Master 2

Econometrics I

Nour Meddahi (Toulouse School of Economics)

Maximum Likelihood Estimator

I. Examples

We consider again the linear model for cross-section data,

$$y_t = x_t'\beta + u_t.$$

However, we now make an additional assumption about the distribution of u_t. For instance, we assume that u_t follows a $\mathcal{N}(0, \sigma^2)$ or a $\mathcal{T}(\nu)$. We know that the OLS estimator of β is BLUE and consistent. However, we do not know whether it is the best estimator, i.e., whether there exists a consistent estimator of β and non-linear in y which is more precise (has a smaller variance).

Consider another example. Suppose that one has a non-linear model for cross-section data where we know that the conditional mean of y_t given x_t is positive. A possible model is

$$y_t = \exp(x_t'\beta) + u_t$$

where u_t follows a $\mathcal{N}(0, \sigma^2)$. We know that the Non Linear Least Squares estimator is consistent. However, we do not know whether it is the best one, i.e., whether there exists a consistent estimator of β which has a smaller variance.

The main goal of this lecture is to study the maximum likelihood estimator. It is the best estimator among all possible estimators.

II. General Theory

Let (y_t, x_t), $t = 1, ..., n$, be a sample of i.i.d. observations. The density function of y_t given x_t depends on an unknown vector θ_0; we denote this density function as $f_{y_t|x_t}(\theta_0; y_t)$. Therefore, the density function of the sample, denoted $f_{Y|X}(\theta_0, \mathbf{y})$ is also a function of θ_0. Observe that

$$f_{Y|X}(\theta_0; \mathbf{y}) = \prod_{t=1}^{n} f_{y_t|X}(\theta_0; y_t) = \prod_{t=1}^{n} f_{y_t|x_t}(\theta_0; y_t).$$

The likelihood function of the sample is the density function of the sample when θ varies and is assumed unknown

$$L(\theta; \mathbf{y}) = f_{Y|X}(\theta; \mathbf{y}) = \prod_{t=1}^{n} f_{y_t|x_t}(\theta; y_t).$$

Take the example where the linear and Gaussian model

$$y_t = x_t'\beta_0 + u_t, \quad u_t \sim \mathcal{N}(0, \sigma_0^2).$$

Observe that now we prefer write β_0 instead of β for convenience. Here, $\theta = (\beta', \sigma^2)' = (\beta_1, \beta_2, ..., \beta_k, \sigma^2)'$. The conditional density function of the sample is

$$f_{Y|X}(y_1, y_2, ..., y_n) = \prod_{t=1}^{n} f_{y_t|X}(y_t) = \prod_{t=1}^{n} f_{y_t|x_t}(y_t) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(y_t - x_t'\beta_0)^2}{2\sigma_0^2}\right).$$

Therefore, the likelihood function of the sample is

$$L(\theta;\mathbf{y}) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - x_t'\beta)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{\sum_{t=1}^{n}(y_t - x_t'\beta)^2}{2\sigma^2}\right).$$

The log-likelihood is the logarithm of the likelihood function and is denoted $l(\theta;\mathbf{y})$

$$l(\theta;\mathbf{y}) \equiv \log L(\theta;\mathbf{y}) = \sum_{t=1}^{n} \log f_{y_t|x_t}(\theta; y_t).$$

The score function denoted $s(\theta;y)$ is the first derivative of the log-likelihood while the Hessian matrix $H(\theta)$ is the second derivative of log-likelihood,

$$s(\theta) = \frac{\partial l}{\partial\theta}(\theta;\mathbf{y}), \quad H(\theta) = \frac{\partial^2 l}{\partial\theta\partial\theta'}(\theta;\mathbf{y}).$$

The maximum likelihood estimator is the vector $\hat{\theta}$ that maximizes the likelihood and the log-likelihood (the same thing since the log function is an increasing function)

$$\hat{\theta}^{MLE} = \hat{\theta} = ArgMax_\theta\ L(\theta;\mathbf{y}) = ArgMax_\theta\ l(\theta;\mathbf{y}).$$

In other words, $\hat{\theta}$ is defined as

$$s(\hat{\theta}) = 0.$$

Properties of the MLE.

