6. MAXIMUM LIKELIHOOD ESTIMATION

[1] Maximum Likelihood Estimator

(1) Cases in which θ (unknown parameter) is scalar.

Notational Clarification:

- From now on, we denote the true value of θ as θ_o.
- Then, view θ as a variable.

Definition: (Likelihood function)

- Let $\{x_1, \ldots, x_T\}$ be a sample from a population.
- It does not have to be a random sample.
  - x_t is a scalar.
- Let $f(x_1,x_2, \ldots, x_T,\theta_o)$ be the joint density function of $x_1, \ldots, x_T$.
- The functional form of f is known, but not θ_o.
- Then, $L_T(\theta) \equiv f(x_1, \ldots, x_T, \theta)$ is called "likelihood function".
- $L_T(\theta)$ is a function of θ given $x_1, \ldots, x_T$.
- The functional form of f is known, but not θ_o.

Definition: (log-likelihood function)

$$l_T(\theta) = \ln[f(x_1, \ldots, x_T,\theta)].$$

Example:

- $\{x_1, \ldots, x_T\}$: a random sample from a population distributed with $f(x,\theta_o)$.
- $f(x_1, \ldots, x_T, \theta_o) = \prod_{t=1}^{T} f(x_t, \theta_o)$.

  $\rightarrow \quad L_T(\theta) = f(x_1, \ldots, x_T, \theta) = \prod_{t=1}^{T} f(x_t, \theta)$.

  $\rightarrow \quad l_T(\theta) = \ln\left(\prod_{t=1}^{T} f(x_t, \theta)\right) = \Sigma_t \ln f(x_t, \theta)$.

Definition: (Maximum Likelihood Estimator (MLE))

MLE $\hat{\theta}_{MLE}$ maximizes $l_T(\theta)$ given data points $x_1, \ldots, x_T$.

Example:

- $\{x_1, \ldots, x_T\}$ is a random sample from a population following a Poisson distribution [i.e., $f(x,\theta) = e^{-\theta}\theta^x/x!$ (suppressing subscript "o" from θ)].
- Note that $E(x) = var(x) = \theta_o$ for Poisson distribution.
- $l_T(\theta) = \Sigma_t \ln[f(x_t,\theta)] = -\theta T + (\ln(\theta))\Sigma_t x_t - \Sigma_t x_t!$
- FOC of max.: $\partial \ell_T / \partial \theta = -T + \frac{1}{\theta}\Sigma_t x_t = 0$.
- Solving this, $\hat{\theta}_{MLE} = \frac{\Sigma_t x_t}{T} = \bar{x}$.

(2) Extension to the Cases with Multiple Parameters.

Definition:

- $\theta = [\theta_1, \theta_2, \ldots, \theta_p]'$.
- $L_T(\theta) = f(x_1, \ldots, x_T, \theta) = f(x_1, \ldots, x_T, \theta_1, \ldots, \theta_p)$.
- $l_T(\theta) = \ln[f(x_1, \ldots, x_T, \theta) = \ln[f(x_1, \ldots, x_T, \theta_1, \ldots, \theta_p)]$.
- x_t could be a vector.
- If $\{x_1, \ldots, x_T\}$ is a random sample from a population with $f(x, \theta_o)$,

$$l_T(\theta) = \ln\left(\prod_{t=1}^{T} f(x_t, \theta)\right) = \Sigma_t \ln f(x_t, \theta).$$

Definition: (MLE)

MLE $\hat{\theta}_{MLE}$ maximizes $l_T(\theta)$ given data (vector) points $x_1, \ldots, x_T$. That is, $\hat{\theta}_{MLE}$ solves

$$\frac{\partial \ell_T(\theta)}{\partial \theta} = \begin{bmatrix} \partial \ell_T(\theta)/\partial \theta_1 \\ \partial \ell_T(\theta)/\partial \theta_2 \\ \vdots \\ \partial \ell_T(\theta)/\partial \theta_p \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{p \times 1}.$$

Example:

- Let $\{x_1, \ldots, x_T\}$ be a random sample from $N(\mu, \sigma^2)$ [suppressing subscript "o".]
- Since $\{x_1, \ldots, x_T\}$ is a random sample, $E(x_t) = \mu_o$ and $var(x_t) = \sigma_o^2$.
- Let $\theta = (\mu, v)'$, where $v = \sigma^2$.

- $$f(x_t,\theta)=\frac{1}{\sqrt{2\pi v}}\exp\left[-\frac{(x_t-\mu)^2}{2v}\right]$$
$$=(2\pi)^{-1/2}(v)^{-1/2}\exp\left[-\frac{(x_t-\mu)^2}{2v}\right].$$

- $\ln[f(x_t,\theta)]=-\frac{1}{2}\ln(2\pi)-\frac{1}{2}\ln(v)-\frac{(x_t-\mu)^2}{2v}$.

