Econometrics Lecture 4: Maximum Likelihood Estimation and Large Sample Tests

R. G. Pierse

1 Large Sample Theory

In the last lecture we saw that when the *feasible GLS* estimator is considered, with the unknown covariance matrix $\mathbf{\Sigma}$ replaced by its estimate $\widehat{\mathbf{\Sigma}}$, then assumption (A4) of the classical model that

$$\text{The matrix of regressors } \mathbf{X} \text{ is } \textit{fixed in repeated sampling} \qquad \text{(A4)}$$

can no longer be maintained and the regressors have to be treated as random variables. This is also true if the matrix $\mathbf{X}$ includes lags of the dependent variable $\mathbf{y}$ since lags of $\mathbf{y}$ are a function of the error process u and must therefore be random.

Once assumption (A4) is abandoned, then the distribution of any coefficient estimator such as the *OLS* estimator

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \end{aligned}$$

will depend on the distribution of $\mathbf{X}$ as well as that of $\mathbf{u}$. Since $\widehat{\boldsymbol{\beta}}$ depends on $\mathbf{X}$ in a nonlinear way, its exact distribution is complicated and in general will not be exactly normal. Thus the exact sample theory developed under the assumption of normality can no longer be used.

Instead, we can use large sample methods which generate results that hold under conditions much weaker than those of the classical model. These results are asymptotic and hold exactly only in the limit when the number of observations tends to infinity. Nevertheless, they continue to be useful in finite samples where they will hold as an approximation. In the absence of exact results, this is the best that we can hope to achieve.

1.1 Convergence of a random variable

Consider the sample mean of a random variable x

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

This can be regarded as a function of the sample size n and we can consider the sequence $\overline{x}_1, \cdots, \overline{x}_n$ as the number of observations n increases. This is an example of a *sequence of random variables*.

Consider the sequence of random variables a_n as the number of observations n increases. This sequence is said to *converge in probability* to a constant c if, as $n \to \infty$,

$$\lim_{n\to\infty} \Pr(|a_n - c| > \delta) = 0 \quad , \quad \forall \delta > 0. \tag{1.1}$$

This condition states that as the sample size increases, the *probability* of a_n being different from c will tend to zero, which implies that the probability distribution of a_n will eventually collapse to the single value c. When a_n is a vector or matrix, then the definition applies to each element. If condition (1.1) holds then we write

$$a_n \to_P c.$$

Convergence in probability is sometimes also known as *weak convergence*. The concept of convergence in probability can be extended to the case where c is not a constant but instead a random variable.

The constant c is called the *probability limit* of the sequence of random variables a_n or *plim* for short and we write

$$\operatorname*{plim}_{n\to\infty}(a_n) = c.$$

Once the meaning is clear, the subscript n on a can be dropped and we will write

$$\operatorname{plim}(a) = c.$$

An estimator $\widetilde{\boldsymbol{\beta}}$ of a parameter vector $\boldsymbol{\beta}$ that satisfies the property that

$$\operatorname{plim}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

is said to be a *consistent* estimator of the parameter. As the sample size increases, the probability distribution of the estimator collapses to the true value of the parameter.

The plim operator has the property that, for any continuous function g,

$$\operatorname{plim} g(a) = g(\operatorname{plim}(a)). \tag{1.2}$$

This important result is known as *Slutsky's theorem* and makes manipulation of plims very easy. In particular, it follows that, if a_n and b_n are random sequences whose plims both exist, then

$$\text{plim}(a+b) = \text{plim}(a) + \text{plim}(b)$$

and

$$\text{plim}(ab) = \text{plim}(a)\,\text{plim}(b).$$

Furthermore, matrix inversion is a continuous function of the matrix elements so that, if $\mathbf{A}$ is a matrix with $\text{plim}(\mathbf{A}) = \overline{\mathbf{A}}$ and if $\overline{\mathbf{A}}$ is nonsingular, then

$$\text{plim}(\mathbf{A}^{-1}) = (\text{plim}(\mathbf{A}))^{-1} = \overline{\mathbf{A}}^{-1}.$$

