EC402: MEI. Spring Term: The Analysis of Time Series*

Lecture Notes #2: MLE of Time Series Data

In what follows we discuss how to implement ML estimation for dependent data and present several practical examples.

1. Ergodic Theorem

If a stochastic process y_t, $t = 1, 2, ...$ is ergodic with mean $\mu < \infty$ then

$$P\lim \frac{1}{T}\sum_{t=1}^{T} y_t = \mu.$$

Ergodicity is a sufficient condition for sample means to converge to their expectations.

The definitions above extend naturally to vector valued stochastic processes. If a vector valued stochastic process is stationary then functions of that stochastic process are also stationary; if (x_t, y_t) are a jointly stationary stochastic process then

$$z_t = f(x_t, y_t)$$

is a stationary stochastic process and in particular

$$z_t = ax_t + by_t$$

is a stationary stochastic process. Similarly, functions of vector valued ergodic processes are ergodic.

---

2. Maximum Likelihood Estimation of Time Series Models

Last term you were introduced to the basic ideas of Maximum Likelihood (ML). However most of the arguments presented rested on the assumption that the observations were independent. In time series, observations are dependent. Do the ML results carry over? The basic answer is yes for *ergodic* processes.

The standard approach to ML that you have seen last term is to obtain the likelihood function by writing down the density for each observation and then (since the observations are independent) producting them. However a moments thought indicates that the standard approach will not work in time series since the observations are generally dependent.

However it is always the case that a joint density can be factored into a conditional times a marginal

For example if you have three observations

$$\begin{aligned} f(y_3, y_2, y_1) &= f\left(y_3 \middle| y_2, y_1\right) \cdot f(y_2, y_1) \\ &= f\left(y_3 \middle| y_2, y_1\right) \cdot f\left(y_2 \middle| y_1\right) \cdot f(y_1). \end{aligned}$$

Hence the likelihood for T observations is

$$L\left(y;\psi\right) = \left[\prod_{t=2}^{T} f\left(y_t | y_{t-1}, ..., y_1\right)\right] \cdot f(y_1)$$

This can be written as

$$L(y;\psi) = \prod_{t=2}^{T} f\left(y_t | I_{t-1}\right) \cdot f(y_1)$$

where I_{t-1} denotes all the information available at time $t-1$.

Taking logs then yields

$$\log L(y;\psi) = \sum_{t=2}^{T} \log f\left(y_t | I_{t-1}\right) + \log f(y_1).$$

This is only useful if the conditional densities are easily written down. But many time series models are actually specified in terms of their conditional distributions.

Most importantly when the likelihood is constructed as shown above we have that, for ergodic processes, the standard asymptotics results you have seen last term still hold. That is, the ML estimator of a vector of parameters ψ is consistent and

$$\sqrt{T}(\hat{\psi}_{MLE} - \psi) \xrightarrow{D} N\left(0, \left(\lim \frac{1}{T} I(\psi)\right)^{-1}\right)$$

where $I(\psi)$ is the information matrix.

This implies that, for example, the variance of $\hat{\psi}$ can be approximated by either $\left[\frac{-\partial^2 \log L(\hat{\psi}_{MLE})}{\partial\psi\partial\psi'}\right]^{-1}$ or the inverse of the empirical information matrix, $I(\psi)^{-1}$, where

$$I(\psi) = -E\left[\frac{\partial^2 \log L(\psi)}{\partial\psi\partial\psi'}\right].$$

In what follows, several examples of ML for time series data are provided.

2.1. MLE of the AR(1) process

Consider the $AR(1)$

$$y_t = \phi y_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim \text{iid } N(0, \sigma^2), \ \ |\phi| < 1.$$

Then $y_t | \, y_{t-1}$ is $N(\phi y_{t-1}, \sigma^2)$ and

$$f\left(y_t | \, I_{t-1}\right) = f\left(y_t | \, y_{t-1}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\underbrace{y_t - \phi y_{t-1}^2}_{\varepsilon_t}\right)^2\right\}$$

and the log likelihood is simply,

$$\log L(y; \phi, \sigma^2) = -\frac{(T-1)}{2}\log 2\pi - \frac{(T-1)}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t - \phi y_{t-1})^2 + \log f(y_1).$$

What do we do about the *initial conditions*? One possibility is to condition on y_1, take it as fixed. In this case the final term can be dropped and the likelihood becomes the likelihood for the linear regression of y_t on y_{t-1} for observations $t = 2, ..., T$.

Thus we have, at the maximum,

$$\frac{\partial \log L}{\partial \phi} = \frac{1}{\sigma^2} \sum_{t=2}^{T} \left(y_t - \phi y_{t-1}\right) y_{t-1} = 0 \Rightarrow \hat{\phi} = \frac{\sum\limits_{t=2}^{T} y_t y_{t-1}}{\sum\limits_{t=2}^{T} y_{t-1}^2}$$

so we end up with the OLS estimator.

