Maximum Likelihood: An Introduction

Christian Julliard

Department of Economics and FMG
London School of Economics

Outline

MLE in a Nutshell

- The likelihood function (often simply the likelihood) is a function of the parameters, ψ, of a statistical model

$$L(x|\psi) = f(x_1, ..., x_n|\psi)$$

where $x_1, ..., x_n$ is the sample of data, $f(.|\psi)$ is a known probability density function (pdf) parameterized by the unknown vector of parameters ψ

- The maximum likelihood estimator (MLE) is

$$\hat{\psi}_{MLE} = \arg\max_{\psi} L(x|\psi) = \arg\max_{\psi} \log L(x|\psi)$$

- That is, the MLE maximizes a conditional probability function considered as a function of its second argument, with its first argument – the data – held fixed.

$\Rightarrow$ MLE answers the question: "What is the most likely value of ψ given the sample we have observed?"

MLE of Independent Data

- If the data are independent, identically distributed (iid) we have

$$f(x_1, ..., x_n|\psi) = f(x_1|\psi) \times f(x_2|\psi) \times ... \times f(x_n|\psi)$$

in the same way as $P(A, B) = P(A) P(B)$ if and only if A and B are independent.

- The likelihood can than be written as the product of n probability densities

$$L(x|\psi) = \prod_{i=1}^{n} f(x_i|\psi) \rightarrow \log L(x|\psi) = \sum_{i=1}^{n} \log f(x_i|\psi)$$

Example: estimating mean and variance

Recall the Normal (Gaussian) distribution $N\left(\mu, \sigma^2\right)$ has pdf

$$f\left(x_i|\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right\}$$

- The corresponding pdf for a sample of n idd Normal random variables – the likelihood – is

$$L\left(x|\mu, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right\}$$
$$= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}\right\}$$

$$\Rightarrow \log L\left(x|\mu, \sigma^2\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}$$

where $\psi = [\mu, \sigma^2]$ are the unknown parameters we want to estimate.

Taking the FOC for a maximum we have

$$\frac{\partial \log L\left(x|\mu, \sigma^2\right)}{\partial \mu} = -\frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2} = 0$$
$$\therefore \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{\partial \log L\left(x|\mu, \sigma^2\right)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^4} = 0$$
$$\therefore \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

Example: OLS as MLE

- Consider the linear standard regression model

$$y_i = x_i'\beta + \varepsilon_i \sim N\left(0, \sigma^2\right);\ i = 1, ..., n;\ E\left[x_i\varepsilon_s\right] = 0\ \forall s, i \qquad (1)$$

- Since $y_i - x_i'\beta = \varepsilon_i \sim N\left(0, \sigma^2\right)\ \forall i$ we have

$$L\left(y, x|\beta, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - x_i'\beta\right)^2\right\}$$

$$\log L\left(y, x|\beta, \sigma\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2$$

- Note that by definition

$$\hat{\beta}_{MLE} = \arg\max \log L\left(y, x|\beta, \sigma\right) = \arg\max -\sum_{n=1}^{n}\left(y_i - x_i'\beta\right)^2$$
$$= \arg\min \sum_{n=1}^{n}\left(y_i - x_i'\beta\right)^2 = \hat{\beta}_{OLS}$$

MLE for Time Series Models

- The standard approach to MLE we have seen so far is to obtain the likelihood function by
  1. writing the density for each observation and then
  2. *since the observations are independent*, write the likelihood as the product of these densities.
- This standard approach will not work in time series since the *observations are generally dependent*.

But: a *joint density* can be always factored into a *conditional* times a *marginal*.

Example: if you have three observations

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

- Hence the likelihood for T observations is

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

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

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

Note: $f(y_1)$ can be either modeled directly or y_1 can be assumed to be a constant (more on this later)

Ergodic Theorem

(An) 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.
- This definition extends to vector valued stochastic processes.
- Moreover, functions of vector valued ergodic processes are ergodic.

