OLS, GLS and ML Estimation

I. Ordinary Least Squares Estimation:

- For a linear model

$$Y_j = \beta_0 + \beta_1 X_{1j} + \cdots + \beta_r X_{rj} + \epsilon_j,$$

the OLS estimator for

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_r \end{bmatrix} \quad \text{is any} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ \vdots \\ b_r \end{bmatrix}$$

that minimizes the sum of squared residuals

$$Q(\mathbf{b}) = \sum_{j=1}^{n} (Y_j - b_0 - b_1 X_{1j} - \cdots - b_r X_{rj})^2.$$

- The estimating equations (normal equations) are

$$\frac{\partial Q(\mathbf{b})}{\partial b_0} = -2\sum_{j=1}^{n}(Y_j - b_0 - b_1 X_{1j} \cdots - \beta_r X_{rj}) = 0$$

and

$$\frac{\partial Q(\mathbf{b})}{\partial b_i} = -2\sum_{j=1}^{n} X_{ij}(Y_j - b_0 - b_1 X_{1j} \cdots - b_r X_{rj}) = 0, \quad \text{for } i = 1, 2, \ldots, r$$

The matrix form of these equations is

$$(X^T X)\mathbf{b} = X^T \mathbf{Y}$$

and a solution is

$$\mathbf{b} = (X^T X)^{-} X^T \mathbf{Y}.$$

The OLS estimator for an **estimable** function $C^T\beta$ is

$$C^T\mathbf{b} = C^T(X^TX)^-X^T\mathbf{Y}$$

for any solution to the normal equations.

- $E(C^T\mathbf{b}) = C^T\beta$
- $Var(C^T\mathbf{b}) = C^T(X^TX)^-X^T\Sigma X[(X^TX)^-]^TC$, where $\Sigma = Var(\mathbf{Y})$.
- The distribution of $\mathbf{Y}$ is not completely specified.

For a Gauss-Markov model with

$$E(\mathbf{Y}) = X\boldsymbol{\beta} \;\text{ and }\; Var(\mathbf{Y}) = \sigma^2 I$$

the OLS estimator of an estimable function $C^T\boldsymbol{\beta}$ is the unique best linear unbiased estimator (b.l.u.e.) of $C^T\boldsymbol{\beta}$.

- $E(C^T\mathbf{b}) = C^T\boldsymbol{\beta}$
- $Var(C^T\mathbf{b}) = \sigma^2 C^T(X^TX)^- C$ is smaller than the variance of any other linear unbiased estimator for $C^T\boldsymbol{\beta}$.
- The distribution of $\mathbf{Y}$ is not completely specified.

II. Generalized Least Squares Estimation

Consider the Aitken model

$$E(\mathbf{Y}) = X\boldsymbol{\beta} \quad \text{and} \quad Var(\mathbf{Y}) = \sigma^2 V$$

where V is a positive definite symmetric matrix of known constants and σ^2 is an unknown variance parameter.

- A GLS estimator for $\boldsymbol{\beta}$ is any $\mathbf{b}$ that minimizes

  $$Q(\mathbf{b}) = (\mathbf{Y} - X\mathbf{b})^T V^{-1}(\mathbf{Y} - X\mathbf{b})$$

  (from Definition 3.8 with $\Sigma = \sigma^2 V$).

- The estimating equations are

$$(X^T V^{-1} X)\mathbf{b} = X^T V^{-1} \mathbf{Y}.$$

- A solution is

$$\mathbf{b}_{GLS} = (X^T V^{-1} X)^{-} X^T V^{-1} \mathbf{Y}.$$

- For any estimable function $C^T \beta$ the unique b.l.u.e. is

$$C^T \mathbf{b}_{GLS} = C^T (X^T V^{-1} X)^{-} X^T V^{-1} \mathbf{Y}$$

for any solution to the normal equations.

- $E(C^T\mathbf{b}) = C^T\beta$ and $Var(C^T\mathbf{b}) = \sigma^2 C^T(X^T V^{-1} X)^- C$.
- The distribution of $\mathbf{Y}$ is not completely specified.
- An unbiased estimator for σ^2 in the Aitken model is

$$\hat{\sigma}^2_{GLS} = \frac{\mathbf{Y}^T\left[V^{-1} - V^{-1}X(X^T V^{-1} X)^- X^T V^{-1}\right]\mathbf{Y}}{n - rank(X)}$$
$$= \frac{(\mathbf{Y} - X\mathbf{b}_{\text{GLS}})^T V^{-1}(\mathbf{Y} - X\mathbf{b}_{\text{GLS}})}{n}$$

- In practice, V may not be known. Then $\mathbf{b}_{GLS}$ and σ^2_{GLS} can be approximated by replacing V with a consistent estimator:
  - The estimator for $C^T\beta$ is not b.l.u.e.
  - The estimator for σ^2 is not unbiased.
  - Both estimators are consistent.

III. Maximum Likelihood Estimation

The model must include a specification of the joint distribution of the observations.