Alternatively, you can use the unconditional distribution for y_1, $N\left(0, \frac{\sigma^2}{(1-\phi^2)}\right)$. Recall that in the $AR(1)$, the unconditional mean, $E(y_t) = 0$, and the unconditional variance, $var(y_t) = \frac{\sigma^2}{(1-\phi^2)}$. This assumption for y_1 is sensible if the process has been going on for a long time at $t = 1$. Under this assumption

$$\log f(y_1) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 + \frac{1}{2}\log(1-\phi^2) - \frac{1}{2\sigma^2}(1-\phi^2)y_1^2$$

and gives the log likelihood

$$\begin{aligned} \log L(y; \phi, \sigma^2) &= -\frac{T}{2}\log 2\pi - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t - \phi y_{t-1})^2 \\ &\quad + \frac{1}{2}\log(1-\phi^2) - \frac{1}{2\sigma^2}(1-\phi^2)y_1^2. \end{aligned}$$

This makes the ML estimator non-linear. Whether you estimate the truncated likelihood, least squares or the likelihood above, the adjustment involves only a single observation. As a result its effect in large samples is negligible and you can approximate the nonlinear estimator by the least squares estimator in large samples and the large sample properties are the same.

To sum up, we have shown that for the stationary $AR(1)$, even though least squares does not have the standard small sample properties, in large samples it is maximum likelihood and hence consistent, asymptotically efficient and asymptotically normal and hence standard inference procedures are valid in large samples.

Again these results can be extended to the stationary $AR(p)$ model and to the regression model with both process independent control regressors and lagged dependent variables.

2.2. MLE of Nonlinear least squares models

Obviously the class of all ML models is huge. An important sub class is that of nonlinear regression models,

$$y_t = g(x_t;\beta) + \varepsilon_t \qquad \varepsilon_t \text{ iid } N(0,\sigma^2), \quad t = 1, ..., T,$$

x_t process independent. So

$$\varepsilon_t(\beta) := y_t - g(x_t;\beta)$$

and

$$f(\varepsilon_t(\beta)) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{\frac{-\varepsilon_t(\beta)^2}{2\sigma^2}\right\}.$$

Hence,

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

It is clear that maximizing $\log L$ (or L) with respect to β is equivalent to minimizing the residual sum of squares with respect to β and hence that the nonlinear least squares estimator is ML. Differentiating the log likelihood,

$$\begin{aligned}
\frac{\partial \log L}{\partial \beta} &= -\frac{1}{\sigma^2}\sum_t \frac{\partial \varepsilon_t(\beta)}{\partial \beta}\varepsilon_t(\beta) = \frac{1}{\sigma^2}\sum_t z_t\varepsilon_t = 0 \\
\frac{\partial \log L}{\partial \sigma^2} &= -\frac{T}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_t \varepsilon_t(\beta)^2 = 0
\end{aligned}$$

at maximum, where

$$z_t := -\frac{\partial \varepsilon_t}{\partial \beta} = \frac{\partial g(x_t;\beta)}{\partial \beta}.$$

The first order conditions with respect to β are generally nonlinear and the ML estimates of β have to be obtained by an iterative maximization algorithm. The first order conditions with respect to σ^2 yield the usual ML estimator for σ^2,

$$\hat{\sigma}^2 = \frac{1}{T}\sum_t \varepsilon_t(\hat{\beta})^2.$$

Recall from your ML notes that we constructed an estimate of the variance-covariance matrix of our estimates based on the empirical information matrix $I(\psi)$,

$$I(\psi) = -E\left[\frac{\partial^2 \log L(\psi)}{\partial\psi\partial\psi'}\right].$$

In the present case $\psi = (\beta', \sigma^2)$. If x_t is process independent, x_t is independent of ε_t and thus $z_t = -\frac{\partial \varepsilon_t}{\partial \beta} = \frac{\partial g(x_t;\beta)}{\partial \beta}$ is also independent of ε_t, as are the second derivatives of ε_t.

