ECON 452* -- NOTE 13

Maximum Likelihood Estimation of the Classical Normal Linear Regression Model

This note introduces the basic principles of maximum likelihood estimation in the familiar context of the multiple linear regression model.

Recall that the multiple linear regression model can be written in either scalar or matrix notation.

- **In scalar notation**, the population regression equation, or PRE, for the linear regression model is written in general as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + u_i \quad (i = 1, \ldots, N) \tag{1.1}$$

or

$$Y_i = \beta_0 + \sum_{j=1}^{j=k} \beta_j X_{ij} + u_i \quad (i = 1, \ldots, N) \tag{1.2}$$

or

$$Y_i = \sum_{j=0}^{j=k} \beta_j X_{ij} + u_i , \qquad X_{i0} = 1 \ \forall i \quad (i = 1, \ldots, N) \tag{1.3}$$

where

$Y_i \equiv$ the i-th sample (observed) value of the regressand, or dependent variable;

$X_{ij} \equiv$ the i-th sample (observed) value of the j-th regressor, or independent variable, j = 1, …, k;

$\beta_j \equiv$ the partial regression coefficient of X_{ij}, j = 1, …, k;

$u_i \equiv$ the i-th value of the unobservable random error term.

- **In vector-matrix notation**, the population regression equation, or PRE, for the linear regression model is written in general as:

$$y = X\beta + u \tag{2}$$

where

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_N \end{bmatrix} = \text{ the N}\times\text{1 regressand vector}$$

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ x_3^T \\ \vdots \\ x_N^T \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ 1 & X_{31} & X_{32} & \cdots & X_{3k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{N1} & X_{N2} & \cdots & X_{Nk} \end{bmatrix} = \text{ the N}\times\text{K regressor matrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = \text{ the K}\times\text{1 regression coefficient vector}$$

$$u = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_N \end{bmatrix} = \text{ the N}\times\text{1 error vector}$$

1. Assumption A6: The Error Normality Assumption

In order to perform statistical inference in the linear regression model, it is necessary to specify the form of the probability distribution of the error vector u in population regression equation (1). The normality assumption does this.

❑ Scalar Formulation of the Error Normality Assumption A6

The random error terms u_i are ***identically*** **and** ***independently*** **distributed** as the ***normal*** **distribution** with

1. zero conditional means

$$E\left(u_i \middle| x_i^T\right) = E(u_i) = 0 \qquad \forall\ i$$

2. constant conditional variances

$$Var\left(u_i \middle| x_i^T\right) = E\left(u_i^2 \middle| x_i^T\right) = E\left(u_i^2 \middle| 1, X_{i1}, X_{i2}, \ldots, X_{ik}\right) = \sigma^2 > 0 \qquad \forall\ i$$

3. zero conditional covariances

$$Cov\left(u_i, u_s \middle| x_i^T, x_s^T\right) = E\left(u_i u_s \middle| x_i^T, x_s^T\right) = 0 \qquad \forall\ i \neq s$$

- A compact way of stating the error normality assumption is:

  condtional on x_i^T, the u_i are iid as $N(0, \sigma^2)$

  where

  "iid" means "*independently* and *identically* distributed"

  $N(0, \sigma^2)$ denotes a normal distribution with zero mean and variance σ^2.

❑ Matrix Formulation of the Error Normality Assumption A6

The N×1 error vector u has a multivariate normal distribution with

1. a zero conditional mean vector

$$E(u \mid X) = \underline{0} \qquad \text{where } \underline{0} \text{ is an N×1 vector of zeros}$$

2. a constant scalar diagonal covariance matrix V(u)

$$V(u \mid X) = E(uu^T \mid X) = \sigma^2 I_N \qquad \text{where } I_N \text{ is the N×N identity matrix}$$

- A compact way of stating the error normality assumption in matrix terms is:

$$u \mid X \sim N(\underline{0}, \sigma^2 I_N)$$

where $N(\cdot, \cdot)$ here denotes the N-variate normal distribution.

❑ Implications of Assumption A6 for Distribution of the Regressand Vector y

- **Linearity Property of Normal Distribution:** Any linear function of a normally distributed random variable is itself normally distributed.

- **y is a linear function of u:** The PRE $y = X\beta + u$ states that the regressand vector y is a linear function of the error vector u.