If $x_1, x_2, \cdots, x_n$ are independent identically distributed random variables with finite mean $\text{E}(x_i) = \mu$, then the sample mean $\overline{x} = \sum_{i=1}^{n} x_i$ converges in probability to μ as $n \to \infty$, or, put alternatively,

$$\text{plim}(\overline{x}) = \mu. \tag{1.3}$$

This result is known as *Khintchine's theorem.*

1.2 Convergence in distribution

Suppose that for the sequence of random variables a_n, the sequence of distribution functions $F_n(a_n)$ converges as $n \to \infty$ to a distribution function $F()$. Then $F()$ is said to be the *limiting distribution* of the sequence a_n. If a is a *random variable* having the distribution function $F()$ then a_n is said to *converge in distribution* to a. This is written as

$$a_n \to_D a.$$

One very important case is where the limiting distribution is the normal distribution. If $\mathbf{a}_n$ is a vector sequence of random variables converging in distribution to a vector $\mathbf{a}$ with normal distribution $\mathbf{a} \sim N(\mathbf{0}, \mathbf{A})$ then we write

$$\mathbf{a}_n \sim_a N(\mathbf{0}, \mathbf{A}) \tag{1.4}$$

where the symbol $\sim_a$denotes the phrase 'has the limiting distribution' or 'is asympototically distributed'.

If $\mathbf{a}_n$ satisfies (1.4) and the matrix sequence $\mathbf{B}_n$ converges in probability such that $\text{plim}(\mathbf{B}_n) = \overline{\mathbf{B}}$, then

$$\mathbf{B}_n\mathbf{a}_n \sim_a N(\mathbf{0}, \overline{\mathbf{B}}\mathbf{A}\overline{\mathbf{B}}'). \tag{1.5}$$

This result is known as *Cramér's theorem.*

Convergence in distribution is less easy to establish than convergence in probability. In general we need to invoke an appropriate *central limit theorem.*

2 Least Squares in Large Samples

It is possible to establish under quite general conditions that the *OLS* estimator $\widehat{\boldsymbol{\beta}}$ in the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

is *consistent*, so that it satisfies

$$\operatorname{plim} \widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} \tag{2.1}$$

and that it is *asymptotically normally distributed* with

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim_a N(\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}) \tag{2.2}$$

where

$$\mathbf{Q} = \operatorname{plim}(\frac{1}{n}\mathbf{X}'\mathbf{X}).$$

This result provides an asymptotic justification for the use of *OLS* even when condition A4 of the classical model does not hold so that it may not possible to show that *OLS* is unbiased or that it is exactly normally distributed.

The consistency result can be shown to follow from the assumptions that

$$\operatorname{plim}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right) = \mathbf{Q} \text{ exists and is non singular} \tag{AA1}$$

and

$$\operatorname{plim}\left(\frac{\mathbf{X}'\mathbf{u}}{n}\right) = \mathbf{0} \tag{AA2}$$

while the asymptotic normally can be proved with the additional assumption that

$$\frac{\mathbf{X}'\mathbf{u}}{\sqrt{n}} \sim_a N(\mathbf{0}, \sigma^2 \mathbf{Q}). \tag{AA3}$$

AA1 is the assumption that the sample moments of the regressors converge in probability to fixed finite values. AA2 is an assumption that the sample covariance between the regressors and the error term tends asymptotically to zero. This assumption does not entirely rule out correlation between the regressors and error term so long as this dies out as the sample size tends to infinity.

Assumption (AA3) is a particular form of *central limit theorem*. For this to hold requires conditions on the error vector $\mathbf{u}$. The conditions A1 and A2 of the classical model, that $\mathbf{u}$ is independently identically distributed with zero mean and constant covariance matrix, or formally,

$$\mathbf{u} \sim iid(\mathbf{0}, \sigma^2 \mathbf{I})$$

are sufficient for AA3 to hold although it can also be proved to hold under considerably weaker conditions. Note that result (2.2) does not require any assumption about the *form* of the distribution of the errors $\mathbf{u}$. In particular, it is not necessary to assume normality of $\mathbf{u}$ to derive the asymptotic normality of $\widehat{\boldsymbol{\beta}}$.

