



## Two types of MLE Estimators

Type I maximizes the log-likelihood over  $\Omega$ .

If the loglikelihood function is differentiable and attains an interior maximum within  $\Omega$  then it must satisfy the f.o.c.

Type II MLE is defined to be the solution of the likelihood equations

$$g(y, \hat{\theta}) = 0$$

where  $g(\cdot)$  is the gradient vector (a score) which has typical element

$$g_i(y, \theta) \equiv \frac{\partial l(\theta|y)}{\partial \theta_i} = \sum_{t=1}^n \frac{\partial l_t(\theta|y)}{\partial \theta_i}$$

The maximum  $\hat{\theta}$  must be a local max and as  $T \rightarrow \infty$  the value of the loglikelihood associated with  $\hat{\theta}$  has to be bigger than the value associated with any other root of  $l(\theta|y)$ .

## Computation

In a few cases you can obtain analytical solutions. Usually, this is not the case. In this event, numerical solutions are used. and are in fact easy to get in most cases.

Recall in NCS we minimized the SSE using Gauss-Newton algorithm. Maximizing  $l(\theta|y)$  works the same way (Gauss-Newton) using the Hessian of  $l()$

$$H(\theta) = \frac{\partial^2 l(\theta|y)}{\partial \theta \partial \theta'}$$

note  $H(\theta) = \frac{\partial^2 g(\theta|y)}{\partial \theta \partial \theta'}$

i.e., the derivatives of the gradient

Let  $\theta_{(j)}$  denote the value of the  
vector at step  $j$  and  $g_{(j)}$  and  $H_{(j)}$

respectively, the gradient and Hessian  
evaluated at  $\theta_{(j)}$ . Then, Gauss Newton  
algorithm is

$$\theta_{(j+1)} = \theta_{(j)} - H_{(j)}^{-1} g_{(j)}$$

This requires starting value  $\theta_0$ .  $\Rightarrow$  different  
ones could lead to different values  
of  $\theta$ . Also,  $H_{(j)}$  has to be  
negative definite or the algorithm  
will head in the wrong direction.

Quasi-Newton methods tend to work  
better since  $H_j$  may not be  
neg def in parts of param. space.

Quasi-Newton method (For Max)

$$\theta_{(j+1)} = \theta_{(j)} + d_{(j)} D_{(j)}^{-1} g_{(j)}$$

where  $d_j$  (step-length) is determined  
at each step.

$D_j$  is a matrix that approximates  $-H_j$  near the maximum, but is positive definite by design

## Asymptotic Properties

Under regularity, MLE is

- (1) consistent
- (2) asymptotically normal
- (3) asymptotically efficient
- (4) invariant to reparameterizations of the model.

## Information Matrix

Define the  $T \times k$  matrix  $G(y, \theta)$  so that its typical element

$$G_{ti}(y, \theta) \equiv \frac{\partial^2 l_t(y_i, \theta)}{\partial \theta_i^2}$$

note  $\frac{\partial}{\partial \theta_i}$  contains  $\frac{\partial}{\partial \theta_i}$   $i^{\text{th}}$  column of sample.

$\partial \theta_i$

gradient =  
sum of the  
 $T$  columns

note: 
$$g_i(y, \theta) = \sum_{t=1}^T G_{ti}(y_i, \theta)$$

$i^{\text{th}}$  element of gradient or score is just the column sum of  $G$

The covariance matrix of  $G_t(y, \theta)$  (which is  $t^{\text{th}}$  row of  $G$ ) is a  $K \times K$  matrix

$I_t(\theta)$  of which the  $ij^{\text{th}}$  element is

$$E_{\theta} (G_{ti}(y, \theta) \cdot G_{tj}(y, \theta))$$

Note:  $E_{\theta}(G_{ti}(y, \theta)) = 0$  so this is a covariance matrix

$$I(\theta) \equiv \sum_{t=1}^T I_t(\theta) = \sum_{t=1}^T E_{\theta} (G_t(y, \theta) G_t^T(y, \theta))$$

$K \times K$     $t=1$     $K \times K$     $t=1$     $K \times 1$     $1 \times K$

which is called the information matrix.  $I_t$  is also

$$E_{\theta} (g(y, \theta) \cdot g^T(y, \theta))$$

Expectation of the outer product of the gradients.

The information matrix measures the curvature of  $l(\cdot)$ . The bigger it is, the more info we have on  $\theta$ .

Under general conditions

$$\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} N\left(0, \lim_{n \rightarrow \infty} \left(\frac{1}{n} I(\theta)\right)^{-1}\right)$$

So, any method that allows consistent estimation of

$$\lim_{n \rightarrow \infty} \left(\frac{1}{n} I(\theta)\right)^{-1} \text{ can be}$$

used to estimate cov of MLE.

3 candidates

$$(1) \quad \text{Var}_{\theta}(\tilde{\theta}) = -H^{-1}(\tilde{\theta})$$

empirical Hessian

$$(2) \quad \text{Var}_{\theta}(\tilde{\theta}) = I^{-1}(\tilde{\theta})$$

$$(3) \quad \text{Var}_{\theta}(\tilde{\theta}) = \left[ G^T(\tilde{\theta}) G(\tilde{\theta}) \right]^{-1} \\ = \sum_{t=1}^T G_t^T G_t$$

## Estimating Asymp. Covariance

Recall,

The Asymptotic <sup>covariance</sup> distribution of MLE

$$\left\{ -E\{H(\theta_T)\} \right\}^{-1} \text{Var}[\bar{g}(\theta_T)] \left\{ -E\{H(\theta_T)\} \right\}^{-1}$$

$$\text{and } \text{Var } g(\theta_T) = -E[H(\theta_T)]$$

Var of MLE is

$$\underline{-E[H(\theta_T)]}^{-1} \equiv I(\theta_T)^{-1}$$

We also showed that

$$-E[H(\theta_T)] = E[g(\theta_T) g(\theta_T)^T]$$

So, we have 2 equivalent expressions for Var of MLE

# Estimators of

$$-E[H(\theta_T)]^{-1}$$

(1)  $-H(\hat{\theta})^{-1}$  Newton-Raphson

(2)  $-E[H(\hat{\theta}_n)]^{-1}$  i.e., Take expectation of H  
 $= I(\hat{\theta})^{-1}$  and plug in  $\hat{\theta}$ .

(3)  $G(\hat{\theta})G(\hat{\theta})^T$   
 $k \times n$   $n \times k$ .

$$G(\hat{\theta}) = \begin{matrix} k \times n & \frac{\partial l_1(\theta)}{\partial \theta_1} & \frac{\partial l_2(\theta)}{\partial \theta_1} & \dots & \frac{\partial l_n(\theta)}{\partial \theta_1} \\ & \frac{\partial l_1(\theta)}{\partial \theta_2} & \frac{\partial l_2(\theta)}{\partial \theta_2} & & \frac{\partial l_n(\theta)}{\partial \theta_2} \\ & \vdots & \vdots & \ddots & \vdots \\ & \frac{\partial l_n(\theta)}{\partial \theta_k} & \dots & \dots & \frac{\partial l_n(\theta)}{\partial \theta_k} \end{matrix}$$

$l_n = \ln L(\theta | y_n)$  log likelihood  
evaluated at obs. n.

NOTE:  $g(\hat{\theta})$  is row sum of  $G(\hat{\theta})$ .  
 $k \times 1$