MLE Asympotics for Time Series Models

- Even if observations are dependent, for *ergodic* processes, the ML estimator of a vector of parameters ψ is generally consistent.
- Moreover, the asymptotic normality results derived for the MLE in the iid setting carry over for *ergodic* processes.
- That is, the ML estimated parameters will be efficient and have an asymptotic Gaussian distribution.

ML Asymptotics

For a vector of parameters ψ and ergodic data, we have the standard asymptotic result

$$\sqrt{T}(\hat{\psi} - \psi_0) \xrightarrow{D} N\left(0, \left(\frac{1}{T} I(\psi_0)\right)^{-1}\right) \tag{2}$$

where $I(\psi_0)$ is the information matrix defined as

$$I(\psi_0) := -E\left[\frac{\partial^2 \log L(\psi_0)}{\partial\psi\partial\psi'}\right] = E\left[\frac{\partial \log L(\psi_0)}{\partial\psi}\frac{\partial \log L(\psi_0)'}{\partial\psi}\right]$$

where the last identity is the so called *information matrix identity*.

- Obviously, $\frac{1}{T} I(\psi_0)$ is in general not observed.
- So to make the asymptotic normality result operational we need a consistent estimator of $\frac{1}{T} I(\psi_0)$
- Two commonly used estimators are:

Asymptotic Variance Estimators

1. The Hessian based estimator
$$-\left[\frac{1}{T}\frac{\partial^2 \log L(\hat{\psi}_{MLE})}{\partial\psi\partial\psi'}\right]^{-1}.$$
2. The empirical information matrix $\frac{1}{T} I\left(\hat{\psi}_{MLE}\right)$ based (more on this shortly).

These are both consistent since $\hat{\psi}_{MLE} \rightarrow \psi_0$

The Delta Method

- Suppose we know that
$$\sqrt{T}\left(\hat{\psi} - \psi_0\right) \xrightarrow{D} N(0, V)$$
and we are interested in making inference about $g\left(\hat{\psi}\right)$, (where g(.) is some differentiable function with continuous first derivative).
- What is the distribution of $g\left(\hat{\psi}\right)$?

- Consider the Taylor expansion around ψ_0
$$g\left(\hat{\psi}\right) - g(\psi_0) \cong G(\psi_0)\left(\hat{\psi} - \psi_0\right)$$
where $G(\psi) := \frac{\partial g(\psi)}{\partial \psi'}$.
- This implies that
$$\begin{aligned} Var\left(g\left(\hat{\psi}\right) - g(\psi_0)\right) &\cong Var\left(G(\psi_0)\left(\hat{\psi} - \psi_0\right)\right) \\ &= G(\psi_0)\, Var\left(\left(\hat{\psi} - \psi_0\right)\right) G(\psi_0)' \\ &= G(\psi_0)\, VG(\psi_0)' \end{aligned}$$
- We can therefore apply a CLT argument to get
$$\sqrt{T}\left(g\left(\hat{\psi}\right) - g(\psi_0)\right) \xrightarrow{D} N\left(0, G(\psi_0)'\, VG(\psi_0)\right).$$
- This result is the so called Delta method.

The linear standard regression model

The linear standard regression model

- Consider the standard model (1)

Recall: a consistent estimator of the asymptotic variance is
$$-\left[\frac{1}{T}\frac{\partial^2 \log L(\hat{\psi}_{MLE})}{\partial\psi\partial\psi'}\right]^{-1}. \quad (3)$$

