TA Session for Econometrics I, 2025

9

Jukina HATAKEYAMA

The University of Osaka, Department of Economics

June 19, 2025

1. M-estimation
2. Introductory Topics of the ML Method
   - MLE
   - The Fisher Information
   - The Cramér–Rao Lower Bound
   - Example of the ML Method
3. Consistency and Asymptotic Normality for the Maximum Likelihood Estimator
   - Consistency for the Maximum Likelihood Estimator
4. Non–linear Optimisation Procedure
   - Newton–Raphson Method
   - Scoring Method
5. The MLE of a Single Regression Model
6. The MLE of a Multiple Regression Model
   - Reminder on Change of Variables
   - Multiple Regression Model

1. M-estimation
2. Introductory Topics of the ML Method
   - MLE
   - The Fisher Information
   - The Cramér–Rao Lower Bound
   - Example of the ML Method
3. Consistency and Asymptotic Normality for the Maximum Likelihood Estimator
   - Consistency for the Maximum Likelihood Estimator
4. Non–linear Optimisation Procedure
   - Newton–Raphson Method
   - Scoring Method
5. The MLE of a Single Regression Model
6. The MLE of a Multiple Regression Model
   - Reminder on Change of Variables
   - Multiple Regression Model

M-estimation

An estimator $\hat{\theta}$ is called an extremum estimator if there is a scalar objective function $Q_n(\theta)$ such that

$$\hat{\theta} = \arg\max_{\theta \in \Theta} Q_n(\theta), \tag{1}$$

where $\Theta \in \mathbb{R}^p$ is the parameter space, the set of possible parameter values.

The objective function $Q_n(\theta)$ depends not only on θ but also the sample $(w_1, w_2, \cdots, w_n)$, where w_i is the i th observation and n is the sample size.

The maximum likelihood method and the generalised method of moments(GMM) estimators are particular extremum estimators.

Although we do not prove, the extremum estimator can be derived under some general conditions[1].

Read Hayashi(2000):446-447 for details.

One of the extremum estimators explained in Econometrics I is M-estimator.

The objective function of it is a sample average:

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(w_i;\theta), \tag{2}$$

where $m(w_i;\theta)$ is a real-valued function of w_i and θ.

An example of this estimator is MLE (explained in this session).

1 M-estimation

2 Introductory Topics of the ML Method
- MLE
- The Fisher Information
- The Cramér–Rao Lower Bound
- Example of the ML Method

3 Consistency and Asymptotic Normality for the Maximum Likelihood Estimator
- Consistency for the Maximum Likelihood Estimator

4 Non–linear Optimisation Procedure
- Newton–Raphson Method
- Scoring Method

5 The MLE of a Single Regression Model

6 The MLE of a Multiple Regression Model
- Reminder on Change of Variables
- Multiple Regression Model

Maximum Likelihood Estimation

Suppose that $X_1, X_2, \cdots, X_n$: i.i.d. random variables.

Here, $f(\theta; x_i)$ implies the probability density function of X, where θ is a parameter.

Then, the maximum likelihood estimator maximises the likelihood function defined as $l(\theta) := \prod_{i=1}^{n} f(\theta; x_i)$.

We can rewrite the likelihood by taking logarithm to the likelihood function as follows:

$$L_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \log[f(\theta; x_i)]. \tag{3}$$

This is called as a log-likelihood function of X.

The maximum likelihood estimator is of θ satisfies the following condition.

Definition

We say that $\hat{\theta} = \hat{\theta}(X)$ is a MLE of θ if it satisfies the following condition:

$$\hat{\theta} = \arg\max_{\theta} L_n(\theta).$$

In other words, MLE satisfies the following conditions as follows:

$$\frac{\partial L_n(\hat{\theta})}{\partial \theta} = 0, \quad \frac{\partial^2 L_n(\hat{\theta})}{\partial \theta \partial \theta'} < 0.$$

The Fisher Information

We establish a remarkable inequality called the Cramér - Rao lower bound, which gives a lower bound on the variance of any unbiased estimator.

