Maximum Likelihood Estimation

So far we have assumed knowledge of the true population parameters:

$\vec{\theta} = (c, \phi_1, \phi_2, \ldots, \phi_p, \theta_1, \ldots, \theta_q \sigma^2)'$

- In reality, the true parameters are unknown.
- We will estimate them with maximum likelihood

Maximum likelihood has two steps:

1. Express the likelihood
   - This is the joint probability density function

     $f_{\overline{y}_T, \ldots, \overline{y}_1}(y_T, \ldots, y_1|\vec{\theta})$ where $(y_1, \ldots, y_T)' = \vec{\overline{y}}_T$ is a realization of $\vec{y}_T = (\overline{y}_1, \ldots, \overline{y}_T)$
   - $f_{\vec{\overline{y}}_T}$ can be viewed (loosely) as the joint probability of observing $\vec{y}_T$ given $\vec{\theta}$.
   - Alternatively,

     $L(\vec{\theta}|y_1, \ldots, y_T) = f_{\overline{y}_T, \ldots, \overline{y}_1}(y_T, \ldots, y_1|\vec{\theta})$ is viewed as the "likelihood" of $\vec{\theta}$ given $\vec{y}_T$.
2. Finding the value of $\vec{\theta}$ that maximizes $L(\vec{\theta}|\vec{y}_T)$.

   Expressing the joint density/likelihood requires making a distributional assumption about ε_t.
   - Typically, we'll assume $\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$.

Exact Likelihood of AR(1):

Recall a Gaussian AR(1) process is:

$Y_t = c + \phi Y_{t-1} + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$.

- In this case $\hat{\theta} = (c, \phi, \sigma^2)$
- Suppose $|\phi| < 1$ (stationary and causal)

We know that the first and second unconditional moments are:

$E[Y_t] = \dfrac{c}{1-\phi}, \quad Var(Y_t) = \dfrac{\sigma^2}{1-\phi^2}, \ \forall t.$

Thus, $Y_1 \sim N\left(\dfrac{c}{1-\phi}, \dfrac{\sigma^2}{1-\phi^2}\right)$, or

$$f_{Y_1}(y_1|\vec{\theta}) = \frac{1}{\sqrt{2\pi \dfrac{\sigma^2}{1-\phi^2}}} \cdot exp\left\{-\frac{1}{2}\frac{y_1 - \frac{c}{1-\phi}}{\frac{\sigma^2}{1-\phi^2}}\right\}$$

Now, conditional of $Y_1 = y_1$,

$Y_2 = c + \phi y_1 + \varepsilon_2$

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

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

The same is true for all $t > 1$

$f_{Y_t|Y_{t-1}=y_t}(y_2|y_{t-1},\vec{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp\left\{-\frac{1}{2}\frac{(y_t - c - \phi y_{t-1})}{\sigma^2}\right\}$

The joint density is the product of marginal + conditionals:

$f_{\vec{Y}}(\vec{y}|\vec{\theta}) = f_{Y_t|Y_{t-1}}(y_t|y_{t-1},\vec{\theta})f_{Y_{t-1}|Y_{t-2}}(y_{t-1}|y_{t-2},\vec{\theta})\cdots f_{Y_2|Y_1}(y_2|y_1,\vec{\theta})f_{Y_1}(y_1|\vec{\theta})$

This is analogous to:

$P(A\bigcap B\bigcap C) = P(A|BC)P(B|C)P(C)$

<u>Bayes Rule</u>

$P(A|B) = \frac{P(A\bigcap B)}{P(B)}$

$P(A\bigcap B) = P(A|B)P(B)$

$$\begin{aligned}
P(\underbrace{A\bigcap B}_{\text{Bob}}\bigcap C) &= P(Bob\bigcap C) \\
&= P(Bob|C)P(C) \\
&= P(A\bigcap B)|C)P(C) \\
&= P(A|B,C)P(B|C)P(C)
\end{aligned}$$

$f_{\vec{Y}}(\vec{y}|\vec{\theta}) = f_{Y_1}(y_1|\vec{\theta})\prod_{t=2}^{T} f_{Y_t|Y_{t-1}}(y_t|y_{t-1},\vec{\theta}) = L(\vec{\theta}|\vec{y})$

