Maximum Likelihood Estimation

Leopoldo Catania
Aarhus University and CREATES

leopoldo.catania@econ.au.dk

Program for today

Today we (briefly) review the maximum likelihood estimator and see its implementation in R

- Writing a log-likelihood function
- Maximizing the log-likelihood function
- Asymptotic covariance matrix of parameters
- Standard errors
- Hypothesis Testing with the LR test

The principle of maximum likelihood

The method of maximum likelihood (ML) is a fundamental and very general estimation method, that encompasses other estimation methods, such as least squares.

Maximum likelihood estimates of parameter vector, θ, are easily obtained under few hypothesis:

- Models are *parametric*, in the sense the probability distribution can be described by a finite number of parameters, θ ;
- The sample $\mathbf{X} = \{X_1, ..., X_T\}$ is draw from a family of probability distributions parametrized by an unknown parameter vector θ;
- This probability distribution is known;
- Data drawn from this probability are *iid*.

The principle of maximum likelihood

The likelihood estimation is obtained, given the specification of the model, by the joint density associated with our dataset $f(X;\theta)$. Given θ fixed, then $f(X;.)$ is the density of $\mathbf{X}$. Then, given that X_t is *iid*, we can compute the joint density of the entire sample as

$$f(\mathbf{X};\theta) = \prod_{t=1}^{T} f(X_t;\theta) \tag{1}$$

The likelihood function is then defined as the joint density function, $f(\mathbf{X};\theta)$, given $\mathbf{X}$ fixed, as a function of θ, i.e. $L(\theta;\mathbf{X}) = f(\mathbf{X};\theta)$. Note that $L(\theta;\mathbf{X})$ is a function of θ, while $\mathbf{X}$ is fixed. We usually work with the log of the likelihood function:

$$\log L(\theta;\mathbf{X}) = \sum_{t=1}^{T} \log f(X_t;\theta) \tag{2}$$

The parameter vector, θ is identified if for any $\theta_1 \neq \theta$, $L(\theta;\mathbf{X}) \neq L(\theta_1;\mathbf{X})$.

The principle of maximum likelihood

- The principle of maximum likelihood means of choosing an asymptotically efficient estimator for a set of parameters, θ, by maximizing the likelihood function with respect to this set of parameters.
- The estimation entails the calculation of the first and second derivatives of the the likelihood function with respect to the parameter vector.
- In some cases, derivatives are in closed form, but, in general, the derivatives must be computed numerically.
- The necessary first-order condition for maximizing $\log L(\theta; \mathbf{X})$ is

$$\left.\frac{\partial \log L(\theta; \mathbf{X})}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0 \tag{3}$$

Writing and maximizing a likelihood function in R

In R the steps are:

- Write a function that computes the negative of log-likelihood function
- The inputs of this function are:
  1. A parameter vector, θ;
  2. The data, X;
  3. Additional optional inputs
- The output is the negative of the sum of the log-likelihood.
- Set initial values for θ.
- Maximize the log-likelihood function using the optimization routines of R such as `optim()`.

Example: Exponential i.i.d. observations

Suppose that we have a sample, $\{x_t\}_{t=1,\dots,T}$, of i.i.d. observations extracted from an exponential distribution. We denote $X_t \overset{iid}{\sim} Exp(\lambda)$ with

$$f_X(x_t, \lambda) = \lambda e^{-\lambda x_t}, \tag{4}$$

where $x_t \geq 0$.

- We want to estimate the coefficient λ by ML using the information coming from the i.i.d. sample, $\{x_t\}_{t=1,\dots,T}$.
- Let's write an R code for this!

ML in the case of dependent observations

In financial time–series observations are rarely *iid* that is:

$$f(X_t, X_{t-s}) \neq f(X_t)f(X_{t-s}). \tag{5}$$

for $s = 1, 2, \ldots$.

