MAXIMUM LIKELIHOOD ESTIMATION OF TIME SERIES MODELS

These notes have been divided into different sections of varying degrees of importance and information for the MLE approach with time series data.

Section 1 is a revision of the general properties of MLE and gives examples, 1(a) through 1(d) of the application of MLE to estimation from different standard distributions. All this is standard statistical inference and is essentially a revision.

Section 2 applies MLE to the classical linear regression model with fixed regressors and NID disturbances.

Section 3 then considered estimation of the AR(p) model with initial conditions assumed fixed and equal to zero. Then under the additional assumption of normality, it is shown that the "conditional" MLE in this situation is the same as solving the Yule Walker equations. The estimator is seen to attain the Cramer Rao lower bound and therefore to be fully asymptotically efficient.

Section 4 describes the numerical solution to finding the MLE in non-linear situations when regression type methods cannot be used. The estimation of ARMA(p, q) models with initial values conditioned to be zero, is also discussed.

The next three sections are concerned with the form of the asymptotic distribution of the MLE for various types of ARMA models. Section 5 illustrates the estimation method for the MA(1) model and also gives details of its asymptotic distribution. General results for the ARMA(p, q) process are then given in Section 6. The next section, 7 describes, without proof, the corresponding results for the MLE of the parameters in the stationary linear regression model with ARMA(p, q) errors. An interesting aspect of the results is that the regression parameter estimates, the ARMA error parameter estimates and the estimate of σ^2 are all asymptotically mutually uncorrelated. This allows the construction of Two Step estimators, which are briefly discussed in section 8. A simple example of a two step estimator is the Cochrane Orcutt estimator of the parameters in a regression with AR(1) errors. These methods can be quite useful in more complex situations.

The final three sections discuss the issue of finding full MLE based on not conditioning the initial observations to be zero, but rather to assume that they are generated by the same process as the observed series. This is slightly esoteric, and is primarily included for interest, rather than being an essential technique. However, the methods do rely on the Prediction Error Decomposition, which is discussed in section 9 and is the basis of the Kalman Filter. Section 10 finds the full MLE for the AR(1) model, and sections 11 and 12 provide analogous results for the AR(p) and ARMA(p, q) models respectively.

1) Properties of Maximum Likelihood Estimation (MLE)

Once an appropriate model or distribution has been specified to describe the characteristics of a set of data, the immediate issue is one of finding desirable parameter estimates. From a frequentist perspective the ideal is the Maximum Likelihood Estimator (MLE) which provides a general method for estimating a vector of unknown parameters in a possibly multivariate distribution. In order to consider as general a situation as possible suppose y is a random variable with probability density function $f(y)$ which is characterized by a set of p unknown parameters $\Theta^{/} = (\Theta_1, \Theta_2, ..., \Theta_p)$. A random sample of T observations $(y_1, y_2, ... y_T)$ is available and the likelihood L, is defined as the joint density of the observations; i.e.

$$L = f(y_1, y_2, ... y_T) = \Pi f(y_t; \Theta)$$

Note that since the y_t are randomly selected from the population distribution, sample values can be considered to have been independently drawn, so the likelihood, which is the joint distribution of $(y_1, y_2, ... y_T)$ is merely the product of the marginal densities. The Maximum Likelihood Estimator (MLE), $\Theta^{/} = (\Theta_1, \Theta_2, ..., \Theta_p)$, is the value of the parameters that is most likely to have generated the observed sample of data. The attraction of MLE is that subject to fairly minor conditions, it has very desirable properties in large samples (i.e., asymptotically). These conditions are that,

(i) the range of y is independent of Θ,

(ii) f(y) possesses third order derivatives with respect to Θ and are bounded by integrable functions of y, and

(iii) the space of admissible values of Θ is a closed subset of p dimensional space.

Subject to these conditions and on denoting the MLE of Θ by Θ it can be shown that

(a) $\hat{\Theta}$ is a consistent estimator of Θ.

(b) $T^{1/2}(\hat{\Theta}-\Theta) \rightarrow N(0,V)$

where $V = \left[\lim_{T\rightarrow\infty} \frac{I(\Theta)}{T}\right]^{-1}$ and $I(\Theta)$ is known as the information matrix and is defined as

$$I(\Theta) = -E\left[\frac{\delta^2 \ln(L)}{(\delta\Theta)(\delta\Theta)^{/}}\right] = -E\left[\frac{\delta^2 \ln(L)}{(\delta\Theta_i \delta\Theta_j)}\right] \text{ for i,j } = 1,2,\ldots\text{p.}$$

$$I(\Theta) = -E\begin{bmatrix} \frac{\delta^2 \ln(L)}{\delta^2\Theta_1^2} & \frac{\delta^2 \ln(L)}{\delta\Theta_1\delta\Theta_2} & . & \frac{\delta^2 \ln(L)}{\delta\Theta_1\delta\Theta_p} \\ \frac{\delta^2 \ln(L)}{\delta\Theta_2\delta\Theta_1} & \frac{\delta^2 \ln(L)}{\delta\Theta_2^2} & . & \frac{\delta^2 \ln(L)}{\delta\Theta_2\delta\Theta_p} \\ . & . & . & . \\ \frac{\delta^2 \ln(L)}{\delta\Theta_p\delta\Theta_1} & \frac{\delta^2 \ln(L)}{\delta\Theta_p\delta\Theta_2} & . & \frac{\delta^2 \ln(L)}{\delta\Theta_p^2} \end{bmatrix}$$

The asymptotic distribution of $\hat{\Theta}$ which is used for hypothesis testing is

$$\hat{\Theta} \sim N\left[0, \left(\frac{V}{T}\right)\right]$$

It should be noted that in most practical situations, it is generally easier to maximize the logarithm of the likelihood, which is a monotonic function of the likelihood.

For any unbiased estimator $\tilde{\Theta}$, with covariance matrix of Ω, the **Cramer Rao Lower Bound** (CRLB), indicates that the matrix $\left[\Omega - I(\Theta)^{-1}\right]$ must be positive semi definite.

If an estimator has covariance matrix $I(\Theta)^{-1}$ then it is efficient.

In some situations the CRLB may not be attainable; although $\tilde{\Theta}$ may still be efficient in this case. To prove such an estimator is efficient, requires showing that the estimator is a function of a completely sufficient statistic. Some examples of MLE and their properties follow before some specific time series models.

Examples of MLE:

1(a): MLE of the Exponential Distribution:

One of the simplest examples of MLE is to consider estimation of the parameter θ, in the exponential distribution,

$$f(y) = \theta e^{-\theta y}$$

Suppose there is a random sample of T observations, $(y_1, y_2, ...y_T)$ where y_t has an exponential distribution with parameter θ. The likelihood is then given by,

$$L = \left[\theta \exp(-\theta y_1)\right]\left[\theta \exp(-\theta y_2)\right].......\left[\theta \exp(-\theta y_T)\right] = \theta^T \exp(-\theta \sum_{t=1}^{T} y_t)$$

Then

$$\ln(L) = T\ln(\theta) - \theta\sum_{t=1}^{T} y_t$$

Since there is only one parameter the score vector is a scalar and is given by,

$$\frac{\delta \ln(L)}{\delta\theta} = \frac{T}{\theta} - \sum_{t=1}^{T} y_t$$

The MLE of θ can be determined by equating the above with zero to obtain

$$\hat{\theta} = \frac{T}{\sum_{t=1}^{T} y_t} = \frac{1}{\bar{y}}$$

so that the MLE of θ is the inverse of the sample mean, which is an intuitively reasonable estimator of θ since the expected value of y_t is $E(y_t) = (1/\theta)$. In this one parameter case, the information matrix also reduces to a scalar and the second derivative of the log likelihood function is,

$$\frac{\delta^2 \ln(L)}{\delta\theta^2} = -\frac{T}{\theta^2}$$

Hence, $I(\theta) = \frac{T}{\theta^2}$ and $V = \theta^2$, so that

$$T^{1/2}(\hat{\theta} - \theta) \rightarrow N\left(0, \theta^2\right)$$

Or rather more informally, the asymptotic distributions of the MLE can be expressed as,

$$\hat{\theta} \to N\left(\theta, \frac{\theta^2}{T}\right)$$

1(b): MLE for the Poisson Distribution:

The discrete distribution is,

$$f(y) = \frac{\theta^y e^{-\theta}}{y!} \text{ for } y = 0,1,2,....$$

The likelihood is,

$$L = \left(\frac{\theta^{y_1} e^{-\theta}}{y_1!}\right)\left(\frac{\theta^{y_2} e^{-\theta}}{y_2!}\right)..........\left(\frac{\theta^{y_T} e^{-\theta}}{y_T!}\right) = \frac{\theta^{\sum y_t} e^{-T\theta}}{\Pi(y_t!)}$$

$$\ln(L) = \left(\sum y_t\right)\ln(\theta) - T\theta - \ln\left[\Pi\left(y_t!\right)\right]$$

$$\frac{\delta \ln(L)}{\delta\theta} = \frac{\left(\sum y_t\right)}{\theta} - T$$

On setting $\frac{\delta \ln(L)}{\delta\theta} = 0$, the MLE is found to be $\hat{\theta} = \frac{\sum y_t}{T} = \bar{y}$.

$$\frac{\delta^2 \ln(L)}{\delta\theta^2} = -\frac{\left(\sum y_t\right)}{\theta^2}$$

$$I(\theta) = -E\left[\frac{\delta^2 \ln(L)}{\delta\theta^2}\right] = \frac{\sum E(y_t)}{\theta^2} = \frac{T\theta}{\theta^2} = \frac{T}{\theta}$$

Then,

$$T^{1/2}(\overset{\wedge}{\theta}-\theta) \rightarrow N(0,\theta)$$

1(c): MLE of a scalar Normal Distribution

One of the simplest examples of MLE is to consider estimation of the mean μ, and variance σ^2, of a scalar normal distribution. Suppose that there exists a random sample of T observations, $(y_1, y_2, ...y_T)$, where $y_t \sim N(\mu, \sigma^2)$ and $\theta' = (\mu, \sigma^2)$. The likelihood is then given by,

$$L = (2\pi)^{-\frac{T}{2}} (\sigma^2)^{-\frac{T}{2}} \exp\left\{-\left(\frac{1}{2\sigma^2}\right)\sum(y_t - \mu)^2\right\},$$

Then

$$\ln(L) = c - \left(\frac{T}{2}\right)\ln(\sigma^2) - \left(\frac{1}{2\sigma^2}\right)\sum_{t=1}^{T}(y_t - \mu)^2$$

and the first derivatives of the logarithm of the likelihood, which are known as the elements of the score vector are defined as,

$$\frac{\delta \ln(L)}{\delta\mu} = \left(\frac{1}{\sigma^2}\right)\sum(y_t - \mu)$$

$$\frac{\delta \ln(L)^2}{\delta\sigma^2} = -\left(\frac{T}{2\sigma^2}\right) + \left(\frac{1}{2\sigma^4}\right)\sum(y_t - \mu)^2,$$

The MLE of μ and σ^2 can be determined by equating the above scores or first derivatives with zero to obtain

$$\hat{\mu} = \left(\frac{1}{T}\right)\sum y_t = \bar{y}$$

and

$$\hat{\sigma}^2 = \left(\frac{1}{T}\right)\sum\left(y_t - \hat{\mu}\right)^2 = \left(\frac{1}{T}\right)\sum\left(y_t - \bar{y}\right)^2$$

To find the information matrix the second derivatives of the log likelihood are found to be,

$$\frac{\delta^2 \ln(L)}{\delta\mu^2} = -\frac{T}{\sigma^2}$$

$$\frac{\delta^2 \ln(L)}{\delta\sigma^4} = \frac{T}{2\sigma^4} - \left(\frac{1}{\sigma^6}\right)\sum(y_t - \mu)^2$$

$$\frac{\delta^2 \ln(L)}{\delta\mu\delta\sigma^2} = -\frac{1}{\sigma^4}\sum(y_t - \mu)$$

To obtain I(θ), where $\theta' = \left(\mu, \sigma^2\right)$, it is necessary to evaluate minus the expectation of the above three expressions. In particular,

$$-E\left(\frac{\delta^2 \ln(L)}{\delta\mu^2}\right) = \frac{T}{\sigma^2}$$

$$-E\left(\frac{\delta^2 \ln(L)}{\delta\sigma^4}\right) = \frac{T}{2\sigma^4} - \left(\frac{1}{\sigma^6}\right)\sum E(y_t - \mu)^2 = \frac{T}{2\sigma^4}$$

$$-E\left(\frac{\delta^2 \ln(L)}{\delta\mu\delta\sigma^4}\right) = -\left(\frac{1}{\sigma^4}\right)\sum E(y_t - \mu) = 0$$

Hence,

$$I(\theta)=\begin{bmatrix}\frac{\delta^2 \ln(L)}{\delta\mu^2} & \frac{\delta^2 \ln(L)}{\delta\mu\delta\sigma^2} \\ \frac{\delta^2 \ln(L)}{\delta\mu\delta\sigma^2} & \frac{\delta^2 \ln(L)}{\delta\sigma^2}\end{bmatrix}=\begin{bmatrix}\frac{T}{\sigma^2} & 0 \\ 0 & \frac{T}{2\sigma^4}\end{bmatrix}$$

$$\frac{I(\theta)}{T}=\begin{bmatrix}\frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4}\end{bmatrix}$$

$$V=\left[\lim_{T\to\infty}\frac{I(\theta)}{T}\right]^{-1}=\left[\frac{I(\theta)}{T}\right]^{-1}=\begin{bmatrix}\sigma^2 & 0 \\ 0 & 2\sigma^4\end{bmatrix}$$

Or, rather more informally, the asymptotic distributions of the MLE can be expressed as,

$$\hat{\mu}\to N\left(\sigma^2,\frac{2\sigma^4}{T}\right)$$

and

$$\hat{\sigma}^2\to N\left(\sigma^2,\frac{2\sigma^4}{T}\right)$$

The diagonality of I(θ) implies that the MLE of μ and σ^2 are asymptotically uncorrelated. The MLE of the disturbance variance will generally have this property in most linear models. As is well known in elementary statistics the MLE of σ^2 is biased in small samples and it is conventional to use the unbiased estimator s^2, where

$$s^2 = \left(\frac{1}{T-1}\right)\sum_{t=1}^{T}(y_t - \bar{y})^2$$

It can be shown that $E\left(s^2\right) = \sigma^2$ and $Var\left(s^2\right) = \frac{2\sigma^4}{T-1}$. Hence,

$$s^2 \to N\left(\sigma^2, \frac{2\sigma^4}{T-1}\right)$$

so that the unbiased estimator does not attain the Cramer Rao lower bound and has a variance that exceeds that of σ^2. Hence the usual trade off between bias and MSE in small samples.

2) MLE of the Classical Linear Regression Model

In the classical linear regression model, the dependent or endogenous variable at time t is y_t and is defined as a linear combination of k explanatory variables contained in the vector x_t plus a disturbance term ε_t. The t'th observation on the model can be expressed as

$$y_t = x_t^{/}\beta + \varepsilon_t,$$

where $E(\varepsilon_t) = 0,$ $E(\varepsilon_t^2) = \sigma^2$ and $E(\varepsilon_t\varepsilon_s) = 0,$ for $s \neq t.$ In observation form the model can be expressed as,

$$Y = X\beta + \varepsilon$$

where Y is a T dimensional vector of observations on the dependent variable, X is an Txk matrix of observations on k independent variables, β is a k dimensional vector of unknown parameters and ε is a T dimensional vector of disturbances. On making the additional assumption of normality so that $\varepsilon \sim N(0,\sigma^2 I)$, the likelihood function is then given by,

$$L = (2\pi)^{-T/2}(\sigma^2)^{-T/2}\exp\left[-(1/2\sigma^2)\varepsilon^{/}\varepsilon\right]$$

The Jacobian of the transformation from ε to Y is unity, and on taking natural logarithms of the above,

$$\ln(L) = -\left(\frac{T}{2}\right)\ln(2\pi) - \left(\frac{T}{2}\right)\ln(\sigma^2) - \left(\frac{1}{2\sigma^2}\right)(Y - X\beta)^{/}(Y - X\beta)$$

