UNIVERSITÀ DEGLI STUDI DI MILANO
Dipartimento di Economia, Management
e Metodi Quantitativi

Academic Year 2019-2020

Time Series Econometics

Fabrizio Iacone

Chapter 6, Parametric Estimation, part 2

Topics: Exact Maximum Likelihood estimation, Conditional Maximum Likelihood estimation, Optimisation of the (Pseudo) Maximum Likelihood

Estimation : maximum likelihood

Let

$$\mathbf{Y} = (Y_1, \ldots, Y_T)'$$

be a Normally distributed vector with

$$E(\mathbf{Y}) = \boldsymbol{\mu}, E\Big((\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})'\Big) = \boldsymbol{\Omega}.$$

The Gaussian density, computed at the point

$$\mathbf{y} = (y_1, .., y_T)'$$

in the support of **Y** is

$$\begin{aligned} &f_{Y_1,..,Y_T}(y_1, .., y_T) \\ &= (2\pi)^{-T/2} |\boldsymbol{\Omega}|^{-1/2} \exp\Big(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \boldsymbol{\mu})\Big) \end{aligned}$$

Now assume that $\mathbf{y} = (y_1, .., y_T)'$ is the realisation of **Y**, and consider $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ and $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a set of parameters of interest. Then,

$$\begin{aligned} f_{Y_1,..,Y_T}(\boldsymbol{\beta}) &= (2\pi)^{-T/2} |\boldsymbol{\Omega}(\boldsymbol{\beta})|^{-1/2} \\ &\times \exp\Big(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}))' \boldsymbol{\Omega}(\boldsymbol{\beta})^{-1} (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}))\Big) \end{aligned}$$

is the likelihood function.

Maximising that function with respect to $\boldsymbol{\beta}$ gives the (exact) maximum likelihood estimate, $\widehat{\boldsymbol{\beta}}_{ML}$, i.e.

$$\widehat{\boldsymbol{\beta}}_{ML} = \arg \max_{\boldsymbol{\beta}} f_{Y_1,..,Y_T}(\boldsymbol{\beta})$$

Some comments about the notation

1. Hamilton uses $\boldsymbol{\theta}$ instead of $\boldsymbol{\beta}$. I put $\boldsymbol{\beta}$ to avoid confusion with the parameter θ.

2. We are often interested in ARMA models: in this case, the parameters of interest are $\boldsymbol{\beta} = (c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)'$.

3. Be careful to distinguish between the density and the likelihood. The density $f_{Y_1,..,Y_T}(y_1, .., y_T)$ is a function of $y_1, .., y_T$, and it is computed for given (known) value of the parameters $\boldsymbol{\beta}$; the likelihood $f_{Y_1,..,Y_T}(\boldsymbol{\beta})$ is a function of the parameters and it is computed for the (given) value of the observations $y_1, .., y_T$. We indicate the likelihood as $f_{Y_1,..,Y_T}(\boldsymbol{\beta})$ (without reference to $y_1, .., y_T$ in the argument) but Hamilton uses $f_{Y_1,..,Y_T}(y_1, .., y_T; \boldsymbol{\beta})$ instead.

4. Hamilton sometimes also uses $f_{Y_1,..,Y_T}(y_1,..,y_T;\boldsymbol{\beta})$ for the density; the difference between density and likelihood must be clear from the context.

5. Hamilton also uses the notation $f_{Y_T,..,Y_1}(y_T,..,y_1;\boldsymbol{\beta})$, i.e. inverting the order of Y_1, ..., Y_T and y_1, ..., y_T in the notation: $f_{Y_T,..,Y_1}(y_T,..,y_1;\boldsymbol{\beta})$ and $f_{Y_1,..,Y_T}(y_1,..,y_T;\boldsymbol{\beta})$ are the same function.

6. We will often refer to the parameters that generated the data (and the values of which we want to estimate) by adding a subscript 0, i.e. $\boldsymbol{\beta}_0$.

Examples:

AR(1) ($|\phi_0| < 1$):

$$Y_t = c_0 + \phi_0 Y_{t-1} + \varepsilon_t,\ \ \varepsilon_t \sim Nid(0, \sigma_0^2)$$

$\boldsymbol{\beta}$=$(c, \phi, \sigma^2)'$, ($|\phi| < 1$) and

$$\boldsymbol{\Omega}(\boldsymbol{\beta}) = \frac{\sigma^2}{1-\phi^2}$$

$$\times \begin{pmatrix} 1 & \phi & \dots & \phi^{T-2} & \phi^{T-1} \\ \phi & 1 & \dots & \phi^{T-3} & \phi^{T-2} \\ \dots & \dots & \dots & \dots & \dots \\ \phi^{T-2} & \phi^{T-3} & \dots & 1 & \phi \\ \phi^{T-1} & \phi^{T-2} & \dots & \phi & 1 \end{pmatrix}$$

MA(1) ($|\theta_0| < 1$):

$$Y_t = \mu_0 + \varepsilon_t + \theta_0 \varepsilon_{t-1}, \;\; \varepsilon_t \sim Nid(0, \sigma_0^2)$$

$\boldsymbol{\beta} = (\mu, \theta, \sigma^2)'$ and

$$\boldsymbol{\Omega}(\boldsymbol{\beta}) = \sigma^2(1+\theta^2)$$

$$\times \begin{pmatrix} 1 & \frac{\theta}{(1+\theta^2)} & \dots & 0 & 0 \\ \frac{\theta}{(1+\theta^2)} & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & \frac{\theta}{(1+\theta^2)} \\ 0 & 0 & \dots & \frac{\theta}{(1+\theta^2)} & 1 \end{pmatrix}$$

The likelihood function may be computed for given set of observations and for any parameter (within the range of the parameter space).

For example, assume that we know that $\mu_0 = 0$ and we observed

time	y_1	y_2	y_3	y_4
observation	0.5	-0.8	-0.2	2

and suppose you want to estimate θ_0 in the MA(1) model when $\mu_0 = 0$ and $\sigma_0^2 = 1$ is also known (so $\boldsymbol{\beta} = \theta$ in this case). Consider five potential values for θ_0: $-0.5,\ -0.25,\ 0,\ 0.25,\ 0.5$. Then, we have to compute $\boldsymbol{\Omega}(\boldsymbol{\beta})$ for each θ: for example, when $\theta = 0.5$,

$$\boldsymbol{\Omega}(0.5) = (1 + 0.5^2)$$

$$\times \begin{pmatrix} 1 & \frac{0.5}{(1+0.5^2)} & 0 & 0 \\ \frac{0.5}{(1+0.5^2)} & 1 & \frac{0.5}{(1+0.5^2)} & 0 \\ 0 & \frac{0.5}{(1+0.5^2)} & 1 & \frac{0.5}{(1+0.5^2)} \\ 0 & 0 & \frac{0.5}{(1+0.5^2)} & 1 \end{pmatrix}$$

and

$$(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\beta}))'\boldsymbol{\Omega}(\boldsymbol{\beta})^{-1}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\beta})) =$$

$$= \begin{pmatrix} 0.5 & -0.8 & -0.2 & 2 \end{pmatrix} \times$$

$$\times \begin{pmatrix} 1.25 & 0.5 & 0 & 0 \\ 0.5 & 1.25 & 0.5 & 0 \\ 0 & 0.5 & 1.25 & 0.5 \\ 0 & 0 & 0.5 & 1.25 \end{pmatrix}^{-1} \begin{pmatrix} 0.5 \\ -0.8 \\ -0.2 \\ 2 \end{pmatrix}$$

$$= 4.6903$$

so

$$(2\pi)^{-T/2}|\boldsymbol{\Omega}(\boldsymbol{\beta})|^{-1/2}$$

$$\times \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\beta}))'\boldsymbol{\Omega}(\boldsymbol{\beta})^{-1}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\beta}))\right)$$