In general it should be noted that the particular assumptions AA1–AA3 are sufficient but not necessary and can be relaxed to some degree, at the cost of involving heavier mathematics.

The consistency of $\widehat{\boldsymbol{\beta}}$ is easy to establish since

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$$

so that

$$\begin{aligned}
\operatorname{plim}(\widehat{\boldsymbol{\beta}}) &= \boldsymbol{\beta} + \operatorname{plim}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\right) \\
&= \boldsymbol{\beta} + \operatorname{plim}\left(\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{u}}{n}\right)\right) \\
&= \boldsymbol{\beta} + \operatorname{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\operatorname{plim}\left(\frac{\mathbf{X}'\mathbf{u}}{n}\right) \\
&= \boldsymbol{\beta} + \mathbf{Q}^{-1}\mathbf{0} = \boldsymbol{\beta}
\end{aligned}$$

by (AA1) and (AA2) and Slutsky's theorem (1.2).

The asymptotic normality follows since

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$$

so that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\frac{\mathbf{X}'\mathbf{u}}{\sqrt{n}}$$

but from assumptions (AA1) and (AA3) and Cramér's theorem (1.5), it follows that

$$\begin{aligned}
\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &\sim {}_aN(\mathbf{0}, \sigma^2\mathbf{Q}^{-1}\mathbf{Q}\mathbf{Q}^{-1}) \\
&\sim {}_aN(\mathbf{0}, \sigma^2\mathbf{Q}^{-1}).
\end{aligned}$$

This result allows the development of asymptotically valid inference on the *OLS* estimator $\widehat{\boldsymbol{\beta}}$ even when the exact distribution of the estimator is unknown.

3 Maximum Likelihood Estimation

Maximum likelihood is a very general statistical method of estimation which has properties that can be justified in terms of large sample theory. It is based on

the joint probability density function of the data observations $\mathbf{y}$ with parameter vector θ:

$$f(\mathbf{y};\theta)$$

where the notation makes explicit that this is interpreted in the conventional way as a function of $\mathbf{y}$ for given values of the parameters θ. However, it is equally possible to interpret this probability density as a function of parameters θ for given values of the data observations $\mathbf{y}$. In this guise it is known as the *likelihood function* and written as

$$L(\theta;\mathbf{y})$$

to make explicit the different interpretation.

The *maximum likelihood estimator* or *ML* estimator is that value of θ that maximises the likelihood function. This estimator is denoted as $\widetilde{\theta}$. Formally

$$\widetilde{\theta} = \max_{\theta} L(\theta;\mathbf{y}). \tag{3.1}$$

The *ML* estimator chooses the parameter values which are the most likely to have generated the data that we actually observe. Intuitively this is a very reasonable principle of estimation.

Solution of the problem (3.1) can be found by solving the set of first order conditions

$$\frac{\partial L}{\partial \theta} = \mathbf{0}$$

along with checking that the matrix of second order conditions

$$\frac{\partial^2 L}{\partial \theta \partial \theta'} = \mathbf{H}$$

is a *negative definite matrix*. This ensures that a *maximum* rather than some other turning point has been found. Sometimes it is easier to maximise the logarithm of the likelihood function, rather than the original likelihood function itself. This is possible since logarithm is a *monotonically increasing function* so that

$$\max_{\theta} \log L(\theta;\mathbf{y}) \equiv \max_{\theta} L(\theta;\mathbf{y})$$

It is not always possible to find an *explicit closed form solution* to the problem (3.1). If not, then the solution has to be found iteratively using numerical optimisation techniques. These methods start from an initial solution θ_0 and search for a new value θ_1 that increases the value of the likelihood function. This process is repeated until a point is found at which it is not possible to increase the function further. This is then the solution. It is possible for a likelihood function to have more than one local maximum. In this case the iterative methods may find a solution that is only a local maximum rather than the global maximum.

