\ & -[(T-1)/2]\log(2\pi) - [(T-1)/2]\log(\sigma^2) - \sum_{t=2}^{T}\left[\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}\right].\end{aligned} \tag{17-6}$$

2.2 Evaluating the Likelihood Function Using (Vector) Joint Density

A different description of the likelihood function for a sample of size T from a Gaussian $AR(1)$ process is some time useful. Collect the full set of observations in a $(T \times 1)$ vector, $\mathbf{y} \equiv (Y_1, Y_2, ..., Y_T)'$. The mean of this $(T \times 1)$ vector is

$$E(\mathbf{y}) = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ . \\ . \\ . \\ E(Y_T) \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ . \\ . \\ . \\ \mu \end{bmatrix} = \boldsymbol{\mu},$$

where $\mu = c/(1-\phi)$. The variance-covariance of $\mathbf{y}$ is

$$\boldsymbol{\Omega} = E[(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})'] = \sigma^2 \frac{1}{(1-\phi^2)} \begin{bmatrix} 1 & \phi & . & . & . & \phi^{T-1} \\ \phi & 1 & \phi & . & . & \phi^{T-2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \phi^{T-1} & . & . & . & . & 1 \end{bmatrix} = \sigma^2 \mathbf{V}$$

where

$$\mathbf{V} = \frac{1}{(1-\phi^2)} \begin{bmatrix} 1 & \phi & . & . & . & \phi^{T-1} \\ \phi & 1 & \phi & . & . & \phi^{T-2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \phi^{T-1} & . & . & . & . & 1 \end{bmatrix}.$$

The sample likelihood function is therefore the multivariate Gaussian density:

$$f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) = (2\pi)^{-T/2}|\boldsymbol{\Omega}^{-1}|^{1/2}\exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right],$$

with log likelihood

$$\mathcal{L}(\boldsymbol{\theta}) = (-T/2)\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Omega}^{-1}| - \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\mu}). \qquad (17\text{-}7)$$

It should be noted that (17-6) and (17-7) must represent the identical likelihood function. It is easy to verify by direct multiplication that $\mathbf{L}'\mathbf{L} = \mathbf{V}^{-1}$, with

$$\mathbf{L} = \begin{bmatrix} \sqrt{1-\phi^2} & 0 & . & . & . & 0 \\ -\phi & 1 & 0 & . & . & 0 \\ 0 & -\phi & 1 & 0 & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & . & . & -\phi & 1 \end{bmatrix}.$$

Then (17-7) becomes

$$\mathcal{L}(\boldsymbol{\theta}) = (-T/2)\log(2\pi) + \frac{1}{2}\log|\sigma^{-2}\mathbf{L}'\mathbf{L}| - \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\sigma^{-2}\mathbf{L}'\mathbf{L}(\mathbf{y}-\boldsymbol{\mu}). \qquad (17\text{-}8)$$

Define the $(T \times 1)$ vector $\tilde{\mathbf{y}}$ to be

$$\begin{aligned}\tilde{\mathbf{y}} &\equiv \mathbf{L}(\mathbf{y}-\boldsymbol{\mu}) \\ &= \begin{bmatrix} \sqrt{1-\phi^2} & 0 & . & . & . & 0 \\ -\phi & 1 & 0 & . & . & 0 \\ 0 & -\phi & 1 & 0 & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & . & . & -\phi & 1 \end{bmatrix} \begin{bmatrix} Y_1-\mu \\ Y_2-\mu \\ Y_3-\mu \\ . \\ . \\ Y_T-\mu \end{bmatrix} \\ &= \begin{bmatrix} \sqrt{1-\phi^2}(Y_1-\mu) \\ (Y_2-\mu)-\phi(Y_1-\mu) \\ (Y_3-\mu)-\phi(Y_2-\mu) \\ . \\ . \\ (Y_T-\mu)-\phi(Y_{T-1}-\mu) \end{bmatrix} \\ &= \begin{bmatrix} \sqrt{1-\phi^2}[Y_1-c/(1-\phi)] \\ Y_2-c-\phi Y_1 \\ Y_3-c-\phi Y_2 \\ . \\ . \\ Y_T-c-\phi Y_{T-1} \end{bmatrix}.\end{aligned}$$

The last term in (17-8) can thus be written

$$\begin{aligned}\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\sigma^{-2}\mathbf{L}'\mathbf{L}(\mathbf{y}-\boldsymbol{\mu}) &= \left[\frac{1}{2\sigma^2}\right]\tilde{\mathbf{y}}'\tilde{\mathbf{y}} \\ &= \left[\frac{1}{2\sigma^2}\right](1-\phi^2)[Y_1-c/(1-\phi)]^2 \\ &\quad+\left[\frac{1}{2\sigma^2}\right]\sum_{t=2}^{T}(Y_t-c-\phi Y_{t-1})^2.\end{aligned}$$

The middle term in (17-8) is similarly

$$\begin{aligned}\frac{1}{2}\log|\sigma^{-2}\mathbf{L}'\mathbf{L}| &= \frac{1}{2}\log\{\sigma^{-2T}\cdot|\mathbf{L}'\mathbf{L}|\} \\ &= -\frac{1}{2}\log\sigma^{2T}+\frac{1}{2}\log|\mathbf{L}'\mathbf{L}| \\ &= -\frac{T}{2}\log\sigma^2+\frac{1}{2}\log\{|\mathbf{L}'||\mathbf{L}|\}\quad(\textit{since } \mathbf{L} \textit{ is triangular}) \\ &= -\frac{T}{2}\log\sigma^2+\log|\mathbf{L}| \\ &= -\frac{T}{2}\log\sigma^2+\frac{1}{2}\log(1-\phi^2).\end{aligned}$$

Thus equation (17-6) and (17-7) are just two different expressions for the same magnitude. Either expression accurately describes the log likelihood function.

2.3 Exact Maximum Likelihood Estimators for the Gaussian $AR(1)$ Process

The *MLE* $\hat{\boldsymbol{\theta}}$ is the value for which (17-6) is maximized. In principle, this requires differentiating (17-6) with respect to c, ϕ and σ^2 and setting the derivatives equal to zero, we obtain

$$c=[2+(T-2)(1-\phi)]^{-1}\left[Y_1+(1-\phi)\sum_{t=2}^{T-1}Y_t+Y_T\right],$$

$$[(Y_1-c)^2-(1-\phi^2)^{-1}\sigma^2]\phi+\sum_{t=2}^{T}[(Y_t-c)-\phi(Y_{t-1}-c)](Y_{t-1}-c)=0,$$

$$\sigma^2=T^{-1}\left\{(Y_1-c)^2(1-\phi^2)+\sum_{t=1}^{T}[(Y_t-c)-\phi(Y_{t-1}-c)]^2\right\}.$$

In practice, when an attempt is made to carry this out, the result is a system of nonlinear equation in $\boldsymbol{\theta}$ and $(Y_1,Y_2,...,Y_T)$ for which there is no simple solution for $\boldsymbol{\theta}$

in terms of $(Y_1, Y_2, ..., Y_T)$. Maximization of (17-6) thus requires iterative or numerical procedure described in p.21 of Chapter 3.