Example: Normal theory Gauss-Markov model:

$$Y_j = \beta_0 + \beta_1 X_{1_j} + \cdots + \beta_r X_{r_j} + \epsilon_j$$

where

$$\epsilon_j \sim \text{NID}(0, \sigma^2), \qquad j = 1, \ldots, n$$

or

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$$

- Find the parameter values that maximize the *likelihood* of the observed data.

  For the normal-theory Gauss-Markov model, the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2; Y_1, \dots, Y_n) = \frac{1}{(2\pi)^{n/2}\sigma^n} \; e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-X\boldsymbol{\beta})^T(\mathbf{Y}-X\boldsymbol{\beta})}$$

  Find values of $\boldsymbol{\beta}$ and σ^2 that maximize this likelihood function.

- This is equivalent to finding values of $\boldsymbol{\beta}$ and σ^2 that maximize the log-likelihood.

$$
\begin{aligned}
\ell(\boldsymbol{\beta}, \sigma^2; Y_1, \ldots, Y_n) &= \log\left(L(\boldsymbol{\beta}, \sigma^2; Y_1, \ldots, Y_n)\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) \\
&\qquad -\frac{1}{2\sigma^2}(\mathbf{Y} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\boldsymbol{\beta}) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) \\
&\qquad -\frac{1}{2\sigma^2}\underline{\sum_{j=1}^{n}(Y_j - \beta_0 - \cdots - \beta_r X_{r_j})^2}
\end{aligned}
$$

$\nearrow$

this is minimized by an OLS estimator
for β regardless of the value of σ^2

Solve the likelihood equations:

$$0 = \frac{\partial \ell(\boldsymbol{\beta}, \sigma^2; \mathbf{Y})}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{j=1}^{n} (Y_j - \beta_0 - \cdots - \beta_r X_{r_j})$$

$$0 = \frac{\partial \ell(\boldsymbol{\beta}, \sigma^2; \mathbf{Y})}{\partial \beta_i} = \frac{1}{\sigma^2} \sum_{j=1}^{n} X_{ij} (Y_j - \beta_0 - \cdots - \beta_r X_{r_j})$$

$$\text{for } i = 1, 2, \ldots, r$$

$$0 = \frac{\partial \ell(\boldsymbol{\beta}, \sigma^2; \mathbf{Y})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{j=1}^{n} (Y_j - \beta_0 - \cdots - \beta_r X_{r_j})^2$$

Solution:

$$\hat{\beta} = \mathbf{b}_{\text{OLS}} = (X^T X)^- X^T \mathbf{Y}$$

and

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n} \sum_{j=1}^{n} (Y_j - \hat{\beta}_0 - \cdots - \hat{\beta}_r X_{r_j})^2 \\
&= \frac{1}{n} \mathbf{Y}^T (I - P_X) \mathbf{Y} = \frac{1}{n} \text{ SSE}
\end{aligned}$$

- This is a biased estimator for σ^2.
- $[n - \text{rank}(X)^{-1} SSE$ is an unbiased estimator for σ^2.
- $n^{-1} SSE$ and $[n - \text{rank}(X)]^{-1} SSE$ are asymptotically equivalent.

Example: Normal-theory Aitken model

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 V)$ and V is a known positive definite matrix.

The multivariate normal likelihood function is

$$L(\boldsymbol{\beta}; \mathbf{Y}) = \frac{1}{(2\pi\sigma^2)^{n/2}|V|^{1/2}}\ e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-X\boldsymbol{\beta})^T V^{-1}(\mathbf{Y}-X\boldsymbol{\beta})}$$

The log-likelihood function is

$$\begin{aligned}\ell(\boldsymbol{\beta}; \mathbf{Y}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|V|) - \frac{n}{2}log(\sigma^2) \\ &\quad - \frac{1}{2\sigma^2}(\mathbf{Y} - X\boldsymbol{\beta})^T V^{-1}(\mathbf{Y} - X\boldsymbol{\beta})\end{aligned}$$

For any value of σ, the log-likelihood is maximized by finding a $\boldsymbol{\beta}$ that minimizes

$$(\mathbf{Y} - X\boldsymbol{\beta})^T V^{-1}(\mathbf{Y} - X\boldsymbol{\beta})$$

The estimating equations are

$$(X^T V^{-1} X)\boldsymbol{\beta} = X^T V^{-1}\mathbf{Y}$$

Solutions are of the form

$$\hat{\boldsymbol{\beta}} = \mathbf{b}_{\text{GLS}} = (X^T V^{-1} X)^{-} X^T V^{-1}\mathbf{Y}$$

When V is known the mle for $\boldsymbol{\beta}$ is also the generalized least squares estimator.