- ***Implication:*** Since u is normally distributed by assumption A6 and y is a linear function of u by assumption A1, the ***linearity property*** of the normal distribution implies that

$$y \mid X \sim N(X\beta, \sigma^2 I_N).$$

That is, the regressand vector y has an N-variate normal distribution with

(1) conditional mean vector equal to $E(y \mid X) = X\beta$

(2) conditional covariance matrix equal to $V(y \mid X) = \sigma^2 I_N$.

❑ **The Classical Normal Linear Regression Model -- the CNLR Model**

The classical normal linear regression model consists of the population regression equation

$$y = X\beta + u \qquad (2)$$

plus Assumptions A1 to A6.

2. Outline of the Method of Maximum Likelihood

ML estimation involves **joint estimation of *all* the unknown parameters** of a statistical model. ML estimation therefore requires that the model in question be completely specified. Complete specification of the model includes specifying the specific form of the probability distribution of the model's random variables.

In the case of the CNLR model, ML estimation involves ***joint* estimation** of the **regression coefficient vector β** and the **scalar error variance σ^2**.

3. ML Estimation of the CNLR Model: Derivation

- Derivation of the ML estimators of K+1 parameters $\theta^T = (\beta^T, \sigma^2)$ of the CNLR model consists of two main steps:

  Step 1: *Formulation* of the sample likelihood function for the CNLR model.

  Step 2: *Maximization* of the sample likelihood function with respect to the unknown parameters β and σ^2.

STEP 1: Formulation of the Sample Likelihood Function

- ***First*, formulate the probability density function (pdf) for each of the individual random error terms u_i** under the error normality assumption A6.

- The **normal pdf**. If the random variable y_i is normally distributed with mean μ and variance σ^2 -- i.e., if $y_i \sim N(\mu, \sigma^2)$, then the individual probability density function for y_i is:

$$f(y_i) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(y_i-\mu)^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{(y_i-\mu)^2}{2\sigma^2}\right]} \quad (3)$$

- The error normality assumption A6 states the i-th random error term u_i has a normal distribution with mean 0 and variance σ^2 -- i.e., $u_i \sim N(0, \sigma^2)$.

- The **pdf for each u_i** is therefore

$$f(u_i) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(u_i-0)^2}{2\sigma^2}\right] = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{u_i^2}{2\sigma^2}\right] \quad (4)$$

- ***Second*, transform the pdf $f(u_i)$ for u_i into an equivalent pdf for the observed regressand Y_i.**

- This step is necessary because the random error terms u_i are unobservable. We therefore need to reformulate the pdf for u_i into the corresponding pdf for Y_i. The pdf for Y_i is expressed in terms of the observed sample data (Y_i, x_i^T) and the unknown parameters $\theta^T = (\beta^T, \sigma^2)$ of the CNLR model.

1. Use the **change-of-variable theorem for probability density functions** to transform the pdf for u_i into the corresponding pdf for Y_i:

$$f(Y_i) = \left|\frac{\partial u_i}{\partial Y_i}\right| f(u_i)$$

where

$\left|\frac{\partial u_i}{\partial Y_i}\right|$ = the absolute value of the partial derivative $\frac{\partial u_i}{\partial Y_i}$

= the Jacobian of the transformation from u_i to Y_i.

2. In the present case, $u_i = Y_i - x_i^T\beta$, so that

$$\frac{\partial u_i}{\partial Y_i} = 1 \qquad \text{and hence} \qquad \left|\frac{\partial u_i}{\partial Y_i}\right| = 1$$

We therefore have the simple result that

$$f(Y_i) = 1 \cdot f(u_i) = f(u_i) = f(Y_i - x_i^T\beta).$$

- **Result:** Setting $u_i = Y_i - x_i^T\beta$ in the pdf for u_i

$$f(u_i) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left[-\frac{u_i^2}{2\sigma^2}\right]$$

yields the following **pdf for Y_i**:

$$f(Y_i) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left[-\frac{(Y_i - x_i^T\beta)^2}{2\sigma^2}\right]. \qquad (5)$$

- ***Third*, construct the *joint* pdf for the sample of N independent observed values of Y**.