3.1 Properties of *ML* estimators

Under quite general conditions it can be shown that *ML* estimators are: *consistent*, *asymptotically normally distributed*, and *asymptotically efficient*. Proofs of these results are beyond the scope of this course. The asymptotic distribution of the *ML* estimator is given by

$$\sqrt{n}(\widetilde{\theta} - \theta) \sim_a N(\mathbf{0}, \mathbf{Q}^{-1})$$

where

$$\mathbf{Q} = \text{plim}\left(-\frac{1}{n}\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right).$$

The matrix $\mathbf{Q}$ can be written as

$$\mathbf{Q} = \lim_{n\to\infty}\frac{1}{n}\ell(\theta)$$

where

$$\ell(\theta) = -\,\text{E}\left(\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right)$$

is known as the *information matrix*.

Under some fairly general conditions, it can be proved that any unbiased estimator has variance covariance matrix greater than or equal to the inverse of the information matrix

$$\ell(\theta)^{-1}.$$

This is known as the *Cramér-Rao bound*. Thus asymptotically, the *ML* estimator attains the Cramér-Rao bound and so is asymptotically efficient.

In order to write down the likelihood function, it is necessary to specify the form of the probability distribution function of the observations $\mathbf{y}$. In econometric applications, the normal distribution is almost always assumed. However, whatever the distribution $\mathbf{y}$, the ML estimator is asymptotically normally distributed. It is possible to use ML estimation using a normal likelihood function, even when the true distribution function is known to be not normal. This is called *quasi-maximum likelihood estimation*. It may be sensible if the true distribution function is unknown or very complicated. In many cases, the large sample distribution of the resulting parameter vector is the same as that of the correct *ML* estimator.

3.2 Concentrating the Likelihood Function

Consider the likelihood function

$$L(\theta_1, \theta_2; \mathbf{y})$$

where the parameter vector θ has been partitioned into two subsets and it is assumed that for the second subset, θ_2, the first order conditions

$$\frac{\partial L}{\partial \theta_2} = \mathbf{0}$$

can be solved to give an explicit expression for θ_2 as a function of the remaining parameters θ_1:

$$\theta_2 = \mathbf{h}(\theta_1).$$

In this case, the likelihood function L can be written as

$$L(\theta_1, \theta_2; \mathbf{y}) = L(\theta_1, \mathbf{h}(\theta_1); \mathbf{y}) = L^*(\theta_1; \mathbf{y})$$

which is a function only of θ_1. Substituting the explicit solution for the parameters θ_2 into L reduces the dimensions of the parameter space, and hence the computational cost of maximising the likelihood function. This process is called *concentrating the likelihood* and the parameters θ_2 are said to have been *concentrated out*. It is often possible to concentrate out some of the parameters, especially those of the variance covariance matrix.

3.3 Maximum Likelihood in the Classical model

As an example, we derive the *ML* estimator in the classical linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where $\mathbf{X}$ is fixed in repeated samples and

$$\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

In this model, as we have previously shown,

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

so that the model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ and the likelihood function $L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})$ is given by the joint normal density function:

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right). \tag{3.2}$$

It is more convenient to work with the logarithm of this density function

$$\log L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{3.3}$$

Since logarithm is a monotonically increasing function, the values of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ that maximise the log-likelihood function (3.3), also maximise the likelihood function (3.2).

Setting the first derivatives of (3.3) to zero gives the first order conditions for a maximum:

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}(-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \tag{3.4}$$

and

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0. \tag{3.5}$$

Solving these equations gives explicit solutions for the maximum likelihood estimators

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{3.6}$$

and

$$\widetilde{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n} \tag{3.7}$$

where $\mathbf{e}$ is the vector of *ML* residuals

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}.$$

From (3.6) it is clear that under the assumptions of the classical linear model, the maximum likelihood estimator of $\boldsymbol{\beta}$ is identical to the *OLS* estimator. On the other hand, the *ML* estimator of σ^2, (3.7), differs from the *OLS* estimator in that it has no correction for degrees of freedom. As the sample size increases, this difference disappears.

