CLR model in matrix form

We begin with the basic model

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} \dots + \beta_k X_{kt} + U_t \leftarrow (1)$$

$t = 1, 2, \dots n$ (for the different data point

$n > k$ for estimation to be possible.

for $t = 1$, equation (1) gives —

$$Y_1 = \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \dots + \beta_k X_{k1} + U_1$$

for $t = 2$, equation (1) gives —

$$Y_2 = \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \dots + \beta_k X_{k2} + U_2$$

- - - - - - - - - - - -

for $t = n$, we have

$$Y_n = \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \dots + \beta_k X_{kn} + U_n$$

Using matrix algebra we can write the forms of (1) for 't' going from $1, 2, \dots n$ as —

$$\underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}_{y_{(n\times1)}} = \underbrace{\begin{bmatrix} \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \dots + \beta_k X_{k1} \\ \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \dots + \beta_k X_{k2} \\ - - - - - - - - \\ \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \dots + \beta_k X_{kn} \end{bmatrix}}_{(n\times1)} + \underbrace{\begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{bmatrix}}_{u_{(n\times1)}}$$

Please observe that here x_i values of different observations (data points) are multiplied by the same Parameters. Hence we can write this as —

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}_{(n\times 1)} = \underbrace{\begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ - & - & - & - & - \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix}}_{X_{(n\times k)}} \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix}}_{\beta_{(k\times 1)}} + \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_n \end{bmatrix}_{n\times 1}$$

Thus, in matrix form equation (1) becomes —

(here small y, β and u are used because in case of vector according to matrix rule we use small letters)

$$y = X\beta + u$$

y — $(n\times 1)$ vector of values of the dependent variable across n observation

X — $(n\times k)$ matrix of values of the k explanatory variables across n observations.

β — $(k\times 1)$ vector of the regression parameters

u — $(n\times 1)$ vector of random disturbance associated with the n different observation.

Please note that we ~~use~~ use small letters to indicate vectors. (eg y, β, u) and capital letters to represent matrices (eg X).

To sum up.

The linear regression model can be written as —

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_k X_{kt} + U_t \,, \quad \cancel{t = 1, 2, \dots k}$$

$t = 1, 2, \dots n > k$

or $$\underset{(n\times 1)}{y} = \underset{(n\times k)}{X}\,\underset{(k\times 1)}{\beta} + \underset{(n\times 1)}{u}$$

Lesson 6 classical linear regression model

Derivation of OLS estimators. Given the model

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_K X_{Kt} + U_t$$

$t = 1, 2, \dots n) > K$

OLS estimators of $\beta_1, \beta_2, \dots \beta_K$ are to be derived by minimizing the sum of squares of the residuals, S.

When $S = \Sigma(\hat{U}_t)^2 = \Sigma(Y_t - \hat{Y}_t)^2$

$$= \Sigma\{Y_t - (\hat{\beta}_1 + \hat{\beta}_2 X_{2t} + \hat{\beta}_3 X_{3t} + \dots + \hat{\beta}_K X_{Kt})\}^2$$

$$= f(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \dots \hat{\beta}_K)$$ given X_{it} and Y_t values.

~~Under~~ The first order conditions fore minimization of S. ~~(w.r.t. $\hat{\beta}_1, \hat{\beta}_2 \dots$) are $\frac{\delta S}{\delta \hat{\beta}_1} = 0$~~ (with respect to $\hat{\beta}_1, \hat{\beta}_2 \dots \hat{\beta}_K$) are —

$$\frac{\delta S}{\delta \hat{\beta}_1} = 0, \quad \frac{\delta S}{\delta \hat{\beta}_2} = 0 \quad \dots \dots \quad \frac{\delta S}{\delta \hat{\beta}_K} = 0.$$

These will give us k linear equations in $\hat{\beta}_1, \hat{\beta}_2, \dots \hat{\beta}_K$, which we shall have to solve to get the OLS estimators of $\beta_1, \beta_2, \dots \beta_K$.