- The N sample values of the regressand Y are contained in the N×1 regressand vector $y = \left(Y_1 \;\; Y_2 \;\; Y_3 \;\; \cdots \;\; Y_N\right)^T$.

- **Assumption A5 of independent random sampling** implies that the ***joint* pdf** of all N sample values of Y_i is simply the **product** of the **pdf's of the *individual* Y_i values**.

- Accordingly, the ***joint* pdf of all N sample values of Y_i** can be written as

$$f(y) = f(Y_1 \;\; Y_2 \;\cdots\; Y_N) = f(Y_1)\cdot f(Y_2)\cdot\cdots\cdot f(Y_N) = \prod_{i=1}^{N} f(Y_i). \qquad (6)$$

- Finally, substitute the right-hand side of the normal pdf (5) for Y_i in (6) to obtain the joint pdf for all N sample values of Y:

$$\begin{aligned} f(y) &= \prod_{i=1}^{N} f(Y_i) \\ &= \prod_{i=1}^{N} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(Y_i - x_i^T\beta)^2}{2\sigma^2}\right] \\ &= (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{\sum_{i=1}^{N}(Y_i - x_i^T\beta)^2}{2\sigma^2}\right] \\ &= (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - x_i^T\beta)^2\right] \\ &= (2\pi\sigma^2)^{-N/2} \exp\left[-\mathrm{RSS}(\beta)/2\sigma^2\right] \qquad \text{where } \mathrm{RSS}(\beta) \equiv \sum_{i=1}^{N}(Y_i - x_i^T\beta)^2 \end{aligned}$$

- **<u>Result</u>:** The **joint pdf for the N sample values** $(Y_1 \;\; Y_2 \;\cdots\; Y_N)$ of Y, conditional on the sample values of the regressors $\{x_i^T : i = 1, \ldots, N\}$, is

$$\begin{aligned} f(y) &= (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - x_i^T\beta)^2\right] \\ &= (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right] \\ &= (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\mathrm{RSS}(\beta)\right] \qquad (7) \end{aligned}$$

where $\mathrm{RSS}(\beta) \equiv \sum_{i=1}^{N}(Y_i - x_i^T\beta)^2 = (y - X\beta)^T(y - X\beta)$.

- ***Fourth***, the **joint pdf f(y) is the *sample likelihood function* for the sample of N independent observations** $(Y_i, x_i^T) = (Y_i \; 1 \; X_{i1} \; X_{i2} \; \cdots \; X_{ik})$ i = 1, …, N.

- The key difference between the joint pdf $f(y)$ and the sample likelihood function is their interpretation, not their form.

  The **joint pdf** $f(y)$ is interpreted as **a function of the observable random variables** $(Y_1 \; Y_2 \; \cdots \; Y_N)$ for given values of the parameters β, σ^2 and the regressors $(X_1 \; X_2 \; \cdots \; X_k)$.

  The **sample likelihood function** is interpreted as **a function of the *parameters*** β **and** σ^2 for given values of the observable variables $(Y \; X_1 \; X_2 \; \cdots \; X_k)$.

- To emphasize the difference between the joint pdf of the sample observations and the sample likelihood function, we denote the joint pdf as $f(y;\beta,\sigma^2)$ and the sample likelihood function as $L(\beta,\sigma^2;y)$. (Both the joint pdf of the sample and the sample likelihood function are conditional on the regressors X, but we suppress this to keep our notation a bit simpler.)

  The **joint pdf of the sample** $f(y;\beta,\sigma^2)$ is **a function of the observed Y-values** y = $(Y_1 \; Y_2 \; \cdots \; Y_N)$, given the parameters β and σ^2.

  The **sample likelihood function** $L(\beta,\sigma^2;y)$ is **a function of the model parameters β and σ^2**, given the sample values y = $(Y_1 \; Y_2 \; \cdots \; Y_N)$ of the regressand.

- Apart from this difference in interpretation, **the sample likelihood function** $L(\beta,\sigma^2;y)$ ***equals* the joint pdf of the sample** $f(y;\beta,\sigma^2)$:

$$L(\beta,\sigma^2;y) = f(y;\beta,\sigma^2)$$

$$= (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - x_i^T\beta)^2\right] \qquad (8.1)$$

$$= (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right] \qquad (8.2)$$

STEP 2: Maximization of the Sample Likelihood Function

- ***First,* take the natural logarithm of the sample *likelihood* function** to obtain the **sample *log-likelihood* function**.