2.4 Conditional Maximum Likelihood Estimation

An alternative to numerical maximization of the exact likelihood function is to regard the value of y_1 as deterministic (that is, $f_{Y_1}(y_1) = 1$) and maximize the likelihood conditioned on the first observation

$$f_{Y_T,Y_{T-1},Y_{T-2},..,Y_2|Y_1}(y_T, y_{T-1}, y_{T-2}, ..., y_2|y_1; \boldsymbol{\theta}) = \prod_{t=2}^{T} f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \boldsymbol{\theta}),$$

the objective then being to maximize

$$\begin{aligned}\mathcal{L}^*(\boldsymbol{\theta}) &= -[(T-1)/2]\log(2\pi) - [(T-1)/2]\log(\sigma^2) - \sum_{t=2}^{T}\left[\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}\right] \\ &= -[(T-1)/2]\log(2\pi) - [(T-1)/2]\log(\sigma^2) - \sum_{t=2}^{T}\left[\frac{\varepsilon_t^2}{2\sigma^2}\right]. \qquad (17\text{-}9)\end{aligned}$$

Maximization of (17-9) with respect to c and ϕ is equivalent to minimization of

$$\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \qquad (17\text{-}10)$$

which is achieved by an ordinary least square (OLS) regression of y_t on a constant and its own lagged value, where

$$\mathbf{y} = \begin{bmatrix} y_2 \\ y_3 \\ . \\ . \\ . \\ y_T \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & y_1 \\ 1 & y_2 \\ . & . \\ . & . \\ . & . \\ 1 & y_{T-1} \end{bmatrix}, \quad \textit{and} \ \boldsymbol{\beta} = \begin{bmatrix} c \\ \phi \end{bmatrix}.$$

The conditional maximum likelihood estimates of c and ϕ are therefore given by

$$\begin{bmatrix} \hat{c} \\ \hat{\phi} \end{bmatrix} = \begin{bmatrix} T-1 & \sum_{t=2}^{T} y_{t-1} \\ \sum_{t=2}^{T} y_{t-1} & \sum_{t=2}^{T} y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=2}^{T} y_{t-1} \\ \sum_{t=2}^{T} y_{t-1} y_t \end{bmatrix}.$$

The conditional maximum likelihood estimator of σ^2 is found by setting

$$\frac{\partial \mathcal{L}^*}{\partial \sigma^2} = \frac{-(T-1)}{2\sigma^2} + \sum_{t=2}^{T}\left[\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^4}\right] = 0 \qquad (17\text{-}11)$$

or

$$\hat{\sigma}^2 = \sum_{t=2}^{T} \left[\frac{(y_t - \hat{c} - \hat{\phi} y_{t-1})^2}{T-1} \right] = \frac{\sum_{t=2}^{T} \hat{\varepsilon}_t^2}{T-1}.$$

It is important to note if you have a sample of size T to estimate an $AR(1)$ process by conditional MLE, you will only use $T-1$ observation of this sample.

2.5 An Example of R code for Conditional Maximum Likelihood $AR(1)$ Estimation

The followings are the R code for estimation of $AR(1)$ process provided by Huo, Wen Wei at NSYSU.

```
# Simulat the AR(1) without mean process dataset
data <- arima.sim(list(order = c(1,0,0), ar = 0.3), n = 1500)
# Set y_lag as independent variable ; y as dependent variable
y_lag <- data[1:length(data)-1] ; y <- data[2:length(data)]
# Set the initial value for parameters
params <- matrix(c(NA,NA),2,1) ; params[1,1] <- 0.2 ; params[2,1] <- 0.2
# Set a condition to stop the program if it equal maximum number of iterations
iteration <- 1 ; max_iter <- 100
# Using loop to approximate conditional maximum likelihood of AR(1) without mean process
repeat{
  phi_1 <- params[1,1] ; sigma_sq  <- abs(params[2,1])
  res <- y-phi_1*y_lag # residuals form
  A <- sum((res)*y_lag)/sigma_sq
  B <- -0.5*(length(data)-1)/sigma_sq+0.5*sum((res)^2)/sigma_sq^2
  C <- -sum(y_lag^2)/sigma_sq
  D <- E <- -sum((res)*y_lag)/sigma_sq^2
  F <- 0.5*(length(data)-1)/sigma_sq^2-sum((res)^2)/sigma_sq^3
# The Newton-Raphson Method
# Some idea about newton method's code learn from statistics textbook ISBN:9789571188454
  params <- (-solve(matrix(c(C,D,E,F),2,2))%*%matrix(c(A,B),2,1))+params
# Set a condition to restrict the iterations times
  iteration <- iteration + 1
  if(iteration == max_iter)
    { break }
  # Compute the log-liklihood
  p <- length(params) - 1
  L <- -(length(data)-1)/2*log(2*pi)-(length(data)-1)/2*log(params[2,1])
         -1/(2*params[2,1])*sum((res)^2)
  # Save the output from the program
  coefficients <- (list("AR(1)" <- round(params[1],digits = 3)        ,
                        "sigma_sq" <- round(params[2],digits = 3)     ,
                        "total_iter" <- iteration                     ,
                        "log-liklihood" <- L))
}
coefficients
```

3 MLE of a Gaussian $AR(p)$ Process

This section discusses the estimation of a Gaussian $AR(p)$ process,

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \varepsilon_t,$$

where all the roots of $1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p = 0$ lie outside the unit circle and $\varepsilon_t \sim i.i.d.\ N(0, \sigma^2)$. In this case, the vector of population parameters to be estimated is $\boldsymbol{\theta} = (c, \phi_1, \phi_2, ..., \phi_p, \sigma^2)'$.