The log likelihood is:

$$\begin{aligned}
l(\vec{\theta}|\vec{y}) &= log(L(\vec{\theta}|\vec{y}) = log(f_{Y_1}(y_1|\vec{\theta})) + \sum_{t=2}^{T} log(f_{Y_t|Y_{t-1}}(y_t|y_{t-1},\vec{\theta})) \\
&= -\frac{1}{2}log(2\pi) - \frac{1}{2}log\left(\frac{\sigma^2}{1-\phi^2}\right) - \frac{1}{2}\frac{\left(y_1 - \frac{c}{1-\phi}\right)^2}{\frac{\sigma^2}{1-\phi^2}} + \sum_{t=2}^{T}\left(-\frac{1}{2}log(2\pi) - \frac{1}{2}log(\sigma^2) - \frac{1}{2}\frac{(y_t - c - \phi y_{t-1})^2}{\sigma^2}\right) \\
&= -\frac{T}{2}log(2\pi) - \frac{T}{2}log(\sigma^2) + \frac{1}{2}(1-\phi^2) - \frac{1}{2}\frac{\left(y_1 - \frac{c}{1-\phi}\right)^2}{\frac{\sigma^2}{1-\phi^2}} - \frac{1}{2}\sum_{t=2}^{T}\frac{(y_t - c - \phi y_{t-1})^2}{\sigma^2}
\end{aligned}$$

Expressing in vector form - we can view $\vec{y}' = (y_1, \ldots, y_T)'$ as a single observation of a vector RV $\vec{Y}' = (Y_1, \ldots, Y_T)'$ with Gaussian errors $Y \sim MVN(\vec{\mu}, \Omega)$

$$\vec{\mu} = \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix}_{Tx1} \qquad \Omega = \begin{bmatrix} \gamma_0 & \cdots & \gamma_{T-1} \\ \vdots & \ddots & \vdots \\ \gamma_{T-1} & \cdots & \gamma_0 \end{bmatrix} = \frac{\sigma^2}{1-\phi^2}\begin{bmatrix} 1 & \phi & \cdots & \phi^{T-1} \\ \vdots & \ddots & \ddots & \vdots \\ \phi^{T-1} & \cdots & \cdots & 1 \end{bmatrix}$$

Thus, the joint density of $\vec{Y}$ is:

$f_{\vec{Y}}(\vec{y}|\vec{\theta}) = (2\pi)^{-T/2}|\Omega|^{1/2}exp\left\{-\frac{1}{2}(\vec{y}-\vec{\mu})'\Omega^{-1}(\vec{y}-\vec{\mu})\right\}$

$\Longrightarrow$ log likelihood is $l(\vec{\theta}|\vec{y}) = -\frac{T}{2}log(2\pi) + \frac{1}{2}log(|\Omega^{-1}|) - \frac{1}{2}(\vec{y}-\vec{\mu})'\Omega^{-1}(\vec{y}-\vec{\mu})$

If we factor $\Omega^{-1} = LL^{-1}$, we can manipulate the vector equation above to show it is equivalent to the scalar equation.

If we differentiate the exact log likelihood w.r.t. μ or θ & set $= 0$, we will have a non-linear system of equations with $\vec{\theta}$ and $\vec{y}$ as variables. This would require a numerical solution.

Conditional Maximum Likelihood for Gaussian AR(1)

Suppose we neglect the marginal density of Y_1 and only consider the conditional joint density:

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

We can approximate the likelihood as:

$$L(\vec{\theta}|\vec{y}) \approx f_{Y_t,\ldots,Y_2|Y_1}(y_t,\ldots,y_2|y_1,\vec{\theta})$$

When $\varepsilon_t \overset{i.i.d.}{\sim} N(0,\sigma^2)$

$$l(\vec{\theta}|\vec{y}) \approx log(f_{Y_t,\ldots,Y_2|Y_1}) = -\frac{T-1}{2}log(2\pi) - \frac{T-1}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2$$

