University of Pavia

Maximum Likelihood Asymptotic Theory

Eduardo Rossi

The MLE is an implicit function of the random sample. MLE is not a function of sample averages of the data.

But the sample log-likelihood is a sum of i.i.d. random variables. Because the $(U_t, V_t) \sim i.i.d.$ so are any such transformations as the $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}; U_t|V_t), t = 1, 2, \ldots, N$. The LLN can apply to the sample average log-likelihood function itself

$$E_N[L(\boldsymbol{\theta})] \xrightarrow{p} E[L(\boldsymbol{\theta})]$$

for any fixed $\boldsymbol{\theta}$.

Under the assumptions

1. Distribution
2. Dominance
3. Global Identification
4. Compactness of Θ

The MLE is consistent

$$\widehat{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_0$$

- The sample average log-likelihood converges to the expected log-likelihood for any value of $\boldsymbol{\theta}$:

$$E_N[L(\boldsymbol{\theta})] \xrightarrow{p} E[L(\boldsymbol{\theta})]$$

-

$$\widehat{\boldsymbol{\theta}}_N = \arg\max_{\boldsymbol{\theta}\in\Theta} E_N[L(\boldsymbol{\theta})] \qquad \text{by construction}$$
$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\Theta} E[L(\boldsymbol{\theta})] \qquad \text{by strict log-likelihood inequality}$$

As a result, $\widehat{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_0$, provided that the relationships are continuous.

The argument of $\arg\max_{\boldsymbol{\theta}\in\Theta}$ is a function of $\boldsymbol{\theta}$, $E_N[L(\boldsymbol{\theta})]$.
$\arg\max_{\boldsymbol{\theta}\in\Theta}$ must be a continuous function of its functional argument.

We use the *Uniform Convergence in Probability* in order to define the probability limit of a sequence of random functions.

Uniform LLN. $g(\boldsymbol{\theta}, U)$ continuous function over $\boldsymbol{\theta} \in \Theta$, where $\Theta \subset \mathbb{R}^K$ is closed and bounded, $\{U_t\}$ is a sequence of i.i.d. r.v. with c.d.f. $F_U(u)$. If $E[sup_{\boldsymbol{\theta}\in\Theta}||g(\boldsymbol{\theta}; U)||]$ exists, then

1. $E[g(\boldsymbol{\theta}; U)]$ is continuous over $\boldsymbol{\theta} \in \Theta$
2. $E_N[g(\boldsymbol{\theta}; U)] \xrightarrow{p} E[g(\boldsymbol{\theta}; U)]$

We apply the uniform LLN to the sample average log-likelihood.

Consistency of Maxima. If there is a sequence of functions $Q_N(\boldsymbol{\theta})$ that converges in probability uniformly to a function $Q_0(\boldsymbol{\theta})$ on the closed and bounded Θ and if $Q_0(\boldsymbol{\theta})$ is continuous and uniquely maximized at $\boldsymbol{\theta}_0$, then

$$\widehat{\boldsymbol{\theta}}_N = \arg\max_{\boldsymbol{\theta}\in\Theta} Q_N(\boldsymbol{\theta}) \xrightarrow{p} \boldsymbol{\theta}_0$$

Compactness and differentiability guarantee that $E_N[L(\boldsymbol{\theta})]$ has a maximum.

- Differentiability implies continuity of $L(\boldsymbol{\theta})$
- Compactness of Θ.
- (U_t, V_t) are i.i.d. with c.d.f. $F_{U|V}(u|v; \boldsymbol{\theta})$
- Dominance states that $E[\sup_{\boldsymbol{\theta}\in\Theta} |L(\boldsymbol{\theta})|]$ exists

Then $E[L(\boldsymbol{\theta})]$ is continuous and

$$E_N[L(\boldsymbol{\theta})] \xrightarrow{p} E[L(\boldsymbol{\theta})]$$

uniformly.

For the Consistency of Maxima, $Q_N(\boldsymbol{\theta}) = E_N[L(\boldsymbol{\theta})]$ and $Q_0(\boldsymbol{\theta}) = E[L(\boldsymbol{\theta})]$.