3.1 Evaluating the Likelihood Function

We first collect the first p observation in the sample $(Y_1, Y_2, ..., Y_p)$ in a $(p \times 1)$ vector $\mathbf{y}_p$, which has mean vector $\boldsymbol{\mu}_p$ with each element

$$\mu = \frac{c}{1 - \phi_1 - \phi_2 - ... - \phi_p}$$

and variance-covariance matrix is given by

$$\sigma^2 \mathbf{V}_p = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdot & \cdot & \cdot & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdot & \cdot & \cdot & \gamma_{p-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdot & \cdot & \cdot & \gamma_{p-3} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \gamma_{p-1} & \gamma_{p-2} & \gamma_{p-3} & \cdot & \cdot & \cdot & \gamma_0 \end{bmatrix}.$$

The density of the first p observations is then

$$\begin{aligned} & f_{Y_p, Y_{p-1}, ..., Y_1}(y_p, y_{p-1}, ..., y_1; \boldsymbol{\theta}) \\ &= (2\pi)^{-p/2} |\sigma^{-2} \mathbf{V}_p^{-1}|^{1/2} \exp\left[-\frac{1}{2\sigma^2} (\mathbf{y}_p - \boldsymbol{\mu}_p)' \mathbf{V}_p^{-1} (\mathbf{y}_p - \boldsymbol{\mu}_p) \right] \\ &= (2\pi)^{-p/2} (\sigma^{-2})^{p/2} |\mathbf{V}_p^{-1}|^{1/2} \exp\left[-\frac{1}{2\sigma^2} (\mathbf{y}_p - \boldsymbol{\mu}_p)' \mathbf{V}_p^{-1} (\mathbf{y} - \boldsymbol{\mu}_p) \right]. \end{aligned}$$

For the remaining observations in the sample $(Y_{p+1}, Y_{p+2}, ..., Y_T)$, conditional on the first $t-1$ observations, the tth observations is Gaussian with mean

$$c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p},$$

and variance σ^2. Only the p most recent observations matter for this distribution. Hence for $t > p$

$$\begin{aligned}
& f_{Y_t|Y_{t-1},...,Y_1}(y_t|y_{t-1},...,y_1;\boldsymbol{\theta}) \\
&= f_{Y_t|Y_{t-1},..,Y_{t-p}}(y_t|y_{t-1},..,y_{t-p};\boldsymbol{\theta}) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - ... - \phi_p y_{t-p})^2}{2\sigma^2}\right].
\end{aligned}$$

The likelihood function for the complete sample is then

$$\begin{aligned}
f_{Y_T,Y_{T-1},...,Y_1}(y_T, y_{T-1},...,y_1;\boldsymbol{\theta}) &= f_{Y_p,Y_{p-1},...,Y_1}(y_p,y_{p-1},...,y_1;\boldsymbol{\theta}) \\
&\quad \times \prod_{t=p+1}^{T} f_{Y_t|Y_{t-1},..,Y_{t-p}}(y_t|y_{t-1},..,y_{t-p};\boldsymbol{\theta}),
\end{aligned}$$

and the loglikelihood is therefore

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \log f_{Y_T,Y_{T-1},...,Y_1}(y_T, y_{T-1},...,y_1;\boldsymbol{\theta}) \\
&= -\frac{p}{2}\log(2\pi) - \frac{p}{2}\log(\sigma^2) + \frac{1}{2}\log|\mathbf{V}_p^{-1}| - \frac{1}{2\sigma^2}(\mathbf{y}_p - \boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y} - \boldsymbol{\mu}_p) \\
&\quad -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) \\
&\quad -\sum_{t=p+1}^{T}\frac{(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - ... - \phi_p y_{t-p})^2}{2\sigma^2}.
\end{aligned}$$

Maximization of this exact log likelihood of an $AR(p)$ process must be accomplished numerically.

3.2 Conditional Maximum Likelihood Estimates

The log of the likelihood conditional on the first p observation assume the simple form

$$\begin{aligned}
\mathcal{L}^*(\boldsymbol{\theta}) &= \log f_{Y_T,Y_{T-1},..,Y_{p+1}|Y_p,...,Y_1}(y_T, y_{T-1},..y_{p+1}|y_p,...,y_1;\boldsymbol{\theta}) \\
&= -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) \\
&\quad -\sum_{t=p+1}^{T}\frac{(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - ... - \phi_p y_{t-p})^2}{2\sigma^2} \\
&= -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) - \sum_{t=p+1}^{T}\frac{\varepsilon_t^2}{2\sigma^2}. \qquad (17\text{-}12)
\end{aligned}$$

The value of $c, \phi_1, ..., \phi_p$ that maximizes (17-11) are the same as those that minimize

$$\sum_{t=p+1}^{T} (y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - ... - \phi_p y_{t-p})^2.$$

Thus, the conditional MLE of these parameters can be obtained from an OLS regression of y_t on a constant and p of its own lagged values. The conditional MLE estimator of σ^2 turns out to be the average squared residual from this regression:

$$\hat{\sigma}^2 = \frac{1}{T-p} \sum_{t=p+1}^{T} (y_t - \hat{c} - \hat{\phi}_1 y_{t-1} - \hat{\phi}_2 y_{t-2} - ... - \hat{\phi}_p y_{t-p})^2.$$

It is important to note if you have a sample of size T to estimate an $AR(p)$ process by conditional MLE, you will only use $T-p$ observation of this sample.

4 MLE of a Gaussian $MA(1)$ Process

This section discusses the estimation of a Gaussian $MA(1)$ process,

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1} \tag{17-13}$$

where $|\theta| < 1$ and $\varepsilon_t \sim i.i.d.\ N(0, \sigma^2)$. In this case, the vector of population parameters to be estimated is $\boldsymbol{\theta} = (\mu, \theta, \sigma^2)'$.