Maximization of the above logarithm of the likelihood with respect of β will be equivalent to the minimization of $(Y - X\beta)^{/}(Y - X\beta)$ and hence it is clear that the MLE under normality will be equivalent to the standard OLS estimator. The maximization of the log likelihood is done by taking derivatives with respect to the unknown parameters $\Theta' = \left(\beta', \sigma^2\right)$ and equating with zero,

$$\frac{\delta \ln(L)}{\delta\beta} = -\left(\frac{1}{\sigma^2}\right)\left(X^{/}Y - X^{/}X\beta\right) = 0$$

$$\frac{\delta \ln(L)}{\delta\sigma^2} = -\left(\frac{T}{2\sigma^2}\right) + \left(\frac{1}{2\sigma^4}\right)\left(Y - X\beta\right)'\left(Y - X\beta\right) = 0,$$

The regularity condition of $E\left(\frac{\delta \ln[L]}{\delta\theta}\right) = 0$ is also satisfied. On solving the above equations the MLEs are found to be,

$$\hat{\beta} = (X^{/}X)^{-1}X^{/}Y$$

$$\hat{\sigma}^2 = \frac{(Y - X\hat{\beta})^{/}(Y - X\hat{\beta})}{T} = \frac{e'e}{T}$$

To form the information matrix,

$$\frac{\delta^2 \ln(L)}{\delta\beta\delta\beta'} = -\sigma^2 \left(X'X\right)^{-1}$$

$$\frac{\delta^2 \ln(L)}{\delta\beta\delta\sigma^2} = \frac{1}{\sigma^4}\left(X'Y - X'X\beta\right)$$

and

$$\frac{\delta^2 \ln(L)}{\delta\sigma^4} = \frac{T}{2\sigma^4} - \frac{1}{\sigma^6}\left(Y - X\beta\right)'\left(Y - X\beta\right)$$

On taking expectations of minus the above quantities gives,

$$-E\left(\frac{\delta^2 \ln(L)}{\delta\beta\delta\beta'}\right) = \sigma^2 \left(X'X\right)^{-1}$$

$$-E\left(\frac{\delta^2 \ln(L)}{\delta\sigma^4}\right) = -\frac{T}{2\sigma^4} + \frac{1}{\sigma^6}E\left(Y - X\beta\right)'\left(Y - X\beta\right) = -\frac{T}{2\sigma^4} + \frac{T\sigma^2}{\sigma^6} = \frac{T}{2\sigma^4}$$

which follows from the fact that $E\left(\varepsilon'\varepsilon\right) = T\sigma^2$. Also,

$$-E\left(\frac{\delta^2 \ln(L)}{\delta\beta\delta\sigma^2}\right) = \frac{1}{\sigma^4}\left[X'Y - X'XE\left(\hat{\beta}\right)\right] = 0$$

Hence,

$$I(\Theta) = \begin{bmatrix} \sigma^{-2}(X^{/}X) & 0 \\ 0 & (T/2\sigma^{4}) \end{bmatrix}$$

$$V = \left[\lim_{T\to\infty}\left(\frac{I(\Theta)}{T}\right)\right]^{-1} = \begin{bmatrix} \sigma^{2}(X^{/}X)^{-1} & 0 \\ 0 & (2\sigma^{4}) \end{bmatrix} = \begin{bmatrix} \sigma^{2}B^{-1} & 0 \\ 0 & (2\sigma^{4}) \end{bmatrix}$$

where $B = \lim_{T\to\infty}(T^{-1}\sum_{t=1}^{T} x_t x_t^{/}) = \lim_{T\to\infty}(T^{-1}X^{/}X)$

$$T^{1/2}(\hat{\beta} - \beta) \to N\left[0, \sigma^{2}(X^{/}X)^{-1}\right]$$

$$T^{1/2}(\hat{\sigma}^{2} - \sigma^{2}) \to N\left(0, 2\sigma^{4}\right).$$

Hence under the normality assumption, the MLE of β is equivalent to the OLS estimator. However the MLE of σ^2 uses a divisor of T rather than (T-k). Hence in small samples the MLE of σ^2 will be biased, although of course asymptotically it is unbiased.

3) Conditional MLE of AR(p) Process

In the case of the AR(p) process $\Theta' = \left(\phi_1, \phi_2, ..., \phi_p, \sigma^2\right)$, and if and the initial observations $y_0, y_{-1}, ..., y_{1-p}$ are equated with zero, then the resulting approximate MLE

$$L = \left(2\pi\right)^{-(T-p)/2}\left(\sigma^{2}\right)^{-(T-p)/2} \exp\left(-\left(\frac{1}{2\sigma^{2}}\right)\sum_{t=p+1}^{T}\{\phi(L)y_t\}^{2}\right),$$

$$\ln(L) = -\left(\frac{T-p}{2}\right)\ln(2\pi) - \left(\frac{T-p}{2}\right)\ln(\sigma^2) - \left(\frac{1}{2\sigma^2}\right)\sum_{t=p+1}^{T}[\phi(L)y_t]^2$$

will be equivalent to minimizing the **Conditional Sum of Squares (CSS)** function,

$$S = \left(\frac{1}{2\sigma^2}\right)\sum_{t=p+1}^{T}[\phi(L)y_t]^2 \,.$$

Partially differentiating the log likelihood with respect to a typical AR parameter ϕ_i and also σ^2 gives the (p+1) dimensional score vector

$$\frac{\delta \ln(L)}{\delta\phi_i} = -\frac{1}{\sigma^2}\sum \varepsilon_t\left(\frac{\partial\varepsilon_t}{\partial\phi_i}\right) = \left(\frac{1}{\sigma^2}\right)\sum \varepsilon_t y_{t-i},$$

$$\frac{\delta \ln(L)}{\delta\sigma^2} = -\left(\frac{T}{2\sigma^2}\right) + \left(\frac{1}{2\sigma^4}\right)\sum[\phi(L)y_t]^2$$

Equating the score vector to zero and solving gives the same Yule Walker equations as with OLS, and hence

$$\sum\left(y_t - \hat{\phi}_1 y_{t-1} - \hat{\phi}_2 y_{t-2} - \ldots - \hat{\phi}_p y_{t-p}\right)y_{t-k} = 0\,,$$

for k = 1,2,….p and

$$\sum y_t y_{t-k} = \hat{\phi}_1\sum y_{t-1}y_{t-k} + \hat{\phi}_2\sum y_{t-2}y_{t-k} + \ldots + \hat{\phi}_p\sum y_{t-p}y_{t-k}\,,$$

which are the same as the Yule Walker equations derived earlier. To derive the information matrix,

$$\frac{\delta^2 \ln(L)}{\delta\phi_i \delta\phi_j} = -\left(\frac{1}{\sigma^2}\right)\sum y_{t-j} y_{t-i} - \left(\frac{1}{\sigma^2}\right)\sum \varepsilon_t y_{t-i-j},$$

$$\frac{\delta^2 \ln(L)}{\delta\phi_i \delta\sigma^2} = -\left(\frac{1}{\sigma^4}\right)\sum \varepsilon_t y_{t-i},$$

$$\frac{\delta^2 \ln(L)}{\delta\sigma^4} = -\left(\frac{T}{2\sigma^4}\right) - \left(\frac{1}{\sigma^6}\right)\sum \varepsilon_t^2$$

Hence,

$$-E\frac{\delta^2 \ln(L)}{\delta\phi_i \delta\phi_j} = \left(\frac{1}{\sigma^2}\right)\sum E(y_{t-j} y_{t-i}) = \Gamma_{p-1}$$

which is the p dimensional Toeplitz autocovariance matrix. Also, since $E(\varepsilon_t y_{t-i}) = 0,$ for $i \geq 1,$ it follows that

$$-E\frac{\delta^2 \ln(L)}{\delta\phi_i \delta\sigma^2} = 0,$$

which is a p dimensional null vector. Finally, the (p+1, p+1) element of the information matrix is,

$$-E\frac{\delta^2 \ln(L)}{\delta\sigma^4} = T^{-1}\left(\frac{T}{2\sigma^4} - \left(\frac{1}{\sigma^6}\right)T\sigma^2\right) = \left(\frac{1}{2\sigma^4}\right)$$

$$T^{\frac{1}{2}}\left(\hat{\Theta} - \Theta\right) \to N(0, V)$$

$$V = \begin{bmatrix} \Gamma_{p-1}^{-1} & 0 \\ 0 & 2\sigma^4 \end{bmatrix}$$

Hence estimation of the AR(p) process by minimizing the CSS is equivalent to MLE under normality and with the initial observations set to zero. There is plenty of simulation evidence to suggest this is a reasonable procedure for at least a moderately large sample, i.e. T > 50 and for a relatively small set of parameters.