Assume that the log-likelihood function is continuously twice differentiable, and the integral of the log-likelihood function is also continuously twice differentiable.

Then, we begin with the identity that $\int f(\theta; x)\, dx = 1$, where $f(\theta; x)$ is a probability density function and dx denotes the Lebesgue measure. Taking the derivative with respect to θ, we obtain:

$$\begin{aligned}
\frac{\partial}{\partial \theta} \int f(\theta; x)\, dx &= \int \frac{\partial f(\theta; x)}{\partial \theta}\, dx \\
&= \int \frac{\partial \log f(\theta; x)}{\partial \theta} f(\theta; x)\, dx \\
&= \mathbb{E}\left[\frac{\partial \log f(\theta; x)}{\partial \theta}\right] = 0.
\end{aligned}$$

The second term on the left-hand side of the previous equation can be rewritten as an expectation. We call this expectation **the Fisher information** and denote it by $\mathrm{I}(\theta)$, defined as follows:

$$\mathrm{I}(\theta) = \mathrm{Var}\left[\nabla_\theta \log f(\theta; x)\right] = -\mathbb{E}\left[\nabla^2_{\theta\theta'} \log f(\theta; x)\right], \tag{4}$$

due to equation above and the formula for the variance, $\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$.

In other words, we use the following relationships to derive (4):

$$\underbrace{\int \frac{\partial \log f(\theta; x)}{\partial \theta} \frac{\partial \log f(\theta; x)}{\partial \theta'} f(\theta; x)\, dx}_{\mathrm{I}(\theta)} = -\int \frac{\partial^2 \log f(\theta; x)}{\partial \theta \partial \theta'} f(\theta; x)\, dx$$

$$= -\mathbb{E}\left[\nabla^2_{\theta\theta'} \log f(\theta; x)\right],$$

$$\underbrace{\int \frac{\partial \log f(\theta; x)}{\partial \theta} \frac{\partial \log f(\theta; x)}{\partial \theta'} f(\theta; x)\, dx}_{\mathrm{I}(\theta)} = \mathbb{E}\left[\left(\nabla_\theta \log f(\theta; x)\right)\left(\nabla_\theta \log f(\theta; x)\right)'\right]$$

$$= \mathrm{Var}\left[\nabla_\theta \log f(\theta; x)\right].$$

Definition

The Fisher information matrix is defined as

$$I(\theta) = -\mathbb{E}[\nabla^2_{\theta\theta'} \log\, f_\theta(X_i)] \tag{5}$$

and we have the equalities

$$I(\theta) = \mathbb{E}[\nabla_\theta \log\, f_\theta(X_i) \nabla_{\theta'} \log\, f_\theta(X_i)] = \mathrm{Var}[\nabla_\theta \log\, f(\theta; x)]. \tag{6}$$

The Cramér–Rao Lower Bound

Note that the following important function is called the score function.

$$\nabla_\theta l(\theta; X) = \nabla_\theta \log\, f(\theta; x)$$

Theorem

Suppose that an unbiased estimator is given by $f(X)$. Then, we can establish a following relationship:

$$\mathrm{Var}(f(X)) \geq \frac{1}{-\mathbb{E}[\nabla^2_{\theta\theta} l(\theta; X)]} = I(\theta)^{-1}. \tag{7}$$

Proof)
At first, taking the derivative with respect to θ to the expectation of $f(X)$ as follows:

$$\begin{aligned}\nabla_\theta \mathbb{E}[f(X)] &= \int f(x)\nabla_\theta f(\theta; X) d\lambda(x) \\ &= \int f(x)\nabla_\theta log(f(\theta; X)) f(\theta; X) d\lambda(x) \\ &= \mathrm{Cov}(f(X), \nabla_\theta l(\theta; X)). \end{aligned} \tag{8}$$