The *ML* estimator of σ^2 in this model is *biased* in finite samples, but it is *consistent* since

$$\begin{aligned} \widetilde{\sigma}^2 &= \frac{\mathbf{e}'\mathbf{e}}{n} = \frac{\mathbf{u}'\mathbf{M}\mathbf{u}}{n} \\ &= \frac{\mathbf{u}'\mathbf{u}}{n} - \frac{\mathbf{u}'\mathbf{X}}{n}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\frac{\mathbf{X}'\mathbf{u}}{n} \end{aligned}$$

so that

$$\operatorname{plim}(\widetilde{\sigma}^2) = \operatorname{plim}\left(\frac{\mathbf{u}'\mathbf{u}}{n}\right) - \operatorname{plim}\left(\frac{\mathbf{u}'\mathbf{X}}{n}\right)\operatorname{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\operatorname{plim}\left(\frac{\mathbf{X}'\mathbf{u}}{n}\right)$$

and by assumptions (AA1) and (AA2) the second term is zero. However, the first term

$$\frac{\mathbf{u}'\mathbf{u}}{n} = \frac{1}{n}\sum_{i=1}^{n} u_i^2$$

is the sample mean of the squared disturbances u_i^2 which are independently identically distributed with $\mathrm{E}(u_i^2) = \sigma^2$. Therefore, by *Khintchine's theorem* (1.3) it follows that

$$\operatorname{plim}\left(\frac{\mathbf{u}'\mathbf{u}}{n}\right) = \sigma^2$$

so that

$$\operatorname{plim}(\widetilde{\sigma}^2) = \sigma^2.$$

3.4 *ML* estimators in the *GLS* model

Consider the likelihood function in the *GLS* model where

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

but where now

$$\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{V}).$$

so that

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V}).$$

The likelihood function has the form of the joint normal distribution

$$L(\boldsymbol{\beta}, \sigma^2, \mathbf{V}; \mathbf{y}) = (2\pi\sigma^2)^{-\frac{n}{2}} |\mathbf{V}|^{-\frac{1}{2}} \exp(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))$$

and the log-likelihood function is given by

$$\log L(\boldsymbol{\beta}, \sigma^2, \mathbf{V}; \mathbf{y}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}).$$

Maximising this likelihood with respect to the unknown parameters $\boldsymbol{\beta}$, σ^2, and $\mathbf{V}$ gives the maximum likelihood estimator. This is a *feasible GLS* estimator.

To take a specific example, consider the first order autoregressive model where

$$\mathbf{V} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \rho \\ \rho^{T-1} & \rho^{T-2} & \cdots & \rho & 1 \end{bmatrix}$$

with $\mathbf{V}^{-1} = \mathbf{L}'\mathbf{L}$ where

$$\mathbf{L} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}$$

and $|\mathbf{L}| = \sqrt{1-\rho^2}$. In this case the log-likelihood function can be written as

$$\log L(\boldsymbol{\beta}, \sigma_\varepsilon^2, \rho; \mathbf{y}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma_\varepsilon^2 - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2\sigma_\varepsilon^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{L}'\mathbf{L}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \tag{3.8}$$

where

$$\log|\mathbf{V}| = -\log\left|\mathbf{V}^{-1}\right| = -\log|\mathbf{L}'\mathbf{L}| = -\log|\mathbf{L}|^2 = -\log(1-\rho^2).$$

Full maximum likelihood estimates are obtained by maximising the function (3.8). A simplification comes from dropping the initial observation which needs to be treated differently from all the others. As long as the number of observations is large, then dropping one will not make much difference.