4.1 Evaluating the Likelihood Function Using (Vector) Joint Density

We collect the observations in the sample $(Y_1, Y_2, ..., Y_T)$ in a $(T \times 1)$ vector $\mathbf{y}$ which has mean vector $\boldsymbol{\mu}$ with each element μ and variance-covariance matrix given by

$$\boldsymbol{\Omega} = E(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})' = \sigma^2 \begin{bmatrix} (1+\theta^2) & \theta & 0 & . \ . \ . & 0 \\ \theta & (1+\theta^2) & \theta & . \ . \ . & 0 \\ 0 & \theta & (1+\theta^2) & . \ . \ . & 0 \\ . & . & . & . \ . \ . & . \\ . & . & . & . \ . \ . & . \\ . & . & . & . \ . \ . & . \\ 0 & 0 & 0 & . \ . \ . & (1+\theta^2) \end{bmatrix}.$$

The likelihood function is then

$$f_{Y_T,Y_{T-1},...,Y_1}(y_T, y_{T-1}, ..., y_1; \boldsymbol{\theta}) \quad = \quad (2\pi)^{-T/2}|\boldsymbol{\Omega}|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right].$$

Using triangular factorization of the variances covariance matrix, the likelihood function can be written

$$f_{Y_T,Y_{T-1},...,Y_1}(y_T, y_{T-1}, ..., y_1; \boldsymbol{\theta}) = (2\pi)^{-T/2}\left[\prod_{t=1}^{T} d_{tt}\right]^{-1/2}\exp\left[-\frac{1}{2}\sum_{t=1}^{T}\frac{\tilde{y}_t^2}{d_{tt}}\right]$$

and the loglikelihood is therefore

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\theta}) = \log f_{Y_T,Y_{T-1},...,Y_1}(y_T, y_{T-1}, ..., y_1; \boldsymbol{\theta}) \\ &= -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log(d_{tt}) - \frac{1}{2}\sum_{t=1}^{T}\frac{\tilde{y}_t^2}{d_{tt}}, \end{aligned}$$

where

$$d_{tt} = \sigma^2\frac{1+\theta^2+\theta^4+...+\theta^{2t}}{1+\theta^2+\theta^4+...+\theta^{2(t-1)}},$$

and

$$\tilde{y}_t = y_t - \mu - \frac{\theta[1+\theta^2+\theta^4+...+\theta^{2t}]}{1+\theta^2+\theta^4+...+\theta^{2(t-1)}}\tilde{y}_{t-1}.$$

Maximization of this exact log likelihood of an $MA(1)$ process must be accomplished numerically.

4.2 Evaluating the Likelihood Function Using (Scalar) Conditional Density

Consider the *p.d.f* of Y_1, $Y_1 = \mu + \varepsilon_1 + \theta\varepsilon_0$, the first observations in the sample. This is a random variable with mean and variance

$$\begin{aligned} E(Y_1) &= \mu \\ Var(Y_1) &= \sigma^2(1+\theta^2). \end{aligned}$$

Since $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ is Gaussian, Y_1 is also Gaussian. Hence,

$$Y_1 \sim N(\mu, (1+\theta^2)\sigma^2)$$

or

$$\begin{aligned} f_{Y_1}(y_1;\boldsymbol{\theta}) &= f_{Y_1}(y_1;\mu,\theta,\sigma^2) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2(1+\theta^2)}}\exp\left[-\frac{1}{2}\cdot\frac{(y_1-\mu)^2}{\sigma^2(1+\theta^2)}\right]. \end{aligned}$$

Next consider the distribution of the second observation Y_2 conditional on the "observing" $Y_1 = y_1$. From (17-12),

$$Y_2 = \mu + \varepsilon_2 + \theta\varepsilon_1. \tag{17-14}$$

Following the method in calculating the joint density of the complete sample of AR process. Conditional on $Y_1 = y_1$ means treating the random variable Y_1 as if it were the deterministic constant y_1. For this case, (17-13) gives Y_2 as the constant $(\mu + \theta\varepsilon_1)$ plus the $N(0,\sigma^2)$ variable ε_2. **However, it is not the case** since observing $Y_1 = y_1$ give no information on the realization of ε_1 because you **can not distinguish** ε_1 **from** ε_0 **even after the first observation on** y_1.

4.2.1 Conditional Maximum Likelihood Estimation

To make the conditional density $f_{Y_2|Y_1}(y_2|y_1;\boldsymbol{\theta})$ feasible,[2] we must impose an additional assumption such as that we know with certainty that $\varepsilon_0 = 0$.

Suppose that we know for certain that $\varepsilon_0 = 0$. Then

$$(Y_1|\varepsilon_0 = 0) \sim N(\mu, \sigma^2)$$

or

$$\begin{aligned} f_{Y_1|\boldsymbol{\varepsilon}_0=0}(y_1|\boldsymbol{\varepsilon}_0 = 0;\boldsymbol{\theta}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\cdot\frac{(y_1-\mu)^2}{\sigma^2}\right] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\varepsilon_1^2}{2\sigma^2}\right]. \end{aligned}$$

Moreover, given observation of y_1, the value of ε_1 is then known with certainty as well:

$$\varepsilon_1 = y_1 - \mu.$$

Hence

$$(Y_2|Y_1 = y_1, \varepsilon_0 = 0) \sim N((\mu + \theta\varepsilon_1), \sigma^2),$$

meaning that

$$\begin{aligned} f_{Y_2|Y_1,\varepsilon_0=0}(y_2|y_1, \varepsilon_0 = 0;\boldsymbol{\theta}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\cdot\frac{(y_2-\mu-\theta\varepsilon_1)^2}{\sigma^2}\right] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\varepsilon_2^2}{2\sigma^2}\right]. \end{aligned}$$

Since ε_1 is know with certainty, ε_2 can be calculated from

$$\varepsilon_2 = y_2 - \mu - \theta\varepsilon_1.$$

Proceeding in this fashion, it is clear that given knowledge that $\varepsilon_0 = 0$, the full sequence $\{\varepsilon_1, \varepsilon_2, ..., \varepsilon_T\}$ can be calculated from $\{y_1, y_2, ..., y_T\}$ by iterating on

$$\varepsilon_t = y_t - \mu - \theta\varepsilon_{t-1}$$

for $t = 1, 2, ..., T$, starting from $\varepsilon_0 = 0$. The condition density of the tth observation can then be calculated as

$$\begin{aligned} f_{Y_t|Y_{t-1},Y_{t-2},...,Y_1,\varepsilon_0=0}(y_t|y_{t-1}, y_{t-2}, ..., y_1, \varepsilon_0 = 0;\boldsymbol{\theta}) &= f_{Y_t|\varepsilon_{t-1}}(y_t|\varepsilon_{t-1};\boldsymbol{\theta}) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\varepsilon_t^2}{2\sigma^2}\right]. \end{aligned}$$

[2] It means to make the information of observation on $Y_1 = y_1$ useful.

