8.1 Example: Mixed Estimation (Theil and Goldberger Model)

A generalization of the restricted OLS $\Longrightarrow$ Stochastic linear restriction:

$$r = R\beta + v, \qquad \mathrm{E}(v) = 0 \text{ and } \mathrm{V}(v) = \sigma^2\Psi$$

$$y = X\beta + u, \qquad \mathrm{E}(u) = 0 \text{ and } \mathrm{V}(u) = \sigma^2 I_n$$

Using a matrix form,

$$\begin{pmatrix} y \\ r \end{pmatrix} = \begin{pmatrix} X \\ R \end{pmatrix}\beta + \begin{pmatrix} u \\ v \end{pmatrix}, \qquad \mathrm{E}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } \mathrm{V}\begin{pmatrix} u \\ v \end{pmatrix} = \sigma^2\begin{pmatrix} I_n & 0 \\ 0 & \Psi \end{pmatrix}$$

For estimation, we do not need normality assumption.

Applying GLS, we obtain:

$$\begin{aligned} b &= \left(\begin{pmatrix} X' & R' \end{pmatrix} \begin{pmatrix} I_n & 0 \\ 0 & \Psi \end{pmatrix}^{-1} \begin{pmatrix} X \\ R \end{pmatrix} \right)^{-1} \left(\begin{pmatrix} X' & R' \end{pmatrix} \begin{pmatrix} I_n & 0 \\ 0 & \Psi \end{pmatrix}^{-1} \begin{pmatrix} y \\ r \end{pmatrix} \right) \\ &= \left(X'X + R'\Psi^{-1}R \right)^{-1} \left(X'y + R'\Psi^{-1}r \right). \end{aligned}$$

Mean and Variance of b: b is rewritten as follows:

$$b = \left(\begin{pmatrix} X' & R' \end{pmatrix} \begin{pmatrix} I_n & 0 \\ 0 & \Psi \end{pmatrix}^{-1} \begin{pmatrix} X \\ R \end{pmatrix} \right)^{-1} \left(\begin{pmatrix} X' & R' \end{pmatrix} \begin{pmatrix} I_n & 0 \\ 0 & \Psi \end{pmatrix}^{-1} \begin{pmatrix} y \\ r \end{pmatrix} \right)$$

$$= \beta + \left(\begin{pmatrix} X' & R' \end{pmatrix} \begin{pmatrix} I_n & 0 \\ 0 & \Psi \end{pmatrix}^{-1} \begin{pmatrix} X \\ R \end{pmatrix} \right)^{-1} \begin{pmatrix} u \\ v \end{pmatrix}$$

Therefore, the mean and variance are given by:

$$\mathrm{E}(b) = \beta \qquad \Longrightarrow \qquad b \text{ is unbiased.}$$

$$\mathrm{V}(b) = \sigma^2 \left(\begin{pmatrix} X' & R' \end{pmatrix} \begin{pmatrix} I_n & 0 \\ 0 & \Psi \end{pmatrix}^{-1} \begin{pmatrix} X \\ R \end{pmatrix} \right)^{-1}$$

$$= \sigma^2 \left(X'X + R'\Psi^{-1}R \right)^{-1}$$

9 Maximum Likelihood Estimation (MLE, 最尤法)

$\longrightarrow$ **Review**

1. The distribution function of $\{X_i\}_{i=1}^n$ is $f(x;\theta)$, where $x = (x_1, x_2, \cdots, x_n)$.

   θ is a vector or matrix of unknown parameters, e.g., $\theta = (\mu, \Sigma)$, where $\mu = \mathrm{E}(X_i)$ and $\Sigma = \mathrm{V}(X_i)$.

   Note that X is a vector of random variables and x is a vector of their realizations (i.e., observed data).

   Likelihood function $L(\cdot)$ is defined as $L(\theta; x) = f(x;\theta)$.

   Note that $f(x;\theta) = \prod_{i=1}^n f(x_i;\theta)$ when X_1, X_2, $\cdots$, X_n are mutually independently and identically distributed.