Under the assumptions:

- From *Likelihood Identification*: if $\forall \boldsymbol{\theta}_1 \in \Theta$, $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_1$ implies $Pr\{L(\boldsymbol{\theta}_0) \neq L(\boldsymbol{\theta}_1)\} > 0$
- we have the *Strict Expected Log-likelihood Inequality*: $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ implies

$$E[L(\boldsymbol{\theta})] < E[L(\boldsymbol{\theta}_0)]$$

Hence $E[L(\boldsymbol{\theta})]$ is uniquely maximized at $\boldsymbol{\theta}_0$. Therefore

$$\widehat{\boldsymbol{\theta}}_N = \arg\max_{\boldsymbol{\theta} \in \Theta} E_N[L(\boldsymbol{\theta})] \xrightarrow{p} \boldsymbol{\theta}_0 = \arg\max E[L(\boldsymbol{\theta})]$$

Assumption: There is an open subset of Θ that contains the population parameter value $\boldsymbol{\theta}_0$.

$\boldsymbol{\theta}_0$ is not on the boundary of Θ.

Assumption:

$$E_N[L_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_N)] = \mathbf{0}$$

the MLE solves the normal equations.

First-order Taylor series expansion:

$$E_N[L_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_N)] = \mathbf{0} = E_N[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] + E_N[L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}_N)](\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)$$

$$\overline{\boldsymbol{\theta}}_N = \alpha_N \widehat{\boldsymbol{\theta}}_N + (1 - \alpha_N)\boldsymbol{\theta}_0 \qquad \alpha_N \in [0, 1]$$

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) = \{-E_N[L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}_N)]\}^{-1}\sqrt{N}E_N[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)]$$

- $\sqrt{N}E_N[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] \xrightarrow{d} N(\mathbf{0}, \mathfrak{I}(\boldsymbol{\theta}_0))$ (by CLT)
- $E_N[L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}_N)] \xrightarrow{p} -\mathfrak{I}(\boldsymbol{\theta}_0)$ (by LLN)

then, $\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathfrak{I}(\boldsymbol{\theta}_0)^{-1})$

$$\sqrt{N} E_N[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] \xrightarrow{d} N(\mathbf{0}, \mathfrak{I}(\boldsymbol{\theta}_0))$$

Proof. Since

$$\sqrt{N} E_N[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] = \sqrt{N} \frac{1}{N} \left. \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \frac{1}{\sqrt{N}} \sum_{t=1}^{N} L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0; U_t | V_t)$$

is a sum of i.i.d. r.v.. Under the assumptions of i.i.d. (U_t, V_t) and differentiability, $E[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)|V = v] = \mathbf{0}$ so that

$$E[\mathbf{c}' L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] = \mathbf{0} \qquad \mathbf{c} \in \mathbb{R}^K$$

$$Var[\mathbf{c}' L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] = \mathbf{c}' \mathfrak{I}(\boldsymbol{\theta}_0) \mathbf{c} \qquad \text{exists} \qquad \forall \mathbf{c} \in \mathbb{R}^K$$

The CLT implies

$$\sqrt{N}E_N[\mathbf{c}'L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] \xrightarrow{d} N(\mathbf{0}, \mathbf{c}'\mathfrak{I}(\boldsymbol{\theta}_0)\mathbf{c})$$

by Cramér-Wold device:

$$\sqrt{N}E_N[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] \xrightarrow{d} N(\mathbf{0}, \mathfrak{I}(\boldsymbol{\theta}_0))$$

$$E_N[L_{\boldsymbol{\theta\theta}}(\overline{\boldsymbol{\theta}}_N)] \xrightarrow{p} -\mathfrak{I}(\boldsymbol{\theta}_0)$$

where $\overline{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_0$.

While $\overline{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_0$ implies $g(\overline{\boldsymbol{\theta}}_N) \xrightarrow{p} g(\boldsymbol{\theta}_0)$ for continuous g. g is a function of the random sample, therefore is random.

Lemma

1. $g_N(\boldsymbol{\theta}) \xrightarrow{p} g_0(\boldsymbol{\theta}), \ \forall \boldsymbol{\theta} \in \Theta$, closed and bounded subset of $\mathbb{R}^K$.
2. $g_0(\boldsymbol{\theta})$ is continuous
3. $\boldsymbol{\theta}_N \xrightarrow{p} \boldsymbol{\theta}_0 \in \Theta$

then

$$g_N(\boldsymbol{\theta}_N) \xrightarrow{p} g_0(\boldsymbol{\theta}_0)$$

Assumption. (Dominance II) $E[\sup_{\boldsymbol{\theta}\in\Theta}|L_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta})|]$ exists.