Let us take the first of these above conditions.

$$\frac{\delta S}{\delta \hat{\beta}_1} = 0$$

$$\Rightarrow \frac{\delta}{\delta \hat{\beta}_1} \sum_{t=1}^{n} \{Y_t - (\hat{\beta}_1 + \hat{\beta}_2 X_{2t} + \dots + \hat{\beta}_K X_{Kt})\}^2 = 0$$

$$\Rightarrow \sum_{t=1}^{n} \frac{\delta}{\delta \hat{\beta}_1} \{Y_t - (\hat{\beta}_1 + \hat{\beta}_2 X_{2t} + \dots + \hat{\beta}_K X_{Kt})\}^2 = 0$$

$$\Rightarrow 2 \sum_{t=1}^{n} \{Y_t - (\hat{\beta}_1 + \hat{\beta}_2 X_{2t} + \dots + \hat{\beta}_K X_{Kt})\}(-1) = 0$$

Dividing both sides by 2 and (-1), we have —

$$\sum\{Y_t - (\hat\beta_1 + \hat\beta_2 X_{2t} + \dots + \hat\beta_k X_{kt})\} = 0$$

$$\Rightarrow \sum Y_t - \sum(\hat\beta_1 + \hat\beta_2 X_{2t} + \dots + \hat\beta_k X_{kt}) = 0$$

$$\Rightarrow \sum Y_t = \sum(\hat\beta_1 + \hat\beta_2 X_{2t} + \dots + \hat\beta_k X_{kt})$$

$$= \sum \hat\beta_1 + \hat\beta_2 \sum X_{2t} + \dots + \hat\beta_k \sum X_{kt}$$

$$= n\hat\beta_1 + \hat\beta_2 \sum X_{2t} + \dots + \hat\beta_k \sum X_{kt}$$ [$\because \sum\beta_1 = n\beta_1$ as β_1 is added n times]

If we follow the same procedure with $\frac{\partial S}{\partial \hat\beta_2} = 0$, we shall get

$$\sum_t Y_t X_{2t} = \hat\beta_1 \sum X_{2t} + \hat\beta_2 \sum X_{2t}^2 + \hat\beta_3 \sum X_{3t} X_{2t} + \dots + \hat\beta_k \sum X_{kt} X_{2t}$$

Similarly from $\frac{\partial S}{\partial \hat\beta_3} = 0$ we get

$$\sum Y_t X_{3t} = \hat\beta_1 \sum X_{3t} + \hat\beta_2 \sum X_{2t} X_{3t} + \hat\beta_3 \sum X_{3t}^2 + \dots + \hat\beta_k \sum X_{kt} X_{3t}$$

and so on and finally

$$\sum Y_t X_{kt} = \hat\beta_1 \sum X_{kt} + \hat\beta_2 \sum X_{2t} X_{kt} + \hat\beta_3 \sum X_{3t} X_{kt} + \dots + \hat\beta_k \sum X_{kt}^2$$

Now collecting all the k equations we have —

$$\begin{bmatrix} \sum Y_t \\ \sum Y_t X_{2t} \\ \sum Y_t X_{3t} \\ \vdots \\ \sum Y_t X_{kt} \end{bmatrix}_{(k \times 1)} = \begin{bmatrix} n\hat\beta_1 + \hat\beta_2 \sum X_{2t} + \hat\beta_3 \sum X_{3t} + \dots + \hat\beta_k \sum X_{kt} \\ \hat\beta_1 \sum X_{2t} + \hat\beta_2 \sum X_{2t}^2 + \hat\beta_3 \sum X_{3t} X_{2t} + \dots + \hat\beta_k \sum X_{kt} X_{2t} \\ \hat\beta_1 \sum X_{3t} + \hat\beta_2 \sum X_{2t} X_{3t} + \hat\beta_3 \sum X_{3t}^2 + \dots + \hat\beta_k \sum X_{kt} X_{3t} \\ \dots \\ \hat\beta_1 \sum X_{kt} + \hat\beta_2 \sum X_{2t} X_{kt} + \hat\beta_3 \sum X_{3t} X_{kt} + \dots + \hat\beta_k \sum X_{kt}^2 \end{bmatrix}_{(k \times 1)}$$