However, recall that every joint distribution can be decomposed in the product of the conditional and marginal distribution:

$$f(X_t, X_{t-s}) = f(X_t|X_{t-s})f(X_{t-s}). \tag{6}$$

ML in the case of dependent observations

Suppose to have a time–series of T observations: $X_1, \dots, X_T$. Its joint distribution can be factorized as:

$$f(X_1, \dots, X_T) = f(X_1) \prod_{t=2}^{T} f(X_t | X_1, \dots, X_{t-1})$$

In general we identify with $\mathcal{F}_t$ with all the information contained in the observations up to time t, that is $\mathcal{F}_t = \{X_1, \dots, X_t\}$, with $F_0 = \{\emptyset\}$. In this way we can write:

$$f(X_1, \dots, X_T) = \prod_{t=1}^{T} f(X_t | \mathcal{F}_{t-1}).$$

In financial econometrics we usually make a parametric assumption on $f(X_t | \mathcal{F}_{t-1})$ and $f(X_1)$ in order to estimate models.

Under the assumption of correct model specification, identification of the parameters, continuity and finiteness of the first three derivatives of $\log(f(\theta, X_t))$ wrt θ, the ML estimator has the following properties:

1. Consistency, $\hat{\theta} \xrightarrow{p} \theta_0$.
2. Asymptotic Normality:

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Sigma(\theta_0))$$

3. $\Sigma(\theta_0) = I(\theta_0)^{-1}$, with

$$I(\theta_0) = -E_{\theta_0}\left[\frac{\partial^2 L(\theta; X)}{\partial\theta\partial\theta'}\bigg|_{\theta=\theta_0}\right]$$

   which is called Fisher information matrix. Detailed derivation in Newey and McFadden (1994), M-estimators.
4. ML estimator is the most efficient unbiased estimator, as the inverse of the Fisher information matrix is the lower bound on the variance of any estimator of θ (Cramer-Rao bound).
5. Invariance: The maximum likelihood estimator of $\gamma_0 = c(\theta_0)$ is $c(\hat{\theta})$ if $c(\cdot)$ is a continuous and continuously differentiable function. (By an application of the continuous mapping theorem)

Asymptotic Variance I

Given the previous results, the asymptotic covariance matrix of the MLE can be computed as

$$\Sigma(\hat{\theta}) = -H(\hat{\theta})^{-1}$$
$$H(\hat{\theta}) = \left. \frac{\partial^2 \log L(\theta; X)}{\partial\theta\partial\theta'} \right|_{\theta=\hat{\theta}}$$

where $H(\hat{\theta})$ is evaluated numerically in $\hat{\theta}$.

Alternative Estimators of the asymptotic covariance matrix

Other estimators of the asymptotic covariance matrix are

- BHHH (Berndt-Hall-Hall-Hausman) or gradient outer-product:

$$\Sigma(\hat{\theta}) = \left[\sum_{i=1}^{N} \hat{g}_i \hat{g}_i'\right]^{-1}$$

  where $\hat{g}_i$ is the gradient for the i-th observation computed in $\hat{\theta}$.

- Sandwich estimator (QML estimator):

$$\Sigma(\hat{\theta}) = [-H(\hat{\theta})^{-1}] \left(\sum_{i=1}^{N} \hat{g}_i \hat{g}_i'\right) [-H(\hat{\theta})^{-1}]$$

Delta Method

Delta method is an approximate method to derive the standard errors of functions of the parameter estimates. Given that:

$$\frac{\hat{\theta} - \theta}{S.E.(\theta)} \rightarrow N(0,1) \tag{7}$$

then

$$\frac{f(\hat{\theta}) - f(\theta)}{f'(\theta)S.E.(\theta)} \rightarrow N(0,1) \tag{8}$$

The multivariate version implies an asymptotic covariance matrix of the parameters, Λ,

$$\Lambda = J(\theta)'\Sigma(\theta)J(\theta) \tag{9}$$

where $J(\theta)'$ is the $p \times p$ matrix with the first derivatives, $f'(\theta)$ with respect to the p parameters in θ, and Σ is the asymptotic covariance matrix of θ.

Hypothesis testing in the ML context

We now focus on one of the most common testing procedure related to ML estimation: the Likelihood-ratio (LR) test.

The LR test is based on the idea that under the null hypothesis values of the log-likelihood of the unrestricted model, $\log L_U$, and the model under $\mathcal{H}_0$, $\log L_R$, must be close. The test takes the form

$$LR = 2 \cdot (\log L_U - \log L_R)$$

As $N \to \infty$,

$$LR \to \chi^2(p)$$

where p is the number of restrictions imposed.