



University of Pavia

Maximum Likelihood Estimation

Eduardo Rossi



LIKELIHOOD FUNCTION

Choosing parameter values that make what one has observed more likely to occur than any other parameter values do.

Assumption(Distribution) The pair $\{U, V\}$ is a random variable and the N variables

$$\{(U_1, V_1), \dots, (U_N, V_N)\}$$

are i.i.d. random sample of (U, V) .

$F_{U|V}(u|v; \theta_0)$ is completely known but θ_0 (true value of the real-valued parameter vector) is unknown, $\theta \in \mathbb{R}^K$.

Support of $F_{U|V}$ is $\mathbb{S}(\theta_0)$

$$\int_{\mathbb{S}(\theta_0)} dF_{U|V}(u|v; \theta_0) = 1 = \begin{cases} \sum_{u \in \mathbb{S}(\theta_0)} f(u|v; \theta_0) & \text{if } U \text{ discrete} \\ \int_{\mathbb{S}(\theta_0)} f(u|v; \theta_0) du & \text{if } U \text{ continuous} \end{cases}$$



LIKELIHOOD FUNCTION

Probability function for $(U_1, \dots, U_N) | (V_1, \dots, V_N)$

$$\prod_{t=1}^N f(u_t | v_t; \boldsymbol{\theta}_0)$$

Normal Linear Regression: $y_t = \mathbf{x}'_t \boldsymbol{\beta}_0 + \epsilon_t$, (y_t, \mathbf{x}_t) i.i.d. normal

$$u_t = y_t, \quad v_t = \mathbf{x}_t$$

$$f(u_t | v_t; \boldsymbol{\theta}_0) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp \left[-\frac{(y_t - \mathbf{x}'_t \boldsymbol{\beta}_0)^2}{2\sigma_0^2} \right]$$

$\mathbb{S}(\boldsymbol{\theta}_0) = \mathbb{R}$. Since the obs are i.i.d. normal. The conditional p.d.f. of the sample is

$$\prod_{t=1}^N f(u_t | v_t; \boldsymbol{\theta}_0) = [2\pi\sigma_0^2]^{-N/2} \exp \left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)}{2\sigma_0^2} \right]$$



The marginal distribution of \mathbf{x}_t does not depend on $\boldsymbol{\theta}_0$.

Student's t Linear Regression

$$\frac{y_t - \mathbf{x}'_t \boldsymbol{\beta}_0}{\sigma_0} | \mathbf{x}_t \sim t_{\nu_0}$$

$$f(y_t | \mathbf{x}_t; \boldsymbol{\theta}_0) = \frac{\Gamma[(\nu_0 + 1)/2]}{\Gamma(\nu_0/2)} \frac{1}{\sqrt{\pi \nu_0 \sigma_0^2}} \left[1 + \frac{(y_t - \mathbf{x}'_t \boldsymbol{\beta}_0)^2}{\nu_0 \sigma_0^2} \right]^{-(\nu_0 + 1)/2}$$



Laplace Linear Regression

$$f(u_t|v_t; \sigma_0^2) = \frac{1}{\sqrt{2\sigma_0^2}} \exp -\sqrt{2} \frac{|y_t - \mathbf{x}'_t \boldsymbol{\beta}_0|}{\sigma_0}$$

$$U = y_t, V = \mathbf{x}_t, \mathbb{S}(\boldsymbol{\theta}_0) = \mathbb{R}, \boldsymbol{\theta}_0 = [\boldsymbol{\beta}'_0, \sigma_0^2]'$$

We can obtain

$$h(\boldsymbol{\theta}_0) \equiv E[g(u)] = \int g(u) dF(u; \boldsymbol{\theta}_0)$$

$$h(v; \boldsymbol{\theta}_0) \equiv E[g(U, V)|V = v] = \int g(u, v) dF(u|v; \boldsymbol{\theta}_0)$$



THE LIKELIHOOD FUNCTION

Unconditional specification: $f(u; \boldsymbol{\theta})$ describes the likely values of every r.v. $U_t, t = 1, 2, \dots, N$ for a specific value of $\boldsymbol{\theta}$.

The sample likelihood function treats the u argument as given and $\boldsymbol{\theta}_0$ as variable.

It describes the likely values of the unknown $\boldsymbol{\theta}_0$ given the realizations of the r.v. U .