The likelihood (conditional on $\varepsilon_0 = 0$) of the complete sample can thus be calculated as the product of these individual densities:

$$\begin{aligned}
&f_{Y_T,Y_{T-1},Y_{T-2},...,Y_1|\varepsilon_0=0}(y_T, y_{T-1}, y_{T-2}, ..., y_1|\varepsilon_0 = 0; \boldsymbol{\theta})\\
= \;& f_{Y_1|\varepsilon_0=0}(y_1|\varepsilon_0 = 0; \boldsymbol{\theta}) \cdot \prod_{t=2}^{T} f_{Y_t|Y_{t-1},Y_{t-2},...,Y_1,\varepsilon_0=0}(y_t|y_{t-1}, y_{t-2}, ..., y_1, \varepsilon_0 = 0; \boldsymbol{\theta}).
\end{aligned}$$

The conditional log likelihood function (denoted $\mathcal{L}^*(\boldsymbol{\theta})$) is therefore

$$\begin{aligned}
\mathcal{L}^*(\boldsymbol{\theta}) &= \log f_{Y_T,Y_{T-1},Y_{T-2},...,Y_1|\varepsilon_0=0}(y_T, y_{T-1}, y_{T-2}, ..., y_1|\varepsilon_0 = 0; \boldsymbol{\theta})\\
&= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T}\frac{\varepsilon_t^2}{2\sigma^2}. \qquad (17\text{-}15)
\end{aligned}$$

In practice, the data implied in the log likelihood function can be calculated from the iteration:

$$(Y_t - \mu) = (1 + \theta L)\varepsilon_t$$

and then we obtain (the reason why invertibility is needed) for $t = 1, 2, \dots, T$,

$$\begin{aligned}
\varepsilon_t &= (1 + \theta L)^{-1}(Y_t - \mu)\\
&= (Y_t - \mu) - \theta(Y_{t-1} - \mu) + \theta^2(Y_{t-2} - \mu) - ... + (-1)^{t-1}\theta^{t-1}(Y_1 - \mu) + (-1)^t\theta^t\varepsilon_0,
\end{aligned}$$

and setting $\varepsilon_i = 0$ for $i \leq 0$, i.e.

$$\begin{aligned}
\varepsilon_0 &= 0;\\
\varepsilon_1 &= (Y_1 - \mu);\\
\varepsilon_2 &= (Y_2 - \mu) - \theta(Y_1 - \mu) = (Y_2 - \mu) - \theta\varepsilon_1;\\
&.\\
&.\\
&.\\
\varepsilon_T &= (Y_T - \mu) - \theta(Y_{T-1} - \mu) + \theta^2(Y_{T-2} - \mu) - ... + (-1)^{T-1}\theta^{T-1}(Y_1 - \mu).
\end{aligned}$$

Although it is simple to program this iteration by computer, the log likelihood function is a fairly complicated nonlinear function of μ and θ, so that an analytical expression for the MLE of μ and θ is not readily calculated. Hence even the conditional MLE for an $MA(1)$ process must be found by numerical optimization.

It is important to note if you have a sample of size T to estimate an $MA(1)$ process by conditional MLE, you will use all the T observation of this sample since it is conditional on $\varepsilon_0 = 0$ and not on first observation Y_1.

4.3 An Example of R code for Conditional Maximum Likelihood $MA(1)$ Estimation

The followings are the R code for estimation of $MA(1)$ process provided by Huo, Wen Wei at NSYSU.

```r
# Simulat the MA(1) without mean process dataset
data <- arima.sim(list(order = c(0,0,1), ma = 0.3), n = 1500)
# Set the initial value for parameters
params <- matrix(c(NA,NA),2,1) ; params[1,1] <- 0.2 ; params[2,1] <- 0.2
# Set a condition to stop the program if it equal maximum number of iterations
breakout <- FALSE ; iteration <- 1 ; max_iter <- 100
# Using loop to approximate conditional maximum likelihood of MA(1) without mean process
repeat{
  rs <- vector()
  l <- length(data)
  theta  <- params[1,1] ; sigma_sq  <- abs(params[2,1])

  for(n in (2:l)) # Using initial valus to get first set of residuals
  {
    rs[1] <- data[1]
    rs[n] <- (data[n]) - theta*rs[n-1]
    output <- list("res" = rs)
    res <- as.matrix(output$res)
  }
  A <- (sum(res*c(0,res[1:c(length(data)-1)])))/sigma_sq
  B <- (-length(data)/sigma_sq)/2 + 0.5*sum(res^2)/sigma_sq^2
  C <- sum((-c(0,res[1:c(length(data)-1)])*c(0,res[1:c(length(data)-1)])))/sigma_sq
  D <- E <- -sum((res)*c(0,res[1:c(length(data)-1)]))/sigma_sq^2
  F <- (length(data)/sigma_sq^2)/2 - sum((res)^2)/sigma_sq^3
# The Newton-Raphson Method
# Some idea about newton method's code learn from statistics textbook ISBN:9789571188454
  params <- (-solve(matrix(c(C,D,E,F),2,2))%*%matrix(c(A,B),2,1))+params
# Set a condition to restrict the iterations times
  iteration <- iteration + 1
  if(iteration == max_iter)
    { break }
  # Compute the log-liklihood
  L <- -length(data)/2*log(2*pi)-length(data)/2*log(params[2,1])-1/(2*params[2,1])*sum(res^2)
  # Save the output from the program
  coefficients <- (list("MA(1)" <- round(params[1] , digits = 3)       ,
                        "sigma_sq" <- round(params[2] , digits = 3)    ,
                        "total_iter" <- iteration                      ,
                        "log-liklihood" <- L))
}
coefficients
```

5 MLE of a Gaussian $MA(q)$ Process

This section discusses the estimation of a Gaussian $MA(q)$ process,

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q} \tag{17-16}$$

where all the roots of $1 + \theta_1 L + \cdots + \theta_q L^q = 0$ lie outside the unit circle and $\varepsilon_t \sim$ $i.i.d.$ $N(0, \sigma^2)$. In this case, the vector of population parameters to be estimated is $\boldsymbol{\theta} = (\mu, \theta_1, \theta_2, .., \theta_q, \sigma^2)'$.

5.1 Evaluating the Likelihood Function

The observations in the sample $(Y_1, Y_2, ..., Y_T)$ in a $(T \times 1)$ vector $\mathbf{y}$ which has mean vector $\boldsymbol{\mu}$ with each element μ and variance-covariance matrix given by $\boldsymbol{\Omega}$. The likelihood function is then

$$f_{Y_T, Y_{T-1}, ..., Y_1}(y_T, y_{T-1}, ..., y_1; \boldsymbol{\theta}) \;=\; (2\pi)^{-T/2} |\boldsymbol{\Omega}|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right].$$

Maximization of this exact log likelihood of an $MA(q)$ process must be accomplished numerically.