- **Rationale:** It is easier to maximize the log-likelihood function than it is to maximize the original likelihood function (8) because of the exponential term in the likelihood function.

- **Equivalence of maximizing the likelihood and log-likelihood functions.**

  Because the natural logarithm is a positive monotonic transformation, the values of β and σ^2 that maximize the likelihood function $L(\beta, \sigma^2; y)$ are the same as those that maximize the log-likelihood function $\ln L(\beta, \sigma^2; y) = \ln[L(\beta, \sigma^2; y)]$, where $L(\beta, \sigma^2; y) > 0$.

  The reason is that, for any individual parameter β_j,

$$\frac{\partial \ln L(\beta, \sigma^2; y)}{\partial \beta_j} = \frac{1}{L} \frac{\partial L(\beta, \sigma^2; y)}{\partial \beta_j} = \frac{\partial L(\beta, \sigma^2; y) / \partial \beta_j}{L} \quad \text{where } L > 0. \qquad (9)$$

  Thus,

$$\frac{\partial L(\beta, \sigma^2; y)}{\partial \beta_j} > 0 \quad \Rightarrow \quad \frac{\partial \ln L(\beta, \sigma^2; y)}{\partial \beta_j} > 0;$$

$$\frac{\partial L(\beta, \sigma^2; y)}{\partial \beta_j} = 0 \quad \Rightarrow \quad \frac{\partial \ln L(\beta, \sigma^2; y)}{\partial \beta_j} = 0;$$

$$\frac{\partial L(\beta, \sigma^2; y)}{\partial \beta_j} < 0 \quad \Rightarrow \quad \frac{\partial \ln L(\beta, \sigma^2; y)}{\partial \beta_j} < 0.$$

- **Derivation of Sample Log-likelihood Function**

❑ ***First*, obtain the sample log-likelihood function.**

- Since $L(\beta,\sigma^2;y) = f(y;\beta,\sigma^2)$ by equation (8),

$$\ln L(\beta,\sigma^2;y) = \ln f(y;\beta,\sigma^2) \tag{10}$$

- Substitute for $f(y;\beta,\sigma^2)$ the expression on the right-hand side of (8.1):

$$\ln L(\beta,\sigma^2;y) = \ln\left\{ \left(2\pi\sigma^2\right)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - x_i^T\beta)^2 \right] \right\}$$

$$= \ln\left\{\left(2\pi\sigma^2\right)^{-N/2}\right\} + \ln\left\{ \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - x_i^T\beta)^2 \right] \right\} \tag{11}$$

- The **natural log of the first term** on the right-hand side of (11) is obtained using the following rule of natural logarithms: $\ln\left(a^b\right) = b\ln(a)$ for a > 0. Thus, since $2\pi\sigma^2 > 0$,

$$\ln\left\{\left(2\pi\sigma^2\right)^{-N/2}\right\} = -\frac{N}{2}\ln\left(2\pi\sigma^2\right) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\left(\sigma^2\right) \tag{12}$$

- The **natural log of the second term** on the right-hand side of (11) is obtained using the following definitional rule of natural logarithms: $\ln\left(\exp[f(x)]\right) = \ln\left(e^{f(x)}\right) = f(x)$. Thus,

$$\ln\left\{ \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - x_i^T\beta)^2 \right] \right\} = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - x_i^T\beta)^2 \tag{13}$$

- Substituting the results in (12) and (13) into the right-hand side of equation (11) for $\ln L(\beta,\sigma^2;y)$ yields the sample log-likelihood function

$$\ln L(\beta,\sigma^2;y) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - x_i^T\beta)^2$$

- **<u>Result</u>: The sample log-likelihood function is**

$$\ln L(\beta,\sigma^2;y) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - x_i^T\beta)^2 \qquad (14.1)$$

$$= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) \qquad (14.2)$$

$$= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}RSS(\beta) \qquad (14.3)$$

❑ ***Second*, derive the ML estimator of the coefficient vector β.**

<u>Step 1</u>: Partially differentiate the log-likelihood function with respect to the coefficient vector β.