Proof of (8)
By definition, the covariance $\mathrm{Cov}(f(X), \nabla_\theta \log f(\theta; X))$ can be rewritten as follows:

$$\mathrm{Cov}(f(X), \nabla_\theta \log f(\theta; X)) = \int [f(x) - \mathbb{E}[f(X)]] \tag{9}$$

$$[\nabla_\theta \log f(\theta; x) - \mathbb{E}[\nabla_\theta \log f(\theta; X)]]\, f(\theta; x)\, dx \tag{10}$$

$$= \int [f(x) - \theta] \cdot \nabla_\theta \log f(\theta; x) \cdot f(\theta; x)\, dx$$

$$= \int f(x) \cdot \nabla_\theta \log f(\theta; x) \cdot f(\theta; x)\, dx, \tag{11}$$

where we have used the following property, which is a special case of the formula for covariance:

$$\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X \mathbb{E}[Y] = \mathbb{E}[XY], \quad \text{if } \mathbb{E}[Y] = 0. \tag{12}$$

In the one dimensional case, we can rewrite (8),

$$\begin{aligned}(\nabla_\theta \mathbb{E}[f(X)])^2 &= [\mathrm{Cov}(f(X), \nabla_\theta l(\theta; X))]^2 \\ &= \rho^2 \mathrm{Var}(f(X))\mathrm{Var}(\nabla_\theta l(\theta; X)) \\ &\leq \mathrm{Var}(f(X))\mathrm{Var}(\nabla_\theta l(\theta; X)), \end{aligned} \tag{13}$$

with ρ, which implies the correlation between $f(X)$ and $\nabla_\theta l(\theta; X)$. Remind that we can say

$$[\mathrm{Cov}(f(X), \nabla_\theta l(\theta; X))]^2 = \rho^2 \mathrm{Var}(f(X))\mathrm{Var}(\nabla_\theta l(\theta; X)),$$

by the definition of the correlation coefficient. Since $|\rho| \leq 1$, we have the inequality. or equivalently, we have following inequality because we have $\mathbb{E}[f(X)] = \theta$ and the derivative of this relationship w.r.t. θ is 1.

$$\mathrm{Var}(f(X)) \geq \frac{1}{-\mathbb{E}[\nabla^2_{\theta\theta} l(\theta; X)]} = I(\theta)^{-1} \tag{14}$$

This inequality also holds for multi dimensional cases. The Cramér–Rao lower bound is a lower bound on the variance of estimators.

Example of the ML Method

Suppose the case of the random variable $X \sim N_{\mathbb{R}}(0, \sigma^2)$. The likelihood of the each observed variable x_i $(i = 1, 2, \cdots, n)$ is given as follows:

$$f(\theta; x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(x_i - \mu)^2}{2\sigma^2}] \tag{15}$$

By taking the logarithm, the above equation is rewritten as follows:

$$\log f(\theta; x_i) = \frac{1}{2}\log 2\pi - \log\sigma - \frac{(x_i - \mu)^2}{2\sigma^2}.$$

Recall that we must minimise $\sum_{i=1}^{n} \log f(\theta; x_i)$ such that:

$$\sum_{i=1}^{n} \log f(\theta; x_i) = (constant) - n\log\sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}.$$

Therefore, when we estimate μ, the first order condition is given as follows:

$$\frac{d\sum_{i=1}^{n}(x_i - \mu)^2}{d\mu} = -2\sum_{i=1}^{n}(x_i - \mu) = 0,$$

we have $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

1. M-estimation
2. Introductory Topics of the ML Method
   - MLE
   - The Fisher Information
   - The Cramér–Rao Lower Bound
   - Example of the ML Method
3. Consistency and Asymptotic Normality for the Maximum Likelihood Estimator
   - Consistency for the Maximum Likelihood Estimator
4. Non–linear Optimisation Procedure
   - Newton–Raphson Method
   - Scoring Method
5. The MLE of a Single Regression Model
6. The MLE of a Multiple Regression Model
   - Reminder on Change of Variables
   - Multiple Regression Model

Consistency and Asymptotic Normality for the MLE

If we set

- setting $m(X_i, \theta) := \log p_\theta(X_i)$, that is,

$$M_n(\theta) := L_n(\theta) = \sum_{i=1}^{n} \log p_\theta(X_i);$$

- $\mathbb{L}(\theta) = \mathbb{E}\left[\mathbb{L}_n(\theta)\right] = \mathbb{E}\left[\log p_\theta(X)\right]$,

Suppose

$$\mathbb{L}(\theta) = \mathbb{E}\left[\mathbb{L}_n(\theta)\right] = \mathbb{E}\left[\log p_\theta(X)\right]$$

exists for $\theta \in \mathbb{R}^d$. Then we obtain the following theorem.