The additional estimating equation corresponding to σ^2 is

$$\begin{aligned} 0 &= \frac{\partial \ell(\boldsymbol{\beta}, \sigma^2; \mathbf{Y})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} \\ &\quad + \frac{1}{2(\sigma^2)^2} (\mathbf{Y} - X\boldsymbol{\beta})^T V^{-1} (\mathbf{Y} - X\boldsymbol{\beta}) \end{aligned}$$

Substituting the solution to the other estimating equations for $\boldsymbol{\beta}$, the solution is

$$\hat{\sigma}^2 = \frac{1}{n} (\mathbf{Y} - X\mathbf{b}_{\text{GLS}})^T V^{-1} (\mathbf{Y} - X\mathbf{b}_{\text{GLS}})$$

This is a biased estimator for σ^2.

When V contains unknown parameters:

- You could maximize the log-likelihood

$$\ell(\boldsymbol{\beta}, \Sigma; \mathbf{Y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|V|) - \frac{n}{2}\ log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y} - X\boldsymbol{\beta})^T\ V^{-1}(\mathbf{Y} - X\boldsymbol{\beta})$$

  with respect to $\boldsymbol{\beta}$, σ^2 and the parameters in V.
- There may be no algebraic formulas for the solutions to the joint likelihood equations.
- The MLE's for σ^2 and the parameters in V are usually biased (too small).
- REML estimates are often used.

General Properties of MLE's

Regularity Conditions:

(i) The parameter space has finite dimension, is closed and compact, and the true parameter vector is in the interior of the parameter space.

(ii) Probability distributions defined by any two different values of the parameter vector are distinct (an identifiability condition).

(iii) First three partial derivatives of the log-likelihood function, with respect to the parameters

1. exist
2. are bounded by a function with a finite expectation.

(iv) The expectation of the negative of the matrix of second partial derivatives of the log-likelihood is

1. finite
2. positive definite

in a neighborhood of the true value of the parameter vector. This matrix is called the *Fisher information matrix*.

Suppose $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are independent vectors of observations, with

$$\mathbf{Y}_j = \begin{bmatrix} Y_{1j} \\ \vdots \\ Y_{pj} \end{bmatrix},$$

and the density function (or probability function) is

$$f(\mathbf{Y}_j; \boldsymbol{\theta})$$

Then, the joint likelihood function is

$$L(\boldsymbol{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = \prod_{j=1}^{n} f(\mathbf{Y}_j; \boldsymbol{\theta})$$

The log-likelihood function is

$$\begin{aligned} \ell(\boldsymbol{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) &= \log\left(L(\boldsymbol{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)\right) \\ &= \sum_{j=1}^{n} \log\left(f(\mathbf{Y}_j; \boldsymbol{\theta})\right). \end{aligned}$$

The score function

$$\mathbf{u}(\boldsymbol{\theta}) = \begin{bmatrix} u_1(\boldsymbol{\theta}) \\ \vdots \\ u_r(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)}{\partial \theta_r} \end{bmatrix}$$

is the vector of first partial derivatives of the log-likelihood function with respect to the elements of

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_r \end{bmatrix}.$$

The likelihood equations are

$$u(\boldsymbol{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = \mathbf{0}$$

The maximum likelihood estimator (MLE)

$$\hat{\boldsymbol{\theta}} = \left[\ \hat{\theta}_1, \cdots, \hat{\theta}_r\ \right]^T$$

is a solution to the likelihood equations, that maximizes the log - likelihood function.

Fisher information matrix:

$$\begin{aligned} I(\boldsymbol{\theta}) &= \mathrm{Var}(u(\boldsymbol{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)) \\ &= E\left(u(\boldsymbol{\theta}; \mathbf{Y}_1 \ldots, \mathbf{Y}_n)[u(\boldsymbol{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)]^T\right) \\ &= -E\left(\left[\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)}{\partial \theta_r \partial \theta_k}\right]\right) \end{aligned}$$

Let

$\boldsymbol{\theta}$ denote the parameter vector

$i(\boldsymbol{\theta})$ denote the Fisher information matrix

$\hat{\boldsymbol{\theta}}$ denote the MLE for $\boldsymbol{\theta}$.

Then, if the Regularity Conditions are satisfied, we have the following results:

Result 8.1: $\hat{\boldsymbol{\theta}}$ is a **consistent** estimator.

$$Pr\left\{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) > \epsilon\right\} \to 0$$

as $n \to \infty$, for any $\epsilon > 0$.

Result 8.2: Asymptotic normality

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\text{dist'n}} N\left(\mathbf{0}, \lim_{n\to\infty} n[I(\boldsymbol{\theta})]^{-1}\right)$$

as $\quad n \to \infty$.

With a slight abuse of notation we may express this as

$$\hat{\boldsymbol{\theta}} \overset{\bullet}{\sim} N\left(\boldsymbol{\theta}, [I(\boldsymbol{\theta})]^{-1}\right)$$

for *large* sample sizes.

Result 8.3: If $\hat{\boldsymbol{\theta}}$ is the mle for $\boldsymbol{\theta}$, then the mle for $g(\boldsymbol{\theta})$ is $g(\hat{\boldsymbol{\theta}})$ for any function $g(\ )$.