- Since

$$\ln L(\beta,\sigma^2;y) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}RSS(\beta)$$

$$\frac{\partial \ln L(\beta,\sigma^2;y)}{\partial \beta} = -\frac{1}{2\sigma^2}\frac{\partial RSS(\beta)}{\partial \beta} \qquad (15)$$

- Since

$$RSS(\beta) = (y - X\beta)^T(y - X\beta) = y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta$$

the K×1 vector of partial derivatives of $RSS(\beta)$ with respect to β is:

$$\frac{\partial RSS(\beta)}{\partial \beta} = -2\frac{\partial \beta^TX^Ty}{\partial \beta} + \frac{\partial \beta^TX^TX\beta}{\partial \beta} \qquad (16)$$

- Using the rules of matrix differentiation, we can show that the two partial derivatives on the right-hand side of equation (16) are:

$$\frac{\partial(\beta^T X^T y)}{\partial \beta} = X^T y \qquad \text{and} \qquad \frac{\partial(\beta^T X^T X \beta)}{\partial \beta} = 2 X^T X \beta \tag{17}$$

- Substitution of the partial derivatives (17) into equation (16) for $\partial \text{RSS}(\beta)/\partial \beta$ yields the following expression for the partial derivatives of $\text{RSS}(\beta)$ with respect to β:

$$\frac{\partial \text{RSS}(\beta)}{\partial \beta} = -2\frac{\partial \beta^T X^T y}{\partial \beta} + \frac{\partial \beta^T X^T X \beta}{\partial \beta} = -2 X^T y + 2 X^T X \beta. \tag{18}$$

- Substitute the right-hand side of (18) for $\partial \text{RSS}(\beta)/\partial \beta$ into equation (15) for $\partial \ln L(\beta, \sigma^2; y)/\partial \beta$:

$$\frac{\partial \ln L(\beta, \sigma^2; y)}{\partial \beta} = -\frac{1}{2\sigma^2}\frac{\partial \text{RSS}(\beta)}{\partial \beta} = -\frac{1}{2\sigma^2}\left(-2 X^T y + 2 X^T X \beta\right). \tag{19}$$

<u>Step 2</u>: Obtain the K = k+1 first-order conditions (FOCs) for the ML estimator of β by setting the partial derivatives $\partial \ln L(\beta, \sigma^2; y)/\partial \beta$ in (19) equal to zero.

$$\frac{\partial \ln L(\beta, \sigma^2; y)}{\partial \beta} = \underline{0} \quad \Rightarrow \quad -\frac{1}{2\sigma^2}\left(-2 X^T y + 2 X^T X \hat{\beta}_{ML}\right) = \underline{0}$$

$$\Rightarrow \qquad -2 X^T y + 2 X^T X \hat{\beta}_{ML} = \underline{0} \tag{20}$$

Note: The K first-order conditions (20) for the ML coefficient estimator $\hat{\beta}_{ML}$ of β are identical to the FOC's for the OLS estimator $\hat{\beta}_{OLS}$.

Implication: The FOC's (20) for $\hat{\beta}_{ML}$ thus imply that, for the CNLR model, the ML coefficient estimator $\hat{\beta}_{ML}$ equals the OLS estimator $\hat{\beta}_{OLS}$: i.e., $\hat{\beta}_{ML} = \hat{\beta}_{OLS}$.

Step 3: Solve the FOCs for the ML estimator $\hat{\beta}_{ML}$ -- that is, obtain an explicit expression for the ML estimator $\hat{\beta}_{ML}$ of β.

- Re-arrange matrix equation (20) by dividing both sides by 2 and then adding the term $X^T y$ to both sides:

$$-2X^T y + 2X^T X\hat{\beta}_{ML} = \underline{0}$$

$$-X^T y + X^T X\hat{\beta}_{ML} = \underline{0} \qquad \text{(dividing both sides by 2)}$$

$$X^T X\hat{\beta}_{ML} = X^T y \qquad \text{(adding } X^T y \text{ to both sides)} \qquad (21)$$