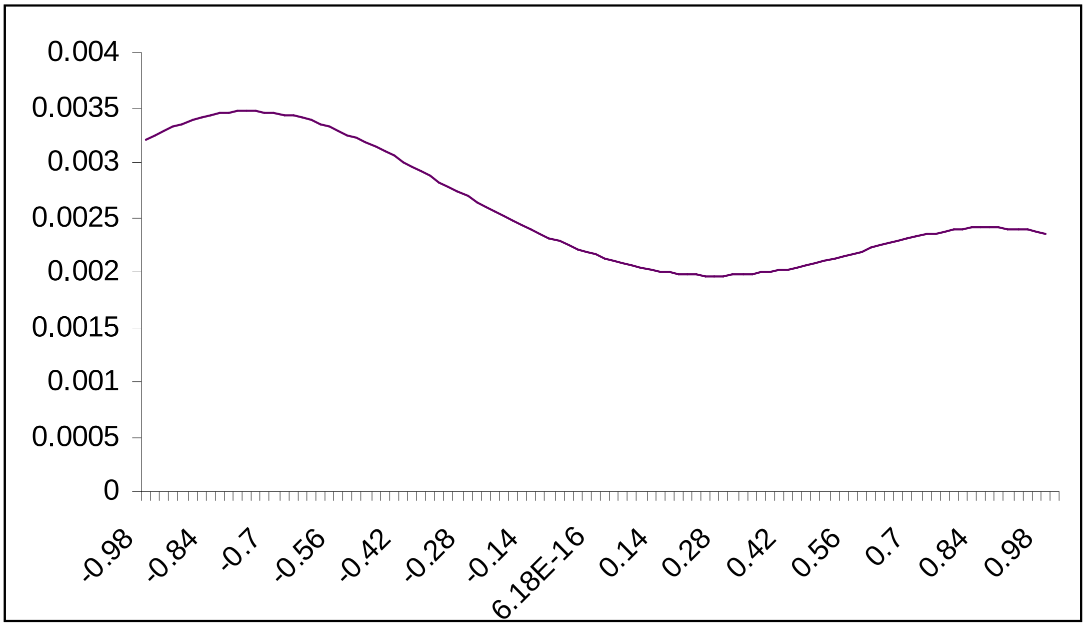
$$= (2\pi)^{-4/2} \times (1.332)^{-1/2} \times \exp(-\frac{1}{2}4.6903)$$

$$= 2.1033 \times 10^{-3}$$

and, for the other values of θ,

θ	-0.5	-0.25	0	0.25	0.5
$1000 \times f$	3.178	2.618	2.153	1.967	2.103

The function may be computed for all the θ, $|\theta| < 1$ $(\widehat{\theta} = -0.76)$



The computation of

$$|\mathbf{\Omega}(\boldsymbol{\beta})|^{-1/2}\exp\Big(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\beta}))'\mathbf{\Omega}(\boldsymbol{\beta})^{-1}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\beta}))\Big)$$

is very heavy, because it requires the inversion of the $T \times T$ matrix $\mathbf{\Omega}(\boldsymbol{\beta})$ for all the admissible values $\boldsymbol{\beta}$.

Luckily, it is sometimes easy to rewrite the likelihood function in a way that does not require the inversion of $\mathbf{\Omega}(\boldsymbol{\beta})$; otherwise, it is also possible to modify the problem so that, again, we can avoid the inversion of of $\mathbf{\Omega}(\boldsymbol{\beta})$.

AR(1)

$$Y_t = c_0 + \phi_0 Y_{t-1} + \varepsilon_t, \ |\phi_0| < 1, \ \ \varepsilon_t \sim Nid(0, \sigma_0^2)$$

Then

$$Y_t \sim N\left(\frac{c_0}{1-\phi_0}, \sigma_0^2 \frac{1}{1-\phi_0^2} \right)$$

so the density of Y_1, $f_{Y_1}(y_1)$, is

$$(2\pi)^{-1/2} \left| \frac{\sigma_0^2}{1-\phi_0^2} \right|^{-1/2} \exp\left(-\frac{1}{2} \frac{\left(y_1 - \frac{c_0}{1-\phi_0}\right)^2}{\frac{\sigma_0^2}{1-\phi_0^2}} \right)$$

Of course, the same density may be expressed for Y_2, however in this case we can also exploit the fact that we observed Y_1 on the period before, and look at the density of Y_2 conditional on Y_1,

$$Y_2 | Y_1 \sim N(c_0 + \phi_0 Y_1, \sigma_0^2),$$

$$\begin{aligned} & f_{Y_2|Y_1}(y_2|y_1) \\ &= (2\pi)^{-1/2} |\sigma_0^2|^{-1/2} \exp\left(-\frac{1}{2} \frac{(y_2 - c_0 - \phi_0 y_1)^2}{\sigma_0^2} \right). \end{aligned}$$

We can then express the joint density $f_{Y_1,Y_2}(y_1, y_2)$ (ie, of $f_{Y_2,Y_1}(y_2, y_1)$) as

$$f_{Y_2,Y_1}(y_2, y_1) = f_{Y_2|Y_1}(y_2|y_1) f_{Y_1}(y_1)$$

and, by the same argument,

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

where

$$f_{Y_t|Y_{t-1},...,Y_1}(y_t|y_{t-1},\ldots,y_1)$$
$$= (2\pi)^{-1/2}|\sigma_0^2|^{-1/2} \exp\left(-\frac{1}{2}\frac{(y_t - c_0 - \phi_0 y_{t-1})^2}{\sigma_0^2}\right)$$

when $t = 2, \ldots T$.

Also notice that, for the AR(1),

$$f_{Y_t|Y_{t-1},...,Y_1}(y_t|y_{t-1},\ldots,y_1) = f_{Y_t|Y_{t-1}}(y_t|y_{t-1})$$

so in what follows we simplify the notation in this way.

So, setting $\boldsymbol{\beta} = (c, \phi, \sigma^2)'$, letting

$$f_{Y_1}(\boldsymbol{\beta})$$

$$= (2\pi)^{-1/2} \left| \frac{\sigma^2}{1-\phi^2} \right|^{-1/2} \exp\left(-\frac{1}{2} \frac{\left(y_1 - \frac{c}{1-\phi}\right)^2}{\frac{\sigma^2}{1-\phi^2}} \right)$$

and

$$f_{Y_t|Y_{t-1}}(\boldsymbol{\beta})$$

$$= (2\pi)^{-1/2} |\sigma^2|^{-1/2} \exp\left(-\frac{1}{2} \frac{(y_t - c - \phi y_{t-1})^2}{\sigma^2} \right)$$

the likelihood is then

$$f_{Y_1,\ldots.,Y_T}(\boldsymbol{\beta}) = \prod_{t=2}^{T} f_{Y_t|Y_{t-1},\ldots,Y_1}(\boldsymbol{\beta})\, f_{Y_1}(\boldsymbol{\beta})$$

and the log-likelihood is

$$\mathcal{L}(\boldsymbol{\beta}) = \ln(f_{Y_1}(\boldsymbol{\beta})) + \sum_{t=2}^{T} \ln(f_{Y_t|Y_{t-1}}(\boldsymbol{\beta}))$$

$$= -\frac{1}{2} \ln\left(2\pi \frac{\sigma^2}{1-\phi^2} \right) - \frac{1}{2} \frac{\left(y_1 - \frac{c}{1-\phi}\right)^2}{\frac{\sigma^2}{1-\phi^2}}$$

$$- \frac{T-1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \sum_{t=2}^{T} \frac{(y_t - c - \phi y_{t-1})^2}{\sigma^2}$$

We then succeded in rewriting the (log) likelihood in a way that does not require the inversion of a $T \times T$ matrix.

Maximising that function gives the "maximum likelihood estimate" when ε_t is normally distributed.

However, althought we eliminated the problem of inverting $\mathbf{\Omega}(\boldsymbol{\beta})$, we still can’t express our estimate $\widehat{\boldsymbol{\beta}}$ as a closed form function of the observations, so we still have to compute the likelihood function on all the admissible parameters in order to find the maximum.

Consider, on the other hand, estimating $\boldsymbol{\beta}_0$ by maximising

$$-\frac{T-1}{2}\ln(2\pi\sigma^2)-\frac{1}{2}\sum_{t=2}^{T}\frac{(y_t - c - \phi y_{t-1})^2}{\sigma^2}$$

That estimate is known as "conditional maximum likelihood estimate", because it is the maximum likelihood estimate if Y_1 is not random (so, the log-likelihood above is called "conditional" log-likelihood). In this case, a closed form solution exists.