Maximizing with respect to (c,ϕ) involves only maximizing:

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

which is equivalent to minimizing:

$$\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2 \text{ (OLS)}$$

The values of c and ϕ that minimize the equation above are identical to the least squares estimates of the following:

$$y_t = c + \phi y_{t-1} + \varepsilon_t$$

These are:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y$$

where

$$\beta = \begin{bmatrix} c \\ \phi \end{bmatrix}, \quad X = \begin{bmatrix} 1 & y_{T-1} \\ \vdots & \vdots \\ 1 & y_1 \end{bmatrix} \quad \text{and } \vec{y} = \begin{bmatrix} y_T \\ \vdots \\ y_2 \end{bmatrix}$$

Differentiating the log likelihood with respect to σ^2 and setting equal to zero:

$$\frac{\partial l}{\partial \sigma^2}|_{\hat{\sigma}^2} = \frac{T-1}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2 = 0$$

$$\implies \frac{1}{2\hat{\sigma}^4}\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2 = \frac{T-1}{2\hat{\sigma}^2}$$

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

This is the average of the squared regression residuals starting at $t = 2$

Exact Likelihood of Gaussian AR(p)

Now $\vec{\theta} = (c, \phi_1, \ldots, \phi_\gamma, \sigma^2)'$. Let:

$\vec{Y}_p = (Y_1, \ldots, Y_p)'$ and $\vec{Y} = (Y_1, \ldots, Y_T)'$.

Where $T > p$. Then $Y_p \sim N(\vec{\mu}, \Omega_p)$

where

$$\vec{\mu}_p = \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix}_{px1} \quad \text{and} \quad \Omega_p = \begin{bmatrix} \gamma_0 & \cdots & \gamma_{p-1} \\ \vdots & \ddots & \vdots \\ \gamma_{p-1} & \cdots & \gamma_0 \end{bmatrix}$$

Thus, the joint marginal density of $\vec{Y}_p$ is

$f_{\vec{y}_p}(\vec{y}_p|\vec{\theta}) = (2\pi)^{\frac{-p}{2}}|\Omega^{-1}|^{\frac{1}{2}} exp\left\{-\frac{1}{2}(y_p - \mu_p)'\Omega^{-1}(y_p\mu_p)\right\}$

Now, given $Y_{t-1} = y_{t-1}, \ldots, Y_{t-p} = y_{t-p} \quad \text{for } t > p$

$Y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t$

This means $Y_t|Y_{t-1} = y_{t-1}, \ldots, Y_{t-p} = y_{t-p} \sim N(c + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p}, \sigma^2)$.

Thus,

$f_{Y_t|Y_{t-1},\ldots,Y_{t-p}} = \dfrac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\dfrac{1}{2\sigma^2}(y_t - c - \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p})\right\}$

The complete joint density is

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

$\implies l(\vec{\theta}|\vec{y}) = -\dfrac{T}{2}log(2\pi) - \dfrac{T-P}{2}log(\sigma 2) + \dfrac{1}{2}log(|\Omega^{-1}|) - \dfrac{1}{2}(\vec{y}_p - \vec{\mu}_p)'\Omega^{-1}(\vec{y}_p - \vec{\mu}_p) - \dfrac{1}{2\sigma^2}\displaystyle\sum_{t=p+1}^{T}(y_t -$

$c - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p})^2$

This involves inverting a PxP matrix, Ω.

As with AR(1), maximizing the exact lg likelihood results in a set of nonlinear equations in $\vec{\theta}$ and $\vec{y}$, which must be solved numerically.