- Solve for $\hat{\beta}_{ML}$ by pre-multiplying both sides of matrix equation (21) by $(X^T X)^{-1}$, the inverse of $X^T X$:

$$(X^T X)^{-1} X^T X\hat{\beta}_{ML} = (X^T X)^{-1} X^T y$$

$$I_K \hat{\beta}_{ML} = (X^T X)^{-1} X^T y \qquad \text{since } (X^T X)^{-1} X^T X = I_K$$

$$\hat{\beta}_{ML} = (X^T X)^{-1} X^T y \qquad \text{since } I_K \hat{\beta}_{ML} = \hat{\beta}_{ML}.$$

- **Result:** The ML coefficient estimator $\hat{\beta}_{ML}$ is identical to the OLS estimator $\hat{\beta}_{OLS}$.

$$\hat{\beta}_{ML} = (X^T X)^{-1} X^T y = \hat{\beta}_{OLS} \qquad (22)$$

- ***Third*, derive the ML estimator of the scalar parameter σ^2, the error variance.**

  <u>Step 1</u>: Partially differentiate the log-likelihood function with respect to the scalar parameter σ^2.

  - Since the log-likelihood function is

    $$\ln L(\beta, \sigma^2; y) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}RSS(\beta)$$

    the partial derivative $\partial \ln L(\beta, \sigma^2; y)/\partial \sigma^2$ is:

    $$\frac{\partial \ln L(\beta, \sigma^2; y)}{\partial \sigma^2} = -\frac{N}{2}\frac{\partial \ln \sigma^2}{\partial \sigma^2} - \frac{RSS(\beta)}{2}\frac{\partial}{\partial \sigma^2}\left(\frac{1}{\sigma^2}\right)$$

    $$= -\frac{N}{2}\frac{\partial \ln \sigma^2}{\partial \sigma^2} - \frac{RSS(\beta)}{2}\frac{\partial (\sigma^2)^{-1}}{\partial \sigma^2} \qquad (23)$$

  - The partial derivative $\partial \ln \sigma^2/\partial \sigma^2$ in the first term on the right-hand side of (23) is evaluated using the following rule for differentiating natural logs: $\partial \ln a/\partial a = 1/a$ for $a > 0$. Thus

    $$\frac{\partial \ln \sigma^2}{\partial \sigma^2} = \frac{1}{\sigma^2} \qquad (24)$$

  - The partial derivative $\partial (\sigma^2)^{-1}/\partial \sigma^2$ in the second term on the right-hand side of (24) is evaluated using the following power rule of differentiation: $\partial a^n/\partial a = n a^{n-1}$. Thus

    $$\frac{\partial (\sigma^2)^{-1}}{\partial \sigma^2} = -1(\sigma^2)^{-2} = -\sigma^{-4} = -\frac{1}{\sigma^4} \qquad (25)$$

- Finally, substitute the partial derivatives (24) and (25) into equation (23) for $\partial \ln L(\beta, \sigma^2; y)/\partial \sigma^2$:

$$\frac{\partial \ln L(\beta, \sigma^2; y)}{\partial \sigma^2} = -\frac{N}{2}\frac{\partial \ln \sigma^2}{\partial \sigma^2} - \frac{RSS(\beta)}{2}\frac{\partial (\sigma^2)^{-1}}{\partial \sigma^2} \tag{23}$$

$$= -\frac{N}{2}\frac{1}{\sigma^2} - \frac{RSS(\beta)}{2}\left(-\frac{1}{\sigma^4}\right)$$

$$= -\frac{N}{2\sigma^2} + \frac{RSS(\beta)}{2\sigma^4} \tag{26}$$

<u>Step 2</u>: Obtain the first-order condition (FOC) for the ML estimator of σ^2 by setting the partial derivative $\partial \ln L(\beta, \sigma^2; y)/\partial \sigma^2$ in (26) equal to zero.

$$\frac{\partial \ln L(\beta, \sigma^2; y)}{\partial \sigma^2} = 0 \quad \Rightarrow \quad -\frac{N}{2\hat{\sigma}^2_{ML}} + \frac{RSS(\hat{\beta}_{ML})}{2\hat{\sigma}^4_{ML}} = 0 \tag{27}$$

where $\hat{\sigma}^2_{ML}$ denotes the ML estimator of σ^2 and $RSS(\hat{\beta}_{ML})$ denotes the residual sum of squares for the ML coefficient estimator:

$$RSS(\hat{\beta}_{ML}) = \left(y - X\hat{\beta}_{ML}\right)^T\left(y - X\hat{\beta}_{ML}\right) = y^T y - 2\hat{\beta}_{ML}{}^T X^T y + \hat{\beta}^T_{ML} X^T X \hat{\beta}_{ML}.$$

<u>Step 3</u>: Solve the first-order condition (FOC) for the ML estimator $\hat{\sigma}^2_{ML}$ -- that is, obtain an explicit expression for the ML estimator $\hat{\sigma}^2_{ML}$ of σ^2.