- Often, we can not say too much about the MLE $\hat{\theta}$ when n is fixed. The properties of the MLE are asymptotic ones.

- The MLE is consistent, $Plim(\hat{\theta}) = \theta_0$.

- The MLE is asymptotically normal, i.e.,

$$\sqrt{n}(\hat{\theta} - \theta_{\mathbf{0}}) \rightarrow_d N(\mathbf{0}, \mathcal{I}^{-\mathbf{1}}(\theta_{\mathbf{0}})),$$

with
$$\mathcal{I}(\theta_0) = \lim_{n \to +\infty} \frac{I(\theta_0)}{n}$$

where $I(\theta)$ is the information matrix of the sample defined as

$$I(\theta) \equiv -E[H(\theta))].$$

Importantly, one has $\quad I(\theta_0) = -E[H(\theta_0)] = Var[s(\theta_0)].$

One could interpret $\mathcal{I}(\theta_0)$ as the information matrix of one observation.

- The MLE is asymptotically efficient, i.e., any other consistent estimator has a larger asymptotic variance: $\sqrt{n}(\bar{\theta} - \theta_0) \rightarrow_d \mathcal{N}(0, \Sigma)$, then $\Sigma - \mathcal{I}^{-1}(\theta_0)$ is positive semi definite. The asymptotic variance of the MLE is called the Cramer-Rao lower bound.

- The MLE is invariant: It means that the MLE of $\mathbf{g}(\theta_{\mathbf{0}})$, for any continuously differentiable function $\mathbf{g}(.)$ is $\mathbf{g}(\hat{\theta})$.

III. Testing Hypotheses.

One can test general hypotheses (linear or non-linear) on θ_0, like $g(\theta_0) = 0$ (for instance $g(\theta) = R\beta - r$ in the linear model). The number of restrictions is q.

In what follows, $\tilde{\theta}$ denotes the constrained maximum likelihood estimator, i.e., the vector that maximizes the likelihood under the restriction $g(\theta) = 0$.

We have three important tests:

1) **The likelihood ratio (LR) test**: This test compares $l(\hat{\theta})$ and $l(\tilde{\theta})$. If the null is true, the difference between $l(\hat{\theta})$ and $l(\tilde{\theta})$ should be small. One can show that under the null,

$$\mathbf{LR} \equiv 2(l(\hat{\theta}) - l(\tilde{\theta})) \rightarrow_d \chi^2(q).$$

Observe that one needs to compute the restricted and unrestricted estimators.

2) **The Wald test (W) test**: This test compares $g(\hat{\theta})$ and the vector of 0 ($q \times 1$). If the null is true, the norm of $g(\hat{\theta})$ should be small. By using the delta method, one can show that under the null,

$$\mathbf{W} \equiv n\ g(\hat{\theta})' \left[\frac{\partial g}{\partial \theta'}(\hat{\theta}) \hat{\mathcal{I}}^{-1} \frac{\partial g'}{\partial \theta}(\hat{\theta}) \right]^{-1} g(\hat{\theta}) \rightarrow_d \chi^2(q).$$

Observe that one needs to compute unrestricted estimator as well as an estimator of the information matrix.

3) **The score test or Lagrange multiplier (LM) test**: Consider the Lagrangian, required to estimate the restricted MLE:

$$\tilde{l}(\theta, \mu) = l(\theta) - \mu' \mathbf{g}(\theta).$$

If the restriction holds in the data, imposing the restriction would be irrelevant and as a consequence the (estimated) Lagrange multiplier $\tilde{\mu}$ should be small, not significantly different from zero. Another way to look at it, this is equivalent to say that the restricted MLE $\tilde{\theta}$ satisfies the FOC of the (unrestricted) MLE: $\mathbf{s}(\tilde{\theta}) \approx \mathbf{0}$. The LM test yields a statistical measure of closeness to zero of the score. In fact it is also called the **score test** and is defined as:

$$LM = s'(\tilde{\theta}) I^{-1}(\tilde{\theta}) s(\tilde{\theta}) \rightarrow_d \chi^2_q,$$

The importance of the LM test consists on the fact it *only* requires estimation of the parameters under the null and this *might be* much easier.