To find the closed form solution for the conditional maximum likelihood estimate, first notice that σ^2 can be estimated and concentrated out:

$$\frac{\partial}{\partial\sigma^2}\left(-\frac{T-1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\sum_{t=2}^{T}\frac{(y_t - c - \phi y_{t-1})^2}{\sigma^2}\right)$$

$$= -\frac{T-1}{2}\frac{1}{\sigma^2} + \frac{1}{2}\sum_{t=2}^{T}\frac{(y_t - c - \phi y_{t-1})^2}{(\sigma^2)^2}$$

and, equating the derivative to 0,

$$\widehat{\sigma^2} = \frac{1}{(T-1)}\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2$$

so replacing this in the log likelihood, the concentrated likelihood is

$$-\frac{T-1}{2}\ln(2\pi)$$

$$-\frac{T-1}{2}\ln\left(\frac{1}{(T-1)}\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2\right)$$

$$-\frac{T-1}{2}$$

which is maximised if $\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2$ is minimised.

This is the standard OLS problem, so the solution is

$$\widehat{c} = \frac{1}{T-1}\sum_{t=2}^{T}(y_t - \phi y_{t-1}) = \bar{y}_{.} - \widehat{\phi}\bar{y}_{.-1}$$

$$\widehat{\phi} = \frac{\sum_{t=2}^{T}(y_t - \bar{y}_{.})(y_{t-1} - \bar{y}_{.-1})}{\sum_{t=2}^{T}(y_{t-1} - \bar{y}_{.-1})^2}$$

where

$$\bar{y}_{.} = \frac{1}{T-1}\sum_{t=2}^{T} y_t,\ \bar{y}_{.-1} = \frac{1}{T-1}\sum_{t=2}^{T} y_{t-1}$$

So for the "conditional maximum likelihood estimate" a closed form solution exists, and it is the OLS estimate in $Y_t = c_0 + \phi_0 Y_{t-1} + \varepsilon_t$.

Notice that this is not the likelihood function of our original stationary AR(1) process, but the likelihood of the process

$$Y_t = c_0 + \phi_0 Y_{t-1} + \varepsilon_t,\ |\phi_0| < 1,$$

$$\varepsilon_t \sim Nid(0, \sigma_0^2) \text{ when } t > 1;$$

$$Y_1 = y_1$$

(hence the name, "conditional maximum likelihood").

AR(p)

$$Y_t = c_0 + \phi_{0;1} Y_{t-1} + \ldots + \phi_{0;p} Y_{t-p} + \varepsilon_t,$$

$$\text{where } \varepsilon_t \sim Nid(0, \sigma_0^2)$$

and the roots of $1 - \phi_{0;1} z - \ldots - \phi_{0;p} z^p = 0$ are outside the unit circle. Using conditioning, we rewrite the density as

$$f_{Y_1,\ldots.,Y_T}(y_1, \ldots, y_T)$$

$$= \prod_{t=p+1}^{T} f_{Y_t|Y_{t-1},\ldots,Y_{t-p}}(y_t|y_{t-1}, \ldots, y_{t-p})$$

$$\times f_{Y_1,\ldots,Y_p}(y_1, \ldots, y_p)$$

Introduce

$$\mathbf{Y}_p = (Y_1, \ldots, Y_p)', \boldsymbol{\mu}_p = E(\mathbf{Y}_p), \mathbf{y}_p = (y_1, \ldots, y_p)'$$

$$\text{and } V_p = (\sigma_0^2)^{-1} E\left[\left(\mathbf{Y}_p - \boldsymbol{\mu}_p \right) \left(\mathbf{Y}_p - \boldsymbol{\mu}_p \right)' \right]$$

and take again the Gaussian density,

$$f_{Y_1,\ldots,Y_p}(y_1, \ldots, y_p) = (2\pi)^{-p/2} |\sigma_0^2 V_p|^{-1/2}$$

$$\times \exp\left(-\frac{1}{2\sigma^2} \left(\mathbf{y}_p - \boldsymbol{\mu}_p \right)' V_p^{-1} \left(\mathbf{y}_p - \boldsymbol{\mu}_p \right) \right)$$

$$f_{Y_t|Y_{t-1},\ldots,Y_{t-p}}(y_t|y_{t-1}, \ldots, y_{t-p}) = (2\pi)^{-1/2} |\sigma_0^2|^{-1/2}$$

$$\times \exp\left(-\frac{1}{2} \frac{(y_t - c_0 - \phi_{0;1} y_{t-1} - \ldots - \phi_{0;p} y_{t-p})^2}{\sigma_0^2} \right)$$

when $t = p + 1, \ldots T$.

Taking logarithms, the log-likelihood is

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}) &= \ln(f_{Y_1,\ldots,Y_p}(\boldsymbol{\beta})) + \sum_{t=p+1}^{T} \ln(f_{Y_t|Y_{t-1},\ldots,Y_{t-p}}(\boldsymbol{\beta})) \\
&= -\frac{p}{2}\ln(2\pi) - \ln\left|\sigma^2 V_p(\boldsymbol{\beta})\right|^{-1/2} \\
&\quad - \frac{1}{2\sigma^2}\Big(\mathbf{y}_p - \boldsymbol{\mu}_p(\boldsymbol{\beta})\Big)' V_p(\boldsymbol{\beta})^{-1}\Big(\mathbf{y}_p - \boldsymbol{\mu}_p(\boldsymbol{\beta})\Big) \\
&\quad - \frac{T-p}{2}\ln(2\pi\sigma^2) \\
&\quad - \frac{1}{2}\sum_{t=p+1}^{T} \frac{(y_t - c - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p})^2}{\sigma^2}
\end{aligned}$$

where the problem of inverting the $T \times T$ matrix $\Omega(\boldsymbol{\beta})$ is reduced to inverting a $p \times p$ matrix $V_p(\boldsymbol{\beta})$.

Maximising the log likelihood yields then the "maximum likelihood estimate".

Again, a "conditional maximum likelihood estimate" can be considered instead: this is obtained by treating $Y_1, \ldots, Y_p$ as given, and maximising $\prod_{t=p+1}^{T} f_{Y_t|Y_{t-1},\ldots,Y_{t-p}}(\boldsymbol{\beta})$ instead. The value of $\boldsymbol{\beta}$ that maximised the (log) likelihood is called "conditional maximum likelihood estimate". This turns out to be the OLS estimate of c_0, $\phi_{0;1}$, ..., $\phi_{0;p}$ in the corresponding regression model.

MA(1)

$$Y_t = \mu_0 + \varepsilon_t + \theta_0 \varepsilon_{t-1}, \; |\theta_0| < 1, \; \varepsilon_t \sim Nid(0, \sigma_0^2)$$

Under the additional assumption that

$$\varepsilon_0 = 0$$

we can also derive a "conditional maximum likelihood estimate" of θ in a MA(1).

In general, since $\varepsilon_t \sim Nid(0, \sigma_0^2)$, then

$$Y_t | \varepsilon_{t-1} \sim N(\mu_0 + \theta_0 \varepsilon_{t-1}, \sigma_0^2)$$

i.e. the density of $Y_t | \varepsilon_{t-1}$ is

$$f_{Y_t|\varepsilon_{t-1}}(y_t | \varepsilon_{t-1}) =$$

$$= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2} \frac{(y_t - \mu_0 - \theta_0 \varepsilon_{t-1})^2}{\sigma_0^2} \right)$$

$$= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2} \frac{\varepsilon_t^2}{\sigma_0^2} \right)$$

Unfortunately ε_{t-1} is not observable.

