Econometrics 1, Class Week 3

Back-up Video: Class week 3 video (click here)

Learning Outcomes

(a) Likelihood function of the simplest normal linear regression model;

Maximum likelihood estimator (MLE) of the parameters in this model.

(b) Information matrix,

asymptotic distribution of the MLE in this model.

(c) Estimation of standard errors in this model.

(d) Comparison with matrix form in Lecture Notes.

Prerequisites

1. Concepts in Mathematical Statistics:
   - probability density function of the normal distribution (see Distributional Handout);
   - joint probability measure of i.i.d. random variables (AMN p.101).

(a) Setting of a simple Normal Linear Regression Model

Actual (and potential) data y_t are assumed to be independently and identically distributed (i.i.d.), with probability density function $f_Y(y)$ given by

$$f_Y(y;\theta_0) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(y-\alpha_0)^2\right), \qquad (1)$$

where $\theta_0 = (\alpha_0, \sigma_0^2)$ are the true, but unknown population parameters that we wish to estimate.

Data for analysis is a random sample $\{y_t, t = 1, \cdots, T\}$ of size T drawn from $f_Y(y;\theta_0)$.

Comparison with Gauss-Markov (GM) Setting

In GM setting, we made assumptions on moments (A1,A2,A5), model linearity (A3) and full rank of $\mathbf{X}$ (A4).

We did not make a distributional assumption.

Here, we assume that data obey a probability model, i.e. we do make a distribution assumption – which implies moments –, but we don't assume model linearity (and today there are no regressors $\mathbf{X}$).

So the GM and ML settings overlap, but each covers cases that the other one does not cover.

General ML Setting

We assume

$$\begin{aligned} y_t &\overset{i.i.d.}{\sim} f_Y(y_t;\theta_0), t = 1, \cdots, T \\ f_Y(y;\theta_0) &\in \mathcal{F} = \{f_Y(y;\theta); \theta \in \Theta\}, \end{aligned} \tag{2}$$

where the pdf f_Y (and the family $\mathcal{F}$ of such densities) is known up to a parameter vector θ that lies in the parameter space Θ. The true population parameter vector θ_0 is unknown.

In the special case of model (1),

$$\begin{aligned} \mathcal{F} &= \Bigg\{ f_Y(y;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\alpha)^2}{2\sigma^2}\right), \\ &\qquad \theta = (\alpha, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \Bigg\} \end{aligned}$$

Estimation idea: Obtain joint density of the sample on the basis of (2) – or (1), resp. – and find the element in Θ that makes the data look most likely, i.e. that maximises joint density of the sample:

$$\prod_{t=1}^{T} f_Y(y_t;\theta) \to \max_{\theta \in \Theta}! \tag{3}$$

Log-likelihood Function

Let $\mathbf{y}_T = (y_1, \cdots, y_T)$.

$L(\theta; \mathbf{y}_T) := \prod_{t=1}^{T} f_Y(y_t; \theta)$ is the likelihood function. It tells us how likely the sample $\mathbf{y}_T$ is if the true, unknown parameter were θ; can evaluate it for any $\theta \in \Theta$.

$l(\theta; \mathbf{y}_T) = \ln L(\theta; \mathbf{y}_T)$ is the log-likelihood function.

$\frac{1}{T} l(\theta; \mathbf{y}_T)$ is the average log-likelihood function; it is a sample average of i.i.d. random variables for each θ.

Note: maximum likelihood estimator (MLE) $\hat{\theta}_T$ satisfies

$$\begin{aligned}
\hat{\theta}_T &= \arg\max_{\theta \in \Theta} L(\theta; \mathbf{y}_T) \\
&= \arg\max l(\theta; \mathbf{y}_T) \\
&= \arg\max \frac{1}{T} l(\theta; \mathbf{y}_T)
\end{aligned}$$

FOCs:

$$\mathbf{0} = \nabla_\theta \left[\frac{1}{T} l(\theta; \mathbf{y}_T) \right]_{\theta = \hat{\theta}_T} =: s(\hat{\theta}_T) \quad \text{(score vector at } \hat{\theta}_T\text{).}$$