Separating out the log-likelihood of the initial observation $\log L(\boldsymbol{\beta}, \sigma_\varepsilon^2, \rho; y_1)$ from the rest,

$$\begin{aligned}\log L(\boldsymbol{\beta}, \sigma_\varepsilon^2, \rho; \mathbf{y}) &= \log L(\boldsymbol{\beta}, \sigma_\varepsilon^2, \rho; y_1) - \frac{n-1}{2}\log 2\pi - \frac{n-1}{2}\log \sigma_\varepsilon^2 \\ &- \frac{1}{2\sigma_\varepsilon^2}(\mathbf{y}^+ - \mathbf{X}^+\boldsymbol{\beta})'(\mathbf{y}^+ - \mathbf{X}^+\boldsymbol{\beta})\end{aligned}$$

where $\mathbf{y}^+ = \mathbf{y} - \rho\mathbf{y}_{-1}$ and $\mathbf{X}^+ = \mathbf{X} - \rho\mathbf{X}_{-1}$are *quasi-differences,* each with $n-1$ rows, corresponding to observations $2, \cdots, n-1$ of the original data.

If the component representing the initial observation is ignored, then, maximising the rest, gives the first order conditions for $\boldsymbol{\beta}$ and ρ

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \mathbf{0} \Rightarrow \widetilde{\boldsymbol{\beta}} = (\mathbf{X}^{+\prime}\mathbf{X}^+)^{-1}\mathbf{X}^{+\prime}\mathbf{y}^+$$

and

$$\frac{\partial \log L}{\partial \rho} = 0 \Rightarrow \widetilde{\rho} = (\mathbf{u}_{-1}'\mathbf{u}_{-1})^{-1}\mathbf{u}_{-1}'\mathbf{u}$$

where $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$.

One method of solving these two simultaneous equations would be to start with an initial value of ρ, say 0, and iterate between the two equations until convergence is achieved. This is the famous *Cochrane-Orcutt* iterative method. Thus we can interpret this as a numerical method of obtaining (approximate) maximum likelihood estimates of $\boldsymbol{\beta}$ and ρ in the *AR(1)* model, where the approximation comes from the fact that the first observation has been dropped. As a numerical method, the Cochrane-Orcutt method may not necessarily be the most efficient way to compute the *ML* estimators.

4 Large sample Tests

Consider the testing of a null hypothesis H_0 concerning the vector of ML parameters θ

$$H_0 : \mathbf{g}(\theta) = \mathbf{0} \tag{4.1}$$

where $\mathbf{g}$ is an $r \times 1$ vector valued function of general nonlinear restrictions. If the restrictions are linear, then they can be written in the simpler form

$$H_0 : \mathbf{H}\theta - \mathbf{h} = \mathbf{0} \tag{4.2}$$

where $\mathbf{H}$ is a $r \times k$ matrix and $\mathbf{h}$ is an $r \times 1$ vector, both of known constants. It is possible to obtain ML estimates both imposing and not imposing the restrictions. Let $\widetilde{\theta}$ represent the unrestricted ML estimates and $\widetilde{\theta}_R$ the restricted estimates. Note that

$$L(\widetilde{\theta}_R) \leq L(\widetilde{\theta})$$

since the unrestricted model includes the restricted as a special case so cannot have a smaller maximum likelihood.

Three general principles of large sample testing have been used in the econometrics literature: the *likelihood ratio test*, the *Wald test* and the *Lagrange multiplier test*.

4.1 The Likelihood ratio test

The ratio

$$\lambda = \frac{L(\widetilde{\theta}_R)}{L(\widetilde{\theta})}$$

is called the *likelihood ratio* or LR. It follows that $0 \leq \lambda \leq 1$. A value close to zero is indication that the restrictions are invalid whereas a value close to one indicates that the restrictions are valid.

It can be proved that minus twice the logarithm of the likelihood ratio is asympototically distributed with a chi-squared distribution, with r degrees of freedom, or algebraically,

$$-2\log\lambda = 2\log L(\widetilde{\theta}) - 2\log L(\widetilde{\theta}_R) \sim_a \chi^2_r.$$

The likelihood ratio test is very easy to compute but involves the estimation of both restricted and unrestricted models.