5.2 Evaluating the Likelihood Function Using (Scalar) Conditional Density

Consider the $p.d.f$ of Y_1,

$$Y_1 = \mu + \varepsilon_1 + \theta_1 \varepsilon_0 + \theta_2 \varepsilon_{-1} + ... + \theta_q \varepsilon_{-q+1}.$$

A simple approach is to condition on the assumption that the first q value of ε were all zero:

$$\varepsilon_0 = \varepsilon_{-1} = ... = \varepsilon_{-q+1} = 0.$$

Let $\boldsymbol{\varepsilon}_0$ denote the $(q \times 1)$ vector $(\varepsilon_1, \varepsilon_{-1}, ..., \varepsilon_{-q+1})'$. Then

$$(Y_1 | \boldsymbol{\varepsilon}_0 = 0) \sim N(\mu, \sigma^2)$$

or

$$\begin{aligned} f_{Y_1|\boldsymbol{\varepsilon}_0=0}(y_1|\boldsymbol{\varepsilon}_0=0;\boldsymbol{\theta}) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\cdot\frac{(y_1-\mu)^2}{\sigma^2}\right] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\varepsilon_1^2}{2\sigma^2}\right]. \; (Because\ \varepsilon_0=\varepsilon_{-1}=...=\varepsilon_{-q+1}=0) \end{aligned}$$

Next consider the distribution of the second observation Y_2 conditional on the "observing" $Y_1=y_1$. From (15),

$$Y_2=\mu+\varepsilon_2+\theta_1\varepsilon_1+\theta_2\varepsilon_0+...+\theta_q\varepsilon_{-q+1}. \tag{17-17}$$

Moreover, given observation of y_1, the value of ε_1 is then known with certainty as well:

$$\varepsilon_1=y_1-\mu\ \ and\ \ \varepsilon_0=\varepsilon_{-1}=...=\varepsilon_{-q+2}=0.$$

Hence

$$(Y_2|Y_1=y_1,\ \boldsymbol{\varepsilon}_0=0)\sim N((\mu+\theta_1\varepsilon_1),\sigma^2),$$

meaning that

$$\begin{aligned} f_{Y_2|Y_1,\,\boldsymbol{\varepsilon}_0=0}(y_2|y_1,\boldsymbol{\varepsilon}_0=0;\boldsymbol{\theta}) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\cdot\frac{(y_2-\mu-\theta_1\varepsilon_1)^2}{\sigma^2}\right] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\varepsilon_2^2}{2\sigma^2}\right]. \end{aligned}$$

Since ε_1 is know with certainty, ε_2 can be calculated from

$$\varepsilon_2=y_2-\mu-\theta_1\varepsilon_1.$$

Proceeding in this fashion, it is clear that given knowledge that $\boldsymbol{\varepsilon}_0=0$, the full sequence $\{\varepsilon_1,\varepsilon_2,...,\varepsilon_T\}$ can be calculated from $\{y_1,y_2,...,y_T\}$ by iterating on

$$\varepsilon_t=y_t-\mu-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-...-\theta_q\varepsilon_{t-q}$$

for $t=1,2,...,T$, starting from $\boldsymbol{\varepsilon}_0=0$. The likelihood (conditional on $\boldsymbol{\varepsilon_0}=0$) of the complete sample can thus be calculated as the product of these individual densities:

$$\begin{aligned} & f_{Y_T,Y_{T-1},Y_{T-2},...,Y_1|\boldsymbol{\varepsilon}_0=0}(y_T,y_{T-1},y_{T-2},...,y_1|\boldsymbol{\varepsilon}_0=0;\boldsymbol{\theta}) \\ = \;& f_{Y_1|\boldsymbol{\varepsilon}_0=0}(y_1|\boldsymbol{\varepsilon}_0=0;\boldsymbol{\theta})\cdot\prod_{t=2}^{T} f_{Y_t|Y_{t-1},Y_{t-2},...,Y_1,\boldsymbol{\varepsilon}_0=0}(y_t|y_{t-1},y_{t-2},...,y_1,\boldsymbol{\varepsilon}_0=0;\boldsymbol{\theta}). \end{aligned}$$

The conditional log likelihood function (denoted $\mathcal{L}^*(\boldsymbol{\theta})$) is therefore

$$\begin{aligned}\mathcal{L}^*(\boldsymbol{\theta}) &= \log f_{Y_T,Y_{T-1},Y_{T-2},...,Y_1|\boldsymbol{\varepsilon}_0=0}(y_T, y_{T-1}, y_{T-2}, ..., y_1|\boldsymbol{\varepsilon}_0 = 0; \boldsymbol{\theta}) \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T}\frac{\varepsilon_t^2}{2\sigma^2}. \qquad (17\text{-}18)\end{aligned}$$

It is important to note if you have a sample of size T to estimate an $MA(q)$ process by conditional MLE, you will also use all the T observation of this sample.

6 MLE of a Gaussian $ARMA(p,q)$ Process

This section discusses a Gaussian $ARMA(p,q)$ process,

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q},$$

where all the roots of $1 - \phi_1 L - \cdots - \phi_p L^p = 0$ and $1 + \theta_1 L + \cdots + \theta_q L^q = 0$ lie outside unit circle and $\varepsilon_t \sim i.i.d.\ N(0, \sigma^2)$. In this case, the vector of population parameters to be estimated is $\boldsymbol{\theta} = (c, \phi_1, \phi_2, ..., \phi_p, \theta_1, \theta_2, ..., \theta_q, \sigma^2)'$.

6.1 Conditional maximum Likelihood estimates

The approximation to the likelihood function for an autoregrssion conditional on initial value of the $y's$. The approximation to the likelihood function for a moving average process conditioned on initial value of the ε's. A common approximation to the likelihood function for an $ARMA(p,q)$ process conditions on both y's and ε's.