- $\ell_T(\theta)=-\frac{T}{2}\ln(2\pi)-\frac{T}{2}\ln(v)-\frac{\Sigma_t(x_t-\mu)^2}{2v}$.

- MLE solves FOC:

  (1) $\frac{\partial \ell_T(\theta)}{\partial \mu}=-\frac{1}{2v}\Sigma_t 2(x_t-\mu)(-1)=\frac{\Sigma_t(x_t-\mu)}{v}=0$;

  (2) $\frac{\partial \ell_T(\theta)}{\partial v}=-\frac{T}{2v}+\frac{\Sigma_t(x_t-\mu)^2}{2v^2}=0$.

- From (1):

  (3) $\Sigma_t(x_t-\mu)=0 \rightarrow \Sigma_t x_t - T\mu = 0 \rightarrow \hat{\mu}_{MLE}=\frac{\Sigma_t x_t}{T}=\bar{x}$.

- Substituting (3) in to (2):

  (4) $-Tv+\Sigma_t(x_t-\hat{\mu}_{MLE})^2=0 \rightarrow \hat{v}_{MLE}=\frac{1}{T}\Sigma_t(x_t-\bar{x})^2$.

- Thus,

$$\hat{\theta}_{MLE}=\begin{pmatrix}\hat{\mu}_{MLE}\\ \hat{v}_{MLE}\end{pmatrix}=\begin{pmatrix}\bar{x}\\ \frac{1}{T}\Sigma_t(x_t-\bar{x})^2\end{pmatrix}.$$

[2] Large Sample Properties of the ML estimator

Definition:

1) Let $g(\theta) = g(\theta_1, \ldots, \theta_p)$ be a scalar function of θ. Let $g_j = \partial\partial g/\partial\theta_j$. Then,

$$\frac{\partial g}{\partial \theta} = \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_p \end{pmatrix}.$$

2) Let $w(\theta) = (w_1(\theta), \ldots, w_m(\theta))'$ be a $m \times 1$ vector of functions of θ. Let $w_{ij} = \partial w_i(\theta)/\partial\theta_j$. Then,

$$\frac{\partial w(\theta)}{\partial \theta'} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1p} \\ w_{21} & w_{22} & \cdots & w_{2p} \\ \vdots & \vdots & & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mp} \end{bmatrix}_{m \times p}.$$

3) Let $g(\theta)$ be a scalar function of θ where $g_{ij} = \partial^2 g(\theta)/\partial\theta_i\partial\theta_j$. Then,

$$\frac{\partial^2 g(\theta)}{\partial\theta\partial\theta'} = \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1p} \\ g_{21} & g_{22} & \cdots & g_{2p} \\ \vdots & \vdots & & \vdots \\ g_{p1} & g_{p2} & \cdots & g_{pp} \end{pmatrix}_{p \times p}.$$

$\rightarrow$ Called Hessian matrix of $g(\theta)$.

Example 1: Let $g(\theta) = \theta_1^2 + \theta_2^2 + \theta_1\theta_2$. Find $\partial g(\theta)/\partial\theta$.

$$\frac{\partial g(\theta)}{\partial \theta} = \begin{pmatrix} 2\theta_1 + \theta_2 \\ 2\theta_2 + \theta_1 \end{pmatrix}.$$

Example 2: Let $w(\theta) = \begin{pmatrix} \theta_1^2 + \theta_2 \\ \theta_1 + \theta_2^2 \end{pmatrix}$.

$$\frac{\partial w(\theta)}{\partial \theta'} = \begin{pmatrix} 2\theta_1 & 1 \\ 1 & 2\theta_2 \end{pmatrix}.$$

Example 3: Let $g(\theta) = \theta_1^2 + \theta_2^2 + \theta_1\theta_2$. Find the Hessian matrix of $g(\theta)$.

$$\frac{\partial^2 g(\theta)}{\partial \theta \partial \theta'} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Some useful results:

1) c': $1\times p$, θ: $p\times 1$ ($c'\theta$ is a scalar) $\rightarrow$ $\partial(c'\theta)/\partial\theta = c$; $\partial(c'\theta)/\partial\theta' = c'$.

2) R: $m\times p$, θ: $p\times 1$ ($R\theta$ is $m\times 1$) $\rightarrow$ $\partial(R\theta)/\partial\theta = R$

3) A: $p\times p$ symmetric, θ: $p\times 1$ ($\theta'A\theta$)

$\rightarrow \partial(\theta'A\theta)/\partial\theta = 2A\theta$.

