1 The maximum likelihood estimator of the linear model

Let $\mathbf{y}$ be a $(n * 1)$ vector of sample observations and $\boldsymbol{\theta}$ be a $(k * 1)$ parameter vector. The joint density is $f(\mathbf{y}; \boldsymbol{\theta})$. The log-likelihood function is defined as

$$l = \ln L(\boldsymbol{\theta}; \mathbf{y}) = \ln f(\mathbf{y}; \boldsymbol{\theta}).$$

Under rather general regularity conditions the maximum likelihood estimator $\widehat{\theta}$ has the following properties:

1. Consistency.

$$p\lim \widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}.$$

2. Asymptotic normality.

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta})^{-1}\right).$$

Where $\mathbf{I}(\boldsymbol{\theta})$ is the information matrix. Can be defined in two ways

$$\begin{aligned} \mathbf{I}(\boldsymbol{\theta}) &= E\left[\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)'\right] \\ \mathbf{I}(\boldsymbol{\theta}) &= -E\left[\frac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] \end{aligned}$$

The vector of first order conditions with respect to the parameters $\mathbf{s}(\boldsymbol{\theta}; \mathbf{y}) = \left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)$ is called the score vector.

The matrix of second order conditions is the Hessian

$$\frac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{bmatrix} \frac{\partial^2 l}{\partial \theta_1^2} & \frac{\partial^2 l}{\partial \theta_1^2 \partial \theta_2} & \cdot & \cdot & \frac{\partial^2 l}{\partial \theta_1^2 \partial \theta_k} \\ \frac{\partial^2 l}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 l}{\partial \theta_2^2} & & & \\ \cdot & & & \cdot & \cdot \\ & & & \cdot & \\ \frac{\partial^2 l}{\partial \theta_k^2 \partial \theta_1} & \cdot & \cdot & & \frac{\partial^2 l}{\partial \theta_k^2} \end{bmatrix}$$

3. Asymptotic Efficiency. It has (asymptotically) the minimum variance, $\mathbf{I}(\boldsymbol{\theta})^{-1}$. The variance of the ML estimator fulfills asymptotically the Cramér-Rao lower bound.

4. Invariance. If $\widehat{\theta}$ is the ML estimator of a parameter θ and $g(\theta)$ is a continuous function of θ, then $g\left(\widehat{\theta}\right)$ is the ML estimator of $g(\theta)$.

If we can show that the Maximum Likelihood estimator of the linear model is equivalent to the OLS, we have shown that the OLS estimator shares the attractive properties of the ML estimator. So let's derive the ML estimator of the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right). \tag{1}$$

The multivariate normal density for $\boldsymbol{\varepsilon}$ is

$$f(\boldsymbol{\varepsilon}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(\frac{-\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{2\sigma^2}\right)$$

The multivariate normal for $\mathbf{y}$ given $\mathbf{X}$ is then

$$f(\mathbf{y} \mid \mathbf{X}) = f(\boldsymbol{\varepsilon})$$

This means that we can use $f(\boldsymbol{\varepsilon})$ to derive the log likelihood function

$$\begin{aligned}
l\left(\boldsymbol{\beta}, \sigma^2; \mathbf{y}\right) &= \ln f(\mathbf{y} \mid \mathbf{X}) = \ln f(\boldsymbol{\varepsilon}) \\
&= \ln\left(\left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(\frac{-\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{2\sigma^2}\right)\right) \\
&= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta})
\end{aligned}$$

The F.O.C. are

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{2\sigma^2}(2\mathbf{X}'\mathbf{y} - \mathbf{2X}'\mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma^2}\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma^2}\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = -\frac{n}{2\sigma^2} + \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{2\sigma^4} = 0$$

Solving for $\boldsymbol{\beta}$ and σ^2 gives us the ML estimators

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{ML} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{b} \\
\widehat{\sigma}^2_{ML} &= \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{ML}\right)'\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{ML}\right)/n = \frac{\mathbf{e}'\mathbf{e}}{n}
\end{aligned}$$

$\widehat{\sigma^2}_{ML}$ is biased since

$$E\left(\widehat{\sigma}^2_{ML}\right) = \frac{1}{n} E\left(\mathbf{e}'\mathbf{e}\right) = \frac{n-k}{n}\sigma^2$$

But it is consistent, since $\frac{n-k}{n} \longrightarrow 1$ as $n \longrightarrow \infty$.

The S.O.C.:s are

$$\begin{aligned}
\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= -\frac{\mathbf{X}'\mathbf{X}}{\sigma^2} \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \sigma^2} &= \frac{\mathbf{X}'\boldsymbol{\varepsilon}}{\sigma^4} \\
\frac{\partial^2 l}{\partial (\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\sigma^6}
\end{aligned}$$

Taking the expectation of these, we have

$$\begin{aligned}
E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right] &= -\frac{\mathbf{X}'\mathbf{X}}{\sigma^2} \\
E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \sigma^2}\right] &= \mathbf{0} \\
E\left[\frac{\partial^2 l}{\partial (\sigma^2)^2}\right] &= \frac{n}{2\sigma^4} - \frac{n\sigma^2}{\sigma^6} = \frac{n}{2\sigma^4} - \frac{2n}{2\sigma^4} = -\frac{n}{2\sigma^4}
\end{aligned}$$

From these we can get the information matrix

$$\mathbf{I}\begin{pmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{pmatrix} = \begin{bmatrix} E\left[\frac{\partial^2 l}{\partial \beta \partial \beta'}\right] & E\left[\frac{\partial^2 l}{\partial \beta \partial \sigma^2}\right] \\ E\left[\frac{\partial^2 l}{\partial \sigma^2 \partial \beta'}\right] & E\left[\frac{\partial^2 l}{\partial (\sigma^2)^2}\right] \end{bmatrix} = \begin{bmatrix} \frac{\mathbf{X}'\mathbf{X}}{\sigma^2} & \mathbf{0} \\ \mathbf{0}' & \frac{n}{2\sigma^4} \end{bmatrix}$$

As it is a block diagonal matrix it is easy to obtain the inverse

$$\mathbf{I}^{-1}\begin{pmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{pmatrix} = \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-1}\sigma^2 & \mathbf{0} \\ \mathbf{0} & \frac{2\sigma^4}{n} \end{bmatrix}$$

These are the Cramér-Rao lower bounds.