Theorem 3.1

Suppose

1. $\mathbb{L}(\theta)$ is uniquely maximised at θ_0, idest

$$\forall \epsilon > 0, \qquad \sup_{\theta:\ \|\theta - \theta_0\| \geq 0} \mathbb{L}(\theta) < \mathbb{L}(\theta_0);$$

2. Θ is compact;
3. $\mathbb{L}(\theta)$ is continuous;
4. $\sup_{\theta \in \Theta} |\mathbb{L}_n(\theta) - \mathbb{L}(\theta)| \xrightarrow[n \to \infty]{\mathbb{P}} 0,$

then $\hat{\theta} \xrightarrow[n \to \infty]{\mathbb{P}} \theta_0$.

Under the same assumption, we can derive the following theorem.

Theorem 3.2

Suppose $\hat{\theta} \xrightarrow{\mathbb{P}} \theta_0$ and

1. θ_0 belongs to the interior of Θ;
2. $\mathbb{L}_n(\theta)$ is twice continuously differentiable;
3. $\sqrt{n}\nabla_\theta \mathbb{L}_n(\theta_0) \xrightarrow[n\to\infty]{d} N_{\mathbb{R}^d}(\mathbf{0}, \Sigma)$;
4. $H(\theta) = \mathbb{E}[\nabla^2_{\theta\theta'} \log p_\theta(X)]$ is continuous at θ_0 and

$$\sup_{\theta:\, \|\theta-\theta_0\|\le\delta} \left|\nabla^2_{\theta\theta'}\mathbb{L}_n(\theta) - H(\theta)\right| \xrightarrow[n\to\infty]{\mathbb{P}} 0, \quad \text{with } \delta > 0;$$

5. $H = H(\theta)$ is non-singular,

then

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow[n\to\infty]{d} N_{\mathbb{R}^d}\left(\mathbf{0}, H^{-1}JH^{-1}\right). \tag{16}$$

1. M-estimation
2. Introductory Topics of the ML Method
   - MLE
   - The Fisher Information
   - The Cramér–Rao Lower Bound
   - Example of the ML Method
3. Consistency and Asymptotic Normality for the Maximum Likelihood Estimator
   - Consistency for the Maximum Likelihood Estimator
4. Non–linear Optimisation Procedure
   - Newton–Raphson Method
   - Scoring Method
5. The MLE of a Single Regression Model
6. The MLE of a Multiple Regression Model
   - Reminder on Change of Variables
   - Multiple Regression Model

Non–linear Optimisation Procedure

From the first–order Taylor series expansion around $\beta = \beta^*$, we have

$$0 = \nabla_\beta \log L(\beta) \approx \nabla_\beta \log L(\beta^*) + \nabla^2_{\beta\beta'} \log L(\beta^*)(\beta - \beta^*)$$

Then, by the **mean–value theorem (expansion)**,

$$\nabla^2_{\beta\beta'} \log L(\overline{\beta})(\beta - \beta^*) = -\nabla_\beta \log L(\beta^*)$$

holds. Thus, assuming that the Hessian matrix is positive definite yields

$$\beta - \beta^* = -\left(\nabla^2_{\beta\beta'} \log L(\overline{\beta})\right)^{-1} \nabla_\beta \log L(\beta^*)$$

This equation yields the following algorithm called **Newton–Raphson Method**.

Mean Value Theorem

Intuition

If a function is smooth and continuous on an interval, then somewhere in that interval, the instantaneous rate of change (i.e. the derivative) must equal the average rate of change.