However, suppose that we know ε_0, then $Y_1 = \mu_0 + \varepsilon_1 + \theta_0\varepsilon_0$, and, given μ_0 and θ_0 we can also compute

$$\varepsilon_1 = y_1 - \mu_0 - \theta_0\varepsilon_0$$

Having computed ε_1 we can also compute $\varepsilon_2 = y_2 - \mu_0 - \theta_0\varepsilon_1$, and, iterating the procedure,

$$\varepsilon_t = y_t - \mu_0 - \theta_0\varepsilon_{t-1}.$$

Then,

$$f_{Y_t|\varepsilon_{t-1}}(y_t|\varepsilon_{t-1}) = f_{Y_t|Y_{t-1},\ldots,Y_1,\varepsilon_0}(y_t|y_{t-1},\ldots,y_1,\varepsilon_0)$$

and

$$\begin{aligned} & f_{Y_1,\ldots,Y_{T-1},Y_T|\varepsilon_0}(y_1,\ldots,y_{T-1},y_T|\varepsilon_0) \\ = & f_{Y_1|\varepsilon_0}(y_1|\varepsilon_0) \\ & \times \prod_{t=2}^{T} f_{Y_t|Y_{t-1},\ldots,Y_1,\varepsilon_0}(y_t|y_{t-1},\ldots,y_1,\varepsilon_0) \\ = & (2\pi)^{-T/2}(\sigma_0^2)^{-T/2} \prod_{t=1}^{T} \exp\left(-\frac{\varepsilon_t^2}{2\sigma_0^2}\right). \end{aligned}$$

Notice that this is not the density of $(Y_1, \ldots, Y_{T-1,} Y_T)'$ when each Y_t has MA(1) representation, but that density (i.e., the density of $(Y_1, \ldots, Y_{T-1}, Y_T)'$ when each Y_t has MA(1) representation) conditional on ε_0.

Morevoer, we cannot compute a likelihood, because we can't observe ε_0.

However, consider the process

$$\begin{aligned} Y_t &= \mu_0 + \varepsilon_t + \theta_0 \varepsilon_{t-1}, \\ \text{with } \varepsilon_t &\sim Nid(0, \sigma_0^2) \text{ when } t > 0; \\ \varepsilon_0 &= 0. \end{aligned}$$

This process is very similar to the stationary MA(1), and it has the density above (setting $\varepsilon_0 = 0$).

Given that we know ε_0, we can compute ε_1 for a given point $\boldsymbol{\beta}$ in the parameter space: this is of course a function of $\boldsymbol{\beta}$, so

$$\varepsilon_1(\boldsymbol{\beta}) = y_1 - \mu$$

and then

$$\varepsilon_2(\boldsymbol{\beta}) = y_1 - \mu - \theta \varepsilon_1(\boldsymbol{\beta})$$

and in general

$$\varepsilon_t(\boldsymbol{\beta}) = y_t - \mu - \theta \varepsilon_{t-1}(\boldsymbol{\beta}).$$

We can then compute the likelihood (which is, then, a "conditional likelihood") as a function of a set of observations $(y_1, \dots, y_T)'$, and of a generic vector of unknown parameters $\boldsymbol{\beta}$,

$$\begin{aligned} & f_{Y_1,\dots,Y_T|\varepsilon_0=0}(\boldsymbol{\beta}) \\ & = f_{Y_1|\varepsilon_0=0}(\boldsymbol{\beta}) \prod_{t=2}^{T} f_{Y_t|Y_{t-1},\dots,Y_1,\varepsilon_0=0}(\boldsymbol{\beta}) \\ & = (2\pi)^{-T/2} |\sigma^2|^{-T/2} \prod_{t=1}^{T} \exp\left(-\frac{1}{2} \frac{\varepsilon_t^2(\boldsymbol{\beta})}{\sigma^2} \right). \end{aligned}$$

Taking logs, the (conditional) log-likelihood is

$$\mathcal{L}(\boldsymbol{\beta}) = -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^{T} \varepsilon_t^2(\boldsymbol{\beta})$$

The value of $\boldsymbol{\beta}$ that maximises the (conditional) (log) likelihood is called "conditional maximum likelihood estimate".

MA(q)

$$Y_t = \mu_0 + \varepsilon_t + \theta_{0;1}\varepsilon_{t-1} + \ldots + \theta_{0;q}\varepsilon_{t-q},\ \varepsilon_t \sim Nid(0, \sigma_0^2)$$

and the roots of $1 + \theta_{0;1}z + \ldots + \theta_{0;q}z^q = 0$ are all outside the unit circle.

Introduce $\boldsymbol{\varepsilon}_0 = (\varepsilon_0, \ldots, \varepsilon_{-q+1})'$.

Again, if $\boldsymbol{\varepsilon}_0 = \mathbf{0}$, we compute, for $\boldsymbol{\beta} = (\mu, \theta_1, \ldots, \theta_q, \sigma^2)'$,

$$\varepsilon_t(\boldsymbol{\beta}) = y_t - \mu - \theta_1\varepsilon_{t-1}(\boldsymbol{\beta}) - \ldots - \theta_q\varepsilon_{t-q}(\boldsymbol{\beta})$$

iteratively, and we can formulate a "conditional maximum likelihood":

$$\begin{aligned} & f_{Y_1,\ldots,Y_T|\boldsymbol{\varepsilon}_0=\mathbf{0}}(\boldsymbol{\beta}) \\ &= f_{Y_1|\boldsymbol{\varepsilon}_0=0}(\boldsymbol{\beta}) \prod_{t=2}^{T} f_{Y_t|Y_{t-1},\ldots,Y_1,\boldsymbol{\varepsilon}_0=\mathbf{0}}(\boldsymbol{\beta}) \\ &= (2\pi)^{-T/2}|\sigma^2|^{-T/2} \prod_{t=1}^{T} \exp\left(-\frac{1}{2}\frac{\varepsilon_t^2(\boldsymbol{\beta})}{\sigma^2}\right) \end{aligned}$$

Taking logarithms, the log-likelihood is

$$\mathcal{L}(\boldsymbol{\beta}) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t^2(\boldsymbol{\beta})$$

The value of $\boldsymbol{\beta}$ that maximises the conditional (log) likelihood is the "conditional maximum likelihood estimate".

ARMA(p, q)

$$\begin{aligned} Y_t &= c_0 + \phi_{0;1} Y_{t-1} + \ldots + \phi_{0;p} Y_{t-p} \\ &\quad + \varepsilon_t + \theta_{0;1}\varepsilon_{t-1} + \ldots + \theta_{0;q}\varepsilon_{t-q}, \\ \varepsilon_t &\sim Nid(0, \sigma_0^2) \end{aligned}$$

the roots of $1 - \phi_{0;1} z - \ldots - \phi_{0;p} z^p = 0$ and of $1 + \theta_{0;1} z + \ldots + \theta_{0;q} z^q = 0$ are all outside the unit circle, and there is no common factor.

Again, assume that $Y_1, \ldots, Y_p$ are given and $\varepsilon_p = \varepsilon_{p-1} = \ldots = \varepsilon_{p-q+1} = 0$.

Then we can compute, for $\boldsymbol{\beta} = (c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)'$

$$\begin{aligned} \varepsilon_t(\boldsymbol{\beta}) &= y_t - c - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p} \\ &\quad - \theta_1 \varepsilon_{t-1}(\boldsymbol{\beta}) - \ldots - \theta_q \varepsilon_{t-q}(\boldsymbol{\beta}) \end{aligned}$$

for $t > p$.

The conditional likelihood is then

$$\begin{aligned} & f_{Y_{p+1},\ldots,Y_T | Y_p,\ldots,Y_1,\varepsilon_p=0,\ldots,\varepsilon_{p-q+1}=0}(\boldsymbol{\beta}) \\ &= (2\pi)^{-(T-p)/2} |\sigma^2|^{-(T-p)/2} \prod_{t=p+1}^{T} \exp\left(-\frac{1}{2}\frac{\varepsilon_t^2(\boldsymbol{\beta})}{\sigma^2}\right) \end{aligned}$$

The conditional log-likelihood is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) &= -\frac{T-p}{2}\ln(2\pi) \\ &- \frac{T-p}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=p+1}^{T}\varepsilon_t^2(\boldsymbol{\beta})\end{aligned}$$

The value of $\boldsymbol{\beta}$ that maximises the conditional (log) likelihood is called "conditional maximum likelihood estimate".

Concentrated likelihood

Consider again

$$\mathcal{L}(\boldsymbol{\beta}) = -\frac{T-p}{2}\ln(2\pi)$$

$$-\frac{T-p}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=p+1}^{T}\varepsilon_t^2(\boldsymbol{\beta})$$

Clearly, this is maximised as long as $\sum_{t=p+1}^{T}\varepsilon_t^2(\boldsymbol{\beta})$ is minimised with respect to $(c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$: although we wrote $\varepsilon_t(\boldsymbol{\beta})$, the parameter σ^2 does not actually enter the recursions to compute $\varepsilon_t(\boldsymbol{\beta})$.