Application to Model (1)

$$
\begin{aligned}
L(\theta; \mathbf{y}_T) &= \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_t - \alpha)^2\right) \\
&= (2\pi\sigma^2)^{-\frac{T}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - \alpha)^2\right). \\
l(\theta; \mathbf{y}_T) &= const. - \frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - \alpha)^2 \quad \text{(log-lik.)} \\
\frac{1}{T} l(\theta; \mathbf{y}_T) &= const. - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\frac{1}{T}\sum_{t=1}^{T}(y_t - \alpha)^2. \\
s(\theta) &= \begin{bmatrix} \frac{\partial}{\partial\alpha}\frac{1}{T} l(\theta; \mathbf{y}_T) \\ \frac{\partial}{\partial\sigma^2}\frac{1}{T} l(\theta; \mathbf{t}_T) \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sigma^2}\frac{1}{T}\sum_{t=1}^{T}(y_t - \alpha) \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\frac{1}{T}\sum_{t=1}^{T}(y_t - \alpha)^2 \end{bmatrix} \quad \text{(score vec.)} \\
s(\hat{\theta}_T) &= \begin{bmatrix} \frac{1}{\sigma^2}\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{\alpha}_T) \\ -\frac{1}{2\hat{\sigma}_T^2} + \frac{1}{2\hat{\sigma}_T^4}\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{\alpha}_T)^2 \end{bmatrix} = \mathbf{0} \qquad (4) \\
\hat{\theta}_T &= \begin{bmatrix} \hat{\alpha}_T \\ \hat{\sigma}_T^2 \end{bmatrix} \quad \text{(MLE)} \\
&= \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T} y_t \\ \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{\alpha}_T)^2 \end{bmatrix} = \begin{bmatrix} \bar{y}_T \\ \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y}_T)^2 \end{bmatrix}
\end{aligned}
$$

SOCs for maximum

Hessian $H(\theta)$ needs to be negative semi-definite at the MLE

$$
\begin{aligned}
H(\theta) &= \nabla_{\theta\theta'} \frac{1}{T} l(\theta; \mathbf{y}_T) \\
&= \nabla_{\theta'} s(\theta) \\
&= \left[\begin{array}{cc} -\frac{1}{\sigma^2} & -\frac{1}{\sigma^4}\frac{1}{T}\sum_{t=1}^T (y_t - \alpha) \\ -\frac{1}{\sigma^4}\frac{1}{T}\sum_{t=1}^T (y_t - \alpha) & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}\frac{1}{T}\sum_{t=1}^T (y_t - \alpha)^2 \end{array} \right]
\end{aligned}
$$

Hessian at the MLE $\hat{\theta}_T$:

$$
H(\hat{\theta}_T) = \left[\begin{array}{cc} -\frac{1}{\hat{\sigma}_T^2} & 0 \\ 0 & -\frac{1}{2\hat{\sigma}_T^4} \end{array} \right], \tag{5}
$$

where off-diagonal terms are zero by FOC (4) w.r.t. α.

So indeed, $H(\hat{\theta}_T)$ is negative definite (not just n.s.d.).

(b) Information Matrix

Information matrix $\mathcal{I}(\theta_0)$= negative expected Hessian of the average log-likelihood function (at the true population parameter θ_0)

$$\mathcal{I}(\theta_0) = -\mathbb{E}[H(\theta_0)] = \begin{bmatrix} \frac{1}{\sigma_0^2} & 0 \\ 0 & \frac{1}{2\sigma_0^4} \end{bmatrix} \tag{6}$$

So the information matrix is positive definite.

Asymptotic distribution of the MLE

$$\begin{aligned} \sqrt{T}(\hat{\theta}_T - \theta_0) &= \sqrt{T}\left(\begin{bmatrix} \hat{\alpha}_T \\ \hat{\sigma}_T^2 \end{bmatrix} - \begin{bmatrix} \alpha_0 \\ \sigma_0^2 \end{bmatrix}\right) \\ &\xrightarrow{d} N(\mathbf{0}, \mathcal{I}(\theta_0)^{-1}) \text{ as } T \to \infty, \end{aligned} \tag{7}$$

so asymptotic variance-covariance matrix of the MLE is the inverse of the information matrix,

$$\mathcal{I}(\theta_0)^{-1} = \begin{bmatrix} \sigma_0^2 & 0 \\ 0 & 2\sigma_0^4 \end{bmatrix} \tag{8}$$

Normality of the asymptotic distribution has nothing to do with assumption (1).

For asymptotics, it is necessary to assume (2) – and (1), resp. – for actual and *potential* data.

(c) Estimation of Standard Error of $\hat{\alpha}_T$

From (7) and (8), asymptotic variance of $\hat{\alpha}_T$ is σ_0^2.

So estimate SE by $\hat{\sigma}_T$.

(d) Comparison with Notes

With regressors $\mathbf{X}$ in (conditional) mean of $\mathbf{y}$,

- MLE for β_0 is OLS estimator;
- corresponding element of the inverse information matrix, conditional on $\mathbf{X}$, is $\sigma_0^2(\mathbf{X'X})^{-1}$;
- it is estimated by $\hat{\sigma}_T^2(\mathbf{X'X})^{-1}$.