Theorem (Mean Value Theorem)

Let f be a function that is

- continuous on the closed interval $[a, b]$, and
- differentiable on the open interval (a, b).

Then, there exists at least one point $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

This means that at some point, the tangent line is parallel to the secant line connecting $f(a)$ and $f(b)$.

Algorithm: Newton–Raphson Method

$$\beta^{(j+1)} = \beta^{(j)} - \left(\nabla^2_{\beta\beta'} \log L(\beta^{(j)}) \right)^{-1} \nabla_\beta \log L(\beta^{(j)})$$

Scoring Method

If we take expectation on second derivative of likelihood function, the method is known as the **method of scoring**.

Algorithm: Scoring Method

$$\beta^{(j+1)} = \beta^{(j)} - \left(\mathbb{E}\left[\nabla^2_{\beta\beta'} \log L(\beta^{(j)})\right]\right)^{-1} \nabla_\beta \log L(\beta^{(j)})$$

Note that

$$I(\theta) := -\mathbb{E}\left[\nabla^2_{\beta\beta'} log L(\beta)\right]$$

is the **Fisher's information** matrix.

1. M-estimation
2. Introductory Topics of the ML Method
   - MLE
   - The Fisher Information
   - The Cramér–Rao Lower Bound
   - Example of the ML Method
3. Consistency and Asymptotic Normality for the Maximum Likelihood Estimator
   - Consistency for the Maximum Likelihood Estimator
4. Non–linear Optimisation Procedure
   - Newton–Raphson Method
   - Scoring Method
5. The MLE of a Single Regression Model
6. The MLE of a Multiple Regression Model
   - Reminder on Change of Variables
   - Multiple Regression Model

The MLE of a Single Regression Model

$$y_i = \alpha + \beta x_i + u_i,$$

where u_i.
Assume the error term, u_i, follows the Gaussian distribution:
$u_i \overset{i.i.d}{\sim} N(0, \sigma^2)$.

Let us denote the probability density function of the error term $f_u(\theta; u_i)$ for all i. In addition, we set $f_y(\theta; y_i)$ as the pdf for y_i for all i. By the Change of Variables, we have:

$$\begin{aligned} f_y(\theta; y_i) &= f_u(\theta; y_i) \left| \frac{\partial u_i}{\partial y_i} \right| \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right) \end{aligned}$$

The parameter vector is $\theta = (\alpha, \beta, \sigma^2)' \in \mathbb{R}^3$. The joint density function, represented as $f_y(\theta; y_1, \cdots, y_n)$ (or simply $f_y(\theta; y)$), is rewritten as:

$$\begin{aligned} f_y(\theta; y_1, \cdots, y_n) &= f(\theta; y_1) \cdots f(\theta; y_n) \\ &= \prod_{i=1}^{n} f(\theta; y_i) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2) \end{aligned}$$

by the i.i.d assumption. This is the likelihood function.

Then, the log-likelihood function is defined as:

$$\begin{aligned} l_n(\theta;(y,x)) &:= l_n(\theta;(y_i,x_i), i=1,2,\cdots,n) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2. \end{aligned}$$

Given the observed data $(y_i, x_i), (i = 1, \cdots, n,)$ we consider the maximisation problem of the log-likelihood function with respect to $(\beta, \alpha, \sigma^2)$ and obtain the following MLE:

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} l_n(\theta; y, x).$$

The first order condition of the maximisation problem is given as follows:

$$\partial_\alpha l_n(\hat{\theta};(y,x)) = \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0,$$

$$\partial_\beta l_n(\hat{\theta};(y,x)) = \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)x_i = 0,$$

$$\partial_{\sigma^2} l_n(\hat{\theta};(y,x)) = -\frac{n}{2}\frac{1}{\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = 0.$$

The solution is denoted as $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)'$, called the MLE. These solutions are given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

Note that the estimator of the variance is not same as that of the OLS. (biased estimator)