The maximum likelihood estimate (MLE) of θ is the θ such that:

$$\max_{\theta} \; L(\theta; x). \qquad \Longleftrightarrow \qquad \max_{\theta} \; \log L(\theta; x).$$

Thus, MLE satisfies the following two conditions:

(a) $\dfrac{\partial \log L(\theta; x)}{\partial \theta} = 0. \quad \Longrightarrow \quad$ Solution of θ: $\tilde{\theta} = \tilde{\theta}(x)$

(b) $\dfrac{\partial^2 \log L(\theta; x)}{\partial \theta \partial \theta'}$ is a negative definite matrix.

2. $x = (x_1, x_2, \cdots, x_n)$ are used as the observations (i.e., observed data).

   $X = (X_1, X_2, \cdots, X_n)$ denote the random variables associated with the joint distribution $f(x; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$.

3. Replacing x by X, we otain the maximum likelihood **estimator** (MLE, which is the same word as the maximum likelihood **estimate**).

   That is, MLE of θ satisfies the following two conditions:

   (a) $\dfrac{\partial \log L(\theta; X)}{\partial \theta} = 0. \qquad \Longrightarrow \qquad$ Solution of θ: $\tilde{\theta} = \tilde{\theta}(X)$

   (b) $\dfrac{\partial^2 \log L(\theta; X)}{\partial \theta \partial \theta'}$ is a negative definite matrix.

4. **Fisher's information matrix (フィッシャーの情報行列)** or simply **information matrix**, denoted by $I(\theta)$, is given by:

$$I(\theta) = -\mathrm{E}\Big(\frac{\partial^2 \log L(\theta; X)}{\partial \theta \partial \theta'}\Big),$$

   where we have the following equality:

$$-\mathrm{E}\Big(\frac{\partial^2 \log L(\theta; X)}{\partial \theta \partial \theta'}\Big) = \mathrm{E}\Big(\frac{\partial \log L(\theta; X)}{\partial \theta}\frac{\partial \log L(\theta; X)}{\partial \theta'}\Big) = \mathrm{V}\Big(\frac{\partial \log L(\theta; X)}{\partial \theta}\Big)$$

   Note that $\mathrm{E}(\cdot)$ and $\mathrm{V}(\cdot)$ are expected with respect to X.

Proof of the above equality:

$$\int L(\theta; x)\mathrm{d}x = 1$$

Take a derivative with respect to θ.

$$\int \frac{\partial L(\theta; x)}{\partial \theta}\mathrm{d}x = 0$$

(We assume that (i) the domain of x does not depend on θ and (ii) the derivative $\dfrac{\partial L(\theta; x)}{\partial \theta}$ exists.)

Rewriting the above equation, we obtain:

$$\int \frac{\partial \log L(\theta; x)}{\partial \theta} L(\theta; x)\mathrm{d}x = 0,$$

i.e.,

$$\mathrm{E}\left(\frac{\partial \log L(\theta; X)}{\partial \theta}\right) = 0.$$

Again, differentiating the above with respect to θ, we obtain:

$$
\begin{aligned}
&\int \frac{\partial^2 \log L(\theta; x)}{\partial\theta\partial\theta'} L(\theta; x)\mathrm{d}x + \int \frac{\partial \log L(\theta; x)}{\partial\theta} \frac{\partial L(\theta; x)}{\partial'\theta}\mathrm{d}x \\
&\quad = \int \frac{\partial^2 \log L(\theta; x)}{\partial\theta\partial\theta'} L(\theta; x)\mathrm{d}x + \int \frac{\partial \log L(\theta; x)}{\partial\theta} \frac{\partial \log L(\theta; x)}{\partial\theta'} L(\theta; x)\mathrm{d}x \\
&\quad = \mathrm{E}\Big(\frac{\partial^2 \log L(\theta; X)}{\partial\theta\partial\theta'}\Big) + \mathrm{E}\Big(\frac{\partial \log L(\theta; X)}{\partial\theta} \frac{\partial \log L(\theta; X)}{\partial\theta'}\Big) = 0.
\end{aligned}
$$