The $(p+1)$th observation is

$$Y_{p+1} = c + \phi_1 Y_p + \phi_2 Y_{p-1} + ... + \phi_p Y_1 + \varepsilon_{p+1} + \theta_1 \varepsilon_p + ... + \theta_q \varepsilon_{p-q+1}.$$

Conditional on $Y_1 = y_1, Y_2 = y_2, ..., Y_p = y_p$ and setting $\varepsilon_p = \varepsilon_{p-1} = ... = \varepsilon_{p-q+1} = 0$ we have

$$Y_{p+1} \sim N((c + \phi_1 Y_p + \phi_2 Y_{p-1} + ... + \phi_p Y_1), \sigma^2).$$

Then the conditional likelihood calculated from $t = p+1, ..., T$ is

$$\begin{aligned} \mathcal{L}^*(\boldsymbol{\theta}) &= \log f(y_T, y_{T-1}, ..y_{p+1} | y_p, ..., y_1, \varepsilon_p = \varepsilon_{p-1} = ... = \varepsilon_{p-q+1} = 0; \boldsymbol{\theta}) \\ &= -\frac{T-p}{2} \log(2\pi) - \frac{T-p}{2} \log(\sigma^2) - \sum_{t=p+1}^{T} \frac{\varepsilon_t^2}{2\sigma^2}, \end{aligned} \quad (17\text{-}19)$$

where the sequence $\{\varepsilon_{p+1}, \varepsilon_{p+2}, ..., \varepsilon_T\}$ can be calculated from $\{y_1, y_2, ..., y_T\}$ by iterating on

$$\begin{aligned} \varepsilon_t &= Y_t - c - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - ... - \phi_p Y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q}, \\ & \quad t = p+1, p+2, ..., T. \end{aligned}$$

It is important to note if you have a sample of size T to estimate an $ARMA(p,q)$ process by conditional MLE, you will only use the $T-p$ observation of this sample.

From (17-9),(17-11),(17-14),(17-17), and (17-18) we see that all the conditional log-likelihood function take a concise form

$$-\frac{T^*}{2}\log(2\pi)-\frac{T^*}{2}\log(\sigma^2)-\sum_{t=t^*}^{T}\left[\frac{\varepsilon_t^2}{2\sigma^2}\right],$$

where T^* and t^* is the appropriate total and first observations used, respectively. The solution to the conditional log-likelihood function $\hat{\boldsymbol{\theta}}$ is also called the **conditional sums of squared estimator, CSS**, denoted as $\hat{\boldsymbol{\theta}}_{CSS}$.

7 A Short Tour to Numerical Optimization

We consider the general problem of maximizing a function of several variables:

$$\text{maximize}_{\boldsymbol{\theta}}\ F(\boldsymbol{\theta}),$$

where $F(\boldsymbol{\theta})$ may be log-likelihood or some other function. An efficient means of solving most nonlinear maximization problem is by an *iterative algorithm*:

> "Beginning from initial value $\boldsymbol{\theta}_0$, at entry to iteration t, if $\boldsymbol{\theta}_t$, is not the optimal value for $\boldsymbol{\theta}$, compute direction vector $\boldsymbol{\Delta}_t$, and step size λ_t, then
>
> $$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t.''$$

The most commonly used algorithm are gradient method and template for most gradient method in common use is the Newton's method.

The basis for Newton's method is a linear Taylor series approximation. Expanding the first-order conditions,

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad (\textit{which may have nonlinear solution})$$

in a linear Taylor series around an arbitrary $\boldsymbol{\theta}^0$ yield

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \simeq \mathbf{g}^0 + \mathbf{H}^0(\boldsymbol{\theta} - \boldsymbol{\theta}^0),$$

where the superscript indicates that the term is evaluated at $\boldsymbol{\theta}^0$ and $\mathbf{g}$ and $\mathbf{H}$ are the gradient vector and Hessian matrix, respectively. If $F(\boldsymbol{\theta})$ attains a local maximum at $\boldsymbol{\theta}_1$, then we must necessarily have $\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}_1} = \mathbf{0}$. Solving for $\boldsymbol{\theta}_1$, we obtain

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 - \mathbf{H}_0^{-1}\mathbf{g}_0.$$

If now we approximate $\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ with another linear function, by again applying Taylor's expansion in a neighborhood of $\boldsymbol{\theta}_1$,[3] and then repeat the same process as before with $\boldsymbol{\theta}_1$ used instead of $\boldsymbol{\theta}_0$, we obtain

$$\boldsymbol{\theta}_2 = \boldsymbol{\theta}_1 - \mathbf{H}_1^{-1}\mathbf{g}_1.$$

Further repetitions of this process lead to the iteration,

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \mathbf{H}_i^{-1}\mathbf{g}_i, \quad i = 0, 1, 2, ...$$

[3] It is noted that here, for a given initial $\boldsymbol{\theta}_0$, we can obtain the value of $\boldsymbol{\theta}_1$.

The iterations stop at the value of $\boldsymbol{\theta}^*$ such that $\mathbf{g}^* \equiv \frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}^*} = \mathbf{0}$. Under this circumstance, $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}^*$ and therefore $\mathbf{g}_{i+1} = \mathbf{g}_i = \mathbf{0}$. In this case, $\lambda_t = 1$ and $\boldsymbol{\Delta}_t = -\mathbf{H}_t^{-1}\mathbf{g}_t$.

The Newton-Raphson method requires finding the inverse of the Hessian matrix $\mathbf{H}$ at each iteration. This can be computationally involved, especially if the number of the variables in $\boldsymbol{\theta}$ is large. Furthermore, the method may fail to converge if $\mathbf{H}_i$ is not positive definite. This can occur, for example, when $\boldsymbol{\theta}_i$ is far from the location $\boldsymbol{\theta}^*$ of the true maximum. If, however, the initial point $\boldsymbol{\theta}_0$, is close to $\boldsymbol{\theta}^*$, then convergence occurs at a rapid rate.

8 Statistical Inference with MLE

8.1 Asymptotic Standard Errors for MLE

Although MLE's enjoy several optimum finite sample properties, their asymptotic properties provide the main justification for the almost universal appeal to the method of maximum likelihood. As argued, under certain regularity conditions, MLE can be shown to be consistent, asymptotically normal and asymptotically efficient.

If the sample size T is sufficiently large, it often turns out that the distribution of the MLE $\hat{\boldsymbol{\theta}}$ can be well approximated by the following distribution:

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}) \xrightarrow{L} N(\mathbf{0}, (\boldsymbol{I}_T(\boldsymbol{\theta}))^{-1}),$$

where $\boldsymbol{\theta}$ denote the true parameter vector. The matrix $\boldsymbol{I}_T(\boldsymbol{\theta})$ is known as the information matrix and can be estimated in either of three ways.

8.1.1 Estimating the Asymptotic Variance of the MLE, $\boldsymbol{\theta}$ is $k \times 1$

The asymptotic covariance matrix of the MLE is a matrix of parameters that must be estimated. The followings are three methods to estimate this variance.