<u>Conditional Likelihood of Gaussian AR(p)</u>

As with the AR(1), we neglect the initial marginal likelihood of $\vec{y}_p$

$l(\vec{\theta}|\vec{y}) \approx -\dfrac{T-P}{2}log(2\pi) - \dfrac{T-P}{2}log(\sigma^2) - \dfrac{1}{2\sigma^2}\displaystyle\sum_{t=p+1}^{T}(y_t - c - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p})^2$

Maximizing the conditional log likelihood with respect to $c, \phi_1, \ldots, \phi_p$ is the same as minimizing

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

These are indentical to the LS estimates of

$\vec{y} = X\vec{\beta} + \vec{\varepsilon}$

where

$$\vec{\beta} = \begin{bmatrix} c \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix}, \quad X = \begin{bmatrix} 1 & y_{T-1} & y_{T-2} & \cdots & y_{T-p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & y_p & y_{p-1} & \cdots & y_1 \end{bmatrix} \text{ and } \quad \vec{\varepsilon} = \begin{bmatrix} \varepsilon_T \\ \vdots \\ \varepsilon_{p+1} \end{bmatrix} \quad \vec{y} = \begin{bmatrix} y_T \\ \vdots \\ y_{p+1} \end{bmatrix}$$

$\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{y}$

Differentiating the log likelihood with respect to σ^2, we get

$\dfrac{\partial l}{\partial \sigma^2}|_{\hat{\sigma}^2} = -\dfrac{T-P}{2\hat{\sigma}^2} + \dfrac{1}{2\hat{\sigma}^4}\displaystyle\sum_{t=p+1}^{T}(y_t - c - \phi y_{t-1} - \ldots - \phi_p y_{t-p})^2 = 0$

$\implies \sigma^2 = \dfrac{1}{T-p}\displaystyle\sum_{t=p+1}^{T}(y_t - c - \phi y_{t-1} - \ldots - \phi_p y_{t-p})^2 \approx \dfrac{1}{T-1}\sum_{t=2}^{T}(y_t - \hat{c} - \hat{\phi} y_{t-1})^2$

So assuming a Gaussian model doesn't impact the consistency of our estimates

That is $\hat{\beta}$ is consistent for the linear projection of Y_t on $Y_{t-1}, \ldots, Y_{t-p}$ even if ε isn't Gaussian.

If $\hat{\varepsilon}$ is not Gaussian, then $\hat{\beta}$ are the Quasi Maximum Likelihood Estimates because the model is misspecified.

Conditional Gaussian MA(1) Likelihood

Recall the form of a Gaussian MA(1) process

$Y_t = \mu + \varepsilon_t + \psi\varepsilon_{t-1},\ \ \varepsilon_t \overset{i.i.d}{\sim} N(0, \sigma^2)$

In this case $\vec{\theta} = (\mu, \psi, \sigma^2)'$.

Suppose we know ε_{t-1}, then

$Y_t|\varepsilon_{t-1} \sim N(\mu + \psi\varepsilon_{t-1}, \sigma^2)$

$$\implies f_{Y_t|\varepsilon_{t-1}(y_t|\varepsilon_{t-1},\vec{\theta})=} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(\overbrace{y_t - \mu - \psi\varepsilon_t}^{\varepsilon_t})\right\}$$

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

With an observation y_1, we can compute,

$\varepsilon_1 = y_1 - \mu$

Thus,

$$f_{Y_2|Y_1,\varepsilon_0} = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(\overbrace{y_t - \mu - \psi\varepsilon_t}^{\varepsilon_t})^2\right\}$$

We can continue to iterate

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

$$\implies f_{Y_t|Y_{T-1},\ldots,Y_1,\varepsilon_0} = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(\overbrace{y_t - \mu - \psi\varepsilon_t}^{\varepsilon_t})^2\right\}$$

The joint density is

$$f_{Y_1|\varepsilon_0}\prod_{t=2}^{T} f_{Y_t|Y_{t-1},\ldots,Y_1,\varepsilon_0} = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{T}{2}} exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t^2\right\}$$

The log likelihood is

$$l(\vec{\theta}|\vec{y}) = -\frac{T}{2}log(2\pi) - \frac{T}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t^2$$

The conditional log likelihood of an MA(1) is:

$$l(\vec{\theta}|\vec{y}) = -\frac{T}{2}log(2\pi) - \frac{T}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t^2$$

where: $\vec{\theta} = (\mu, \psi, \sigma^2)'$.

In this case the conditional log likelihood cannot be solved analytically for μ, ψ, σ^2.