Therefore, we can derive the following equality:

$$
-\mathrm{E}\left(\frac{\partial^2 \log L(\theta; X)}{\partial\theta\partial\theta'}\right) = \mathrm{E}\left(\frac{\partial \log L(\theta; X)}{\partial\theta} \frac{\partial \log L(\theta; X)}{\partial\theta'}\right) = \mathrm{V}\left(\frac{\partial \log L(\theta; X)}{\partial\theta}\right),
$$

where the second equality utilizes $\mathrm{E}\left(\frac{\partial \log L(\theta; X)}{\partial\theta}\right) = 0$.

5. **Cramer-Rao Lower Bound (クラメール・ラオの下限)** is given by: $(I(\theta))^{-1}$.

   Suppose that an estimator of θ is given by $s(X)$.

The expectation of $s(X)$ is:

$$\mathrm{E}(s(X)) = \int s(x)L(\theta; x)\mathrm{d}x.$$

Differentiating the above with respect to θ,

$$\begin{aligned}\frac{\partial \mathrm{E}(s(X))}{\partial \theta} &= \int s(x)\frac{\partial L(\theta; x)}{\partial \theta}\mathrm{d}x = \int s(x)\frac{\partial \log L(\theta; x)}{\partial \theta}L(\theta; x)\mathrm{d}x \\ &= \mathrm{Cov}\left(s(X), \frac{\partial \log L(\theta; X)}{\partial \theta}\right)\end{aligned}$$

For simplicity, let $s(X)$ and θ be scalars.

Then,

$$\begin{aligned}\left(\frac{\partial \mathrm{E}(s(X))}{\partial \theta}\right)^2 &= \left(\mathrm{Cov}\left(s(X), \frac{\partial \log L(\theta; X)}{\partial \theta}\right)\right)^2 = \rho^2 \mathrm{V}\left(s(X)\right)\mathrm{V}\left(\frac{\partial \log L(\theta; X)}{\partial \theta}\right) \\ &\leq \mathrm{V}\left(s(X)\right)\mathrm{V}\left(\frac{\partial \log L(\theta; X)}{\partial \theta}\right),\end{aligned}$$

where ρ denotes the correlation coefficient between $s(X)$ and $\dfrac{\partial \log L(\theta; X)}{\partial \theta}$, i.e.,

$$\rho = \frac{\operatorname{Cov}\left(s(X), \dfrac{\partial \log L(\theta; X)}{\partial \theta}\right)}{\sqrt{\mathrm{V}\left(s(X)\right)}\sqrt{\mathrm{V}\left(\dfrac{\partial \log L(\theta; X)}{\partial \theta}\right)}}.$$

Note that $|\rho| \leq 1$.

Therefore, we have the following inequality:

$$\left(\frac{\partial \mathrm{E}(s(X))}{\partial \theta}\right)^2 \leq \mathrm{V}(s(X))\, \mathrm{V}\left(\frac{\partial \log L(\theta; X)}{\partial \theta}\right),$$

i.e.,

$$\mathrm{V}(s(X)) \geq \frac{\left(\dfrac{\partial \mathrm{E}(s(X))}{\partial \theta}\right)^2}{\mathrm{V}\left(\dfrac{\partial \log L(\theta; X)}{\partial \theta}\right)}$$

Especially, when $\mathrm{E}(s(X)) = \theta$, i.e., when $s(X)$ is an unbiased estimator of θ, the numerator of the right-hand side leads to one.