- Subtract the second term in equation (27) from both sides:

$$-\frac{N}{2\hat{\sigma}^2_{ML}} + \frac{RSS(\hat{\beta}_{ML})}{2\hat{\sigma}^4_{ML}} = 0$$

$$-\frac{N}{2\hat{\sigma}^2_{ML}} = -\frac{RSS(\hat{\beta}_{ML})}{2\hat{\sigma}^4_{ML}} \qquad (28)$$

- Multiply both sides of equation (28) by −2:

$$\frac{N}{\hat{\sigma}^2_{ML}} = \frac{RSS(\hat{\beta}_{ML})}{\hat{\sigma}^4_{ML}} \qquad (29)$$

- Multiply both sides of equation (29) by $\hat{\sigma}^4_{ML}$ and then divide by N:

$$\frac{\hat{\sigma}^4_{ML}}{\hat{\sigma}^2_{ML}} = \frac{RSS(\hat{\beta}_{ML})}{N} \qquad \Rightarrow \qquad \hat{\sigma}^2_{ML} = \frac{RSS(\hat{\beta}_{ML})}{N}$$

This is the expression for the ML estimator $\hat{\sigma}^2_{ML}$ of σ^2.

- **<u>Result</u>:** The ML estimator of the error variance σ^2 is the residual sum of squares divided by sample size N:

$$\hat{\sigma}^2_{ML} = \frac{RSS(\hat{\beta}_{ML})}{N} = \frac{\hat{u}^T\hat{u}}{N} = \frac{\sum_{i=1}^{N}\hat{u}_i^2}{N} \qquad (30)$$

where

$\hat{u} = y - X\hat{\beta}_{ML}$ = the K×1 ML residual vector
$\hat{u}_i = Y_i - x_i^T\hat{\beta}_{ML}$ = the i-th ML residual (i = 1, …, N).

4. Second Order Conditions for the OLS Coefficient Estimator

Although the second-order conditions for a minimum of the residual sum of squares function $\text{RSS}(\hat{\beta})$ with respect to $\hat{\beta}$ are a bit of a technical fine point, it is quite easy to state them.

- Recall that the vector of first-order partial derivatives of $\text{RSS}(\hat{\beta})$ with respect to $\hat{\beta}$ are

$$\frac{\partial \text{RSS}(\hat{\beta})}{\partial \hat{\beta}} = -2X^{T}y + 2X^{T}X\hat{\beta}. \tag{18}$$

  For given sample data (y, X), these partial derivatives are simply a linear function of the coefficient estimator $\hat{\beta}$.

- The second-order derivatives of $\text{RSS}(\hat{\beta})$ with respect to $\hat{\beta}$ are given by the K×K matrix

$$\frac{\partial^2 \text{RSS}(\hat{\beta})}{\partial \hat{\beta} \partial \hat{\beta}^{T}} = \frac{\partial}{\partial \hat{\beta}^{T}}\left(\frac{\partial \text{RSS}(\hat{\beta})}{\partial \hat{\beta}}\right) = 2\frac{\partial (X^{T}X\hat{\beta})}{\partial \hat{\beta}^{T}}. \tag{31}$$

  We need to evaluate the paritial derivative $\partial(X^{T}X\hat{\beta})/\partial\hat{\beta}^{T}$ in (31), where the K×1 vector $X^{T}X\hat{\beta}$ is a set of K linear functions of the coefficient estimator $\hat{\beta}$.