4) Computation of the MLE for ARMA(p, q) Models

For simplicity consider the problem where maximization of the log likelihood function is equivalent to minimization of the CSS (conditional sum of squares) function. Then a Taylor series expansion of $f(\Theta)$ around the minimum of Θ, which is denoted by $\tilde{\Theta}$ gives,

$$f(\Theta) \approx f(\tilde{\Theta}) + (\Theta - \tilde{\Theta})' \left(\frac{\delta f(\Theta)}{\delta \Theta} \right) + (1/2)(\Theta - \tilde{\Theta})' \left(\frac{\delta^2 f(\Theta)}{\delta \Theta' \delta \Theta} \right) (\Theta - \tilde{\Theta})$$

On differentiating the above expression with respect to θ,

$$\frac{\delta f(\Theta)}{\delta \Theta} \approx \frac{\delta f(\tilde{\Theta})}{\delta \Theta} + (\Theta - \tilde{\Theta})' \left(\frac{\delta^2 f(\tilde{\Theta})}{\delta \Theta' \delta \Theta} \right)$$

and neglecting higher order terms. However, at the optimum $\frac{\delta f(\tilde{\Theta})}{\delta \Theta} = 0$ by definition, then

$$\tilde{\Theta} = \Theta - \left(\frac{\delta^2 f(\tilde{\Theta})}{\delta \Theta' \delta \Theta} \right)^{-1} \left(\frac{\delta f(\Theta)}{\delta \Theta} \right)$$

Suppose $\tilde{\Theta}^k$ is the optimum estimate (i.e. minimizing CSS, or MLE), from the k'th iteration. Then the updating formula is,

$$\tilde{\Theta}^k = \tilde{\Theta}^{k-1} - \left(\frac{\delta^2 f(\tilde{\Theta}^{k-1})}{\delta\Theta'\delta\Theta} \right)^{-1} \left(\frac{\delta f(\tilde{\Theta}^{k-1})}{\delta\Theta} \right)$$

On denoting the Hessian as $H(\Theta) = \left(\frac{\delta^2 f(\Theta)}{\delta\Theta'\delta\Theta} \right)$ and the score vector as $s(\Theta) = \left(\frac{\delta f(\Theta)}{\delta\Theta} \right)$; then

$$\tilde{\Theta}^k = \tilde{\Theta}^{k-1} - H(\Theta)^{-1} s(\Theta)$$

so that the at the k'th iteration the estimate of the parameter vector Θ, is equal to the estimate at the previous iteration plus an adjustment due to the Hessian post multiplied by the score vector. The Hessian and score vector are both evaluated at the parameter estimates obtained at the previous iteration. The quality of good starting values is hard to over estimate. For ARMA models that do not have problems with parameters being close to the unit circle and are thus close to violating stationarity or invertibility conditions, then even setting initial parameter estimates equal to zero, is adequate.

The Hessian can either be evaluated as the second numerical derivatives of the log likelihood, or as the outer product gradient (opg) as the product of the score vector times its transpose.

The conditioning method can generally be applied to many types of dynamic models; for example for the stationary and invertible ARMA(p, q) process,

$$\phi(L) y_t = \theta(L) \varepsilon_t,$$

where $\varepsilon_t \sim NID(0, \sigma^2)$, it is necessary to estimate the (p+q+1) parameters $\Theta' = (\phi_1, \phi_2, ... \phi_p, \theta_1, \theta_2, ... \theta_q, \sigma^2)$. The conditional MLE can be obtained by neglecting the initial observations $y_0, y_{-1}, ..., y_{-(p+q)}$, and $\varepsilon_0, \varepsilon_{-1}, ..., \varepsilon_{-(q-1)}$, in the sense of equating them to zero. The objective is to then maximize the function given by,

$$\ln(L) = -\left(\frac{T}{2}\right)\ln(2\pi) - \left(\frac{T}{2}\right)\ln(\sigma^2) - \left(\frac{1}{2\sigma^2}\right)\sum_{t=p+q+1}^{T} \varepsilon_t^2$$

Again the maximization of the logarithm of the likelihood is equivalent to the minimization of the conditional sum of squares function,

$$S = \left(\frac{1}{2\sigma^2}\right)\sum_{t=p+q+1}^{T} \varepsilon_t^2 = \left(\frac{1}{2\sigma^2}\right)\sum_{t=p+q+1}^{T} [\phi(L)\theta(L)^{-1} y_t]^2,$$

$$S = \left(\frac{1}{2\sigma^2}\right)\sum_{t=p+q+1}^{T} [\pi(L) y_t]^2 = \left(\frac{1}{2\sigma^2}\right)\sum_{t=p+q+1}^{T} \left(\sum_{j=0}^{\infty} \pi_j y_{t-j}\right)$$

That is the last term has replaced ε_t with $\left(\sum_{j=0}^{\infty} \pi_j y_{t-j}\right)$, which in practice must be truncated at (t-1). Hence the CSS becomes,

$$S = \left(\frac{1}{2\sigma^2}\right)\sum_{t=p+q+1}^{T} \left(\sum_{j=0}^{t-1} \pi_j y_{t-j}\right)^2,$$

where the π_j coefficients are obtained from the usual AR representation $\pi(L) = \phi(L)\theta(L)^{-1}$ and hence each π_j is restricted to be functions of the original ARMA parameters. The maximization of the log likelihood then becomes a problem in numerical analysis.

5) The MA(1) Process in Detail:

The simplest example of non linear estimation is really the MA(1) process given by,

$$y_t = \varepsilon_t - \theta\varepsilon_{t-1},$$

so that the parameter vector is $\Theta^{/} = (\theta, \sigma^2)$. The log likelihood is

$$\ln(L) = -\left(\frac{T}{2}\right)\ln(2\pi) - \left(\frac{T}{2}\right)\ln(\sigma^2) - \left(\frac{1}{2\sigma^2}\right)\sum_{t=2}^{T}[(1-\theta L)^{-1}y_t]^2$$

so that the function has to be maximized by the usual software. The problem can also be expressed in terms of minimizing the CSS function,

$$S = \left(\frac{1}{2\sigma^2}\right)\sum_{t=1}^{T}\left(\sum_{j=0}^{t-1}\theta^j y_{t-j}\right)^2 .$$

The score vector is

$$\frac{\delta \ln(L)}{\delta\theta_i} = -\left(\frac{1}{\sigma^2}\right)\sum_{t=2}^{T}\varepsilon_t\left(\frac{\delta\varepsilon_t}{\delta\theta}\right) = \left(\frac{1}{\sigma^2}\right)\sum_{t=2}^{T}\varepsilon_t[(1-\theta L)^{-2}y_{t-1}]$$

which is easily expressed as,

$$\frac{\delta \ln(L)}{\delta\theta_i} = \left(\frac{1}{\sigma^2}\right)\sum_{t=2}^{T}\varepsilon_t v_{t-1},$$

where $v_t = (1-\theta L)^{-2}y_t = (1-\theta L)^{-2}(1-\theta L)\varepsilon_t = (1-\theta L)^{-1}\varepsilon_t$. Hence v_t is an AR(1) process of the form,

$$(1-\theta L)v_t = \varepsilon_t.$$

$$\frac{\delta \ln(L)}{\delta\sigma^2} = -\left(\frac{T}{2\sigma^2}\right) + \left(\frac{1}{2\sigma^4}\right)\sum[\phi(L)y_t]^2$$

The MLE then requires solution of the above two score vectors. The information matrix is found from,

$$\frac{\delta^2 \ln(L)}{\delta\theta^2} = \left(\frac{1}{\sigma^2}\right)\sum\left[\left(\frac{\delta\varepsilon_t}{\delta\theta}\right)v_{t-1} + \varepsilon_t\left(\frac{\delta v_{t-1}}{\delta\theta}\right)\right]$$

$$\frac{\delta^2 \ln(L)}{\delta\theta^2} = \left(\frac{1}{\sigma^2}\right)\sum\left[(-v_{t-1})v_{t-1} + \varepsilon_t\{-(1-\theta L)^{-2}\varepsilon_{t-1}\right]$$

$$\frac{\delta^2 \ln(L)}{\delta\theta^2} = -\left(\frac{1}{\sigma^2}\right)\sum v_{t-1}^2 - \left(\frac{1}{\sigma^2}\right)\sum\varepsilon_t(1+2\theta L+3\theta^2 L^2+....)\varepsilon_{t-1}$$

$$-E\left(\frac{\delta^2 \ln(L)}{\delta\theta^2}\right) = -\left(\frac{1}{\sigma^2}\right)E(v_{t-1}^2) - \left(\frac{1}{\sigma^2}\right)\sum E[\varepsilon_t(\varepsilon_{t-1}+2\theta\varepsilon_{t-2}+3\theta^2\varepsilon_{t-3}+.....)]$$