4.2 The Wald test

The *Wald* test principle W is based upon estimates from the unrestricted model only. Since

$$\sqrt{n}(\widetilde{\theta} - \theta) \sim_a N(\mathbf{0}, \mathbf{Q}^{-1})$$

it follows that, under the linear null hypothesis (4.2),

$$\sqrt{n}(\mathbf{H}\widetilde{\theta} - \mathbf{h}) \sim_a N(\mathbf{0}, \mathbf{H}\mathbf{Q}^{-1}\mathbf{H}')$$

and that

$$n(\mathbf{H}\widetilde{\theta} - \mathbf{h})'(\mathbf{H}\mathbf{Q}^{-1}\mathbf{H}')^{-1}(\mathbf{H}\widetilde{\theta} - \mathbf{h}) \sim_a \chi^2_r.$$

On the null hypothesis this quadratic form will be close to zero, whereas on the alternative hypothesis it will be greater than zero. In practice, the asymptotic covariance matrix $\mathbf{Q}^{-1}$ is replaced by the finite sample estimator $n\ell(\widetilde{\theta})^{-1}$.

The classical exact F and t tests in the standard regression model follow from the general principle of the Wald test.

For nonlinear restrictions (4.1), an analogous result holds approximately so that the quadratic form

$$\mathbf{g}(\widetilde{\theta})'(\mathbf{G}(\widetilde{\theta})\ell(\widetilde{\theta})^{-1}\mathbf{G}(\widetilde{\theta})')^{-1}\mathbf{g}(\widetilde{\theta}) \approx_a \chi^2_r$$

where $\approx_a$denotes 'is approximately asymptotically distributed' and

$$\mathbf{G}(\widetilde{\theta}) = \frac{\partial \mathbf{g}(\widetilde{\theta})}{\partial \theta'}$$

is the $r \times k$ matrix of derivatives of the restrictions with respect to the parameter vector. It is a feature of nonlinear Wald tests that the test statistic is not invariant to the way that the nonlinear restrictions are formulated.

4.3 The Lagrange Multiplier test

The *Lagrange multiplier* or *LM* test principle is based on estimates from the restricted model only. The test statistic is

$$\frac{\partial \log L(\widetilde{\theta}_R)}{\partial \theta}' \ell(\widetilde{\theta}_R)^{-1} \frac{\partial \log L(\widetilde{\theta}_R)}{\partial \theta} \sim_a \chi^2_r$$

It can be shown that this statistic is related to the lagrange multipliers in the solution of the constained problem of the maximisation of the likelihood function subject to the restrictions (4.1)

$$\max_\theta L(\theta; \mathbf{y}) \quad s.t. \quad \mathbf{g}(\theta) = \mathbf{0}.$$

Diagnostic tests based on examining regression residuals for evidence of misspecification generally take the form of an *LM* test.

4.4 Relationship between the test procedures

It can be shown that the three test procedures: *Likelihood Ratio*, *Wald* and *Lagrange Multiplier* are all asymptotically equivalent. Thus it is possible to choose the test principle that is most convenient. If it is easier to estimate the restricted model under the null hypothesis than the unrestricted model on the alternative hypothesis, then the *LM* test will generally be the preferred choice. On the other hand, it may be easier to estimate the unrestricted model, in which case a *Wald* test will be more convenient. The likelihood ratio test has the advantage that it treats both models symmetrically.

Despite the asymptotic equivalence of the three testing principles, in many important cases including the linear regression model, it is possible to establish the inequality that

$$LM \leq LR \leq W.$$

4.5 The Classical Linear Model

Consider testing the linear hypothesis

$$H_0 : \mathbf{H}\boldsymbol{\beta} - \mathbf{h} = \mathbf{0}$$

in the classical linear model. The three testing principles result in the statistics

$$LR : n \log(\widetilde{\sigma}_R^2/\widetilde{\sigma}^2) = n \log(SSR_R/SSR)$$

$$W : (\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{h})'(\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}')^{-1}(\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{h})/\widetilde{\sigma}^2$$

$$LM : \mathbf{e}_R'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}_R/\widetilde{\sigma}_R^2$$

where $\widetilde{\sigma}^2$ and $\widetilde{\sigma}_R^2$ are the estimated *ML* error variances in the unrestricted and restricted models respectively, SSR and SSR_R are sum of squared residuals in unrestricted and restricted models, and $\mathbf{e}_R$ is the vector of OLS residuals in the resricted model. The numerators of the *Wald* and *LM* statistics can be shown to be identical so that these two tests differ only in the estimator of the error variance. Asymptotically, this makes no difference since both are consistent under the null hypothesis.