The likelihood function of $\boldsymbol{\theta}$ for a random variable U with p.f.

$f(u; \boldsymbol{\theta}_0)$ is defined to be

$$l(\boldsymbol{\theta}; U) = f(u; \boldsymbol{\theta})$$

$$L(\boldsymbol{\theta}; U) = \log l(\boldsymbol{\theta}; U)$$



THE LIKELIHOOD FUNCTION

Likelihood function: we evaluate the p.f. at a random variable and consider the result as a function of the variable θ :

$$\begin{aligned} L(\theta; U_1, \dots, U_N) &= \log \left[\prod_{t=1}^N f(U_t; \theta) \right] \\ &= \sum_{t=1}^N L(\theta; U_t) \end{aligned}$$

The *conditional likelihood function* of θ for a r.v. U with p.f. $f(u|v; \theta_0)$ given the r.v. V is

$$\begin{aligned} l(\theta, U|V) &= f(u|v; \theta) \\ L(\theta; U|V) &= \log l(\theta; U|V) \end{aligned}$$

$\theta_0 \in \Theta$, Θ parameter space, the set of permitted values of the model.



Assumption (Dominance condition)

$$E \left[\sup_{\boldsymbol{\theta} \in \Theta} |L(\boldsymbol{\theta}; U|V)| \right] \text{ exists.}$$

This means that $|L(\boldsymbol{\theta}; U|V)|$ is dominated by

$$h(U, V) \equiv \sup_{\boldsymbol{\theta} \in \Theta} |L(\boldsymbol{\theta}; U|V)|$$

where $h(U, V)$ does not depend on $\boldsymbol{\theta}$.

The existence of $E[h(U)]$ implies the existence of $E[L(\boldsymbol{\theta}; U|V)]$, $\boldsymbol{\theta} \in \Theta$.

Lemma. If $L(\boldsymbol{\theta}; U|V)$ is the conditional log-likelihood for $\boldsymbol{\theta}$, the *Dominance* condition holds, then

$$E [L(\boldsymbol{\theta}; U|V)|V] \leq E[L(\boldsymbol{\theta}_0; U|V)|V].$$



EXPECTED LOG-LIKELIHOOD INEQUALITY

Unconditional case:

$$E[L(\boldsymbol{\theta}_0; U)] \geq E[L(\boldsymbol{\theta}; U)]$$

The specification of p.f. of U determines expected values of functions of U .

Therefore

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \equiv E[L(\boldsymbol{\theta}; U)]$$

which depends on $\boldsymbol{\theta}$ because the L does and depends on $\boldsymbol{\theta}_0$ because Q is the expected value of a function of U . The expected loglikelihood inequality states that

$$Q(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = \max_{\boldsymbol{\theta} \in \Theta} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$$



NORMAL LINEAR REGRESSION MODEL

$$y_t | \mathbf{x}_t \sim N(\mathbf{x}'_t \boldsymbol{\beta}_0, \sigma_0^2)$$

$$\begin{aligned} E [L(\boldsymbol{\theta}, y_t | \mathbf{x}_t) | \mathbf{x}_t] &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{E[(y_t - \mathbf{x}'_t \boldsymbol{\beta})^2 | \mathbf{x}_t]}{2\sigma^2} \\ &= -\frac{1}{2} \log(2\pi\sigma^2) + \\ &\quad - \frac{1}{2} \frac{E[(y_t - \mathbf{x}'_t \boldsymbol{\beta}_0 + \mathbf{x}'_t \boldsymbol{\beta}_0 - \mathbf{x}'_t \boldsymbol{\beta})^2 | \mathbf{x}_t]}{\sigma^2} \\ &= -\frac{1}{2} \left[\log(2\pi\sigma^2) + \frac{\sigma_0^2 + (\mathbf{x}'_t \boldsymbol{\beta} - \mathbf{x}'_t \boldsymbol{\beta}_0)^2}{\sigma^2} \right] \end{aligned}$$

which is uniquely maximized at $\mathbf{x}'_t \boldsymbol{\beta} = \mathbf{x}'_t \boldsymbol{\beta}_0$ and $\sigma^2 = \sigma_0^2$.