(a). If the form of the expected value of the second derivative of the log-likelihood is known, we can evaluate the information matrix at $\hat{\boldsymbol{\theta}}$ to estimate the covariance matrix for the MLE,

$$[\widehat{\boldsymbol{I}_T(\boldsymbol{\theta})}]^{-1} = \left\{ -E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right\}^{-1}.$$

■

If the expected value of the second derivative of the log-likelihood is complicated, two alternative estimators is

(b).

$$[\widetilde{\widehat{\boldsymbol{I}_T(\boldsymbol{\theta})}}]^{-1} = \left\{ -\left[\frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}\,\partial\hat{\boldsymbol{\theta}}'}\right] \right\}^{-1}, \tag{17-20}$$

■

and

(c). the BHHH estimator

$$\overline{[\boldsymbol{I}_T(\boldsymbol{\theta})]^{-1}} = \left[\sum_{i=1}^{n} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i'\right]^{-1}, \; where \; \hat{\mathbf{g}}_i = \frac{\partial \ln f(y_t|y_{t-1}, y_{t-2}, ...; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \; is \; k \times 1.$$

■

Example.

Let the log likelihood be

$$\ln L(\boldsymbol{\theta}) = -1.5\theta_1^1 - 2\theta_2^2.$$

We can easily see analytically for this example that the MLE is given by $\hat{\boldsymbol{\theta}} = (0,0)'$. For this case, one can see analytically that

$$\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \left[\begin{array}{cc} -3 & 0 \\ 0 & -4 \end{array}\right],$$

and so results (17-19) suggest that the variance of the MLE $\hat{\theta}_2$ can be approximated by 1/4. The MLE for this example was $\hat{\theta}_2 = 0$. Thus an approximated 95% confidence interval for θ_2 is given by

$$0 \pm 2\sqrt{\frac{1}{4}} = \pm 1.$$

8.2 CSS estimators will be equivalent to MLE.

For an large number of observations the CSS estimators will be equivalent to MLE, i.e.

$$\hat{\boldsymbol{\theta}}_{CSS} - \hat{\boldsymbol{\theta}}_{MLE} \xrightarrow{p} \mathbf{0}.$$

See Pierce (1971), " Least square estimation of a mixed autoregressive-moving average process", *Biometrika* 58: pp. 299-312.

9 Bias in the OLS (CSS) Estimation of AR(1) Model

For the $AR(1)$ model, $Y_t = c + \phi Y_{t-1} + \varepsilon_t$, the OLS or CSS estimators (see eq. (17.11)) will be

$$\hat{\phi} = \frac{\sum_{t=2}^{T}(Y_t - \bar{Y}_1)(Y_{t-1} - \bar{Y}_2)}{\sum_{t=2}^{T}(Y_{t-1} - \bar{Y}_2)^2}, \tag{17-21}$$

with $\bar{Y}_1 = \frac{\sum_{t=2}^{T} Y_t}{T-1}$ and $\bar{Y}_2 = \frac{\sum_{t=1}^{T-1} Y_t}{T-1}$. Kendall (1954) show that

$$E(\hat{\phi}) \approx \phi - \frac{1+3\phi}{T-1},$$

therefore $\hat{\phi}$ is bias in finite sample.[4] With relatively small T and large ϕ, which are typical of macroeconomic time series, the first order bias could be quite large.

9.1 Bias-Reduction

9.1.1 Kendall(1954)

Replacing $E(\hat{\phi})$ with the estimate $\hat{\phi}$, leads to Kendall's bias-corrected estimate:[5]

$$\hat{\phi}^K = \frac{T-1}{T-4}\hat{\phi} + \frac{1}{T-4}.$$

9.1.2 Andrews (1993)

The condition of median-unbiasedness is often more useful than that of mean-unbiasedness when the parameter space is bounded or when the distribu- tions of estimators are skewed and/or kurtotic. When the parameter space is bounded and closed and estimators take values in the parameter space, it is impossible to have a mean-unbiased estimator because all estimators are biased at extreme boundary point.....

9.1.3 Rudebusch (1992)

Although the OLS estimates of the coefficients of the TS model are consistent and asymptotically normal, they are biased in small samples because of the presence of

[4] The bias can be seen from the following: Let $\mathbf{X} = (Y_0, Y_1, ..., Y_{T-1})'$, and that $E(\varepsilon_t \mathbf{X}) = E_{\mathbf{X}} E(\varepsilon_t|\mathbf{X})$. Because $E(\varepsilon_t \mathbf{X}) \neq 0$, hence $E(\varepsilon_t|\mathbf{X}) \neq 0$, which violate the requirement of assumption for unbiasedness in p.11 of Ch.7.

[5] Such that $E(\hat{\phi}^K) = \frac{T-1}{T-4}\left(\phi - \frac{1+3\phi}{T-1}\right) + \frac{1}{T-4} = \phi$.

lagged dependent variables. An arguably more plausible TS alternative would correct the coefficient estimates for this bias.

The small-sample bias of the OLS estimates of autoregressive model coefficients is most easily documented for an AR(1) process. The middle column of Table 8 provides the median value of the OLS estimate $\hat{\rho}_1$, based on repeated samples from the first-order process $Y_t = \mu + \gamma t + \rho_1 Y_{t-1} + \varepsilon_t$, for a variety of values of ρ_1 (with $\mu = \gamma = 0$). This estimate is downwardly biased over a wide range of ρ_1, with the deviation of the median estimate from the true value being particularly pronounced for values of ρ_1 that are just less than one. The third column of Table 8 gives the probability of obtaining an OLS estimate equal to or greater than ρ_1; these probabilities also indicate that a given estimate is more likely to be below rather than above the true value of the autoregressive parameter.

Based on Table 8, it is likely that the OLS estimation employed in a TS model $\hat{\rho}_1$ would be lower than the true value of ρ_1. A more plausible TS model of the data-generating process would correct for this downward small-sample bias.

Rudebusch I define the 'median-unbiased' TS model as the one that, across repeated simulations, has a median OLS estimate of each parameter that is equivalent to the actual sample estimate of that parameter. Formally, let the vector of median-unbiased TS model coefficients be denoted as $\Phi_{MUE} = (\mu, \gamma, \rho_1)$; across repeated samples the median OLS estimates of these coefficients is median $(\hat{\Phi}_{MUE})$. Let $\hat{\Phi}_s$ be the vector of OLS parameters estimated from the true data sample under consideration (which formed the parameters of the TS data-generating processes2). The vector (Φ_{MUE}) is defined by the equality of median $(\hat{\Phi}_{MUE} = \hat{\Phi}_s)$; that is, the median-unbiased model has median OLS parameter estimates equal to the sample estimates.

Spartan Stadium, MSU.

End of this Chapter