$\rightarrow \partial(\theta'A\theta)/\partial\theta' = 2\theta'A$

$\rightarrow \partial(\theta'A\theta)/\partial\theta\partial\theta' = 2A$.

Definition: (Hessian matrix of log-likelihood function)

$$H_T(\theta) = \frac{\partial^2 l_T}{\partial\theta\partial\theta'} = \left[\frac{\partial^2 l_T}{\partial\theta_i\partial\theta_j}\right]_{p\times p}.$$

Theorem:

Let $\hat{\theta}$ be MLE. Then, under suitable regularity conditions,

$\hat{\theta}$ is consistent, and,

$$\sqrt{T}(\hat{\theta}-\theta_o) \to_d N\left(0_{p\times 1}, \left[-p\lim_{T\to\infty}\frac{1}{T}H_T(\theta_o)\right]^{-1}\right).$$

Further, $\hat{\theta}$ is asymptotically efficient.

Implication:

$\hat{\theta} \approx N(\theta_o, [-H_T(\theta_o)]^{-1}) \to \hat{\theta} \approx N(\theta_o, [-H_T(\hat{\theta})]^{-1})$.

Example: $\{x_1, \ldots, x_T\}$ is a random sample from $N(\mu_o,\sigma_o^2)$.

Let $\theta = [\mu, v]'$ and $v = \sigma^2$.

$$l_T = -\frac{1}{2}\ln(\pi) - \frac{T}{2}\ln(v) - \frac{1}{2v}\Sigma_t(x_t-\mu)^2.$$

The first derivatives:

$$\frac{\partial l_T(\theta)}{\partial\mu} = \frac{\Sigma_t(x_t-\mu)}{v}; \quad \frac{\partial l_T(\theta)}{\partial v} = -\frac{T}{2v} + \frac{1}{2v^2}\Sigma_t(x_t-\mu)^2.$$

The second derivatives:

$$\frac{\partial^2 l_T(\theta)}{\partial\mu\partial\mu} = \frac{1}{v}\Sigma_t(-1) = -\frac{T}{v};$$

$$\frac{\partial^2 l_T(\theta)}{\partial\mu\partial v} = -\frac{\Sigma_t(x_t-\mu)}{v^2};$$

$$\frac{\partial^2 l_T(\theta)}{\partial v\partial v} = \frac{T}{2v^2} + \frac{0\times 2v^2 - 1\times 4v}{(2v^2)^2}\Sigma_t(x_t-\mu)^2 = \frac{T}{2v^2} - \frac{1}{v^3}\Sigma_t(x_t-\mu)^2$$

Therefore,

$$-H_T(\theta) = \begin{pmatrix} \frac{T}{v} & \frac{\Sigma_t(x_t-\mu)}{v^2} \\ \frac{\Sigma_t(x_t-\mu)}{v^2} & -\frac{T}{2v^2} + \frac{\Sigma_t(x_t-\mu)^2}{v^3} \end{pmatrix}$$

$$-H_T(\hat{\theta}_{ML}) = \begin{bmatrix} \frac{T}{\hat{v}_{ML}} & 0 \\ 0 & \frac{T}{2\hat{v}_{ML}^2} \end{bmatrix}.$$

Hence,

$$\hat{\theta} = \begin{pmatrix} \hat{\mu}_{ML} \\ \hat{v}_{ML} \end{pmatrix} \approx N\left(\begin{pmatrix} \mu_o \\ v_o \end{pmatrix}, \begin{pmatrix} \frac{\hat{v}_{ML}}{T} & 0 \\ 0 & \frac{2\hat{v}_{ML}^2}{T} \end{pmatrix} \right).$$

[3] Testing Hypotheses Based on MLE

General form of hypotheses:

- Let $w(\theta) = [w_1(\theta), w_2(\theta), \ldots, w_m(\theta)]'$, where $w_j(\theta) = w_j(\theta_1, \theta_2, \ldots, \theta_p)$ = a function of $\theta_1, \ldots, \theta_p$.
- H_o: The true θ (θ_o) satisfies the m restrcitions, $w(\theta) = 0_{m\times 1}$ ($m \leq p$).

Definition: (Restricted MLE)

Let $\tilde{\theta}$ be the restricted ML estimator which maximizes

$$l_T(\theta) \text{ s.t. } w(\theta) = 0.$$

Wald Test:

$$W_T = w(\hat{\theta})'[W(\hat{\theta})Cov(\hat{\theta})W(\hat{\theta})']^{-1}w(\hat{\theta}).$$

If $\hat{\theta}$ is a (unrestricted) ML estimator,

$$W_T = w(\hat{\theta})'[W(\hat{\theta})\{-H_T(\hat{\theta})\}^{-1}W(\hat{\theta})']^{-1}w(\hat{\theta}).$$

Note: Can be computed with any consistent estimator $\hat{\theta}$ and $Cov(\hat{\theta})$.