- Note that
$$\left.\frac{\partial \log L\left(y, z|\beta, \sigma^2\right)}{\partial\beta}\right|_{MLE} = -\frac{1}{\sigma^2}\sum_{t=1}^{T} x_t\left(y_t - x_t'\beta\right)\Big|_{MLE} = 0$$
$$\left.\frac{\partial^2 \log L\left(y, z|\beta, \sigma^2\right)}{\partial\beta\partial\sigma^2}\right|_{MLE} = \frac{1}{\sigma^4}\sum_{t=1}^{T} x_t\left(y_t - x_t'\beta\right)\Big|_{MLE} = 0$$
$$\left.\frac{\partial^2 \log L\left(y, z|\beta, \sigma^2\right)}{\partial\beta\partial\beta'}\right|_{MLE} = -\frac{\sum_{t=1}^{T} x_t x_t'}{\sigma^2}\Big|_{MLE} = -\frac{\sum_{t=1}^{T} x_t x_t'}{\hat{\sigma}^2}$$

$\Rightarrow$ we have the usual result for the variance of the OLS coefficients $\left(\sum_{t=1}^{T} x_t x_t'\right)^{-1}\hat{\sigma}^2$.

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 $N(\phi y_{t-1}, \sigma^2)$, therefore

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

- And the log likelihood is simply,

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

What do we do about the *initial condition*?

- One possibility is to condition on y_1, i.e. 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,

$$\begin{aligned} \frac{\partial \log L}{\partial \phi} &= \frac{1}{\sigma^2} \sum (y_t - \phi y_{t-1}) y_{t-1} = 0 \\ \Rightarrow \quad \hat{\phi} &= \frac{\sum y_t y_{t-1}}{\sum y_{t-1}^2} \Rightarrow \hat{\phi} = \hat{\phi}_{OLS} \end{aligned}$$

- Alternatively, you can use the *unconditional distribution* for y_1,

Recall: in the $AR(1)$, the unconditional mean, $E(y_t) = 0$, and the unconditional variance, $var(y_t) = \frac{\sigma^2}{(1-\phi^2)}$.

- so the unconditional distribution is $N\left(0, \frac{\sigma^2}{(1-\phi^2)}\right)$.
- 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 this gives the log likelihood

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

Note: these results can be extended to:

1. the stationary $AR(p)$ model
2. the regression model with both process independent regressors and lagged dependent variables.

MLE of Nonlinear least squares models

- An important sub class of MLE 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.
- Note that

$$\begin{aligned} \varepsilon_t(\beta) &= y_t - g(x_t;\beta) \\ f(\varepsilon_t(\beta)) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-\varepsilon_t(\beta)^2}{2\sigma^2}\right\}. \end{aligned}$$

- Hence,

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

- So, maximizing $\log L$ wrt β is equivalent to minimizing the residual sum of squares with respect to β.

- Differentiating the log likelihood,

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

  where

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

Note: the first order conditions with respect to β are nonlinear and the ML estimates of β have to be obtained by numerical maximization.

- 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 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 $\psi = (\beta', \sigma^2)$.

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

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

- Hence, 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

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

that is equivalent to (2)

MLE of the MA(1) process

- Consider the MA(1)

$$\begin{aligned} y_t &= \varepsilon_t + \psi\varepsilon_{t-1} \qquad \varepsilon_t \text{ iid } N(0,\sigma^2) \\ &\rightarrow \quad y_t|\,\varepsilon_{t-1} \sim N\left(\psi\varepsilon_{t-1}, \sigma^2\right) \end{aligned}$$

- Assume we start from $\varepsilon_0 = 0$, then we may define $\varepsilon_t(\psi)$ by using the recursive equation

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

- Since $\varepsilon_0 = 0$,

$$\begin{aligned} \varepsilon_1(\psi) &= y_1 \\ \varepsilon_2(\psi) &= y_2 - \psi y_1 \\ \varepsilon_3(\psi) &= y_3 - \psi y_2 + \psi^2 y_1 \\ \varepsilon_t(\psi) &= y_t - \psi y_{t-1} + \psi^2 y_{t-2} + ... + (-\psi)^{t-1} y_1. \end{aligned}$$

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

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

- So the log likelihood is

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

- As before we have

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

- So the $\hat{\psi}_{MLE}$ satisfies

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

Furthermore, using the empirical $I(\psi)$ we can show as before that the variance of $\hat{\psi}$ is given by

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

where

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