The likelihood ratio test in this model can be transformed by monotone transformation into

$$n(\exp(LR/n) - 1) = \frac{SSR_R - SSR}{\widetilde{\sigma}^2}$$

which can be shown to be identical to the Wald test statistic. Note that a degrees of freedom correction can be applied to the variance estimators $\widetilde{\sigma}^2$ and $\widetilde{\sigma}_R^2$ without affecting the asymptotic results.

5 Appendix: The Cramér-Rao bound

Since the likelihood function is a probability density function, it follows that it integrates to one or that

$$\int L(\theta;\mathbf{y})d\mathbf{y} = 1$$

Under certain regularity conditions it is possible to differentiate under the integral sign so that

$$\begin{aligned}\frac{\partial}{\partial\theta}\int L(\theta;\mathbf{y})d\mathbf{y} &= \int \frac{\partial}{\partial\theta}L(\theta;\mathbf{y})d\mathbf{y} \\ &= \int L\frac{\partial \log L}{\partial\theta}d\mathbf{y} = \mathbf{0} \end{aligned} \tag{5.1}$$

where the last line follows from the standard derivative result for a logarithm that

$$\frac{\partial \log L}{\partial\theta} = \frac{1}{L}\frac{\partial L}{\partial\theta}.$$

However, since $L(\theta;\mathbf{y})$ is a probability density function, by definition it is true that

$$\int L\frac{\partial \log L}{\partial\theta}d\mathbf{y} = E\left(\frac{\partial \log L}{\partial\theta}\right)$$

so that (5.1) says that

$$E\left(\frac{\partial \log L}{\partial\theta}\right) = \mathbf{0}. \tag{5.2}$$

Differentiating (5.1) again,

$$\begin{aligned}&\frac{\partial}{\partial\theta'}\int L\frac{\partial \log L}{\partial\theta}d\mathbf{y} \\ =\ &\int L\frac{\partial^2 \log L}{\partial\theta\delta\theta'}d\mathbf{y} + \int \frac{\partial \log L}{\partial\theta}\frac{\partial L}{\partial\theta'}d\mathbf{y} \\ =\ &\int L\frac{\partial^2 \log L}{\partial\theta\partial\theta'}d\mathbf{y} + \int \frac{\partial \log L}{\partial\theta}\frac{\partial \log L}{\partial\theta'}Ld\mathbf{y} = \mathbf{0} \end{aligned} \tag{5.3}$$

where the last line makes use of the standard logarithm derivative result again.

The first term in (5.3) is equal to the expectation of the matrix of second log derivatives

$$E\left(\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right)$$

whereas the second term is equal to

$$E\left(\frac{\partial \log L}{\partial\theta}\frac{\partial \log L}{\partial\theta'}\right) = \text{var}\left(\frac{\partial \log L}{\partial\theta}\right)$$

from (5.2). Thus (5.3) states that

$$\text{var}\left(\frac{\partial \log L}{\partial \theta}\right) = -E\left(\frac{\partial^2 \log L}{\partial \theta \partial \theta'}\right) = \ell(\theta) \tag{5.4}$$

where the right hand side of (5.4) is called the *information matrix*, denoted by $\ell(\theta)$.

It can be shown that for any *unbiased* estimator $\overline{\theta}$, its variance covariance matrix

$$\text{var}(\overline{\theta}) = \text{E}(\overline{\theta}-\theta)(\overline{\theta}-\theta)'$$

has the property that

$$\text{var}(\overline{\theta}) - \ell(\theta)^{-1} \geq \mathbf{0}$$

is a *positive semi-definite* matrix. This result is known as the *Cramér-Rao inequality*. It implies that the inverse of the information matrix $\ell(\theta)^{-1}$ is a *lower bound* for the variance covariance matrix of an unbiased estimator.