The conditional expectation of the conditional log-likelihood of the entire sample is the sum of such terms

$$E [L(\boldsymbol{\theta}; \mathbf{y}|\mathbf{X})|\mathbf{X}] = -\frac{N}{2} \log (2\pi\sigma^2) - \frac{N\sigma_0^2 + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \mathbf{X}' \mathbf{X} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)}{2\sigma^2}$$

which is uniquely maximized at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}_0$ and $\sigma^2 = \sigma_0^2$ if \mathbf{X} is full-column rank.



The expected log-likelihood is analytically intractable. We show that $E[L(\boldsymbol{\theta}; U|V)]$ exists, for $\nu_0 > 2$, because the concavity of the logarithmic function

$$\log(1 + z^2) \leq z^2$$

$$\begin{aligned} E \left[\log \left[1 + \frac{(y_t - \mathbf{x}'_t \boldsymbol{\beta})^2}{\nu \sigma^2} \right] \middle| \mathbf{x}_t \right] &\leq E \left[\frac{(y_t - \mathbf{x}'_t \boldsymbol{\beta})^2}{\nu \sigma^2} \middle| \mathbf{x}_t \right] \\ &= \frac{\nu_0 \sigma_0^2 + (\mathbf{x}'_t \boldsymbol{\beta}_0 - \mathbf{x}'_t \boldsymbol{\beta})^2}{\nu \sigma^2 (\nu_0 - 2)} \end{aligned}$$

provided that $E[\mathbf{x}_t \mathbf{x}'_t]$ exists, the expected log-lik exists.



UNCONDITIONAL INEQUALITY

The expected log-likelihood inequality implies the unconditional inequality

$$E[L(\boldsymbol{\theta}; U|V)] \leq E[L(\boldsymbol{\theta}_0; U|V)]$$

starting from

$$E[L(\boldsymbol{\theta}; U|V)|V] \leq E[L(\boldsymbol{\theta}_0; U|V)|V]$$

we can take the $E[\cdot]$ over V

$$\begin{aligned} E[L(\boldsymbol{\theta}; U|V)] &= E[E[L(\boldsymbol{\theta}; U|V)|V]] \\ &\leq E[E[L(\boldsymbol{\theta}_0; U|V)|V]] \\ &= E[L(\boldsymbol{\theta}_0; U|V)] \end{aligned}$$



Because θ_0 maximizes $E[L(\theta; U|V)]$ it is natural to construct an estimator of θ_0 from the value of θ that maximizes the sample: the average log-likelihood functions of the N observations

$$\frac{1}{N} \sum_t L(\theta; U_t|V_t) \equiv E_N[L(\theta; U|V)]$$

$$E[L(\theta; U|V)] = \int L(\theta; u|v) dF(u|v; \theta_0)$$

ML estimator: the MLE is a value of the parameter vector that maximizes the sample average log-lik function

$$\hat{\theta}_N \equiv \arg \max_{\theta \in \Theta} E_N[L(\theta)]$$



NORMAL LINEAR REGRESSION MODEL

The empirical expectation of the log-likelihood

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

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

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

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



NORMAL LINEAR REGRESSION MODEL

Solutions:

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

The MLE of σ^2 is

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

with

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

The Hessian matrix:

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



NORMAL LINEAR REGRESSION MODEL

$$\begin{aligned} E_N[L_{\theta\theta}(\hat{\theta})] &= \begin{bmatrix} -\frac{1}{\hat{\sigma}^2 N} \mathbf{X}'\mathbf{X} & -\frac{\mathbf{X}'(\mathbf{y}-\mathbf{X}\hat{\beta})}{\hat{\sigma}^4 N} \\ -\frac{(\mathbf{y}-\mathbf{X}\hat{\beta})'\mathbf{X}}{\hat{\sigma}^4 N} & \frac{1}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6 N} (\mathbf{y}-\mathbf{X}\hat{\beta})'(\mathbf{y}-\mathbf{X}\hat{\beta}) \end{bmatrix} \\ &= \begin{bmatrix} -\frac{1}{\hat{\sigma}^2 N} \mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6 N} (\mathbf{y}-\mathbf{X}\hat{\beta})'(\mathbf{y}-\mathbf{X}\hat{\beta}) \end{bmatrix} \end{aligned}$$

which is negative definite.

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.