$X'y$

$$= \underbrace{\begin{bmatrix} n & \Sigma X_{2t} & \Sigma X_{3t} & \cdots & \Sigma X_{kt} \\ \Sigma X_{2t} & \Sigma X_{2t}^2 & \Sigma X_{3t}X_{2t} & \cdots & \Sigma X_{kt}X_{2t} \\ \Sigma X_{3t} & \Sigma X_{2t}X_{3t} & \Sigma X_{3t}^2 & \cdot & \Sigma X_{kt}X_{3t} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \Sigma X_{kt} & \Sigma X_{2t}X_{kt} & \Sigma X_{3t}X_{kt} & & \Sigma X_{kt}^2 \end{bmatrix}_{(k\times k)}}_{X'X} \underbrace{\begin{bmatrix} \hat\beta_1 \\ \hat\beta_2 \\ \hat\beta_3 \\ \vdots \\ \hat\beta_k \end{bmatrix}_{(k\times 1)}}_{\hat\beta}$$

or in short

$$\boxed{X'y = X'X\hat\beta}$$

Now if we pre multiply both sides by $(X'X)^{-1}$ we get

$$(X'X)^{-1}(X'y) = \underbrace{(X'X)^{-1}(X'X)}_{I_{k\times k}}\hat\beta$$

$$= \hat\beta$$

Thus OLS estimates of β is —

$$\boxed{\hat\beta = (X'X)^{-1}X'y.}$$

For existence of $\hat\beta$; $(X'X)^{-1}$ must exist i.e. $X'X$ must be non-singular i.e. $|X'X| \neq 0$. This requires that none of the columns of X should be expressible as linear combination of the other columns of X. Technically this is stated as —

$\rho X = k$

X has full column rank

Lesson 7 Properties of OLS estimatores

Before going to the above subject matter let us revisit the properties of a good estimatore. Basically we shall talk about unbiasedness, efficiency and consistency.

UNBIASEDNESS: Let θ be the parameter to be estimated and $\hat{h}$ is an estimatore of θ. Now $\hat{h}$ is technically an unbiased estimatore of θ if $E(\hat{h})=\theta$

What does this condition mean?

If we calculate $\hat{h}$ for all possible alternative samples, the mean of all these estimates should be equal to population parameter θ.

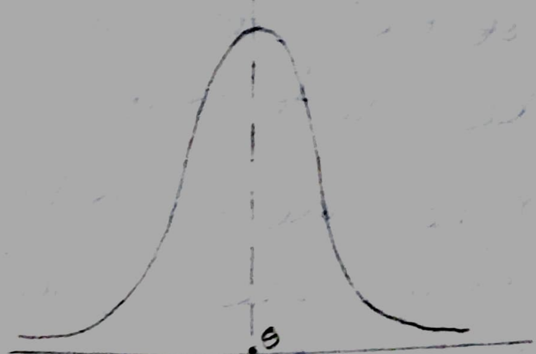
Since in practice, we work with only one

sample, what purpose does this feature serve?
If the estimatore is unbiased we have no priore reason to believe that the estimate from a given random sample is necessarily likely to be under or over estimated.