So, looking at the components of $I(\psi)$, we have

$$\begin{aligned}
-E\left[\frac{\partial^2 \log L}{\partial\beta\partial\beta'}\right] &= \frac{1}{\sigma^2}\left[E\sum_t \frac{\partial^2 \varepsilon_t}{\partial\beta\partial\beta'}\cdot\varepsilon_t + E\sum_t \frac{\partial\varepsilon_t}{\partial\beta}\frac{\partial\varepsilon_t}{\partial\beta'}\right] \\
&= \frac{1}{\sigma^2}\left[\sum_t E\frac{\partial^2 \varepsilon_t}{\partial\beta\partial\beta'}\cdot E(\varepsilon_t) + E\sum_t \frac{\partial\varepsilon_t}{\partial\beta}\frac{\partial\varepsilon_t}{\partial\beta'}\right] \\
&= \frac{1}{\sigma^2}E\sum_t z_t z_t' \qquad \text{since } E(\varepsilon_t) = 0. \\
-E\left[\frac{\partial^2 L}{\partial(\sigma^2)^2}\right] &= -\frac{T}{2(\sigma^2)^2} + \frac{2}{2(\sigma^2)^3}\sum_t E(\varepsilon_t^2) \\
&= -\frac{T}{2(\sigma^2)^2} + \frac{2T}{2(\sigma^2)^3}\sigma^2 \\
&= \frac{T}{2(\sigma^2)^2} \\
-E\left[\frac{\partial^2 \log L}{\partial\beta\partial\sigma^2}\right] &= \frac{1}{(\sigma^2)^2}E\sum_t z_t\varepsilon_t = \frac{1}{(\sigma^2)^2}\sum_t E(z_t)E(\varepsilon_t) \\
&= 0.
\end{aligned}$$

Hence by previous results, the information matrix is

$$I(\psi) = \frac{1}{\sigma^2}\begin{bmatrix} E\sum_t z_t z_t' & 0 \\ 0 & \frac{T}{2\sigma^2} \end{bmatrix}.$$

Inverting, and substituting the consistent ML estimates of β and σ^2 for unknown parameters, and the sample moment $\sum_t z_t z_t'$ for $E\sum_t z_t z_t'$, we approximate the distribution of $(\hat{\beta}', \hat{\sigma}^2)$ by a Normal distribution, mean (β', σ^2) and variance covariance matrix,

$$\begin{bmatrix} \hat{\sigma}^2\left(\sum_t z_t z_t'\right)^{-1} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{T} \end{bmatrix}.$$

i.e.

$$\begin{bmatrix} \hat{\beta}' - \beta' \\ \hat{\sigma}^2 - \sigma^2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}; \begin{bmatrix} \hat{\sigma}^2\left(\sum_t z_t z_t'\right)^{-1} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{T} \end{bmatrix}\right)$$

In particular, therefore, $var(\hat{\beta}) = \hat{\sigma}^2\left(\Sigma z_t z_t'\right)^{-1}$, where $z_t = \partial g(x_t;\beta)/\partial\beta$ is evaluated at $\hat{\beta}$. Furthermore, it may be shown that this procedure is valid in large samples even if x_t is only contemporaneously independent of ε_t.

2.3. MLE of the MA(1) process

$$y_t = \varepsilon_t + \theta\varepsilon_{t-1} \qquad \varepsilon_t \text{ iid } N(0,\sigma^2)$$

So $y_t|\,\varepsilon_{t-1} \sim N\left(\theta\varepsilon_{t-1},\sigma^2\right)$. Assume that we start from $\varepsilon_0 = 0$, then we may define $\varepsilon_t(\theta)$ by using the recursive equation

$$\varepsilon_t(\theta) := y_t - \theta\varepsilon_{t-1}(\theta), \qquad t = 1, 2, ..., T.$$

Since $\varepsilon_0 = 0$,

$$\begin{aligned}\varepsilon_1(\theta) &= y_1\\ \varepsilon_2(\theta) &= y_2 - \theta y_1\\ \varepsilon_3(\theta) &= y_3 - \theta y_2 + \theta^2 y_1\end{aligned}$$

$$-----------$$

$$\varepsilon_t(\theta) = y_t - \theta y_{t-1} + \theta^2 y_{t-2} + ... + (-\theta)^{t-1} y_1.$$

Given $y_t|\,\varepsilon_{t-1} \sim N\left(\theta\varepsilon_{t-1},\sigma^2\right)$, then

$$f\left(y_t|\,I_{t-1}\right) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp -\frac{\left(y_t - \theta\varepsilon_{t-1}(\theta)\right)^2}{2\sigma^2}.$$

So the log likelihood is

$$\begin{aligned}\log L(\theta,\sigma^2) &= -\frac{T}{2}\log 2\pi - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_t - \theta\varepsilon_{t-1}(\theta)\right)^2\\ &= -\frac{T}{2}\log 2\pi - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t(\theta)^2.\end{aligned}$$

Following the results in the previous section, we have

$$\frac{\partial \log L}{\partial\theta} = \frac{1}{\sigma^2}\sum_t z_t(\theta)\varepsilon_t(\theta) \text{ where } z_t(\theta) = -\frac{\partial\varepsilon_t(\theta)}{\partial\theta}.$$

So the ML (or nonlinear least squares) estimator, $\hat{\theta}$ satisfies

$$\sum_t z_t(\theta)\varepsilon_t(\theta) = 0$$

where $\varepsilon_t(\theta)$ is defined above.

Furthermore, the variance of $\hat{\theta}$ is given by

$$var(\hat{\theta}) = \hat{\sigma}^2 \left(\sum_t z_t^2(\hat{\theta}) \right)^{-1}$$

where $\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} \varepsilon_t^2(\hat{\theta})$.