Lemma. Under the assumptions

1. Differentiability
2. Finite Information: $Var[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)]$ exists.
3. Compactness: Θ is bounded and closed.
4. Dominance II: $E[\sup_{\boldsymbol{\theta}\in\Theta}|L_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta})|]$ exists.

$$E_N\left[-L_{\boldsymbol{\theta\theta}}(\overline{\boldsymbol{\theta}}_N)\right] \xrightarrow{p} \mathfrak{I}(\boldsymbol{\theta}_0)$$

Proof

1. To establish that $E_N[L_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta})] \xrightarrow{p} E[L_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta})]$ uniformly. We note that Differentiability implies that $L_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta})$ is continuous. (U_t, V_t) are i.i.d. Θ is compact. $E[\sup_{\boldsymbol{\theta}\in\Theta} |L_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta})|]$ exists. Then we can invoke the uniform LLN: $E_N[L_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta})] \xrightarrow{p} E[L_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta})]$ uniformly in $\boldsymbol{\theta} \in \Theta$.

2. $\overline{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_0$. Applying the Lemma $E_N[L_{\boldsymbol{\theta\theta}}(\overline{\boldsymbol{\theta}}_N)] \xrightarrow{p} E[L_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta}_0)]$. The assumptions of *Distribution, Differentiability and Finite Information* are met so $E[L_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta}_0)] = -\mathfrak{I}(\boldsymbol{\theta}_0)$. Then

$$E_N[L_{\boldsymbol{\theta\theta}}(\overline{\boldsymbol{\theta}}_N)] \xrightarrow{p} -\mathfrak{I}(\boldsymbol{\theta}_0)$$

Now,

$$\begin{aligned}\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) &= \{-E_N[L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}_N)]\}^{-1}\sqrt{N}E_N[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] \\ &\xrightarrow{d} \mathfrak{I}(\boldsymbol{\theta}_0)^{-1}N[\mathbf{0}, \mathfrak{I}(\boldsymbol{\theta}_0)] \\ &\sim N(\mathbf{0}, \mathfrak{I}(\boldsymbol{\theta}_0)^{-1}) \qquad \text{by Slutsky}\end{aligned}$$

Square roots of nonsingular matrices are continuous functions of the elements of the matrix

$$\{-E_N[L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_N)]\}^{1/2} \xrightarrow{p} \mathfrak{I}(\boldsymbol{\theta}_0)^{1/2}$$

Then by Slutsky's Theorem

$$\{-E_N[L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_N)]\}^{1/2}\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \xrightarrow{d} \mathfrak{I}(\boldsymbol{\theta}_0)^{1/2}N[\mathbf{0}, \mathfrak{I}(\boldsymbol{\theta}_0)^{-1}] \sim N[\mathbf{0}, \mathbf{I}_K].$$

Three consistent estimators of the $\Im(\boldsymbol{\theta}_0)$:

- $E_N[-L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_N)]$
- $Var_N[L_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_N)]$
- $E_N[\Im(\widehat{\boldsymbol{\theta}}_N)]$

To use the empirical information estimator the population information function $\Im(\boldsymbol{\theta})$ is known only when the log-likelihood is unconditional.

When the log-likelihood is conditional then the conditional information function is

$$\Im(\boldsymbol{\theta}_0|V) = E[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0; U|V)L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0; U|V)'|V]$$

because the marginal distribution of V is unspecified $\Im(\boldsymbol{\theta})$ is unknown.

The unconditional information matrix is obtained using the LIE:

$$\mathfrak{I}(\boldsymbol{\theta}_0) = E[\mathfrak{I}(\boldsymbol{\theta}_0|V)]$$

the empirical information matrix estimator is $E_N[\mathfrak{I}(\widehat{\boldsymbol{\theta}}_N|V)]$.

Consistency of the estimators

$\widehat{\mathfrak{I}}(\boldsymbol{\theta}) \xrightarrow{p} \mathfrak{I}(\boldsymbol{\theta})$ uniformly.

Because $\widehat{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_0$

$$\widehat{\mathfrak{I}}(\widehat{\boldsymbol{\theta}}_N) \xrightarrow{p} \mathfrak{I}(\boldsymbol{\theta}_0)$$

since the inverse is a continuous function

$$\widehat{\mathfrak{I}}(\widehat{\boldsymbol{\theta}}_N)^{-1} \xrightarrow{p} \mathfrak{I}(\boldsymbol{\theta}_0)^{-1}.$$

The asymptotic distribution of MLE has no bias and the variance matrix of its asymptotic distribution equals the Cramér-Rao Lower Bound.