Is the DGP sufficiently informative about the parameters of the model? If

$$f(u|v; \boldsymbol{\theta}_0) = f(u|v; \boldsymbol{\theta}_1)$$

data drawn from these two distributions will have the same sampling properties. There is no way to distinguish whether $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ or $\boldsymbol{\theta} = \boldsymbol{\theta}_1$.



GLOBAL IDENTIFICATION

The parameter θ_0 is *globally identified* in Θ if, for every $\theta_1 \in \Theta$, $\theta_0 \neq \theta_1$, implies that

$$Pr\{f(U|V; \theta_0) \neq f(U|V; \theta_1)\} > 0$$

Assumption (Global identification): Every parameter vector $\theta_0 \in \Theta$ is globally identified.

Lemma (Strict expected log-likelihood inequality): Under the *Distribution, Dominance and Global identification* assumptions:

$$\theta \neq \theta_0$$

implies

$$E[L(\theta)] < E[L(\theta_0)].$$



EXAMPLE

Exact multicollinearity among explanatory variables in a linear regression $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}_0$ is a failure of global identification.

If $\text{rank}(\mathbf{X}) < K$ then

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

still holds. The normal log-likelihood still attains its maximum in $\boldsymbol{\beta}$ at $\boldsymbol{\beta}_0$ because

$$-(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \mathbf{X}' \mathbf{X} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \leq 0$$

but inequality is not strict for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$.

If $\text{rank}(\mathbf{X}) = K$ then $\boldsymbol{\beta}_0$ is the unique maximum of $E[L(\boldsymbol{\theta})]$.



EXAMPLE

Identification concerns $E[L(\boldsymbol{\theta})]$ and not the $E_N[L(\boldsymbol{\theta})]$.

One can discover failures of identification in the sample log-likelihood.

But if a sample log-likelihood function fails to have a unique global maximum this does not always imply a failure of global identification.



EXAMPLE

Exact multicollinearity among explanatory variables in a LRM

$$E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}_0$$

is a failure of global identification. Note that if

$$\text{rank}(\mathbf{X}) < K$$

the expected log-likelihood inequality

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

still holds.



DIFFERENTIABILITY

When the support of the distribution depends on the unknown parameter values the MLE cannot be found with simple calculus. In such cases the log-lik cannot be differentiable everywhere in the parameter space.

Assumption (Differentiability): The p.f. $f(u|v; \boldsymbol{\theta})$ is twice continuously differentiable in $\boldsymbol{\theta}$, $\forall \boldsymbol{\theta} \in \Theta$. The $\mathbb{S}(\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$, and differentiation and integration are interchangeable in the sense that

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int_{\mathbb{S}(\boldsymbol{\theta})} dF(u|v; \boldsymbol{\theta}) = \int_{\mathbb{S}(\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} dF(u|v; \boldsymbol{\theta})$$
$$\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \int_{\mathbb{S}(\boldsymbol{\theta})} dF(u|v; \boldsymbol{\theta}) = \int_{\mathbb{S}(\boldsymbol{\theta})} \frac{\partial^2}{\partial \boldsymbol{\theta}^2} dF(u|v; \boldsymbol{\theta})$$



$$\frac{\partial E[L(\boldsymbol{\theta})|V = v]}{\partial \boldsymbol{\theta}} = E \left[\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \middle| V = v \right]$$

$$\frac{\partial^2 E[L(\boldsymbol{\theta})|V = v]}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = E \left[\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \middle| V = v \right]$$

The interchange of differentiation and integration is ensured in part by $\mathbb{S}(\boldsymbol{\theta}) = \mathbb{S}$.

$$\boldsymbol{\theta}_0 = \arg \max_{\boldsymbol{\theta} \in \Theta} E[L(\boldsymbol{\theta})]$$

translates into the conditions

$$\frac{\partial E[L(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathbf{0}$$

and the second order conditions that the Hessian matrix

$$\frac{\partial^2 E[L(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \text{ is a n.d. matrix.}$$



THE SCORE FUNCTION

The MLE $\hat{\theta}$ is an implicit function of the data u

$$\hat{\theta} = \arg \max_{\theta \in \Theta} E_N[L(\theta)] \in \arg \text{zero}_{\theta \in \Theta} E_N[L_{\theta}(\theta)]$$

The F.O.C. *Normal equations* or *likelihood equations*

$$E_N[L_{\theta}(\hat{\theta})] = \mathbf{0}$$

where the *score function*

$$L_{\theta} \equiv \frac{\partial L(\theta)}{\partial \theta}$$

$\hat{\theta}$ must be calculated by numerical methods for maximizing differentiable functions.