Since v_t is an AR(1) process it follows that $E(v_t^2) = \sigma^2(1-\theta^2)^{-1}$. Therefore,

$$-E\left(\frac{\delta^2 \ln(L)}{\delta\theta^2}\right) = -\left(\frac{1}{\sigma^2}\right)\frac{\sigma^2}{(1-\theta^2)} = \frac{1}{(1-\theta^2)}$$

$$T^{1/2}\begin{bmatrix}\hat{\theta}-\theta \\ \hat{\sigma}^2-\sigma^2\end{bmatrix} \rightarrow N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}(1-\theta^2) & 0 \\ 0 & 2\sigma^4\end{bmatrix}\right)$$

6) Properties of MLE for ARMA(p, q) Processes

The MLE of the parameters of ARMA processes can be estimated by either exact or approximate MLE and will satisfy the standard properties discussed earlier. As shown previously, the information matrix is a particularly simple form for the pure AR(p) process and is just the inverse of the autocovariance matrix, Γ_{p-1}. There is an interesting duality between the information matrix of the AR(p) process and that of the MA(q) process; see

Pierce (1975). For the ARMA(p,q) process,

$$\phi(L)y_t = \theta(L)\varepsilon_t,$$

$$\ln(L) = -\left(\frac{T}{2}\right)\ln(2\pi) - \left(\frac{T}{2}\right)\ln(\sigma^2) - \left(\frac{1}{2\sigma^2}\right)\sum_{t=p+q+1}^{T}\varepsilon_t^2$$

$$\ln(L) = -\left(\frac{T}{2}\right)\ln(2\pi) - \left(\frac{T}{2}\right)\ln(\sigma^2) - \left(\frac{1}{2\sigma^2}\right)\sum_{t=p+q+1}^{T}[\phi(L)\theta(L)^{-1}y_t]^2$$

The score vector will have (p+q+1) elements and will be of the form,

$$\frac{\delta\ln(L)}{\delta\phi_i} = -\left(\frac{1}{\sigma^2}\right)\sum\varepsilon_t\left(-\theta(L)^{-1}y_{t-i}\right) = \left(\frac{1}{\sigma^2}\right)\sum\varepsilon_t u_{t-i}\,; \quad \text{for i = 1,2,...p}$$

where, $\phi(L)u_t = \varepsilon_t$.

$$\frac{\delta\ln(L)}{\delta\theta_i} = -\left(\frac{1}{\sigma^2}\right)\sum\varepsilon_t\left(\phi(L)\theta(L)^{-2}y_{t-i}\right) = \left(\frac{1}{\sigma^2}\right)\sum\varepsilon_t v_{t-i}\,; \quad \text{i = 1,,2,...q}$$

where$\theta(L)v_t = -\varepsilon_t$.

Also,

$$\frac{\delta\ln(L)}{\delta\sigma^2} = -\left(\frac{T}{2\sigma^2}\right) + \left(\frac{1}{2\sigma^4}\right)\sum\varepsilon_t^2 .$$

Which completes the (p+q+1) elements of the score vector, and are used to solve the first order conditions $\delta\ln(L) = 0$. However, there will be no closed form solution for the AR and MA parameter estimates and numerical methods will have to be used iteratively to obtain the MLEs. To find the information matrix it is necessary to consider the second derivatives,

$$\frac{\delta^2 \ln(L)}{\delta\phi_i\delta\phi_j} = \left(\frac{1}{\sigma^2}\right)\sum u_{t-i}\left(\frac{\delta\varepsilon_t}{\delta\phi_j}\right) + \left(\frac{1}{\sigma^2}\right)\sum \varepsilon_t\left(\frac{\delta u_{t-i}}{\delta\phi_j}\right) = \left(\frac{1}{\sigma^2}\right)\sum u_{t-i}u_{t-j} + \left(\frac{1}{\sigma^2}\right)\sum \varepsilon_t u_{t-i-j}$$

Note that $E(\varepsilon_t u_{t-i-j}) = E[\varepsilon_t \phi(L)^{-1}\varepsilon_{t-i-j}] = 0,$ which implies that

$$p\lim_{T\to\infty}\left(-E\frac{\delta^2 \log L}{\delta\phi_i\delta\phi_j}\right) = -\left(\frac{1}{\sigma^2}\right)\sum u_{t-i}u_{t-j} = \gamma^u_{i-j},$$

which is the autocovariance of the process u_t, where $\phi(L)u_t = \varepsilon_t$ and hence the upper north west corner of the information matrix will be a square (p-1) dimensional autocovariance matrix of the u_t process and is the same structure as for the pure AR(p) process, only for u_t, rather than y_t. The other block matrices of the information matrix are constructed in a similar manner; in particular

$$\frac{\delta^2 \ln(L)}{\delta\theta_i\delta\theta_j} = \left(\frac{1}{\sigma^2}\right)\sum\left(\frac{\delta\varepsilon_t}{\delta\theta_j}\right)v_{t-i} + \left(\frac{1}{\sigma^2}\right)\sum \varepsilon_t\left(\frac{\delta v_{t-i}}{\delta\theta_j}\right)$$

$$\frac{\delta^2 \ln(L)}{\delta\theta_i\delta\theta_j} = \left(\frac{1}{\sigma^2}\right)\sum v_{t-j}v_{t-i} + \left(\frac{1}{\sigma^2}\right)\sum \varepsilon_t v_{t-i-j}$$

Note that $E(\varepsilon_t v_{t-i-j}) = E[\varepsilon_t \theta(L)^{-1}\varepsilon_{t-i-j}] = 0,$ which implies that

$$p\lim_{T\to\infty}\left(-E\frac{\delta^2\log L}{\delta\theta_i\delta\theta_j}\right)=-\left(\frac{1}{\sigma^2}\right)\sum v_{t-i}v_{t-j}=\gamma^{v}_{i-j}$$

Also,

$$\frac{\delta^2\ln(L)}{\delta\phi_i\delta\theta_j}=\left(\frac{1}{\sigma^2}\right)\sum\left(\frac{\delta\varepsilon_t}{\delta\theta_j}\right)u_{t-i}+\left(\frac{1}{\sigma^2}\right)\sum\varepsilon_t\left(\frac{\delta u_{t-i}}{\delta\theta_j}\right)=\left(\frac{1}{\sigma^2}\right)\sum v_{t-j}u_{t-i}+\left(\frac{1}{\sigma^2}\right)\sum\varepsilon_t\left(\frac{\delta u_{t-i}}{\delta\theta_j}\right)$$

$$p\lim_{T\to\infty}\left(-E\frac{\delta^2\ln L}{\delta\varphi_i\delta\theta_j}\right)=\left(\frac{1}{\sigma^2}\right)\sum E(v_{t-j}u_{t-i})=\left(\frac{1}{\sigma^2}\right)\sum v_{t-j}u_{t-i}=\gamma^{vu}_{i-j};$$

for i = 1,2,…p and j = 1,2,…q.

Also, $\frac{\delta^2\ln(L)}{\delta\sigma^4}=-\frac{1}{\sigma^6}\sum\varepsilon_t^2+\frac{T}{2\sigma^4}$ and $-E\left(\frac{\delta^2\ln(L)}{\delta\sigma^4}\right)=\frac{T}{2\sigma^4}$, as usual in the linear type models. Finally,

$$E\left(-\frac{\delta^2\ln(L)}{\delta\phi_i\delta\sigma^2}\right)=-\left(\frac{1}{\sigma^4}\right)\sum\varepsilon_t u_{t-i}=0$$

and

$$E\left(-\frac{\delta^2\ln L}{\delta\theta_i\delta\sigma^2}\right)=-\left(\frac{1}{\sigma^4}\right)\sum E(\varepsilon_t v_{t-j})=0$$

On defining (p+q+1) dimensional vector of parameters as

$$\Theta^{/}=(\phi_1,\phi_2,....\phi_p,\theta_1,\theta_2,....,\theta_q,\sigma^2),$$

then

$$T^{1/2}(\hat{\Theta}-\Theta) \to N(0,V),$$

where,

$$V = \begin{bmatrix} \begin{bmatrix} C & D \\ D^{/} & F \end{bmatrix}^{-1} & 0 \\ 0 & 2\sigma^4 \end{bmatrix}$$

where, $C = (\gamma^{u}_{i-j})$, $F = (\gamma^{v}_{i-j})$ and $D = (\gamma^{uv}_{i-j})$.