The MLE and the efficient (but infeasible) Cramer-Rao estimator are asymptotically equivalent:

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) - \sqrt{N}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0) = \sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*) \xrightarrow{p} \mathbf{0}$$

because

$$-\{E_N[L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}_N)]\}^{-1} \xrightarrow{p} \mathfrak{I}(\boldsymbol{\theta}_0)^{-1}$$

$$\sqrt{N}E_N[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] \xrightarrow{d} N(\mathbf{0}, \mathfrak{I}(\boldsymbol{\theta}_0))$$

$$\boldsymbol{\theta}^* \equiv \boldsymbol{\theta}_0 + [\mathfrak{I}(\boldsymbol{\theta}_0)]^{-1} E_N[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)]$$

$$\begin{aligned}
\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) - \sqrt{N}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0) &= \sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*) \\
&= \sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) - \sqrt{N}\{E_N[\mathfrak{I}(\boldsymbol{\theta}_0)]\}^{-1} \times \\
& E_N[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] \\
&= \left[-\{E_N[L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}_N)]\}^{-1} - \mathfrak{I}(\boldsymbol{\theta}_0)^{-1}\right] \times \\
& \sqrt{N} E_N[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] \\
& \xrightarrow{p} \mathbf{0}
\end{aligned}$$

using the equivalence of convergence in distribution to a constant and convergence in probability.

ML estimator is invariant to nonsingular transformations of the parameters. If $\boldsymbol{\gamma} = g(\boldsymbol{\theta})$ is a one-to-one reparameterization then the MLE for $\boldsymbol{\gamma}$ is

$$\widehat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma} \in \Gamma} E_N\{L[g^{-1}(\boldsymbol{\gamma})]\}$$

where Γ is the parameter space $\{\boldsymbol{\gamma} = g(\boldsymbol{\theta}) | \boldsymbol{\theta} \in \Theta\}$. Because the reparameterization is one-to-one

$$\max_{\boldsymbol{\gamma} \in \Gamma} E_N\{L[g^{-1}(\boldsymbol{\gamma})]\} = \max_{\boldsymbol{\theta} \in \Theta} E_N[L(\boldsymbol{\theta})]$$

$$\widehat{\boldsymbol{\gamma}} = g(\widehat{\boldsymbol{\theta}})$$

Invariance: Reparameterization does not alter the location of the MLE.

For finding the asymptotic distribution of a trasformation. Given a consistent estimator of the approximate variance of $\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)$, $\widehat{\boldsymbol{\Omega}}$, the approximate variance of

$$\sqrt{N}(g(\widehat{\boldsymbol{\theta}}_N) - g(\boldsymbol{\theta}_0))$$

is

$$g_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_N)\widehat{\boldsymbol{\Omega}}g_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_N)'.$$

Delta Method: If $\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega})$ and $g(\boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta}_0$ then

$$\sqrt{N}(g(\widehat{\boldsymbol{\theta}}_N) - g(\boldsymbol{\theta}_0)) \xrightarrow{d} N(\mathbf{0}, \mathbf{J}_0\boldsymbol{\Omega}\mathbf{J}_0')$$

$$\mathbf{J}_0 \equiv \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \quad \mathbf{J}_0 \equiv \mathbf{J}(\boldsymbol{\theta}_0).$$

Proof: Expanding $g(\widehat{\boldsymbol{\theta}}_N)$ around $\boldsymbol{\theta}_0$ we obtain

$$\sqrt{N}(g(\widehat{\boldsymbol{\theta}}_N) - g(\boldsymbol{\theta}_0)) = \mathbf{J}(\overline{\boldsymbol{\theta}}_N)\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)$$

for some

$$\overline{\boldsymbol{\theta}}_N = \alpha_N \widehat{\boldsymbol{\theta}}_N + (1 - \alpha_N)\boldsymbol{\theta}_0 \qquad \alpha_N \in [0, 1]$$

$$\overline{\boldsymbol{\theta}}_N \xrightarrow{\text{p}} \boldsymbol{\theta}_0$$

$$\mathbf{J}(\overline{\boldsymbol{\theta}}_N) \xrightarrow{\text{p}} \mathbf{J}_0 \quad \text{(Slutsky's Theorem)}$$

$$\sqrt{N}(g(\widehat{\boldsymbol{\theta}}_N) - g(\boldsymbol{\theta}_0)) \xrightarrow{\text{d}} \mathbf{J}_0 N(\mathbf{0}, \boldsymbol{\Omega}) \sim N(\mathbf{0}, \mathbf{J}_0 \boldsymbol{\Omega} \mathbf{J}_0').$$

