Illustrations of Maximum Likelihood Estimation

MAXIMUM LIKELIHOOD ESTIMATION - Examples

Maximum Likelihood Estimation (MLE) is a systematic technique for estimating parameters in a probability model from a data sample. Suppose a sample $x_1, \ldots, x_n$ has been obtained from a probability model specified by mass or density function $f_X(x;\theta)$ depending on parameter(s) θ lying in parameter space Θ.

MAXIMUM LIKELIHOOD ESTIMATION - Examples

The **maximum likelihood estimate** or **MLE** is produced as indicated in the next 4 STEPS;

STEP 1 Write down the likelihood function, $L(\theta)$, where

$$L(\theta) = \Pi_{i=1}^{n} f_X(x_i; \theta)$$

that is, the product of the n mass/density function terms (where the ith term is the mass/density function evaluated at x_i) viewed as a function of θ.

STEP 2 Take the natural log of the likelihood, collect terms involving θ.

MAXIMUM LIKELIHOOD ESTIMATION - Examples

STEP 3 Find the value of $\theta \in \Theta$, for which $logL(\theta)$ is maximized, for example by differentiation. If θ is a single parameter, find θ by solving

$$\frac{dlogL(\theta)}{d\theta} = 0$$

in the parameter space Θ. If θ is vector-valued, say $\theta = (\theta_1, ..., \theta_k)$, then find $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_k)$, by simultaneously solving the k equations given by

$$\frac{\partial logL(\theta)}{\partial \theta_j} = 0, \qquad j = 1, ..., k$$

in parameter space Θ. Note that, if parameter space Θ is a bounded interval, then the maximum likelihood estimate may lie on the boundary of Θ.

MAXIMUM LIKELIHOOD ESTIMATION - Examples

STEP 4 Check that the estimate θ obtained in **STEP 3** truly corresponds to a maximum in the (log) likelihood function by inspecting the second derivative of $logL(\theta)$ with respect to θ. In the single parameter case, if the second derivative of the log-likelihood is negative at $\theta = \hat{\theta}$, then θ is confirmed as the MLE of θ (other techniques may be used to verify that the likelihood is maximized at θ).

MAXIMUM LIKELIHOOD ESTIMATION - Examples

EXAMPLE Suppose a sample $x_1, ..., x_n$ is modelled by a Poisson distribution with parameter denoted λ, so that

$$f_X(x;\theta) \equiv f_X(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}, \qquad x = 0, 1, 2, ...$$

for some $\lambda > 0$. To estimate λ by maximum likelihood, proceed as follows.

STEP 1 Calculate the likelihood function $L(\lambda)$.

$$L(\lambda) = \Pi_{i=1}^n f_X(x;\lambda) = \Pi_{i=1}^n \left[\frac{\lambda^x}{x!}e^{-\lambda}\right] = \frac{\lambda^{x_1+x_2+...+x_n}}{x_1!....x_n!}e^{-n\lambda}$$

for $\lambda \in \Theta = R^+$.

MAXIMUM LIKELIHOOD ESTIMATION - Examples

STEP 2 Calculate the log-likelihood function $logL(\lambda)$.

$$logL(\lambda) = \sum_{i=1}^{n} x_i \log \lambda - n\lambda - \sum_{i=1}^{n} log(x_i!)$$

.

STEP 3 Differentiate $logL(\lambda)$ with respect to λ, and equate the derivative to zero to find the MLE.

$$\frac{dlogL(\lambda)}{d\lambda} = 0 \Leftrightarrow \quad (1)$$

$$\sum_{i=1}^{n} \frac{x_i}{\lambda} - n = 0 \Leftrightarrow \quad (2)$$

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \quad (3)$$

Thus the maximum likelihood estimate of λ is $\hat{\lambda} = \bar{x}$.

MAXIMUM LIKELIHOOD ESTIMATION - Examples

STEP 4 Check that t the second derivative of the log-likelihood $logL(\lambda)$ is negative at $\lambda = \hat{\lambda}$.

$$\frac{d^2 logL(\theta)}{d\lambda^2} = -\frac{1}{\lambda}\sum_{i=1}^{n} x_i < 0 \quad \text{at } \lambda = \hat{\lambda}$$

.

MAXIMUM LIKELIHOOD ESTIMATION - Examples

The following data are the observed frequencies of occurrence of domestic accidents: we have n = 647 data as follows

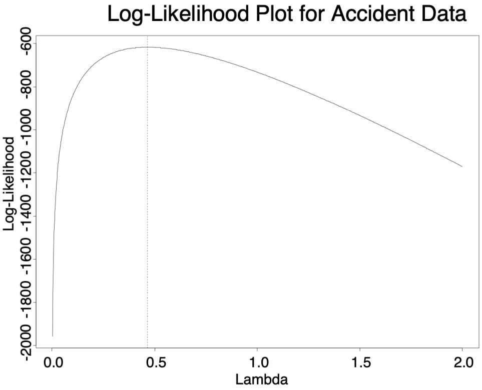
Number of accidents	Frequency
0	447
1	132
2	42
3	21
4	3
5	2

.