If we are not interested in σ_0^2, we can estimate $(c_0, \phi_{0;1}, \ldots, \phi_{0;p}, \theta_{0;1}, \ldots, \theta_{0;q})$ by minimising the (conditional) Residual Sum of Squares (RSS) $\sum_{t=p+1}^{T}\varepsilon_t^2(\boldsymbol{\beta})$, i.e.

$$\widehat{c}, \widehat{\phi}_1, \ldots, \widehat{\phi}_p, \widehat{\theta}_1, \ldots, \widehat{\theta}_q = \underset{c,\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q}{\arg\min}\sum_{t=p+1}^{T}\varepsilon_t^2(\boldsymbol{\beta})$$

To estimate σ_0^2 notice that

$$\frac{\partial\mathcal{L}(\boldsymbol{\beta})}{\partial\sigma^2} = -\frac{T-p}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{t=p+1}^{T}\varepsilon_t^2(\boldsymbol{\beta})$$

so, from the first order condition $\frac{\partial\mathcal{L}(\boldsymbol{\beta})}{\partial\sigma^2} = 0$,

$$\widehat{\sigma^2} = \frac{1}{T-p}\sum_{t=p+1}^{T}\varepsilon_t^2\left(\widehat{c}, \widehat{\phi}_1, \ldots, \widehat{\phi}_p, \widehat{\theta}_1, \ldots, \widehat{\theta}_q\right).$$

Pseudo Maximum Likelihood (PML)

When ε_t is not normally distributed, the density is different and then the maximum likelihood estimate is different as well.

If we use the gaussian density even if ε_t is not normally distributed, then, our estimate is no longer the maximum likelihood one. In this case it usually known as Pseudo (or Quasi) maximum likelihood instead.

We already saw that there many variants of the maximum likelihood estimates; in the same way, there are many variants of the PML ones. In what follows, we will discuss the asymptotic properties of a minRSS type of PML estimate, i.e.

$$\left(\widehat{c},\widehat{\phi}_1,\ldots,\widehat{\phi}_p,\widehat{\theta}_1,\ldots,\widehat{\theta}_q\right)_{PML} = \underset{c,\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q}{\arg\min} \sum_{t=p+1}^{T} \varepsilon_t^2(\boldsymbol{\beta})$$

$$\widehat{\sigma^2}_{PML} = \frac{1}{T-p} \sum_{t=p+1}^{T} \varepsilon_t^2\left(\widehat{c},\widehat{\phi}_1,\ldots,\widehat{\phi}_p,\widehat{\theta}_1,\ldots,\widehat{\theta}_q\right)_{PML}.$$

Notice here that this estimate "generalises" the ML one in the same way as the OLS regression estimate generalises the ML estimate in a regression model.

Optimisation of the objective function

In general, it is not always possible to obtain a closed form formula for the estimate, and it may be extremely time consuming to compute the log-likelihood function (even the conditional log-likelihood) for all the potential $\boldsymbol{\beta}$.

The optimisation of the log-likelihood may be carried using a numerical algorithm, such as the Newton-Raphson one.

Introduce

$$g(\boldsymbol{\beta}^{(0)}) = \left.\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(0)}} \text{ (gradient)}$$

$$H(\boldsymbol{\beta}^{(0)}) = -\left.\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(0)}} \text{ (Hessian)}$$

for a generic $\boldsymbol{\beta}^{(0)}$, and consider an approximate second order Taylor expansion of $\mathcal{L}(\boldsymbol{\beta})$,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) \approx \mathcal{L}(\boldsymbol{\beta}^{(0)}) &+ \left[g(\boldsymbol{\beta}^{(0)})'\right][\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}] \\ &- \frac{1}{2}[\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}]' H(\boldsymbol{\beta}^{(0)})[\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}]\end{aligned}$$

Recall that $\mathscr{L}(\boldsymbol{\beta})$ is maximised at $\widehat{\boldsymbol{\beta}}$ if

$$\left.\frac{\partial \mathscr{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = 0.$$

Now, consider the approximation of the derivative around $\boldsymbol{\beta}^{(0)}$:

$$\frac{\partial \mathscr{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \approx [g(\boldsymbol{\beta}^{(0)})] - H(\boldsymbol{\beta}^{(0)})[\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}].$$

If the approximation was perfect, we could have just computed $\widehat{\boldsymbol{\beta}}$ solving for $\boldsymbol{\beta}$

$$[g(\boldsymbol{\beta}^{(0)})] - H(\boldsymbol{\beta}^{(0)})[\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}] = 0,$$

i.e.,

$$\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)} + H(\boldsymbol{\beta}^{(0)})^{-1}[g(\boldsymbol{\beta}^{(0)})].$$

However, this may be a rather poor estimate, because the approximation is not exact (there is a remainder, in this case of the third order, in the Taylor expansion of $\mathscr{L}(\boldsymbol{\beta})$). Let’s call this possibly poor estimate $\boldsymbol{\beta}^{(1)}$, then, where

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + H(\boldsymbol{\beta}^{(0)})^{-1}[g(\boldsymbol{\beta}^{(0)})] :$$

clearly, this is (in a certain probabilistic sense) better than a generic $\boldsymbol{\beta}^{(0)}$.

Next, we can improve, by considering a second order approximation of $\mathcal{L}(\boldsymbol{\beta})$ in $\boldsymbol{\beta}^{(1)}$, and compute

$$\boldsymbol{\beta}^{(2)} = \boldsymbol{\beta}^{(1)} + H(\boldsymbol{\beta}^{(1)})^{-1}[g(\boldsymbol{\beta}^{(1)})].$$

The procedure can then be iterated until convergence (which gives $\widehat{\boldsymbol{\beta}}$).

Example

ARMA(1, 1) (assuming $\mu_0 = 0$, $\sigma_0^2 = 1$ known), $\boldsymbol{\beta} = (\theta, \phi)'$. Recall

$$\varepsilon_t(\boldsymbol{\beta}) = y_t - \phi y_{t-1} - \theta \varepsilon_{t-1}(\boldsymbol{\beta})$$

so for $t \geq 2$,

$$\begin{aligned}
\frac{\partial \varepsilon_t(\boldsymbol{\beta})}{\partial \theta} &= -\varepsilon_{t-1}(\boldsymbol{\beta}) - \theta \frac{\partial \varepsilon_{t-1}(\boldsymbol{\beta})}{\partial \theta} \\
\frac{\partial \varepsilon_t(\boldsymbol{\beta})}{\partial \phi} &= -y_{t-1} - \theta \frac{\partial \varepsilon_{t-1}(\boldsymbol{\beta})}{\partial \phi}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \varepsilon_t(\boldsymbol{\beta})}{\partial \theta^2} &= -2 \frac{\partial \varepsilon_{t-1}(\boldsymbol{\beta})}{\partial \theta} - \theta \frac{\partial^2 \varepsilon_{t-1}(\boldsymbol{\beta})}{\partial \theta^2} \\
\frac{\partial^2 \varepsilon_t(\boldsymbol{\beta})}{\partial \phi^2} &= -\theta \frac{\partial^2 \varepsilon_{t-1}(\boldsymbol{\beta})}{\partial \phi^2} \\
\frac{\partial^2 \varepsilon_t(\boldsymbol{\beta})}{\partial \theta \partial \phi} &= -\frac{\partial \varepsilon_{t-1}(\boldsymbol{\beta})}{\partial \phi} - \theta \frac{\partial^2 \varepsilon_{t-1}(\boldsymbol{\beta})}{\partial \theta \partial \phi}
\end{aligned}$$