Likelihood Ratio Test: (LR)

$$LR_T = 2[l_T(\hat{\theta}) - l_T(\tilde{\theta})].$$

Lagrangean Multiplier (LM) test

Define $s_T(\theta) = \dfrac{\partial l_T(\theta)}{\partial \theta}$. Then, $LM_T = s_T(\tilde{\theta})'[-H_T(\tilde{\theta})]^{-1}s_T(\tilde{\theta})$.

Theorem:

Under H_o: $w(\theta) = 0$, W_T, LR_T, $LM_T \rightarrow_d \chi^2(m)$.

Implication:

- Given significance level (α), find a critical value from χ^2 table.
- Usually, $\alpha = 0.05$ or $\alpha = 0.01$.
- If $W_T > c$, reject H_o. Otherwise, do not reject H_o.

Comments:

1) Wald needs only $\hat{\theta}$; LR needs both $\hat{\theta}$ and $\tilde{\theta}$; and LM needs $\tilde{\theta}$ only.
2) In general, $W_T \geq LR_T \geq LM_T$.
3) W_T is not invariant to how to write restrictions. That is, W_T for H_o: $\theta_1 = \theta_2$ may not be equal to W_T for H_o: $\theta_1/\theta_2 = 1$.

Example:

(1) $\{x_1, \ldots, x_T\}$: RS from $N(\mu_o, v_o)$ with v_o known. So, $\theta = \mu$.

H_o: $\mu = 0$.

- $w(\mu) = \mu$.
- $l_T(\mu) = -(T/2)\ln(2\pi) - (T/2)\ln(v_o) - \{1/(2v_o)\}\Sigma_t(x_t-\mu)^2$.
- $s_T(\mu) = (1/v_o)\Sigma_t(x_t-\mu)$.
- $-H_T(\mu) = \dfrac{T}{v_o}$.

[Wald Test]

Unrestricted MLE:

- FOC: $\partial l_T(\mu)/\partial\mu = (1/v)\Sigma_t(x_t-\mu) = 0$.
- $\hat{\mu} = \bar{x}$.
- $W(\mu) = 1 \rightarrow W(\hat{\mu}) = 1$.
- $-H_T(\hat{\mu}) = T/v_o$.

[LR Test]

Restricted MLE: $\tilde{\mu} = 0$.

$$l_T(\hat{\mu}) = -(T/2)\ln(2\pi) - (T/2)\ln(v_o) - \{1/(2v_o)\}\Sigma_t(x_t-\bar{x})^2$$

$$l_T(\tilde{\mu}) = -(T/2)\ln(2\pi) - (T/2)\ln(v_o) - \{1/(2v_o)\}\Sigma_t x_t^2$$

[LM Test]

$s_T(\tilde{\mu}) = (1/v_o)\Sigma_t x_t = (T/v_o)\bar{x}$; $-H_T(\tilde{\mu}) = T/v_o$

With this information, can show that $W_T = LR_T = LM_T = \dfrac{T\bar{x}^2}{v_o}$.

(2) Both μ and v unknown: $\theta = (\mu, v)'$.

H_o: $\mu = 0$.

$\rightarrow w(\theta) = \mu$

$\rightarrow W(\theta) = \partial w(\theta)/\partial\theta' = [\partial\mu/\partial\mu, \partial\mu/\partial v] = [1, 0]$.

$\rightarrow l_T(\theta) = -(T/2)\ln(2\pi) - (T/2)\ln(v) - \{1/(2v)\}\Sigma_t(x_t-\mu)^2$

$$\rightarrow s_T(\theta) = \begin{pmatrix} \frac{1}{v}\Sigma_t(x_t - \mu) \\ -\frac{T}{2v} + \frac{1}{2v^2}\Sigma_t(x_t - \mu)^2 \end{pmatrix};$$

$$-H_T(\theta) = \begin{pmatrix} \frac{T}{v} & \frac{\Sigma_t(x_t - \mu)}{v^2} \\ \frac{\Sigma_t(x_t - \mu)}{v^2} & -\frac{T}{2v^2} + \frac{\Sigma_t(x_t - \mu)^2}{v^3} \end{pmatrix}.$$

$\rightarrow$ Unrestricted MLE: $\hat{\mu} = \bar{x}$ and $\hat{v} = \frac{1}{T}\Sigma_t(x_t - \bar{x})^2$.