1. M-estimation
2. Introductory Topics of the ML Method
   - MLE
   - The Fisher Information
   - The Cramér–Rao Lower Bound
   - Example of the ML Method
3. Consistency and Asymptotic Normality for the Maximum Likelihood Estimator
   - Consistency for the Maximum Likelihood Estimator
4. Non–linear Optimisation Procedure
   - Newton–Raphson Method
   - Scoring Method
5. The MLE of a Single Regression Model
6. The MLE of a Multiple Regression Model
   - Reminder on Change of Variables
   - Multiple Regression Model

Reminder on Change of Variables

Consider the case that we change the density function $f_X(x)$ of a random variable X into another density function of $f_Z(z)$ of another random variable Z. In this subsection, we learn the properties of the changed random variable Z.

Theorem

Let $f_X(x)$ the density function of the random variable X. Let $z = \phi(x)$ with $\phi(\cdot)$ continuous and strictly monotone real value function. When the inverse function of $z = \phi(x)$ is given as $x = \phi^{-1}(z) = h(z)$, the following relationship is established:

$$f_Z(z) = |h'(z)| f_X(h(z)).$$

Proof)
Suppose that the distribution function of X as $F_X(x)$ and the distribution function of Z as $F_Z(z)$.

(i) In the case of $h'(x) > 0$, $F_Z(z)$ is rewritten as follows:

$$\begin{aligned} F_Z(z) &= \text{Prob}(Z \leq z) = \text{Prob}(\phi(X) \leq z) \\ &= \text{Prob}(X \leq h(z)) = F_X(h(z)) \end{aligned}$$

By differentiating both sides of the above equation, we can get $f_Z(z) = h(z)' f_X(h(z))$.

(ii) In the case of $h'(x) < 0$, $F_Z(z)$ is rewritten as

$$\begin{aligned} F_Z(z) &= \text{Prob}(Z \leq z) = \text{Prob}(\phi(X) \leq z) \\ &= \text{Prob}(X \geq h(z)) = 1 - \text{Prob}(X \leq h(z)) \\ &= 1 - F_X(h(z)), \end{aligned}$$

because $\phi(\cdot)$ is strictly monotone. By differentiating both sides of the above equation, we can get $f_Z(z) = -h'(z) f_X(h(z))$.

In the multivariate case, if $Z = H(X)$ with H a bijective and differentiable function, the density of Z is

$$f_Z(z) = f_X(x)|\det(\nabla_z x)|,$$

where the differential is the Jacobian of the inverse of H, evaluated at y.

Multiple Regression Model

Multivariate Normal Distribution

Let X a n-dimensional random vector. When X follows **a multivariate normal distribution**, denoted as $X \sim N_{\mathbb{R}^{dim(X)}}(\mu, \Sigma)$, its pdf is defined as:

$$f(X) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)] \tag{17}$$

.

Suppose the regression model such that $y = x\beta + u$, where $u \sim N_{\mathbb{R}^n}(0, \sigma^2 I_n)$. Then, the density function of u is

$$f_u(u) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\frac{1}{2\sigma^2} u'u),$$

By the change of variables from u to y, we have:

$$\begin{aligned} f_Y(y) &= f_u(y - x\beta)\det(\nabla_y u) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\frac{1}{2\sigma^2}(y - x\beta)'(y - x\beta)), \end{aligned}$$

since we have $\nabla_y u = I_n$.

Remind that we can calculate the joint density as the products of individual densities like the case of the single regression, because conditionally on x_i, $y_i|x_i$ are iid.

Assume that the case of $\theta = (\beta', \sigma^2)' \in \mathbb{R}^{K+1}$. The statistical criterion is

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathbb{L}_n(\theta; y, x),$$

with the log-likelihood function

$$\mathbb{L}_n(\theta; y, x) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - x\beta)'(y - x\beta).$$

Then, by optimising the above equation, we have MLEs as follows:

$$\hat{\beta} = (x'x)^{-1}(x'y), \quad \hat{\sigma}^2 = \frac{1}{n}(y - x\hat{\beta})'(y - x\hat{\beta}).$$