These iterations may be initialised setting $\varepsilon_1 = 0$, $\varepsilon_0 = 0$ (which also implies $\frac{\partial \varepsilon_1(\boldsymbol{\beta})}{\partial \theta} = 0$, $\frac{\partial \varepsilon_1(\boldsymbol{\beta})}{\partial \phi} = 0$), and taking $Y_1 = y_1$ as given. So,

$$\mathcal{L}(\boldsymbol{\beta}) = -\frac{T-1}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=2}^{T}\varepsilon_t^2(\boldsymbol{\beta})$$

$$g(\boldsymbol{\beta}) = \frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\sum_{t=2}^{T}\left\{\begin{array}{c} \varepsilon_t(\boldsymbol{\beta})\frac{\partial \varepsilon_t(\boldsymbol{\beta})}{\partial \theta} \\ \varepsilon_t(\boldsymbol{\beta})\frac{\partial \varepsilon_t(\boldsymbol{\beta})}{\partial \phi} \end{array}\right.$$

$$H(\boldsymbol{\beta}^{(0)}) = -\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = -\sum_{t=2}^{T}\left\{\begin{array}{cc} H_{11} & H_{12} \\ H_{12} & H_{22} \end{array}\right.$$

$$where\left\{\begin{array}{l} H_{11} = \left(\frac{\partial \varepsilon_t(\boldsymbol{\beta})}{\partial \theta}\right)^2 + \varepsilon_t(\boldsymbol{\beta})\frac{\partial^2 \varepsilon_t(\boldsymbol{\beta})}{\partial \theta^2} \\ H_{12} = \frac{\partial \varepsilon_t(\boldsymbol{\beta})}{\partial \phi}\frac{\partial \varepsilon_t(\boldsymbol{\beta})}{\partial \theta} + \varepsilon_t(\boldsymbol{\beta})\frac{\partial^2 \varepsilon_t(\boldsymbol{\beta})}{\partial \theta \partial \phi} \\ H_{22} = \left(\frac{\partial \varepsilon_t(\boldsymbol{\beta})}{\partial \phi}\right)^2 + \varepsilon_t(\boldsymbol{\beta})\frac{\partial^2 \varepsilon_t(\boldsymbol{\beta})}{\partial \phi^2} \end{array}\right.$$

In many cases, you may start the optimisation with any set of starting values, but this may result in a rather slow optimisation, or even in an "incorrect" solution (you may end up picking a local maximum, rather than the maximum).
It is then advisable to start from a "good" point, that is, from a consistent estimate of $\boldsymbol{\beta}$ (tipically, an estimate that you may compute easily, even if it is less efficient than maximum likelihood): the correlogram based estimate is a good starting point (given certain regularity conditions, properties as in the pseudo-maximum likelihood estimate may be obtained after just one step).

Appendix

- Maximum likelihood estimation for independent bernoulli trials
- Maximum likelihood estimation in the linear model with iid gaussian innovations
- Step-by-step derivation of the variancecovariance matrix for a stationary time series
- A numerical example of the computation of the LS /CRSS estimate for MA(1)

Maximum likelihoood estimation for independent bernoulli trials

Consider an experiment that may result in two outcomes, success or failure, with probability of success p, and suppose that we repeat the experiment n times. Then, letting X the number of successes, X is binomially distributed with parameters n and p,

$$P(X = x) = \frac{n!}{(n-x)!x!}p^x(1-p)^{n-x}$$

Suppose that we run the experiment 7 times. What is the probability of observing 5 successes? Setting

$$n = 7, X = 5,$$

we can compute $P(X = 5)$ for various values of p:

$$P(X = 5; p = 0.4) = \frac{7!}{2!5!}0.4^5 0.6^2 \approx 0.077$$

$$P(X = 5; p = 0.6) = \frac{7!}{2!5!}0.6^5 0.4^2 \approx 0.261$$

$$P(X = 5; p = 0.8) = \frac{7!}{2!5!}0.8^5 0.2^2 \approx 0.275$$

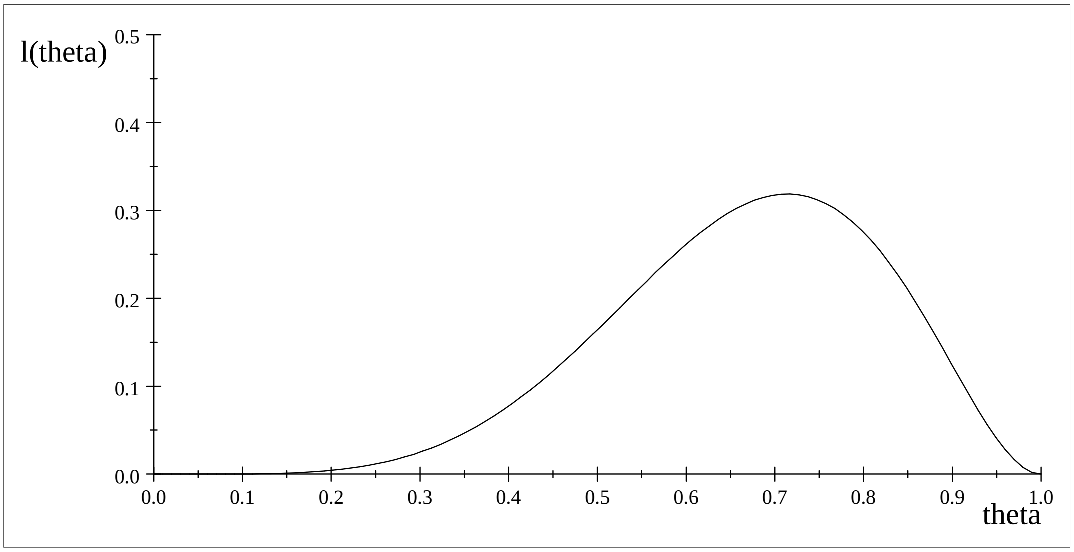
and, for the generic value $p = \theta$,

$$P(X = 5; p = \theta) = \frac{7!}{(7-5)!5!}\theta^5(1-\theta)^{7-5}$$

Suppose now that p is unknown, and that we did run the experiment ($n = 7$) and we observed $X = 5$. Then, n and X are fixed, and

$$l(\theta) = \frac{7!}{(7-5)!5!}\theta^5(1-\theta)^{7-5}$$

is a function of θ only, and it is called likelihood function.



As we have seen before, $P(X = 5; p = 0.8) \approx 0.275$ while $P(X = 5; p = 0.4) \approx 0.077$, so the probability that $X = 5$ occurrs when $p = 0.8$ is higher than the probability that $X = 5$ occurrs when $p = 0.4$. As we do not know p but we observed $X = 5$, the idea behind maximum likelihood is that, therefore, I should rather think that $p = 0.8$ generated this result of $X = 5$, and not $p = 0.4$.

As p may actually be any value in $(0, 1)$, by the same argument the estimate is $\widehat{\theta} = 5/7 \approx 0.714$.

Maximum likelihood estimation in the linear model with iid gaussian innovations

We are interested in estimating α_0 and β_0 in

$$Y_t = \alpha_0 + \beta_0 X_t + u_t$$

assuming that

1) X_t is deterministic

2) u_t is normally, independently distributed with $E(u_t) = 0$, $Var(u_t) = \sigma_0^2$, i.e. u_t is $\mathrm{Nid}(0, \sigma_0^2)$.

(notice: X_t is deterministic means for example that X_t may be some known function of time, such as $X_t = t$, or $X_t = t^2$ or $X_t = \cos(t)$; $X_t = c$ constant is also a deterministic function, but the constant is already in α_0).

Then,

$$Y_t \sim N(\alpha_0 + \beta_0 x_t, \sigma_0^2)$$

with density

$$\begin{aligned} f_{Y_t}(y_t) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2}\left(\frac{y_t - \alpha_0 - \beta_0 x_t}{\sigma_0}\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2}\left(\frac{u_t}{\sigma_0}\right)^2\right) \end{aligned}$$

Morever, the density of $Y_1, \ldots, Y_T$ is

$$f_{Y_1,\ldots,Y_T}(y_1, \ldots, y_T) = \prod_{t=1}^{T} f_{Y_t}(y_t)$$

because u_t is independently distributed.

Note: the density $f_{Y_t}(y_t)$ is a function of $y_t \in \mathbb{R}$, and it is computed for given (known) value of the parameters $\alpha_0, \beta_0, \sigma_0^2$; $f_{Y_1,..,Y_T}(y_1, .., y_T)$ is a function of $y_1, .., y_T$, and it is computed for the same given (known) value of the parameters $\alpha_0, \beta_0, \sigma_0^2$.