Lemma (Score identity): Under *Distribution* and *Differentiability* assumptions

$$E[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)|V = v] = \mathbf{0}$$

Proof: Continuous random variables case

$$1 = \int_{\mathbb{S}} dF(u|v; \boldsymbol{\theta}) = \int_{\mathbb{S}} f(u|v; \boldsymbol{\theta}) du$$



SCORE IDENTITY

we can differentiate both sides of this equality w.r.t. $\boldsymbol{\theta}$

$$\begin{aligned} \mathbf{0} &= \int_{\mathbb{S}} \frac{\partial}{\partial \boldsymbol{\theta}} f(u|v; \boldsymbol{\theta}) du \\ &= \int_{\mathbb{S}} f_{\boldsymbol{\theta}}(u|v; \boldsymbol{\theta}) du \\ &= \int_{\mathbb{S}} \frac{1}{f(u|v; \boldsymbol{\theta})} f_{\boldsymbol{\theta}}(u|v; \boldsymbol{\theta}) f(u|v; \boldsymbol{\theta}) du \end{aligned}$$

consider

$$L_{\boldsymbol{\theta}}(\boldsymbol{\theta}; U|V) = \frac{1}{f(u|v; \boldsymbol{\theta})} f_{\boldsymbol{\theta}}(u|v; \boldsymbol{\theta})$$

$$E[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}; U|V) | V = v] = \int_{\mathbb{S}} \frac{1}{f(u|v; \boldsymbol{\theta})} f_{\boldsymbol{\theta}}(u|v; \boldsymbol{\theta}) f(u|v; \boldsymbol{\theta}_0) du$$



SCORE IDENTITY

The $E[\cdot|V = v]$ is evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. For $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$

$$E[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}; U|V)|V = v] \neq 0$$

But if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ then

$$E[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0; U|V)|V = v] = \int_{\mathbb{S}} \frac{1}{f(u|v; \boldsymbol{\theta}_0)} f_{\boldsymbol{\theta}}(u|v; \boldsymbol{\theta}_0) f(u|v; \boldsymbol{\theta}_0) du = 0.$$



In the Normal Linear Regression Model

$$E[L_{\beta}(\boldsymbol{\theta})] = \frac{1}{\sigma^2} E[\mathbf{x}_t \mathbf{x}_t'] (\boldsymbol{\beta}_0 - \boldsymbol{\beta})$$

$$E[L_{\sigma^2}(\boldsymbol{\theta})] = -\frac{1}{2\sigma^4} (\sigma^2 - \{\sigma_0^2 + E[(\mathbf{x}_t' \boldsymbol{\beta}_0 - \mathbf{x}_t' \boldsymbol{\beta})^2]\})$$

$$\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \sigma_0^2)'$$

$$E[L_{\beta}(\boldsymbol{\theta}_0)] = \frac{1}{\sigma_0^2} E[\mathbf{x}_t \mathbf{x}_t'] (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0) = \mathbf{0}$$

$$E[L_{\sigma^2}(\boldsymbol{\theta}_0)] = -\frac{1}{2\sigma_0^4} (\sigma_0^2 - \{\sigma_0^2 + E[(\mathbf{x}_t' \boldsymbol{\beta}_0 - \mathbf{x}_t' \boldsymbol{\beta}_0)^2]\}) = 0$$



THE INFORMATION MATRIX

If there exists $\tilde{\boldsymbol{\theta}}$ such that

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

we must check that we have a global maximum. Otherwise our solution cannot be the MLE ($\hat{\boldsymbol{\theta}}_N$).

In general, a **sufficient condition** for $\tilde{\boldsymbol{\theta}}_N$ to be a local maximum is that the Hessian matrix

$$E_N[L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}_N)] \equiv \left. \frac{\partial^2 E_N[L(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_N}$$

evaluated at $\tilde{\boldsymbol{\theta}}_N$ is **negative definite**: $\forall \mathbf{c} \in \mathbb{R}^K, \mathbf{c} \neq \mathbf{0}$

$$\mathbf{c}' E_N[L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}_N)] \mathbf{c} < 0$$

it guarantees that $E_N[L(\boldsymbol{\theta})]$ is **strictly concave** in a neighborhood of $\tilde{\boldsymbol{\theta}}$.