In practice, often people use the Wald test for its simplicity. However, the LR test has better finite sample properties. When you can use the LR test, do it.

IV. The Linear and Gaussian Model

The conditional density function of the sample $(y_1, y_2, ..., y_n)$ is

$$f_{Y|X}(y_1, y_2, ..., y_n) = \prod_{t=1}^{n} f_{y_t|X}(y_t) = \prod_{t=1}^{n} f_{y_t|x_t}(y_t) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(y_t - x_t'\beta_0)^2}{2\sigma_0^2}\right).$$

Therefore, the likelihood function of the sample is

$$L(\theta; \mathbf{y}) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - x_t'\beta)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{\sum_{t=1}^{n}(y_t - x_t'\beta)^2}{2\sigma^2}\right).$$

Likewise, the log-likelihood function is

$$l(\theta; \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\sum_{t=1}^{n}(y_t - x_t'\beta)^2}{2\sigma^2}.$$

The first order conditions are

$$\frac{\partial l}{\partial \beta}(\theta; \mathbf{y}) = \frac{\sum_{t=1}^{n} x_t(y_t - x_t'\beta)}{\sigma^2} \quad \text{and} \quad \frac{\partial l}{\partial \sigma^2}(\theta; \mathbf{y}) = -\frac{n}{2\sigma^2} + \frac{\sum_{t=1}^{n}(y_t - x_t'\beta)^2}{2\sigma^4}.$$

By defining the MLE as

$$s(\hat{\theta}) = \begin{pmatrix} \frac{\partial l}{\partial \beta}(\hat{\theta}; \mathbf{y}) \\ \frac{\partial l}{\partial \sigma^2}(\hat{\theta}; \mathbf{y}) \end{pmatrix} = 0$$

one gets

$$\hat{\beta} = \left[\sum_{t=1}^{n} x_t x_t'\right]^{-1} \left[\sum_{t=1}^{n} x_t y_t\right] = \left[\mathbf{X}'\mathbf{X}\right]^{\mathbf{-1}} \mathbf{X}'\mathbf{y} = \hat{\beta}^{\mathbf{OLS}}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(y_t - x_t'\hat{\beta})^2 = \frac{1}{n}\sum_{t=1}^{n} e_t^2.$$

Observe that in this example, $\hat{\beta}$ is unbiased while $\hat{\sigma}^2$ is biased but asymptotically unbiased. In general, $\hat{\theta}$ is biased in finite sample but the bias disappears asymptotically. Likewise, if one assumes that distribution of u_t is Student $\mathcal{T}(\nu)$, then one can prove $\hat{\beta} \neq \hat{\beta}^{OLS}$.

One should compute the second derivative of the log-likelihood function to be sure that $\hat{\theta}$ is really a maximum. One can show that

$$\frac{\partial^2 l}{\partial\beta\partial\beta'}(\theta, \mathbf{y}) = -\frac{1}{\sigma^2}\sum_{t=1}^{n} x_t x_t' = -\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}, \quad \text{with } E\left[\frac{\partial^2 l}{\partial\beta\partial\beta'}(\theta_{\mathbf{0}}, \mathbf{y})\right] = -\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}$$

$$\frac{\partial^2 l}{\partial\beta\partial\sigma^2}(\theta, \mathbf{y}) = -\frac{1}{\sigma^4}\sum_{t=1}^{n} x_t(y_t - x_t'\beta), \quad \text{with} \quad E\left[\frac{\partial^2 l}{\partial\beta\partial\sigma^2}(\theta_{\mathbf{0}}, \mathbf{y})\right] = 0$$

$$\frac{\partial^2 l}{\partial^2\sigma^2}(\theta, \mathbf{y}) = \frac{n}{2\sigma^4} - \frac{\sum_{t=1}^{n}(y_t - x_t'\beta)^2}{\sigma^6}, \quad \text{with} \quad E\left[\frac{\partial^2 l}{\partial^2\sigma^2}(\theta_{\mathbf{0}}, \mathbf{y})\right] = -\frac{n}{2\sigma^4}$$

and that the matrix $\dfrac{\partial s}{\partial \theta'}(\hat{\theta})$ is negative-definite, i.e. $\hat{\theta}$ is a maximum.