Clearly the AR(p) process which was previously examined is clearly a special case of the above with $\theta(L) = 1$, and $C = \Gamma_{p-1}$.

7) MLE of Regression Models with ARMA Disturbances

Economic theory frequently provides information on the inclusion of particular exogenous variables in a regression equation but fails to give much insight into the dynamic specification of the disturbances. In such situations it is generally reasonable to model the disturbances by an ARMA(p, q) process. The form of the model is then

$$y_t = \sum_{j=1}^{k} \beta_j x_{jt} + u_t,$$

$$\phi(L)u_t = \theta(L)\varepsilon_t$$

where $\phi(L) = (1-\phi_1 L - .... - \phi_p L^p)$, and $\theta(L) = (1-\theta_1 L - .... - \theta_q L^q)$. It is necessary for estimation to assume that all the roots of $\phi(L)$ and θ(L) lie outside the unit circle and the exogenous variables satisfy the standard regularity conditions. On assuming $\varepsilon_t \sim$ NID$(0,\sigma^2)$, then the conditional MLE will be asymptotically equivalent to the minimization of $\sum \varepsilon_t^2$, which can be achieved by standard non linear methods. As before the (k+p+q+1) vector of parameters are now,

$$\Theta^{/} = (\beta_{1,}....\beta_k, \phi_1, \phi_2, ...\phi_p, \theta_1, \theta_2, ..., \theta_q, \sigma^2),$$

then

$$T^{1/2}(\hat{\Theta} - \Theta) \rightarrow N(0, V),$$

where it is straightforward to show that

$$V = \begin{bmatrix} B^{-1} & 0 & \\ 0 & \begin{bmatrix} C & D \\ D^{/} & F \end{bmatrix}^{-1} & 0 \\ & 0 & 2\sigma^4 \end{bmatrix}$$

where

$$B = (\lim T^{-1} \sum z_{it} z_{jt}), \text{ for } i, j = 1, 2, ..., k \text{ and } z_{it} = \phi(L)^{-1} \theta(L) x_{it}.$$

Note that the z_{it} are just the GLS transformation of the original x_{it} variables in the model. The asymptotic covariance matrix of the regression parameters, $(\sigma^2 / T)B^{-1}$ is an equivalent expression to the GLS notation result of $\sigma^2(p \lim_{T\rightarrow\infty} T^{-1}(X'\Omega X)^{-1}$. A necessary condition is that the GLS transformed explanatory variables, z_{it} have a bounded second moment matrix. The restriction that the explanatory variables are weakly stationary is a sufficient, but not necessary condition for this to occur.

The other elements of the asymptotic covariance matrix are identical to the information matrix for the pure ARMA(p, q) process. One interesting feature is that the

MLE of regression parameters, the ARMA disturbance parameters and the disturbance variance are all asymptotically mutually uncorrelated.

8) Two Step Estimators

Sometimes the iterated MLE can be fully efficient after just one iteration, such a situation is known as a Two Step Estimator. The approach can be illustrated by the linear regression model with AR disturbances.

$$Y = X\beta + U ,$$

where E(U) = 0 and $E(UU') = \sigma^2\Omega$, Ω is a positive definite covariance matrix, Y is a T dimensional vector of observations on the dependent variable y_t, X is Txk matrix of observations on k explanatory variables and β is a vector of unknown parameters. For the β parameters it is straightforward to use the observation form of the likelihood,

$$\log L = -\left(\frac{T}{2}\right)\log(2\pi) - \left(\frac{T}{2}\right)\log(\sigma^2) - \left(\frac{1}{2\sigma^2}\right)(Y - X\beta)'\Omega^{-1}(Y - X\beta)$$

As an example consider the regression with AR(1) disturbance,

$$y_t = x_t'\beta + u_t ,$$
$$u_t = \phi u_{t-1} + \varepsilon_t ,$$

$E(\varepsilon_t) = 0, E(\varepsilon_t)^2 = \sigma^2$ and $E(\varepsilon_t\varepsilon_s) = 0$ for s = t, and x_t' is a k dimensional row vector of observations on the k explanatory variables at time t. Then the likelihood is,

$$\log L = -\left(\frac{T}{2}\right)\log(2\pi) - \left(\frac{T}{2}\right)\ln(\sigma^2) - \left(\frac{1}{2\sigma^2}\right)\sum_{t=2}^{T}[(1-\phi L)(y_t - x_t'\beta)]^2$$

so that

$$\frac{\delta \log L}{\delta \phi} = \left(\frac{1}{\sigma^2}\right) \sum [(1-\phi L)(y_t - x_t'\beta)](y_{t-1} - x_{t-1}'\beta)$$

$$\frac{\delta \log L}{\delta \beta} = \left(\frac{1}{\sigma^2}\right) \sum \varepsilon_t u_{t-1}$$

Using the fact the information matrix is block diagonal so that the estimate of σ^2 can be separated from the estimates of β and ϕ ; then the score vector can be expressed as

$$s(\Theta) = \begin{bmatrix} \dfrac{\delta \log L}{\delta \beta} \\ \dfrac{\delta \log L}{\delta \phi} \\ \dfrac{\delta \log L}{\delta \sigma^2} \end{bmatrix} = \begin{bmatrix} \left(\dfrac{1}{\sigma^2}\right) X'\Omega^{-1}(Y - X\beta) \\ \left(\dfrac{1}{\sigma^2}\right) \sum \varepsilon_t u_{t-1} \\ \left(\dfrac{1}{2\sigma^4}\right)(Y - X\beta)'\Omega^{-1}(Y - X\beta) \end{bmatrix}$$

$$I(\Theta) = \begin{bmatrix} \left(\dfrac{1}{\sigma^2}\right)(X'\Omega^{-1}X) & 0 & 0 \\ 0 & \dfrac{1}{1-\phi^2} & 0 \\ 0 & 0 & \dfrac{T}{2\sigma^4} \end{bmatrix}$$

$$I(\Theta)^{-1} = \begin{bmatrix} \sigma^2(X'\Omega^{-1}X)^{-1} & 0 & 0 \\ 0 & \dfrac{1-\phi^2}{T} & 0 \\ 0 & 0 & \dfrac{2\sigma^4}{T} \end{bmatrix}$$

$$V = \begin{bmatrix} \sigma^2 B^{-1} & 0 & 0 \\ 0 & 1-\phi^2 & 0 \\ 0 & 0 & 2\sigma^4 \end{bmatrix}$$

The initial estimate is

$$\overset{\wedge}{\beta}_{ols} = (X'X)^{-1} X'Y$$

and $\phi = 0$, so that OLS is applied assuming that the regression disturbances are serially uncorrelated. Because of the block diagonality of the information matrix, the estimation of the regression parameters, β, the autocorrelation parameter ϕ, and the disturbance variance σ^2, can all be treated independently. Once again the score vector and the Information matrix are both evaluated under the initial estimates of OLS estimates of β with ϕ = 0.

$$\overset{\wedge}{\beta}^{(k)} = \overset{\wedge}{\beta}^{(k-1)} + \sigma^2 (X'\Omega^{-1}X)^{-1} \left(\frac{1}{\sigma^2}\right)(X'\Omega^{-1})(Y - X\,\hat{\beta}^{(k-1)})$$

$$\hat{\beta}^{(k)} = \hat{\beta}_{ols} + (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1})(Y - X\hat{\beta}_{ols})$$

$$\hat{\beta}^{(k)} = \hat{\beta}_{ols} + (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}Y) - (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}X)\hat{\beta}_{ols}$$

$$\hat{\beta}^{(k)} = \hat{\beta}_{ols} + (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}Y) - \hat{\beta}_{ols}$$

$$\hat{\beta}^{(k)} = +(X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}Y) = \hat{\beta}_{gls}$$

$$\hat{\phi}^{(k)} = 0 + \left[\frac{1}{\sum \hat{u}_t^2}\right] \sum \hat{u}_t \hat{u}_{t-1} = r_1$$

which is the first order sample autocorrelation coefficient of the OLS residuals. This simple procedure is the Cochrane Orcutt estimator.