We investigate the second-order conditions for the maximum of $E[L(\boldsymbol{\theta})]$.

Assumption (Finite Information): $Var[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)]$ exists.

Lemma (Information Identity): Under *Distribution, Differentiability, Finite Information* assumptions

$$E[L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0)|V = v] = -Var[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)|V = v]$$

and this matrix is *negative semidefinite*.



Proof:

$$\mathbf{0} = \int_{\mathbb{S}} L_{\boldsymbol{\theta}}(\boldsymbol{\theta}; u|v) f(u|v; \boldsymbol{\theta}) du$$

Differentiating both sides

$$\begin{aligned} \frac{\partial(L_{\boldsymbol{\theta}}(\boldsymbol{\theta})f(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}'} &= \frac{\partial L_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}'} f + L_{\boldsymbol{\theta}} \frac{\partial f}{\partial\boldsymbol{\theta}'} \\ &= L_{\boldsymbol{\theta}\boldsymbol{\theta}} f + L_{\boldsymbol{\theta}}(f\boldsymbol{\theta})' \\ &= (L_{\boldsymbol{\theta}\boldsymbol{\theta}} + L_{\boldsymbol{\theta}}L'_{\boldsymbol{\theta}})f \end{aligned}$$

$$f \equiv f(u|v; \boldsymbol{\theta}).$$

$$\mathbf{0} = \int_{\mathbb{S}} [L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}; u|v) + L_{\boldsymbol{\theta}}(\boldsymbol{\theta}; u|v)L_{\boldsymbol{\theta}}(\boldsymbol{\theta}; u|v)'] dF(u|v; \boldsymbol{\theta})$$



$$\int_{\mathcal{S}} L_{\theta\theta}(\boldsymbol{\theta}; u|v) dF(u|v; \boldsymbol{\theta}) = - \int_{\mathcal{S}} [L_{\theta}(\boldsymbol{\theta}; u|v) L_{\theta}(\boldsymbol{\theta}; u|v)'] dF(u|v; \boldsymbol{\theta})$$

Setting $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

$$\begin{aligned} E[L_{\theta\theta}(\boldsymbol{\theta}_0; U|V)|V = v] &= -E[L_{\theta}(\boldsymbol{\theta}_0; U|V) L_{\theta}(\boldsymbol{\theta}_0; U|V)'|V = v] \\ &= -Var[L_{\theta}(\boldsymbol{\theta}_0; U|V)|V = v] \end{aligned}$$

because $E[L_{\theta}(\boldsymbol{\theta}_0; U|V)|V] = 0$. The Hessian is negative semidefinite since is the negative of a variance matrix.



CONDITIONAL INFORMATION

The *conditional variance matrix* of the score vector $L_{\boldsymbol{\theta}}(\boldsymbol{\theta}; U|V)$ given $V = v$ and evaluated at $\boldsymbol{\theta}_0$

$$\mathfrak{I}(\boldsymbol{\theta}_0|v) \equiv E[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)'|V = v] = \text{Var}[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)|V = v]$$

we can always find *the conditional information matrix function*

$$\mathfrak{I}(\boldsymbol{\theta}|v) \equiv \int_{\mathbb{S}} L_{\boldsymbol{\theta}}(\boldsymbol{\theta}; u|v)L_{\boldsymbol{\theta}}(\boldsymbol{\theta}; u|v)'dF(u|v; \boldsymbol{\theta})$$



The marginal expectation

$$\mathfrak{J}(\boldsymbol{\theta}_0) \equiv E[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}; U|V)L_{\boldsymbol{\theta}}(\boldsymbol{\theta}; U|V)']$$

is the *population information matrix*.