$\rightarrow$ Restricted MLE: $\tilde{\mu} = 0$, but need to compute $\tilde{v}$.

$\rightarrow l_T(\tilde{\mu}, v) = -(T/2)\ln(2\pi) - (T/2)\ln(v) - \{1/(2v)\}\Sigma_t(x_t - \tilde{\mu})^2$

$\rightarrow l_T(0, v) = -(T/2)\ln(2\pi) - (T/2)\ln(v) - \{1/(2v)\}\Sigma_t x_t^2$

$\rightarrow$ FOC: $\partial l_T(0,v)/\partial v = -T/(2v) + (1/(2v^2))/\Sigma_t x_t^2 = 0$

$\rightarrow \tilde{v} = (1/T)\Sigma_t x_t^2$.

[Wald Test]

$$w(\hat{\theta}) = \hat{\mu} = \bar{x};\ W(\hat{\theta}) = \begin{pmatrix} 1 & 0 \end{pmatrix};\ -H_T(\hat{\theta}) = \begin{pmatrix} \frac{T}{\hat{v}} & 0 \\ 0 & \frac{T}{2\hat{v}^2} \end{pmatrix}.$$

$$\rightarrow \quad W_T = w(\hat{\theta})'[W(\hat{\theta})\{-H_T(\hat{\theta})\}^{-1}W(\hat{\theta})']^{-1}w(\hat{\theta}) = \frac{T\bar{x}^2}{\hat{v}}.$$

[LR Test]

$l_T(\hat{\theta}) = -(T/2)\ln(2\pi) - (T/2)\ln(\hat{v}) - \{1/(2\hat{v})\}\Sigma_t(x_t - \bar{x})^2$

$l_T(\tilde{\theta}) = -(T/2)\ln(2\pi) - (T/2)\ln(\tilde{v}) - \{1/(2\tilde{v})\}\Sigma_t x_t^2$

[LM Test]

$$s_T(\tilde{\theta}) = \begin{pmatrix} \frac{1}{\tilde{v}}\Sigma_t x_t \\ -\frac{T}{2\tilde{v}} + \frac{1}{2\tilde{v}^2}\Sigma_t x_t^2 \end{pmatrix} = \begin{pmatrix} \frac{T\bar{x}}{\tilde{v}} \\ -\frac{T}{2\tilde{v}} + \frac{T}{2\tilde{v}} \end{pmatrix} = \begin{pmatrix} \frac{T\bar{x}}{\tilde{v}} \\ 0 \end{pmatrix};$$

$$-H_T(\tilde{\theta}_{ML}) = \begin{pmatrix} \frac{T}{\tilde{v}} & \frac{\Sigma_t x_t}{\tilde{v}^2} \\ \frac{\Sigma_t x_t}{\tilde{v}^2} & \frac{T}{2\tilde{v}^2} \end{pmatrix}.$$

$$\rightarrow \quad \mathrm{LM_T} = s_T(\tilde{\theta})'[-H_T(\tilde{\theta})]^{-1} s_T(\tilde{\theta}) = \frac{T\bar{x}^2}{\tilde{v} - 2\bar{x}^2}.$$

[4] Efficiency of OLS estimator under Ideal Conditions

- Assume that y_t is iid $N(x_t.'\beta, v)$ conditional on $x_t.$.
- $f(y_t|x_t.,\beta,v) = \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{1}{2v}(y_t - x_t.'\beta)^2 \right)$.

$$l_T(\beta, v) = \Sigma_t \ln f(y_t \mid \beta, v, x_t.)$$

- $$= -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln v - \frac{1}{2v}\Sigma_t (y_t - x_t.'\beta)^2 \quad .$$

$$= -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln v - \frac{1}{2v}(y - X\beta)'(y - X\beta)$$

- Therefore, we have the following likelihood function of y.
- FOC:

(i) $\partial l_T(\beta,v)/\partial\beta = -(1/2v)[-2X'y + 2X'X\beta] = 0_{k\times 1}$.

(ii) $\partial l_T(\beta,v)/\partial v = -(T/2v) + (1/2v^2)(y-X\beta)'(y-X\beta) = 0$.

- From (i),

$$X'y - X'X\beta = 0_{k\times 1} \rightarrow \quad \hat{\beta}_{MLE} = (X'X)^{-1}X'y = \hat{\beta}.$$

From (ii), $\hat{v}_{MLE} = SSE/T$.

- Thus, we can conclude that $\hat{\beta}$ and $s^2 = SSE/(T-k)$ are asymptotically efficient.