The rough numerical algorithim would be:

1. Guess values for $\vec{\theta} = (\mu, \psi, \sigma^2)'$
2. Assume $\varepsilon_0 = 0$
3. Iteratively compute ε_t
4. Evalutate the likelihood for $\{\varepsilon_t\}_{t=1}^{T}$
5. Update $\vec{\theta}$ and return to step 2 until convergence

Using backward recursion, we can express

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

$$
\begin{aligned}
&= y_t - \mu - \psi(y_{t-1} - \mu - \psi\varepsilon_{t-2}) \\
&= (y_t - \mu) - \psi(y_{t-1} - \mu) + \psi^2\varepsilon_{t-2} \\
&= (y_t - \mu) - \psi(y_{t-1} - \mu) + \psi^2(y_{t-2} - \mu) - \psi^3\varepsilon_{t-3} \\
&= (y_t - \mu) - \psi(y_{t-1} - \mu) + \psi^2(y_{t-2} - \mu) - \ldots + (-1)^{t-1}\psi^{t-1}(y_1 - \mu) + \underbrace{(-1)^t\psi^t\varepsilon_0}_{=0}
\end{aligned}
$$

If $|\psi| < 1$, the effect of assuming $\varepsilon_0 = 0$ will die out with a large sample size.

- Conditional likelihood will be a good approximation of exact likelihood
- The approximation will be better for smaller $|\psi|$.
- If $|\psi| > 1$, the assumption of ε_0 is explosive as the sample size increases
- We can re-express as and invertible MA(1) and compute the new log likelihood

Gaussian MA(q) Conditional Likelihood

Recall the form of a Gaussian MA(q) is:

$y_t = \mu + \varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2\varepsilon_{t-2} + \ldots + \psi_q\varepsilon_{t-q}, \;\; \varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma^2)$

Now, $\vec{\theta} = (\mu, \psi_1, \ldots, \psi_q, \sigma^2)'$.

We will now assume $\varepsilon_0 = \varepsilon_{-1} = \varepsilon_{-2} = \ldots = \varepsilon_{-q+1} = 0$ and iteratively compute $\{\varepsilon_t\}_{t=1}^T$ using $\vec{y}$.

- The conditional likelihood will have the exact same form as for the MA(1), but with these new $\{\varepsilon_t\}_{t=1}^T$
- It must be computed numerically, in a fashion that is almost identical to the MA(1).
- The conditional likelihood can only be used with the invertible form of the MA(q).

Gaussian ARMA(p,q) Conditional Likelihood

Recall the form of a Gaussian ARMA(p,q):

$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t + \psi_1\varepsilon_{t-1} + \ldots + \psi_q\varepsilon_{t-q}$

To form the conditonal likelihood, we combine the methods of AR(p) and MA(q)

- Condition on $y_0 = y_{-1} = \ldots = y_{-p+1} = \dfrac{c}{1 - \phi_1 - \phi_2 - \ldots - \phi_p}$
- Condition on $\varepsilon_0 = \varepsilon_{-1} = \ldots = \varepsilon_{-q+1} = 0$
- Recursively compute $\{\varepsilon_t\}_{t=1}^T$ using $\{y_t\}_{t=1}^T, \{\varepsilon_t\}_{t=-q+1}^0, \{y_t\}_{t=-p+1}^0$
- Form the log likelihood as:

  $$l(\vec{\theta}|\vec{y}) = -\frac{T}{2}log(2\pi) - \frac{T}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t^2$$

Alternatively, we could start the recursions at $t = p + 1$ without an initial condition on $\{y_t\}_{t=-p+1}^0$.

- Condition on $\varepsilon_p = \varepsilon_{p-1} = \ldots = \varepsilon_{p-q+1} = 0$
- Recursively compute $\{\varepsilon_t\}_{t=p+1}^{T}$ using past values of $y_t + \varepsilon_t$
- Form the log likelihood as:

  $l(\vec{\theta}|\vec{y}) = -\frac{T-p}{2} log(2\pi) - \frac{T-p}{2} log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=p+1}^{T} \varepsilon_t^2$
- The MA polynomial must be invertible for us to use the conditional log likelihood in estimation
- The Kalman filter can be used to compute the exact likelihood