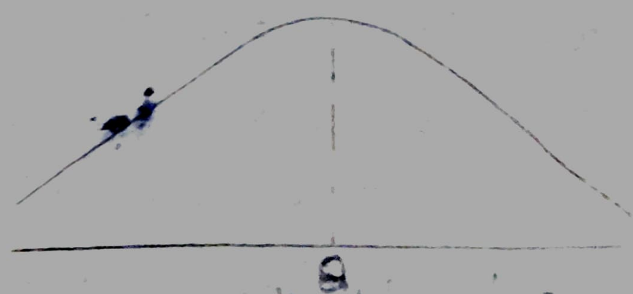
<u>Efficiency</u>: Among the unbiased estimators of a parameter, which ever has the smallest variance of its standard error, is the most efficient. Let 'h' and 'g' be both unbiased estimators of θ i.e.

$$E(h) = \theta \quad \text{and} \quad E(g) = \theta$$

Let the sampling distribution of h and g have the following pattern: [By the sampling distribution we denote the relative frequency distribution of the different possible values of an estimatore across all possible samples]



sampling distribution of h



sampling distribution of g

one can see that in case of h, the relative frequencies are mostly concentrated around θ. In most of the samples the estimatore 'h' will give estimates close to θ.

But in case of 'g', the relative frequencies are more widly distributed. 'g' will give estimates away from θ more often than 'h' will give.

If we use the efficient estimatore 'h' rather than inefficient estimatore 'g', we are more likely to get an estimate closer to θ.

∴ Though both 'g' and 'h' are unbiased (means of theire sampling distributions are equal to θ), h is more reliable.

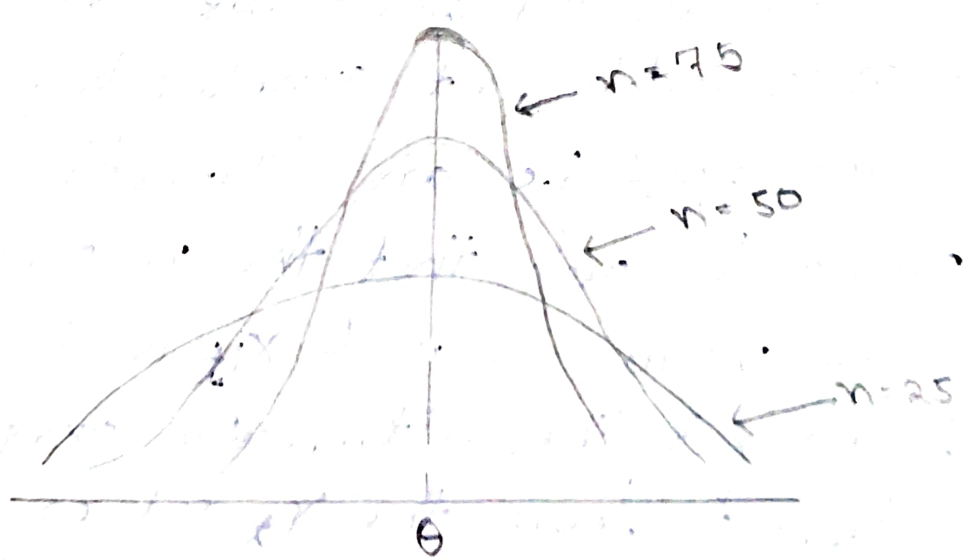
The large sample property of consistency:

- Both unbiasedness and efficiency are called small sample properties. These can be used fore both small and large samples. But consistency is a large sample property i.e. it is relevent only when sample size is large enough (30 ore more)

But consistency is a very usefull property. Because fore small samples, sometimes the pattern of the sampling distributions of estimators are not distinct. Hence we may not be able to verify unbiasedness and efficiency of the estimatore. But in most cases as the sample size is gradually increased, a visible sampling distribution do emerge. In such situations, we can verify its consistency. Consistency is therefore said to be an assymptotic property i.e. this is a property that becomes relevent as the sample size tends to infinity.

But when shall we say that 'h' is consistent?

when the sample size is increased, it becomes more and more representative. Therefore the spread of the values of the estimate from sample to sample gets reduced. Technically this amounts to signify that the concentration of the sampling distribution increases as sample size increases. Now the question is about which value the sampling distribution tends to concentrate. If the sampling distributions tend to concentrate around the population parameter we are estimating, then the estimatore is efficient.



Lesson 7 (Part II)

Having recalled the properties of a good estimator we are now poised to derive the properties of the