The information matrix $\mathbf{I}(\theta_\mathbf{0})$ of the sample and its inverse are

$$\mathbf{I}(\theta_\mathbf{0}) = \begin{pmatrix} \dfrac{1}{\sigma_0^2}\mathbf{X'X} & 0 \\ 0 & \dfrac{n}{2\sigma_0^4} \end{pmatrix} \quad \text{and} \quad \mathbf{I^{-1}}(\theta_\mathbf{0}) = \begin{pmatrix} \sigma_0^2(\mathbf{X'X})^{\mathbf{-1}} & 0 \\ 0 & \dfrac{2\sigma_0^4}{n} \end{pmatrix}$$

Consequently, one has

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow^d N(0, V)$$

$$\text{where} \qquad V = \lim_{n \rightarrow +\infty} \frac{\mathbf{I^{-1}}(\theta_\mathbf{0})}{n} = \begin{bmatrix} \sigma_0^2 Q^{-1} & 0 \\ 0 & 2\sigma_0^4 \end{bmatrix}, \quad Q = Plim_{n \rightarrow +\infty} \frac{1}{n}(\mathbf{X'X}).$$

A consistent estimator of V is

$$\hat{V} = \begin{bmatrix} \hat{\sigma}^2 \left[\dfrac{1}{n}\displaystyle\sum_{t=1}^{n} x_t x_t'\right]^{-1} & 0 \\ 0 & 2\hat{\sigma}^4 \end{bmatrix}.$$

Observe that one could build a 95% confidence interval of $\theta_{i,0}$ by using

$$\sqrt{n}\,\frac{(\hat{\theta}_i - \theta_{i,0})}{\sqrt{\hat{V}_{ii}}} \sim \mathcal{N}(0, 1).$$

Assume now that wants to test a hypotheses

$$H_0 : \;\; \mathbf{R}\beta_{\mathbf{0}} = \mathbf{r},$$

where $\mathbf{R}$, $\mathbf{r}$ are respectively a known constant matrix of dimension $q \times k$ with rank q, where $q < k$, and a known constant vector of dimension $q \times 1$.

We already seen how to implement hypothesis testing for linear restrictions for the linear regression model. We now present the main techniques available when using MLE.

In what follows, $\tilde{\theta}$ denotes the constrained maximum likelihood estimator, i.e., the vector that maximizes the likelihood under the restriction $R\beta - r = 0$. Likewise, $\tilde{e}_t$ denotes the residuals of the restricted model, $\tilde{e}_t = y_t - x_t'\tilde{\beta}$.

Likelihood Ratio Test. We know from the general theory that

$$LR = 2(l(\hat{\theta}) - l(\tilde{\theta})) \rightarrow_d \chi^2_q.$$

One can show that

$$LR = n \log\left(1 + \frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}} - \mathbf{e}'\mathbf{e}}{\mathbf{e}'\mathbf{e}}\right) = n \log\left(\frac{\mathbf{1}}{\mathbf{1} - (\tilde{\mathbf{e}}'\tilde{\mathbf{e}} - \mathbf{e}'\mathbf{e}')/\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}\right).$$

Wald Test.

The Wald test only requires estimation of the unrestricted MLE $\hat{\beta}$ and is based on evaluating whether $\mathbf{R}\hat{\beta} - \mathbf{r}$ is small or not, assuming H_0 holds true. One has

$$W \equiv \frac{\sqrt{T}(\mathbf{R}\hat{\beta} - \mathbf{r})'(\mathbf{R}(\mathbf{X}'\mathbf{X}/\mathbf{T})^{-\mathbf{1}}\mathbf{R}')^{-1}\sqrt{T}(\mathbf{R}\hat{\beta} - \mathbf{r})}{\hat{\sigma}^2} \rightarrow_d \chi^2_q.$$

Lagrange multiplier (LM) or score Test.

One can show that

$$LM = n\frac{\tilde{\mathbf{e}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\tilde{\mathbf{e}}}{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}.$$

An important relationship It can be shown that although the three tests share the same identical asymptotic distribution, they will in general differ in finite samples and, in particular,

$$W \geq LR \geq LM.$$