Topic 9: Maximum Likelihood Estimation

There are many other estimation methodologies besides OLS. For example: GMM, Bayesian, non-parametric, and maximum likelihood (ML). In some of these methodologies, the OLS estimator is just a special case.

- ML proposed by R. A. Fisher, 1921-1925
- MLE is a parametric method.
- That is, we assume each sample data is generated from a known probability distribution function (p.d.f.), $p(y_i|\boldsymbol{\theta})$. i.e. y_i comes from a "family".

Consider:

Random data	$\boldsymbol{y} = \{y_1, \ldots, y_n\}$
Parameter vector	$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$

Objective: estimate $\boldsymbol{\theta}$.

The probability of jointly observing the data is

$$p(y_1, \ldots, y_n|\boldsymbol{\theta}) \qquad \text{"joint p.d.f."}$$

We can view $p(y_1, \ldots, y_n|\boldsymbol{\theta})$ in two different ways:

i. As a function of $\{y_1, \ldots, y_n\}$, given $\boldsymbol{\theta}$.
ii. As a function of $(\theta_1, \ldots, \theta_k)$, given $\boldsymbol{y}$. i.e., the data is *given*, the parameters *vary*.

The latter is called the **likelihood function**.

Note: $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|y_1, \ldots, y_n) = p(y_1, \ldots, y_n|\boldsymbol{\theta})$

Definition: The Maximum Likelihood Estimator (MLE) of $\boldsymbol{\theta}$ (say, $\widetilde{\boldsymbol{\theta}}$) is that value of $\boldsymbol{\theta}$ such that $L(\widetilde{\boldsymbol{\theta}}) > L(\widehat{\boldsymbol{\theta}})$, for all other $\widehat{\boldsymbol{\theta}}$.

Idea: "given the y_i's, what is the most likely $\boldsymbol{\theta}$ to have generated such a sample?"

Note:

i. $\widetilde{\boldsymbol{\theta}}$ need not be unique.

ii. $\tilde{\boldsymbol{\theta}}$ should locate the global max. of $L(\boldsymbol{\theta})$.

iii. If the sample data are independent then $L(\boldsymbol{\theta}|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta})$

iv. Any monotonic transformation of $L(\boldsymbol{\theta})$ leaves location of extremum unchanged (e.g. $\log L(\boldsymbol{\theta})$)

Some Basic Concepts and Notation:

i. "Gradient/Score Vector": $\left[\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$ $(k \times 1)$

ii. "Hessian Matrix": $\left[\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]$ $(k \times k)$

iii. "Likelihood Equations": $\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$ $(k \times 1)$

The optimization problem is:

$$\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} L(\boldsymbol{\theta}|y_i).$$

So, to obtain the MLE, $\tilde{\boldsymbol{\theta}}$, we solve the likelihood equation(s) and then check the second-order condition(s) to make sure we have maximized (not minimized) $L(\boldsymbol{\theta})$. If the Hessian matrix is at least n.s.d., then $\log L(\boldsymbol{\theta})$ is concave, and this is sufficient for a maximum.

So, MLE is accomplished by:

1) Specifying the likelihood function.
   - This involves writing down an equation which states the joint likelihood (or joint probability) of observing the sample data, conditional on the unknown parameter values of the probability distribution function.
   - Independence of the $\boldsymbol{y}$ data is usually assumed (and will be for the purposes of this course).
   - Given independence, the likelihood function is obtained by multiplying together the probability of each y_i occurring.
2) Taking the natural log of the likelihood function. This usually simplifies the next step. The location of the maximum will not change.

3) Taking the first derivative of the log-likelihood function with respect to all parameters, setting each derivative equal to zero, and solving for the parameter values. The solution of the FOCs provides the formulas for the MLEs.
4) Checking to make sure the estimator in (3) attains a maximum (not a minimum). This involves taking the second derivatives of the log-likelihood function with respect to all parameters, so as to construct the Hessian matrix. If the Hessian is *n.s.d.*, then the MLE achieves a global max.
5) Obtaining the variance of the MLEs for use in hypothesis testing. A variance-covariance matrix can be found by inverting the negative of the expected Hessian.

Properties of MLE

- MLE has very desirable asymptotic properties.
- Namely, MLE is Best Asymptotically Normal.
- That is, under mild assumptions, ML estimators are consistent, asymptotically efficient, and asymptotically Normally distributed.
- These properties are obtained by examining the asymptotic distribution of the MLE (which we will not derive in class):

$$\sqrt{n}\left(\tilde{\theta} - \theta_0\right) \xrightarrow{d} N[0, IA^{-1}(\theta)],$$

where

$$IA^{-1}(\theta) = \lim_{n \to \infty} \left(\frac{1}{n}\left[-E[H(\theta)]\right]^{-1}\right)$$

- $IA^{-1}(\theta)$ is the asymptotic information matrix, and $H(\theta)$ is the Hessian.
- The statement of the asymptotic distribution shows that the MLEs are consistent, asymptotically normal, and asymptotically efficient.
- The efficiency result relies on the Cramer-Rao lower bound. The Cramer-Rao lower bound is a theoretical minimum variance that any estimator can obtain. The MLE attains this minimum, that is, $IA^{-1}(\theta)$ is equal to the asymptotic Cramer-Rao lower bound.

The asymptotic distribution also allows us to see the variance of the MLEs in finite samples. The variance-covariance of $\tilde{\theta}$ for finite samples can be solved from the asymptotic variance:

$$var\left[\sqrt{n}\left(\tilde{\theta}\right)\right] = n \times var\left(\tilde{\theta}\right) = \frac{1}{n}\left[-E[H(\theta)]\right]^{-1}, \text{ so}$$

$$var\left(\tilde{\theta}\right) = \left[-E[H(\theta)]\right]^{-1}.$$

The matrix $-E[H]$ is termed the "Information Matrix" and is denoted by $I(\theta)$.

A very useful property of MLEs is their "invariance." That is, the estimator for $g(\theta)$ is $g(\tilde{\theta})$. Hence, an estimator for the variance-covariance of $\tilde{\theta}$ is:

$$\widetilde{var(\tilde{\theta})} = \left[-E\left[H\left(\tilde{\theta}\right)\right]\right]^{-1}.$$

Note that if misspecification occurs (if we have selected the wrong probability density function to begin with), we are not assured of any of the asymptotic properties.

Finite sample properties of MLEs

MLEs can be biased in finite samples (and typically are). We can evaluate bias much like we have done in previous parts of the course; by taking $E\left(\tilde{\theta}\right)$. This knowledge can be used to correct for any bias (as in the case of $\tilde{\sigma}^2$). However, in most cases, there is no closed-form solution for the MLE itself, and numerical methods must be used to solve for the estimate. When the estimator does not have a closed form solution, we cannot take $E\left(\tilde{\theta}\right)$, and we will not be able to "see" whether or not the estimator is biased. In this case, approximations or Monte Carlo experiments may be used to evaluate bias.