MAXIMUM LIKELIHOOD ESTIMATION - Examples

The estimate of λ if a Poisson model is assumed is:

$$\hat{\lambda} = \bar{x} = \frac{(447 * 0) + (132 * 1) + (42 * 2) + (21 * 3) + (3 + 4) + (2 * 5)}{647} = 0.465$$



The (Quasi) Maximum Likelihood Method

Likelihood function and the ML estimator

- Fisher presented the concept of maximum likelihood (ML) around 1925. Since then, this is the most popular estimation method in the time-series analysis because of its flexibility. The price for this flexibility is having to make an explicit distributional assumption.
- The ML estimator is obtained by maximizing the **likelihood function** of the data. If $x_t \sim iid \; f\left(x_t, \theta\right), \; \theta \in \Theta$, then the likelihood function is the joint density function of the data given θ, $i.e.$,

$$L(\theta; x_1, ..., x_T) = L(\theta; \mathbf{x}) = \prod_{t=1}^{T} f\left(x_t; \theta\right) \tag{4}$$

We define

$$\widehat{\theta}_{ML} : \arg\max_{\theta \in \Theta} L(\theta; \mathbf{x}) = \arg\max_{\theta \in \Theta} \prod_{t=1}^{T} f\left(x_t; \theta\right) \tag{5}$$

Likelihood function and the ML estimator

- Noting

$$\log\left(\prod_{i=1}^{T} a_i\right) = \sum_{i=1}^{T} \log\left(a_i\right),$$

then we define the **log-likelihood function** as

$$\mathcal{L}(\theta;\mathbf{x}) \equiv \log L(\theta;\mathbf{x}) = \sum_{t=1}^{T} \log f\left(x_t;\theta\right) \tag{6}$$

noting that

$$\widehat{\theta}_{ML} : \arg\max_{\theta\in\Theta} \mathcal{L}(\theta;\mathbf{x})) = \arg\max_{\theta\in\Theta} L(\theta;\mathbf{x}) \tag{7}$$

because the logarithmic function is a monotonic transformation and preserves the optimum.

Likelihood function and the ML estimator

Example

Let $\{x_t\}_{t=1}^{T}$ with $x_t \sim iid\mathcal{N}(\mu, \sigma^2)$ and $\theta = (\mu, \sigma^2)'$. The likelihood function for each observation is

$$f(x_t, \theta) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{(x_t - \mu)^2}{2\sigma^2}\right). \tag{8}$$

Therefore,

$$L(\mathbf{x};\theta) = \left(2\pi\sigma^2\right)^{-T/2} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\frac{(x_t - \mu)^2}{\sigma^2}\right) \tag{9}$$

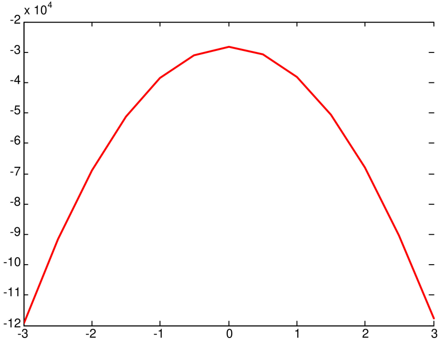
so the Gaussian log-likelihood is

$$\mathcal{L}(\mathbf{x};\theta) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\left(\sigma^2\right) - \frac{1}{2}\sum_{t=1}^{T}\left(\frac{x_t - \mu}{\sigma}\right)^2 \tag{10}$$

Likelihood function and the ML estimator

Example

Let $\{x_t\}_{t=1}^{T}$ with $x_t \sim iid\mathcal{N}(\mu, 1)$. We know $\sigma = 1$ and the (unknown) true mean is $\mu = 0$. The log-likelihood function for μ in the range is $[-3, 3]$ for a random sample with T=20,000 is shown below.



Likelihood function and the ML estimator

Example

Consider the AR(1) model with Gaussian innovations $Y_t = c + \rho Y_{t-1} + \varepsilon_t, \varepsilon_t \sim iid\mathcal{N}\left(0, \sigma^2\right)$. Since $\varepsilon_t = Y_t - c - \rho Y_{t-1}$, the log-likelihood of the AR(1) model can be written as

$$
\begin{aligned}
\mathcal{L}(\theta; \mathbf{y}_t) &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{t=2}^{T}\varepsilon_t^2 \\
&= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{t=2}^{T}\left(Y_t - c - \rho Y_{t-1}\right)^2
\end{aligned}
$$

Asymptotic properties

Under suitable regularity conditions, the CLT applies and $\widehat{\theta}_{ML}$ has the following asymptotic properties:

- **Asymptotically normality.** From the CLT,

$$\left(\widehat{\theta}_{ML} - \theta\right) \sim \mathcal{N}\left(0, \mathbf{V}_{\theta}\right) \tag{11}$$

where $\mathbf{V}_{\theta} < \infty$ is a well-defined matrix. Hence, we can carry out inference as

$$t_i = \frac{\left(\widehat{\theta}_{ML,i} - \theta_i\right)}{\sqrt{\left[\mathbf{V}_{\theta}\right]_{ii}}} \sim \mathcal{N}\left(0, 1\right)$$