9) Treatment of the Initial Observations and Prediction Error Decomposition

A major issue in the derivation of MLE for dynamic and time series models in general concerns the treatment of the initial conditions. In order to obtain the full MLE, it is necessary to assume that the starting values have the same data generating process as the process itself. In the following the full likelihood function is given by,

$$L = f(y_{T,}y_{T-1,}..., y_1;\ \Theta)$$

Clearly, following the derivation of the MLE in the classical linear regression model, one of the simplest ways of expressing the likelihood is in terms of matrix notation. If the vector of observations is denoted by Y, i.e. $Y' = (y_{T,}y_{T-1},..., y_1)$, is Normally distributed, then

$$Y \sim N(\mu,\ \sigma^2\Omega),$$

then the likelihood is given by,

$$L = (2\pi)^{-\frac{T}{2}}(\sigma^2)^{-\frac{T}{2}}|\Omega|^{-\frac{1}{2}}\exp[-\left(\frac{1}{2\sigma^2}\right)(Y-\mu)'\Omega^{-1}(Y-\mu)].$$

When dealing with dynamic models, an equally convenient representation is to use the *prediction error decomposition*, which allows dependent observations to be readily handled. Consider just two observations, y_2 and y_1. Then on using the laws of conditional probability, the likelihood can be represented as,

$$f(y_2 y_1) = f(y_2|y_1)f(y_1),$$

and on generalizing to three observations;

$$f(y_3 y_2 y_1) = f(y_3 | y_2 y_1)\, f(y_2 | y_1),$$

$$f(y_3 y_2 y_1) = f(y_3 | y_2 y_1)\, f(y_2 | y_1)\, f(y_1),$$

Suppose the first p values of Y depend on unobservable quantities. Denote these observations as,

$$Y_p' = (y_p, y_{p-1}, ..., y_1).$$

On continuing with the above process of recursively taking conditional densities, the complete likelihood function can be expressed as,

$$L = f(y_t y_{t-1,} ..., y_1; \Theta) = f(y_1 y_{2,} ..., y_p;\ \Theta) \prod_{\mathrm{t=p+1}}^{\mathrm{T}} \mathrm{f(y_t} | y_{t-1})$$

10) Full MLE for the AR(1) Model

The technique is quite powerful and is best illustrated with some examples. First consider the AR(1) process with a mean μ,

$$y_t = \mu + \phi y_{t-1} + \varepsilon_t,$$

where $\varepsilon_t \sim NID(0, \sigma^2)$. Since, $E(y_t) = \dfrac{\mu}{1-\phi}$ and $Var(y_t) = \dfrac{\sigma^2}{1-\phi^2}$; then the density of the first observation is given by

$$f(y_1; \Theta) = (2\pi)^{-1/2} \left(\frac{\sigma^2}{1-\phi^2} \right)^{-1/2} \exp\left(-\frac{[y_1 - (\mu/(1-\phi))]^2}{2\sigma^2/(1-\phi^2)} \right).$$

where $\Theta^{/} = (\mu, \phi, \sigma^2)$. The distribution of the other observations are

$$f(y_{T,} y_{T-1,} \ldots, y_1;\ \Theta) = \mathrm{f}(\mathrm{y}_1)\prod_{t=2}^{T} f(y_t \left| y_{t-1}\right.)$$

$$f\left(y_t \left| y_{t-1}\right.\right) \sim N(\mu + \phi y_{t-1,} \sigma^2)$$

$$f\left(y_t \left| y_{t-1}; \Theta\right.\right) = (2\pi)^{-1/2}(\sigma^2)^{-1/2} \exp[-\left(\frac{1}{2\sigma^2}\right)(y_t - \mu - \phi y_{t-1})^2$$

and

$$f\left(y_t \left| y_{t-1}, y_{t-2}, \ldots, y_1; \Theta\right.\right) = (2\pi)^{-1/2}(\sigma^2)^{-1/2} \exp[-\left(\frac{1}{2\sigma^2}\right)(y_t - \mu - \phi y_{t-1})^2 .$$

$$\begin{aligned} \ln(L) &= \ln[f(y_1;\Theta)] + \sum_{t=2}^{T} \ln(y_t \left| y_{t-1}\right.), \\ &= -(1/2)\log(2\pi) - (1/2)\ln\left(\frac{\sigma^2}{1-\phi^2}\right) - \left(\frac{2\sigma^2}{1-\phi^2}\right)^{-1}\left(y_1 - \frac{\mu}{1-\phi}\right)^2 \\ &\quad -\left(\frac{T-1}{2}\right)\ln(2\pi) - \left(\frac{T-1}{2}\right)\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=2}^{T}\left(y_t - \mu - \phi y_{t-1}\right)^2 \end{aligned}$$

The first line of the above represents the contribution of the first observation to the log likelihood, while the final three terms are represent the contribution of the last (T-1) observations to the log likelihood.

An alternative method for determining the effects of the initial observations, which is to write the likelihood in a form corresponding to the conventional GLS estimator, for the AR(1) process. Then the autocovariance matrix, Ω is

$$\Omega = \frac{\sigma^2}{1-\phi^2}\begin{bmatrix} 1 & \phi & \phi^2 & . & . & . & \phi^{T-1} \\ \phi & 1 & \phi & & & & \phi^{T-2} \\ \phi^2 & \phi & 1 & \phi & & & \phi^{T-3} \\ . & & & . & & & . \\ . & & & & . & & . \\ . & & & & & . & . \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & . & . & . & 1 \end{bmatrix}$$

Then $\Omega = \sigma^2 (M'M)^{-1}$, and hence $\Omega^{-1} = \sigma^{-2} M'M$, where M is a T dimensional square matrix and for the AR(1) process is defined as

$$M = \begin{bmatrix} (1-\phi^2)^{\frac{1}{2}} & 0 & 0 & 0 & . & . & . & 0 \\ -\phi & 1 & 0 & 0 & . & . & . & 0 \\ 0 & -\phi & 1 & 0 & . & . & . & 0 \\ 0 & 0 & -\phi & 1 & & & & 0 \\ . & & & & . & & & . \\ . & & & & & . & & . \\ . & & & & & & . & . \\ 0 & 0 & 0 & 0 & . & . & -\phi & 1 \end{bmatrix}$$

The log likelihood is then given by,

$$\ln(L) = -\left(\frac{T}{2}\right)\ln(2\pi) + \left(\frac{1}{2}\right)\ln\left|\sigma^{-2}M'M\right| - \left(\frac{1}{2}\right)(Y-\mu)'\sigma^{-2}(M'M)(Y-\mu)$$

The transformed Y vector is $Y^* = MY$,

$$Y^* = \begin{bmatrix} (1-\phi^2)^{1/2}\left(y_1 - \mu/(1-\phi^2)\right) \\ y_2 - \phi y_1 - \mu \\ . \\ y_T - \phi y_{T-1} - \mu \end{bmatrix}$$

Then,

$$\left(\frac{1}{2}\right)(Y-\mu)'\sigma^{-2}M'M(Y-\mu)=(2\sigma^{2})^{-1}Y^{*\prime}Y^{*}$$

$$\ln(L)=\left(\frac{1}{2\sigma^{2}}\right)(1-\phi^{2})\left(y_t-\frac{\mu}{1-\phi}\right)^{2}+\left(\frac{1}{2\sigma^{2}}\right)\sum_{t=2}^{T}(y_t-c-\phi y_{t-1})^{2}$$

Also the middle term in the likelihood is,

$$\left(\frac{1}{2}\right)\ln(\sigma^{2}M'M)=\left(\frac{1}{2}\right)\ln(\sigma^{-2T})\left|M'M\right|$$

$$=-\left(\frac{T}{2}\right)\ln(\sigma^{2})+\left(\frac{1}{2}\right)\ln\left|M'M\right|$$

$$=-\left(\frac{T}{2}\right)\ln(\sigma^{2})+\left(\frac{1}{2}\right)\ln\left(\left|M'\right|\left|M\right|\right)$$

$$=-\left(\frac{T}{2}\right)\ln(\sigma^{2})+\left(\frac{1}{2}\right)\ln\left(\left|M\right|\left|M\right|\right)$$

$$=-\left(\frac{T}{2}\right)\ln(\sigma^{2})+\ln\left|M\right|$$

and since M is lower triangular, $\left|M\right|=(1-\phi^{2}),$ it follows that

$$\left(\frac{1}{2}\right)\ln(\sigma^{-2}M'M)=-\left(\frac{1}{2}\right)\ln(\sigma^{2})+\left(\frac{1}{2}\right)\ln(1-\phi^{2})$$

It is clear that the same likelihood is obtained as from the prediction error decomposition formula.