- **Rule : Vector differentiation of a set of linear functions**

  If f(x) is a set of K linear functions of the K×1 vector x given by

$$f(x) = Ax = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1K} \\ A_{21} & A_{22} & \cdots & A_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ A_{K1} & A_{K2} & \cdots & A_{KK} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{j=K} A_{1j}x_j \\ \sum_{j=1}^{j=K} A_{2j}x_j \\ \vdots \\ \sum_{j=1}^{j=K} A_{Kj}x_j \end{bmatrix} = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_K(x) \end{bmatrix}$$

where A is a K×K matrix, then the K×K matrix vector of partial derivatives of f(x) = Ax with respect to the vector x is

$$\frac{\partial(\mathrm{Ax})}{\partial \mathrm{x}^{\mathrm{T}}} = \mathrm{A} \quad \textit{or} \quad \frac{\partial(\mathrm{Ax})}{\partial \mathrm{x}} = \mathrm{A}^{\mathrm{T}}. \tag{32}$$

- Apply Rule 3 to the term $\mathrm{X}^{\mathrm{T}}\mathrm{X}\hat{\beta}$. Let $\mathrm{x} = \hat{\beta}$ and $\mathrm{A} = \mathrm{X}^{\mathrm{T}}\mathrm{X}$.

$$\frac{\partial(\mathrm{Ax})}{\partial \mathrm{x}^{\mathrm{T}}} = \mathrm{A} \quad \Rightarrow \quad \frac{\partial(\mathrm{X}^{\mathrm{T}}\mathrm{X}\hat{\beta})}{\partial \hat{\beta}^{\mathrm{T}}} = \mathrm{X}^{\mathrm{T}}\mathrm{X}.$$

- The K×K matrix of second-order derivatives of $\mathrm{RSS}(\hat{\beta})$ with respect to $\hat{\beta}$ is therefore

$$\frac{\partial^2 \mathrm{RSS}(\hat{\beta})}{\partial \hat{\beta}\,\partial \hat{\beta}^{\mathrm{T}}} = 2\frac{\partial(\mathrm{X}^{\mathrm{T}}\mathrm{X}\hat{\beta})}{\partial \hat{\beta}^{\mathrm{T}}} = 2\,\mathrm{X}^{\mathrm{T}}\mathrm{X}. \tag{33}$$

- The second-order condition for $\hat{\beta}_{\mathrm{OLS}}$ to correspond to a minimum of the $\mathrm{RSS}(\hat{\beta})$ function is that

$\mathrm{X}^{\mathrm{T}}\mathrm{X}$ be a positive definite matrix. (34)

- How do we know that $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is a positive definite matrix? It turns out that this is ensured by the full rank assumption A5:

$\mathrm{rank}(\mathrm{X}) = \mathrm{K} \quad \Rightarrow \quad \mathrm{X}^{\mathrm{T}}\mathrm{X}$ is positive definite.

5. Statistical Properties of the ML Parameter Estimators

- The **ML estimators of β and σ^2** share the **three large sample properties** of all ML estimators: consistency, asymptotic efficiency, and asymptotic normality.

  1. **Consistency**: the probability limit of $\hat{\beta}_{ML} = \beta$; $\text{plim}(\hat{\beta}_{ML}) = \beta$.

  2. **Asymptotic efficiency**: Asy $\text{Var}(\hat{\beta}_{j,ML}) \leq$ Asy $\text{Var}(\tilde{\beta}_j)$, the asymptotic variance of any other consistent estimator $\tilde{\beta}_j$ of β_j.

  3. **Asymptotic normality**: $\hat{\beta}_{ML} \overset{a}{\sim} N\left[\beta, \sigma^2\left(X^T X\right)^{-1}\right]$.

- In addition, the ML coefficient estimator $\hat{\beta}_{ML}$ shares the **small sample properties** of the OLS coefficient estimator $\hat{\beta}_{OLS}$, since

$$\hat{\beta}_{ML} = \hat{\beta}_{OLS} = \left(X^T X\right)^{-1} X^T y.$$

  4. **Unbiasedness**: the mean of $\hat{\beta}_{ML} = \beta$; $E(\hat{\beta}_{ML}) = \beta$.

  5. **Efficiency**: In the class of all linear unbiased estimators of β, $\hat{\beta}_{ML}$ is the ***minimum variance*** **estimator of β**. If $\tilde{\beta}$ denotes any other linear unbiased estimator of β,

$$\text{Var}(\hat{\beta}_{j,ML}) = \text{Var}(\hat{\beta}_{j,OLS}) \leq \text{Var}(\tilde{\beta}_j) \qquad j = 0, 1, 2, \ldots, k.$$