Therefore, we obtain:

$$\mathrm{V}(s(X)) \geq \frac{1}{-\mathrm{E}\left(\dfrac{\partial^2 \log L(\theta; X)}{\partial\theta^2}\right)} = (I(\theta))^{-1}.$$

Even in the case where $s(X)$ is a vector, the following inequality holds.

$$\mathrm{V}(s(X)) \geq (I(\theta))^{-1},$$

where $I(\theta)$ is defined as:

$$\begin{aligned} I(\theta) &= -\mathrm{E}\left(\frac{\partial^2 \log L(\theta; X)}{\partial\theta\partial\theta'}\right) \\ &= \mathrm{E}\left(\frac{\partial \log L(\theta; X)}{\partial\theta}\frac{\partial \log L(\theta; X)}{\partial\theta'}\right) = \mathrm{V}\left(\frac{\partial \log L(\theta; X)}{\partial\theta}\right). \end{aligned}$$

The variance of any unbiased estimator of θ is larger than or equal to $(I(\theta))^{-1}$.

Thus, $(I(\theta))^{-1}$ results in the lower bound of the variance of any unbiased estimator of θ.

6. Asymptotic Normality of MLE:

   Let $\tilde{\theta}$ be MLE of θ.

   As n goes to infinity, we have the following result:

$$\sqrt{n}(\tilde{\theta} - \theta) \longrightarrow N\left(0, \lim_{n\to\infty}\left(\frac{I(\theta)}{n}\right)^{-1}\right),$$

   where it is assumed that $\lim_{n\to\infty}\left(\frac{I(\theta)}{n}\right)$ converges.

   $\longrightarrow$ The proof will be shown later.

That is, when n is large, $\tilde{\theta}$ is approximately distributed as follows:

$$\tilde{\theta} \sim N\Big(\theta, (I(\theta))^{-1}\Big).$$

Suppose that $s(X) = \tilde{\theta}$.

When n is large, $\mathrm{V}(s(X))$ is approximately equal to $(I(\theta))^{-1}$.

7. **Optimization (最適化):**

MLE of θ results in the following maximization problem:

$$\max_{\theta} \quad \log L(\theta; x).$$

We often have the case where the solution of θ is not derived in closed form.

$\Longrightarrow$ Optimization procedure

$$0 = \frac{\partial \log L(\theta; x)}{\partial \theta} = \frac{\partial \log L(\theta^*; x)}{\partial \theta} + \frac{\partial^2 \log L(\theta^*; x)}{\partial \theta \partial \theta'}(\theta - \theta^*).$$

Solving the above equation with respect to θ, we obtain the following:

$$\theta = \theta^* - \left(\frac{\partial^2 \log L(\theta^*; x)}{\partial\theta\partial\theta'}\right)^{-1} \frac{\partial \log L(\theta^*; x)}{\partial\theta}.$$

Replace the variables as follows:

$$\theta \longrightarrow \theta^{(i+1)}$$

$$\theta^* \longrightarrow \theta^{(i)}$$

Then, we have:

$$\theta^{(i+1)} = \theta^{(i)} - \left(\frac{\partial^2 \log L(\theta^{(i)}; x)}{\partial\theta\partial\theta'}\right)^{-1} \frac{\partial \log L(\theta^{(i)}; x)}{\partial\theta}.$$

$\Longrightarrow$ **Newton-Raphson method (ニュートン・ラプソン法)**

Replacing $\dfrac{\partial^2 \log L(\theta^{(i)}; x)}{\partial\theta\partial\theta'}$ by $\mathrm{E}\left(\dfrac{\partial^2 \log L(\theta^{(i)}; x)}{\partial\theta\partial\theta'}\right)$, we obtain the following optimization algorithm:

$$\begin{aligned}
\theta^{(i+1)} &= \theta^{(i)} - \left(\mathrm{E}\left(\frac{\partial^2 \log L(\theta^{(i)}; x)}{\partial\theta\partial\theta'} \right) \right)^{-1} \frac{\partial \log L(\theta^{(i)}; x)}{\partial\theta} \\
&= \theta^{(i)} + \left(I(\theta^{(i)}) \right)^{-1} \frac{\partial \log L(\theta^{(i)}; x)}{\partial\theta}
\end{aligned}$$

$\Longrightarrow$ **Method of Scoring (スコア法)**