We cannot observe u_t but, using observations $(y_1, \ldots, y_T)'$ (and recall that X_t is deternisitic, so x_t it is known) we can compute

$$u_t(\alpha, \beta) = y_t - \alpha - \beta x_t$$

for each admissible value of α and β, so we compute

$$f_{Y_t}(\alpha, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{u_t(\alpha, \beta)}{\sigma}\right)^2\right)$$

and the likelihood

$$f_{Y_1,\ldots,Y_T}(\alpha, \beta, \sigma^2) = \prod_{t=1}^{T} f_{Y_t}(\alpha, \beta, \sigma^2)$$

Note: $u_t(\alpha, \beta)$ is a function of α and β; it is not a function of y_t or x_t because these are known at the moment we compute $u_t(\alpha, \beta)$. The likelihood $f_{Y_1,..,Y_T}(\alpha, \beta, \sigma^2)$ is a function of the parameters $(\alpha, \beta, \sigma^2)$ and it is computed for the (given) value of the observations $y_1, .., y_T$ (and using the fact that $x_1, \ldots, x_T$ are known). We indicate the likelihood as $f_{Y_1,..,Y_T}(\alpha, \beta, \sigma^2)$ (without reference to $y_1, .., y_T$ in the argument).

The logarithm of the likelihood is

$$\mathcal{L}(\alpha,\beta,\sigma^2) = -\frac{T}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T} u_t(\alpha,\beta)^2.$$

The maximum likelihood estimates of $\alpha_0, \beta_0, \sigma_0^2$ are the values of α, β, σ^2 that maximised the likelihood (and therefore, the log-likelihood), i.e.

$$\widehat{\alpha}, \widehat{\beta}, \widehat{\sigma}^2 = \underset{\alpha,\beta,\sigma^2}{\arg\max}\ \mathcal{L}(\alpha,\beta,\sigma^2)$$

For the maximum likelihood estimation of α, β, first order conditions are $\frac{\partial \mathcal{L}(\alpha,\beta,\sigma^2)}{\partial \alpha} = 0$ and $\frac{\partial \mathcal{L}(\alpha,\beta,\sigma^2)}{\partial \beta} = 0$, so

$$\frac{\partial u_t(\alpha,\beta)}{\partial \alpha} = -1, \ \frac{\partial u_t(\alpha,\beta)}{\partial \beta} = -x_t$$

$$\frac{\partial u_t^2(\alpha,\beta)}{\partial \alpha} = -2u_t(\alpha,\beta), \ \frac{\partial u_t^2(\alpha,\beta)}{\partial \beta} = -2x_t u_t(\alpha,\beta),$$

$$\frac{\partial \mathcal{L}(\alpha,\beta,\sigma^2)}{\partial \alpha} = \frac{1}{2\sigma^2} 2 \sum_{t=1}^{T} u_t(\alpha,\beta)$$

$$\frac{\partial \mathcal{L}(\alpha,\beta,\sigma^2)}{\partial \beta} = \frac{1}{2\sigma^2} 2 \sum_{t=1}^{T} x_t u_t(\alpha,\beta).$$

Equating these two derivatives to 0, and substituting $u_t(\alpha,\beta)$, we have

$$\sum_{t=1}^{T} \left(y_t - \widehat{\alpha} - \widehat{\beta} x_t \right) = 0$$

$$\sum_{t=1}^{T} x_t \left(y_t - \widehat{\alpha} - \widehat{\beta} x_t \right) = 0.$$

Solving the first equation for $\widehat{\alpha}$, we have

$$T\widehat{\alpha} = \sum_{t=1}^{T}\left(y_t - \widehat{\beta}x_t\right)$$

and finally, letting $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$, $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}.$$

Replacing this in the second equation,

$$\sum_{t=1}^{T} x_t\left(y_t - \bar{y} + \widehat{\beta}\bar{x} - \widehat{\beta}x_t\right) = 0$$

$$\sum_{t=1}^{T} x_t(y_t - \bar{y}) = \widehat{\beta}\sum_{t=1}^{T} x_t(x_t - \bar{x})$$

$$\widehat{\beta} = \frac{\sum_{t=1}^{T} x_t(y_t - \bar{y})}{\sum_{t=1}^{T} x_t(x_t - \bar{x})}$$

Notice that $\sum_{t=1}^{T}(y_t - \bar{y}) = 0$, so $\sum_{t=1}^{T}\bar{x}(y_t - \bar{y}) = 0$.

In the same way, $\sum_{t=1}^{T}\bar{x}(x_t - \bar{x}) = 0$, so

$$\widehat{\beta} = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

Notes:

★ if $\alpha_0 = 0$ and we know it, i.e. we want to estimate β_0 in

$$Y_t = \beta_0 X_t + u_t$$

then

$$\widehat{\beta} = \frac{\sum_{t=1}^{T} x_t y_t}{\sum_{t=1}^{T} {x_t}^2}.$$

★ if we have many types of x_t (for example, p) i.e. we have $x_{1,t}$, $x_{2,t}$, ..., $x_{p,t}$ and we want to estimate α_0 and $\beta_{0;1}$ $\beta_{0;2}$, ..., $\beta_{0;p}$ in

$$Y_t = \alpha_0 + \beta_{0;1} X_{1;t} + +\beta_{0;2} X_{2;t} + \ldots + \beta_{0;p} X_{p;t} + u_t$$

then, stacking

$$\boldsymbol{\beta}_0 = (\alpha_0, \beta_{0;1}, \beta_{0;2}, \ldots, \beta_{0;p})'$$

(this is a $((p+1) \times 1)$ vector), and

$$\begin{aligned} y &= (y_1, \ldots, y_T)' && (T \times 1) \\ x_t &= (1, x_{1;t}, \ldots, x_{p;t})' && ((p+1) \times 1) \\ x &= (x_1, \ldots, x_T)' && (T \times (p+1)) \end{aligned}$$

then

$$\widehat{\boldsymbol{\beta}} = (x'x)^{-1}(x'y)$$

★ $\widehat{\boldsymbol{\beta}} = (x'x)^{-1}(x'y)$ is a function of the observations. As we change our observations, we also change the value of $\widehat{\boldsymbol{\beta}}$. As the observations are realizations of the random variables, we can also consider the function

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}(X'Y).$$

This is a random variable and it has a proper distribution. In particular, it is possible to prove that, if there is M such that $\frac{1}{T}X'X \rightarrow M$, then

$$\widehat{\boldsymbol{\beta}} \rightarrow_p \boldsymbol{\beta}$$
$$\sqrt{T}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \rightarrow_d N(0, M^{-1}\sigma^2)$$

Moreover, in this particular case (X_t deterministic and u_t $Nid(0, \sigma^2)$) $\widehat{\boldsymbol{\beta}}$ is the "best" estimator in a certain class

(the class of linear and unbiased estimators; "best" here means "minimum variance" if only one variable is used in X_t; in general, it means that the difference between $M^{-1}\sigma^2$ and the variance-covariance matrix of any other linear unbiased estimator, is negative semidefinite).

★ The estimator $\widehat{\boldsymbol{\beta}} = (X'X)^{-1}(X'Y)$ is also known as the Ordinary Least Squares estimator, because it is obtained minimising the sum of squares of the residuals $\sum_{t=1}^{T} u_t(\alpha, \beta)^2$. It is also used in many other contexts, in which X_t may be not deterministic and u_t is not normally distributed, or indeed not even independently distributed. In all these cases of course the properties are not as stated above, and they must be studied case by case.

Step-by-step derivation of the variance-covariance matrix for a stationary time series

Let $\{Y_t\}_{t=-\infty}^{\infty}$ be a process and each Y_t is identically distributed and

$$\mathbf{Y} = (Y_1, \ldots, Y_T)'$$

with

$$E(\mathbf{Y}) = \boldsymbol{\mu}, E\Big((\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})'\Big) = \boldsymbol{\Omega}$$

How are **Y**, **μ** and **Ω** done?

