which is called GLS (Generalized Least Squares) estimator.

b is rewritten as follows:

$$b = \beta + (X^{\star\prime}X^{\star})^{-1}X^{\star\prime}u^{\star} = \beta + (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}u$$

The mean and variance of b are given by:

$$\mathrm{E}(b) = \beta,$$

$$\mathrm{V}(b) = \sigma^2(X^{\star\prime}X^{\star})^{-1} = \sigma^2(X'\Omega^{-1}X)^{-1}.$$

6. Suppose that the regression model is given by:

$$y = X\beta + u, \qquad u \sim N(0, \sigma^2\Omega).$$

In this case, when we use OLS, what happens?

$$\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'u$$

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$$

Compare GLS and OLS.

(a) Expectation:

$$E(\hat{\beta}) = \beta, \quad \text{and} \quad E(b) = \beta$$

Thus, both $\hat{\beta}$ and b are unbiased estimator.

(b) Variance:

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$$

$$V(b) = \sigma^2(X'\Omega^{-1}X)^{-1}$$

Which is more efficient, OLS or GLS?.

$$
\begin{aligned}
\mathrm{V}(\hat{\beta}) - \mathrm{V}(b) &= \sigma^2 (X'X)^{-1} X'\Omega X (X'X)^{-1} - \sigma^2 (X'\Omega^{-1}X)^{-1} \\
&= \sigma^2 \Big((X'X)^{-1}X' - (X'\Omega^{-1}X)^{-1}X'\Omega^{-1} \Big) \Omega \\
&\qquad\qquad \times \Big((X'X)^{-1}X' - (X'\Omega^{-1}X)^{-1}X'\Omega^{-1} \Big)' \\
&= \sigma^2 A\Omega A'
\end{aligned}
$$

Note that A is $k \times n$ and Ω is $n \times n$.

Ω is the variance-covariance matrix of u, which is a positive definite matrix.

Therefore, except for $\Omega = I_n$, $A\Omega A'$ is also a positive definite matrix.

(From $\Omega = PP'$ and $A\Omega A' = AP(AP)'$, we have $xAP(xAP)' = \sum_{i=1}^{k} z_i^2 > 0$ for $x \neq 0$, where x is $1 \times k$, $z = xAP$ is $1 \times k$ and $z = (z_1, z_2, \cdots, z_k)$.)

This implies that $\mathrm{V}(\hat{\beta}_i) - \mathrm{V}(b_i) > 0$ for the ith element of β.

Accordingly, b is more efficient than $\hat{\beta}$.

7. If $u \sim N(0, \sigma^2\Omega)$, then $b \sim N(\beta, \sigma^2(X'\Omega^{-1}X)^{-1})$.

   Consider testing the hypothesis $H_0 : R\beta = r$.

   $R : \; G \times k, \quad \text{rank}(R) = G \leq k.$

   $Rb \sim N(R\beta, \sigma^2 R(X'\Omega^{-1}X)^{-1}R')$.

   Therefore, the following quadratic form is distributed as:

$$\frac{(Rb - r)'(R(X'\Omega^{-1}X)^{-1}R')^{-1}(Rb - r)}{\sigma^2} \sim \chi^2(G)$$

8. Because $(y^\star - X^\star b)'(y^\star - X^\star b)/\sigma^2 \sim \chi^2(n - k)$, we obtain:

$$\frac{(y - Xb)'\Omega^{-1}(y - Xb)}{\sigma^2} \sim \chi^2(n - k)$$

9. Furthermore, from the fact that b is independent of $y - Xb$, the following F distribution can be derived:

$$\frac{(Rb - r)'(R(X'\Omega^{-1}X)^{-1}R')^{-1}(Rb - r)/G}{(y - Xb)'\Omega^{-1}(y - Xb)/(n - k)} \sim F(G, n - k)$$

10. Let b be the unrestricted GLSE and $\tilde{b}$ be the restricted GLSE.

    Their residuals are given by e and $\tilde{u}$, respectively.

$$e = y - Xb, \qquad \tilde{u} = y - X\tilde{b}$$

    Then, the F test statistic is written as follows:

$$\frac{(\tilde{u}'\Omega^{-1}\tilde{u} - e'\Omega^{-1}e)/G}{e'\Omega^{-1}e/(n - k)} \sim F(G, n - k)$$

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.)

(*) **Differentiation of Composite Functions (合成関数の微分)** or **Chain rule (連鎖律)**:

$$\frac{\partial \log L(\theta; x)}{\partial \theta} = \frac{\partial \log L(\theta; x)}{\partial L(\theta; x)} \frac{\partial L(\theta; x)}{\partial \theta} = \frac{1}{L(\theta; x)} \frac{\partial L(\theta; x)}{\partial \theta}$$

i.e.,

$$\frac{\partial L(\theta; x)}{\partial \theta} = \frac{\partial \log L(\theta; x)}{\partial \theta} L(\theta; x)$$

Rewriting the above equation, we obtain:

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

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

5. **Cramer-Rao inequality (クラメール・ラオの不等式)** is given by:

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

where $s(X)$ denotes an unbiased estimator of θ.

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

Proof:

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,

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

$$= \mathrm{Cov}\left(s(X), \frac{\partial \log L(\theta; X)}{\partial \theta}\right)$$

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

Then,

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

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

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

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