The population information matrix is the unconditional variance matrix of the conditional score vector because

$$E[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0; U|V)|V] = \mathbf{0}$$

$$\begin{aligned} \text{Var}[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0; U|V)] &= E[\text{Var}[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0; U|V)]] + \text{Var}[E[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0; U|V)|V]] \\ &= E[\mathfrak{J}(\boldsymbol{\theta}_0|V)] = \mathfrak{J}(\boldsymbol{\theta}_0) \end{aligned}$$



NORMAL LINEAR REGRESSION MODEL

The conditional information matrix for the normal linear regression model:

$$\mathfrak{I}(\boldsymbol{\theta}_0|\mathbf{x}_t) = \begin{bmatrix} \frac{1}{\sigma_0^2} \mathbf{x}_t \mathbf{x}_t' & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\sigma_0^4} \end{bmatrix}$$

The Hessian of the conditional normal regression log-likelihood function

$$L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}; y_t|\mathbf{x}_t) = \begin{bmatrix} -\frac{1}{\sigma^2} \mathbf{x}_t \mathbf{x}_t' & -\frac{1}{\sigma^4} \mathbf{x}_t (y_t - \mathbf{x}_t' \boldsymbol{\beta}) \\ -\frac{1}{\sigma^4} (y_t - \mathbf{x}_t' \boldsymbol{\beta}) \mathbf{x}_t' & \frac{1}{2\sigma_0^4} - (y_t - \mathbf{x}_t' \boldsymbol{\beta})^2 / \sigma^6 \end{bmatrix}$$
$$-E[L_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0; y_t|\mathbf{x}_t)|V] = \mathfrak{I}(\boldsymbol{\theta}_0|\mathbf{x}_t)$$



It is possible that information matrix can be singular even θ_0 is globally identifiable and the expected log-lik is uniquely maximized at θ_0 .

The second order condition that the Hessian be negative definite is sufficient but not necessary for a local maximum.

We assume this condition explicitly.

Assumption (Nonsingular Information) The information matrix $\mathfrak{J}(\theta_0)$ is nonsingular for all possible $\theta_0 \in \Theta$.



THE CRAMÉR - RAO LOWER BOUND

Information matrix: measure of how much we can learn about θ_0 from the random sample $\{(U_1, V_1), \dots, (U_N, V_N)\}$.

Theorem: $\tilde{\theta}$ unbiased estimator of θ_0 , with finite variance matrix with *interchangeability between differentiation and integration*

$$\begin{aligned}\frac{\partial E[\tilde{\theta}|v_1, \dots, v_N]}{\partial \theta_0} &= \frac{\partial}{\partial \theta_0} \int_{\mathbb{S}} \tilde{\theta} \prod_{t=1}^N dF(u_t|v_t; \theta_0) \\ &= \int_{\mathbb{S}} \tilde{\theta} \frac{\partial}{\partial \theta_0} \prod_{t=1}^N dF(u_t|v_t; \theta_0)\end{aligned}$$

if *Distribution, Differentiability, Finite Information Nonsingularity* assumptions also hold then that for any $\mathbf{a} \in \mathbb{R}^K$

$$\mathbf{a}' \text{Var}[\tilde{\theta}|v] \mathbf{a} \geq \mathbf{a}' (NE_N[\mathcal{J}(\theta_0)|v])^{-1} \mathbf{a}.$$



THE CRAMÉR - RAO LOWER BOUND

In some cases we can find estimators with variances equal to the Cramér-Rao lower bound.

The OLS estimator $\hat{\beta}$ is efficient relative to all unbiased estimators of β_0 .

Proof: Using

$$\mathcal{J}(\theta_0 | \mathbf{x}_t) = \begin{bmatrix} \frac{1}{\sigma_0^2} \mathbf{x}_t \mathbf{x}_t' & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\sigma_0^4} \end{bmatrix}$$

$$(N \cdot E_N[\mathcal{J}(\theta_0 | \mathbf{x}_t)])^{-1} = \begin{bmatrix} \frac{1}{\sigma_0^2} (\mathbf{X}'\mathbf{X}) & \mathbf{0} \\ \mathbf{0} & \frac{N}{2\sigma_0^4} \end{bmatrix}^{-1} = \begin{bmatrix} \sigma_0^2 (\mathbf{X}'\mathbf{X}) & \mathbf{0} \\ \mathbf{0} & \frac{2\sigma_0^4}{N} \end{bmatrix}$$

because

$$\text{Var}[\hat{\beta} | \mathbf{X}] = \sigma_0^2 (\mathbf{X}'\mathbf{X})^{-1}$$

The OLS/MLE estimator attains the Cramér-Rao lower bound.