The empirical expectation of the log-likelihood

$$E_N[L(\boldsymbol{\theta})] = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{E_N[(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2]}{2\sigma^2}$$
$$= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/N}{2\sigma^2}$$

The log-lik is differentiable. F.O.C's:

$$E_N[L_{\boldsymbol{\beta}}(\boldsymbol{\theta})] = \frac{1}{\sigma^2}E_N[\mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta})]$$
$$= \frac{1}{N\sigma^2}[\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]$$

$$E_N[L_{\sigma^2}(\boldsymbol{\theta})] = -\frac{1}{2\sigma^4}\{\sigma^2 - E_N[(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2]\}$$
$$= -\frac{1}{2\sigma^4}\left[\sigma^2 - \frac{1}{N}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]$$

Solutions:

$$\frac{1}{N\sigma^2}[\mathbf{X}'(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})] = 0$$
$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\widehat{\sigma}^2 = \frac{1}{N}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

The Hessian matrix:

$$E_N[L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta})] = \begin{bmatrix} -\frac{1}{\sigma^2 N}\mathbf{X}'\mathbf{X} & -\frac{\mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^4 N} \\ -\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{X}}{\sigma^4 N} & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6 N}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \end{bmatrix}$$

$$E_N[L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_N)] = \begin{bmatrix} -\frac{1}{\widehat{\sigma}^2 N}\mathbf{X}'\mathbf{X} & -\frac{\mathbf{X}'(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})}{\widehat{\sigma}^4 N} \\ -\frac{(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})'\mathbf{X}}{\widehat{\sigma}^4 N} & \frac{1}{2\widehat{\sigma}^4} - \frac{1}{\widehat{\sigma}^6 N}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{1}{\widehat{\sigma}^2 N}\mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2\widehat{\sigma}^4} - \frac{1}{\widehat{\sigma}^6 N}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) \end{bmatrix}$$

which is negative definite.

The MLE of σ^2 is

$$\widehat{\sigma}_N^2 = \frac{\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}}{N} = \frac{N-K}{N}s^2$$

The **second-order necessary condition** for a point to be the local maximum of a twice continuously differentiable function is that the Hessian be negative semidefinite at the point.

$$E_N[-L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_N)] = \begin{bmatrix} \frac{1}{\widehat{\sigma}_N^2 N}\mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2\widehat{\sigma}_N^4} \end{bmatrix} = E_N[\Im(\widehat{\boldsymbol{\theta}}_N|\mathbf{x}_t)]$$

$$Var_N[L_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_N)] = \begin{bmatrix} \frac{1}{\widehat{\sigma}_N^4}E_N[\mathbf{x}_t(y_t-\mathbf{x}_t'\widehat{\boldsymbol{\beta}}_N)^2\mathbf{x}_t'] & \frac{1}{2\widehat{\sigma}_N^6}E_N[\mathbf{x}_t(y_t-\mathbf{x}_t'\widehat{\boldsymbol{\beta}}_N)^3] \\ \cdot & \frac{1}{\widehat{\sigma}_N^8}\{E_N[(y_t-\mathbf{x}_t'\widehat{\boldsymbol{\beta}}_N)^4]-\widehat{\sigma}_N^4\} \end{bmatrix}$$

$$\begin{bmatrix} \frac{1}{\widehat{\sigma}_N}\left(\frac{\mathbf{X}'\mathbf{X}}{N}\right)^{1/2}\sqrt{N}(\widehat{\boldsymbol{\beta}}_N-\boldsymbol{\beta}) \\ \frac{1}{\sqrt{2\widehat{\sigma}_N^4}}\sqrt{N}(\widehat{\sigma}_N^2-\sigma^2) \end{bmatrix} \xrightarrow{\text{d}} N(\mathbf{0},\mathbf{I}_{K+1})$$

and

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_N-\boldsymbol{\beta}) \xrightarrow{\text{d}} N(\mathbf{0},\sigma^2 E(\mathbf{x}_t\mathbf{x}_t')^{-1})$$