**CHAPTER THREE: MAXIMUM LIKELIHOOD ESTIMATION**

A method of point estimation with some stronger theoretical properties than the method of OLS is the method of maximum likelihood (ML). To use this method, however, one must make an assumption about the probability distribution of the disturbance term *ui*. In the regression context, the assumption most popularly made is that *ui* follows the normal distribution. Under the normality assumption, the ML and OLS estimators of the intercept and slope parameters of the regression model are identical. However, the OLS and ML estimators of the variance of *ui* are different. In large samples, however, these two estimators converge. Thus, the ML method is generally called a *large-sample method.* It is of broader application in that it can also be applied to regression models that are nonlinear in the parameters. In this case, OLS is generally not used.

Assume a two-variable model:

 Yi = βo + β1Xi + Ui where the Yi are normally and independently distributed with

mean = βo + β1Xi and variance = σ2.

As a result, the joint probability density function of Y1, Y2, …, Yn given the preceding mean and variance, can be written as

 f (Y1, Y2, …, Yn/ βo + β1Xi , σ2)

But in view of the independence of the Y’s, this joint probability density function can be written as a product of n individual density functions as

f (Y1, Y2, …, Yn/ βo + β1Xi , σ2)

 = f(Y1/ βo + β1Xi , σ2) f(Y2/βo + β1Xi , σ2)… f(Yn/βo + β1Xi, σ2) …………………………….(1)

where,

f(Yi) = ………………(2)

 is the density function of a normally distributed variable with the given mean and variance.

Substituting (2) for each Yi into (1) gives

f(Y1, Y2, …, Yn/ βo + β1Xi, σ2) = …………….(3)

If Y1, Y2, …. Yn are known or given, but βo, β1 and σ2 are not known, the function in (3) is called a *likelihood function*, denoted by LF( βo, β1, σ2) , and written as

 LF (βo, β1, σ2) = ………………………...(4)

The method of maximum likelihood, as the name indicates, consists in estimating the unknown parameters in such a manner that the probability of observing the given Y’s is as high (or maximum) as possible. Therefore, we have to find the maximum of the function in (4).

 lnLF = - nlnσ - ln(2π) - ………….………(5)

 = lnσ2 - ln(2π) - …………….…(6)

Differentiating partially with respect to βo, β1, and σ2, setting the result equal to zero we obtain

 ∑(Yi – βo – β1Xi) (-1) = 0 …………………….……(7)

 ∑(Yi – βo– β1Xi) (-Xi) = 0 …………………….……(8)

 ∑(Yi – βo – β1Xi)2 = 0 …………….………..(9)

The above equations can be rewritten as (letting , and denote the ML estimators)

 ∑(Yi -  - Xi) = 0 ………………….…………….……..…(10)

 ∑(Yi -  - Xi)Xi = 0 ………………………………………(11)

 ∑(Yi -  - Xi)2 = 0 ……………………..……….(12)

After simplifying ∑Yi = n + ∑Xi ……………………………..…(13)

 ∑YiXi = ∑Xi + ∑Xi2…………………….……….(14)

Equations (13) and (14) are precisely the *normal equations* of the least squares theory. Therefore, the ML estimators are the same as the OLS estimators.

Moreover, substituting the ML ( = OLS) estimators into (12) and simplifying, we obtain the ML estimator of 

  = ∑(Yi - - Xi)2

 =∑(Yi -  - Xi)2

 = ∑

It is obvious that the ML estimator  differs from the OLS estimator = which is an unbiased estimator of σ2. Thus, the ML estimator of σ2 is biased. The magnitude of this bias can be easily determined as follows:

 E() = (∑) = = σ2 – 

which shows that is biased downward (i.e., it underestimates the true σ2) in small samples. But notice that as n, the sample size, increases indefinitely, the bias factor, tends to be zero. Therefore, asymptotically (i.e., in a very large sample),  is unbiased too because *lim* E() =σ2 as *n*→∞.