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ \cdots \\ Y_{T-1} \\ Y_T \end{pmatrix} \boldsymbol{\mu} = \begin{pmatrix} \mu \\ \mu \\ \cdots \\ \cdots \\ \mu \\ \mu \end{pmatrix} \mathbf{Y} - \boldsymbol{\mu} = \begin{pmatrix} Y_1 - \mu \\ Y_2 - \mu \\ \cdots \\ \cdots \\ Y_{T-1} - \mu \\ Y_T - \mu \end{pmatrix}$$

So, **Y**, **μ** and **Y** – **μ** are $T \times 1$ vectors.

$$(\mathbf{Y}-\boldsymbol{\mu})(\mathbf{Y}-\boldsymbol{\mu})'$$

$$= \begin{pmatrix} Y_1 - \mu \\ Y_2 - \mu \\ \dots \\ Y_T - \mu \end{pmatrix} \begin{pmatrix} Y_1 - \mu & Y_2 - \mu & \dots & Y_T - \mu \end{pmatrix}$$

$(\mathbf{Y}-\boldsymbol{\mu})(\mathbf{Y}-\boldsymbol{\mu})'$ is a $(T \times 1) \times (1 \times T) = (T \times T)$ matrix.

To ease the notation, assume that

$$\mu = 0$$

Then we are interested in

$$\mathbf{Y}\,\mathbf{Y}' = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \dots \\ Y_T \end{pmatrix} \begin{pmatrix} Y_1 & Y_2 & Y_3 & \dots & Y_T \end{pmatrix}$$

$$= \begin{pmatrix} Y_1\,Y_1 & Y_1\,Y_2 & Y_1\,Y_3 & \dots & Y_1\,Y_T \\ Y_2\,Y_1 & Y_2\,Y_2 & Y_2\,Y_3 & \dots & Y_2\,Y_T \\ Y_3\,Y_1 & Y_3\,Y_2 & Y_3\,Y_3 & \dots & Y_3\,Y_T \\ \dots & \dots & \dots & \dots & \dots \\ Y_T\,Y_1 & Y_T\,Y_2 & Y_T\,Y_3 & \dots & Y_T\,Y_T \end{pmatrix}$$

Recall that $\mu = 0$, then
$E\left(Y_1\, Y_1\right) = \gamma_0$,
$E\left(Y_1\, Y_2\right) = \gamma_1$, ...,
$E\left(Y_1\, Y_T\right) = \gamma_{T-1}$;

in the same way, $E\left(Y_2\, Y_1\right) = \gamma_1$,
$E\left(Y_2\, Y_2\right) = \gamma_0$, ...,
$E\left(Y_2\, Y_T\right) = \gamma_{T-2}$, ... so

$$
\begin{aligned}
& E\left(\mathbf{Y}\,\mathbf{Y}'\right) \\
&= E\begin{pmatrix} Y_1\, Y_1 & Y_1\, Y_2 & Y_1\, Y_3 & \ldots & Y_1\, Y_T \\ Y_2\, Y_1 & Y_2\, Y_2 & Y_2\, Y_3 & \ldots & Y_2\, Y_T \\ Y_3\, Y_1 & Y_3\, Y_2 & Y_3\, Y_3 & \ldots & Y_3\, Y_T \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ Y_T\, Y_1 & Y_T\, Y_2 & Y_T\, Y_3 & \ldots & Y_T\, Y_T \end{pmatrix} \\
&= \begin{pmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \ldots & \gamma_{T-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \ldots & \gamma_{T-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \ldots & \gamma_{T-3} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \gamma_{T-1} & \gamma_{T-2} & \gamma_{T-3} & \ldots & \gamma_0 \end{pmatrix}
\end{aligned}
$$

Assume for example that $\{Y_t\}_{t=-\infty}^{\infty}$ is a MA(1),

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1} \text{ with } \varepsilon_t \text{ iid}(0, \sigma^2)$$

Then, $\mathbf{\Omega}$ is the $T \times T$ matrix

$$\mathbf{\Omega} = \sigma^2 \begin{pmatrix} (1+\theta^2) & \theta & 0 & \dots & 0 \\ \theta & (1+\theta^2) & \theta & \dots & 0 \\ 0 & \theta & (1+\theta^2) & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & (1+\theta^2) \end{pmatrix}$$

$$= \sigma^2(1+\theta^2) \begin{pmatrix} 1 & \frac{\theta}{(1+\theta^2)} & 0 & \dots & 0 \\ \frac{\theta}{(1+\theta^2)} & 1 & \frac{\theta}{(1+\theta^2)} & \dots & 0 \\ 0 & \frac{\theta}{(1+\theta^2)} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

Conditional ML for MA(1)
a numerical example

We observed

$$y_1 = -0.4, \quad y_2 = 0.8, \quad y_3 = 0.6, \quad y_4 = -0.2$$

and we want to compute the Conditional RSS for three values of θ: -0.5, 0, 0.5.
Using $\varepsilon_0 = 0$ for any θ, $\varepsilon_t(\theta) = y_t - \varepsilon_{t-1}(\theta)$

$\varepsilon_t(\theta)$	$t=1$	$t=2$
$\theta = 1/2$	$-0.4 - 1/2 * 0 = -0.4$	$0.8 - 1/2 * (-0.4) = 1.0$
$\theta = 0$	$-0.4 - 0 * 0 = -0.4$	$0.8 - 0 * (-0.4) = 0.8$
$\theta = -1/2$	$-0.4 + 1/2 * 0 = -0.4$	$0.8 + 1/2 * (-0.4) = 0.6$

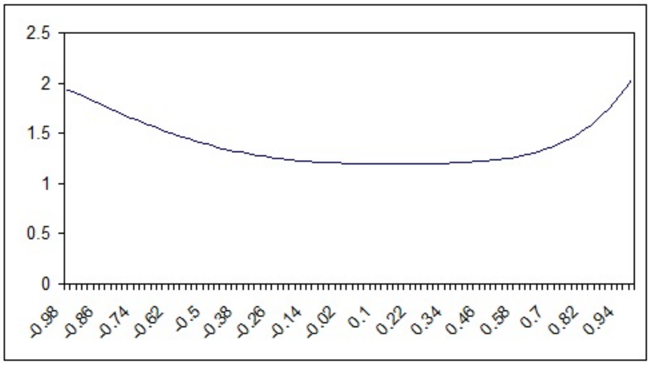
$\varepsilon_t(\theta)$	$t=3$	$t=4$
$\theta = 1/2$	$0.6 - 1/2 * 1 = 0.1$	$-0.2 - 1/2 * 0.1 = -0.25$
$\theta = 0$	$0.6 - 0 * 0.8 = 0.6$	$-0.2 - 0 * 0.6 = -0.2$
$\theta = -1/2$	$0.6 + 1/2 * 0.6 = 0.9$	$-0.2 + 1/2 * 0.9 = 0.25$

$\varepsilon_t^2(\theta)$	$t=1$	$t=2$	$t=3$	$t=4$
$\theta = 1/2$	$(-0.4)^2 = 0.16$	$1^2 = 1$	$0.1^2 = 0.01$	$(-0.25)^2 = 0.062$
$\theta = 0$	$(-0.4)^2 = 0.16$	$0.8^2 = 0.64$	$0.6^2 = 0.36$	$(-0.2)^2 = 0.04$
$\theta = -1/2$	$(-0.4)^2 = 0.16$	$0.6^2 = 0.36$	$0.9^2 = 0.81$	$0.25^2 = 0.0625$

	$\sum_{t=1}^{T} \varepsilon_t^2(\theta)$
$\theta = 1/2$	$0.16 + 1 + 0.01 + 0.0625 = 1.2325$
$\theta = 0$	$0.16 + 0.64 + 0.36 + 0.04 = 1.2$
$\theta = -1/2$	$0.16 + 0.36 + 0.81 + 0.0625 = 1.3925$

This means that if we were to pick a conditional maximum likelihood estimate $\widehat{\theta}$ between the three candidates $1/2, 0, -1/2$, we would pick $\widehat{\theta} = 0$.

If we used the whole $[-0.98, 0.98]$ the estimate $\widehat{\theta}$ would be 0.14. The function $\sum_{t=1}^{T} \varepsilon_t^2 (\theta)$ is


2.5
2
1.5
1
0.5
0
-0.98
-0.86
-0.74
-0.62
-0.5
-0.38
-0.26
-0.14
-0.02
0.1
0.22
0.34
0.46
0.58
0.7
0.82
0.94