Asymptotic properties

- **Efficiency.** If the model is correctly specified, and the regularity conditions hold, the covariance matrix $\mathbf{V}_\theta$ equals the inverse of the **information matrix** , *i.e.*, achieves the **Cramer-Rao bound**.

$$\mathbf{V}_\theta = \underbrace{\left[\underbrace{-E\left[\frac{\partial^2 \mathcal{L}(\mathbf{x};\theta)}{\partial\theta\partial\theta'}\right]}_{\text{Information Matrix}} \equiv \Omega_\theta \right]^{-1}}_{\text{Cramer-Rao bound}}$$

Asymptotic properties

- In general terms, we need to estimate the two matrices that define the covariance matrix in the limit. These matrices are determined numerically and provided by most statistical packages.
  1. **(Hessian matrix)**: $\mathbf{A}_\theta$ equals (minus) the expectation of the Hessian matrix. We can estimate this matrix consistently by its sample analog:

$$\mathbf{A}_\theta = -E\left(\frac{\partial \mathcal{L}(\mathbf{x}_t;\theta)}{\partial\theta\partial\theta'}\right) \Rightarrow \hat{\mathbf{A}}_{\theta T} = -\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \mathcal{L}(\mathbf{x}_t;\theta)}{\partial\theta\partial\theta'}\right)\Big|_{\theta=\hat{\theta}_{ML}} \tag{12}$$

     That is, the (numerical) Hessian evaluated at the estimated value.
  2. **(Outter product of the score vector)** $\mathbf{B}_\theta$ is the variance of the score vector, which has zero expectation. Hence, the sample analog of the covariance matrix is:

$$\hat{\mathbf{B}}_{\theta T} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial \mathcal{L}(\mathbf{x}_t;\theta)}{\partial\theta}\right)\left(\frac{\partial \mathcal{L}(\mathbf{x}_t;\theta)}{\partial\theta}\right)'\Big|_{\theta=\hat{\theta}_T} \tag{13}$$

Asymptotic properties

NOTE I. When the model is correctly specified, it can be shown that

$$\hat{\mathbf{A}}_{\theta T} \xrightarrow{p} \Omega_\theta^{-1}$$

and

$$\hat{\mathbf{B}}_{\theta T} \xrightarrow{p} \Omega_\theta^{-1},$$

where Ω_θ denotes the Information matrix. Hence, both estimators are asymptotically equivalent and hence

$$\mathbf{V}_\theta = \left[\hat{\mathbf{A}}_{\theta T}^{-1}\hat{\mathbf{B}}_{\theta T}\hat{\mathbf{A}}_{\theta T}^{-1}\right] \xrightarrow{p} \Omega_\theta^{-1}.$$

Because $\hat{\mathbf{B}}_{\theta T}\hat{\mathbf{A}}_{\theta T}^{-1} \xrightarrow{p} \mathbf{I}$, statistical packages estimate the covariance matrix on the basis of either the Hessian or the outter product, *e.g.*, $\hat{\mathbf{V}}_\theta = \hat{\mathbf{A}}_{\theta T}^{-1}$. However, it should be remarked once more that this approximation only holds when the specification is correctly specified.

QML estimation

- When the true distribution is NOT normal, then:
  - $\widehat{\theta}_{ML}$ is still consistent and asymptotically normally distributed,
  - $\widehat{\theta}_{ML}$ is no longer efficient, because it has a larger covariance matrix than the inverse of the information matrix. In particular,

$$\sqrt{T}\left(\hat{\theta}_{ML}-\theta\right) \xrightarrow{d} \mathcal{N}\left(0,\ \left[\mathbf{A}_{\theta}^{-1}\mathbf{B}_{\theta}\mathbf{A}_{\theta}^{-1}\right]\right) \tag{14}$$

- We can estimate consistently $\theta = \left(\theta_{\mu}', \theta_{\sigma}'\right)'$ by assuming normality **EVEN** if the true distribution is not normal. The resultant estimator is called the **Q**uasi- (or pseudo-) **M**aximum **L**ikelihood (QML) estimator: $\hat{\theta}_{QML}$.

QML estimation

Theorem

Under general regularity conditions, including the cases in which the analyst specifies the conditional mean of the model, $E(Y_t|\mathcal{F}) = \mu(X_t;\theta)$, and $Var(Y_t|\mathcal{F}) = \sigma^2(X_t;\theta)$, the quasi-maximum likelihood procedure yields a consistent estimator of θ_0, asymptotically distributed as a normal, if and only if the quasi-likelihood function is based on a probability density function family in the quadratic exponential class.

REMARK 1. The primary example of a PDF family encompassed by the quadratic exponential family is the normal distribution.

REMARK 2. This is a crucial theoretical result for many empirical applications: We can estimate parameters consistently through QML even if the true distribution is not normal. The ML and QML parameter estimates are the same, and only differ in the covariance matrix

$$\hat{\mathbf{V}}_{\theta,QML} = \left[\hat{\mathbf{A}}_{\theta T}^{-1}\hat{\mathbf{B}}_{\theta T}\hat{\mathbf{A}}_{\theta T}^{-1}\right].$$