11) Full MLE for the AR(p) Process

The above prediction error decomposition, with conditioning used on previous observations can also be implemented for the AR(p) process with mean of μ,

$$\phi(L)y_t = \mu + \varepsilon_t$$

and since, $E(y_t) = \mu / \left(1 - \sum_{j=1}^{p} \phi_j\right)$. The joint density of the first p observations is given by

$$f(y_{p,}y_{p-1},...,y_1;\Theta) = (2\pi)^{-p/2} \left|\sigma^2 \Gamma_p^{-1}\right|^{1/2} \exp[-\left(\frac{1}{2\sigma^2}\right)(Y_p - \mu)' \Gamma_p^{-1} (Y_p - \mu)],$$

where $Y_p' = [y_{p,}y_{p-1},...,y_1]$ is the vector of the first p observations and $Y_p \sim N(\mu,\ \Gamma_\text{p})$. The distribution of the other observations can be derived conditional on the first observation, since

$$f(y_t \left| y_{t-1},...,y_{t-p};\Theta\right.) = (2\pi)^{-1/2} \left(\sigma^2\right)^{1/2} \exp[-\left(\frac{1}{2\sigma^2}\right)(y_t - \mu - \sum_{j=1}^{p} \phi_j y_{t-j})^2]$$

and

$$f(y_t \left| y_{t-1}, y_{t-2},...,y_1;\Theta\right.) = (2\pi)^{-1/2} \left(\sigma^2\right)^{1/2} \exp[-\left(\frac{1}{2\sigma^2}\right)(y_t - \mu - \sum_{j=1}^{p} \phi_j y_{t-j})^2]$$

Then the full likelihood function for the complete sample is,

$$f(y_t, y_{t-1},...,y_1;\Theta) = f(y_p, y_{p-1},...,y_1) \prod_{t=p+1}^{T} f(y_t \left| y_{t-1}, y_{t-2},...,y_{t-p};\Theta\right.)$$

and

$$\ln(L) = \ln[f(y_T, y_{T-1}, ..., y_1; \Theta)]$$

$$= \ln[f(y_p, y_{p-1}, ..., y_1)] + \ln[\prod_{t=p+1}^{T} f(y_t \left| y_{t-1}, y_{t-2}, ..., y_{t-p}; \Theta)]$$

$$= -\left(\frac{p}{2}\right)\ln(2\pi) - \left(\frac{p}{2}\right)\ln(\sigma^2) + \left(\frac{1}{2}\right)\ln\left|\Gamma_p\right|^{-1} - \left(\frac{1}{2\sigma^2}\right)(Y_p - \mu)'\Gamma_p^{-1}(Y_p - \mu)$$

$$-\left(\frac{T-p}{2}\right)\ln(2\pi) - \left(\frac{T-p}{2}\right)\ln(\sigma^2) - \left(\frac{1}{2\sigma^2}\right)\sum_{t=p+1}^{T}(y_t - \mu - \sum_{j=1}^{p}\phi_j y_{t-j})^2$$

$$= -\left(\frac{T}{2}\right)\ln(2\pi) - \left(\frac{T}{2}\right)\ln(\sigma^2) + \left(\frac{1}{2}\right)\ln\left|\Gamma_p\right|^{-1} - \left(\frac{1}{2\sigma^2}\right)(Y_p - \mu)'\Gamma_p^{-1}(Y_p - \mu)$$

$$-\left(\frac{1}{2\sigma^2}\right)\sum_{t=p+1}^{T}(y_t - \mu - \sum_{j=1}^{p}\phi_j y_{t-j})^2$$

12) Full MLE for ARMA(p, q) Models

The same technique can be extended to obtain full MLEs of the ARMA(p, q) process. The treatment is based around the articles by Osborn (1974), Newbold (1974) and Galbraith and Galbraith (1974). In the following, the (p + q) dimensional vector of starting values is denoted by v, where

$$v' = [y_{1-p}, ..., y_0,\ \varepsilon_{1-q}, ..., \varepsilon_0]$$

It is assumed that

$$v \sim N(0,\ \sigma^2 D^{-1})$$

While the T disturbances $\varepsilon \sim N(0,\ \sigma^2 I)$ and the starting values and disturbances have the distribution,

$$\begin{bmatrix} v \\ \varepsilon \end{bmatrix} = \begin{bmatrix} 0 \\ A \end{bmatrix} y + \begin{bmatrix} I \\ H \end{bmatrix} \varepsilon = Jy + K\varepsilon$$

where A is TxT, H is $T\times$ (p+q) , 0 is (p+q)$\times$ T matrix of zeros and I is the identity matrix of order (p+q). The total vector of innovations is $\begin{bmatrix} v\ \varepsilon \end{bmatrix}' \approx N(0,\ \sigma^2 C^{-1})$, where

$$C = \begin{bmatrix} D & 0 \\ 0 & I \end{bmatrix}$$

The likelihood of the complete sample and starting values is then,

$$f(v\ \mathrm{y}_{\mathrm{T}},\ \mathrm{y}_{\mathrm{T\text{-}1}},...,\ \mathrm{y}_1;\Theta) = (2\pi\sigma^2)^{-\frac{T+p+q}{2}} |C|^{\frac{1}{2}} \exp[-\left(\frac{1}{2\sigma^2}\right)(Jy+K\varepsilon)'C(Jy+K\varepsilon)$$

where $v' = [y_0, y_{-1},..., y_{-p+1}, \varepsilon_0,...,\varepsilon_{-q+1}]$. The above joint density and the sum of squares can be factorized to give,

$$f(v\ \mathrm{Y}) = f(\mathrm{v}|\mathrm{Y}) f(Y)$$

$$S(\Theta, v) = S(\Theta) + (v-v)'K'CK(v-v),$$

where

$$S(\Theta) = (JY + K\varepsilon)'C(JY + K\varepsilon)$$

and v is the OLS estimate of the vector of starting values and is given by,

$$v = -(K'CK)^{-1}(K'CJY)$$

Then,

$$f(v)=(2\pi\sigma^2)^{-\frac{p+q}{2}}\left|K'CK\right|^{-\frac{1}{2}}\exp[-\left(\frac{1}{2\sigma^2}\right)(v-v)'K'CK(v-v)]$$

and

$$f(v)=(2\pi\sigma^2)^{-\frac{T}{2}}\left|C\right|^{\frac{1}{2}}\left|K'CK\right|^{-\frac{1}{2}}\exp[-\left(\frac{1}{2\sigma^2}\right)S(\Theta)]$$

The MLE of σ^2 is obtained from differentiating the above with respect to σ^2 and equating with zero to obtain,

$$\overset{\wedge}{\sigma}^2=\frac{S(\Theta)}{T}$$

The corresponding concentrated log likelihood becomes.

$$f(Y;\Theta)=-\left(\frac{T}{2}\right)\ln(2\pi)-\left(\frac{T}{2}\right)\ln\left(\frac{S(\Theta)}{T}\right)+\left(\frac{1}{2}\right)\ln(C)-\left(\frac{1}{2}\right)\ln\left|K'CK\right|-\left(\frac{T}{2}\right)$$

Hence maximization of the MLE is equivalent to the minimization of the expression,

$$T\ln[S(\Theta)]+\ln\left|K'CK\right|-\ln\left|C\right|$$

so that computation of the full MLE for ARMA(p, q) models is quite straightforward. However, the method is often difficult to extend to more complicated models with exogenous variables, or forms of non linearity.

References

Osborn, Annals of Economic and Social Measuement, 1974;

Newbold, Biometrika, 1974

Galbraith and Galbraith, Journal of Applied Probability, 1974.