LECTURE 09

SIMPLE REGRESSION MODEL - I

Outline of today's lecture:

I. Simple Linear Regression Model

- In econometrics we deal exclusively with *stochastic* relations.
  - Stochastic is a term for *random* or *uncertain*.
  - It is the opposite of *deterministic*.
- The simplest form of stochastic relation between two variables X and Y is called a *simple linear regression* model.
  - $Y_t = \beta_0 + \beta_1 X_t + u_t \qquad t=1,..., n$

    where
    - Y is *dependent variable*
    - X is *independent* (explanatory) variable[1]
    - u is the stochastic *disturbance term, or error term.*
    - β_0 and β_1 are the regression *parameters* which are *unknown*.
    - Subscript t refers to the t^{th} observation.

[1] The terms *regressor, regressand* and *covariate* are also used.

- The values of the variables *X* and *Y* are observable.

- However, the values of *u* are not observable.

- Observations on *X* and *Y* can be made over time, in which case we speak of having "*time series*" data.

    - For example we may have data on Turkey's GDP over 30 years (data collected over a period of time).

- Or they can be made over

    – individuals, or groups of individuals,
    – firms, or group of firms,
    – countries, or group of countries,
    – objects,
    – geographical areas, etc.,

  in which we speak of having "*cross section data*".

    - For example we may have data on several countries' GDP for one year (data collected at one point in time).

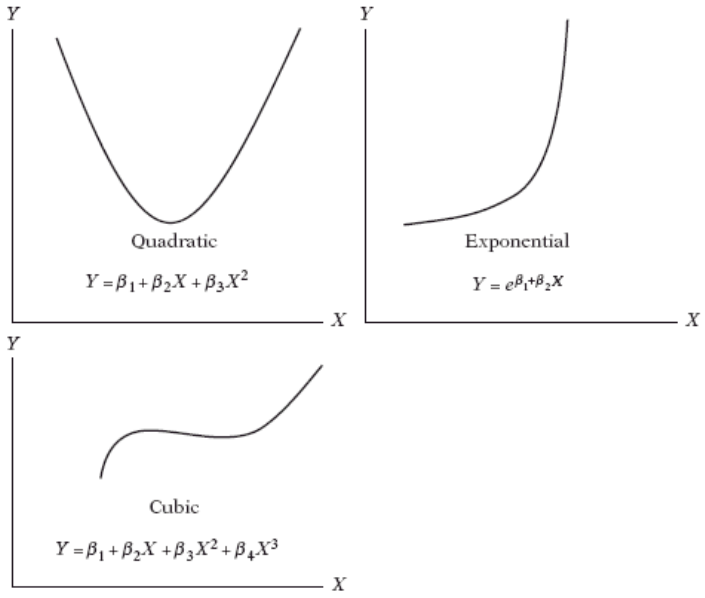
- Hence, the subscript *t* may refer to the t^{th} year (quarter, month, day, etc) or to the t^{th} individual or group (such as countries; Turkey, Germany, France, USA, Japan, etc..).

- Data of both kinds can be combined to obtain "*pooled times and cross-section data*".
    - For example we may have data on 20 countries' GDP over 30 years.

II. Linearity Issue

- $Y_t = \beta_0 + \beta_1{}^3 X_t + u_t$ is nonlinear in the parameter β_1.

- This model is an example of a nonlinear (in the parameter) regression model.



Linear-in-parameter functions.

LINEAR REGRESSION MODELS

Model linear in parameters?	Model linear in variables?	
	Yes	No
Yes	LRM	LRM
No	NLRM	NLRM

Note: LRM = linear regression model
NLRM = nonlinear regression model

III. Stochastic Nature of Linear Regression Model

- The stochastic nature of the regression model implies that for every value of X there is a whole probability distribution of values of Y.

- This means that the value of Y can *never* be forecast exactly.

- *Uncertainty* concerning Y arises because of the presence of the stochastic disturbance u which, being random, imparts randomness to Y.

    - For example, consider a production function of a firm.

- Suppose that output depends in some specified way on the quantity of labor (L) in accordance with firm's technology.

$$Output_t = \beta_0 + \beta_1 Labor_t + u_t$$

  - This function may apply in the short run when the quantities of other factors are fixed.

- In general, the same quantity of labor will lead to different quantities of output because of variations in *weather*, *human performance*, *frequency of machine breakdowns*, and many other factors.

- Output, which is the dependent variable in this case, will depend not only on the quantity of labor input, but also on the large number of random causes,

  - *which we summarize in the form of the stochastic disturbance* (*u*).

IV. Assumptions of Classical Linear Model

- **Assumption 1:** Linear regression model [the regression model is linear in the parameters]

- **Assumption 2:** *X* values are fixed in repeated sampling

  This means that in taking a large number of samples on *Y* and *X*, the X_t values are the same in all samples, but the u_t values differ from sample to sample, and so of course do the values of Y_t. For example assume that every day in market we choose the same prices $X_1, X_2, \ldots, X_t$, and we record the quantities of Y_t's sold each day at these prices. The *X*'s do not vary; they are a set of fixed values, while the Y_t's vary for each day due to different random influences.

- **Assumption 3:** Zero mean value of disturbance u_t $[E(u_t)=0]$

    - This condition is that the expected value of the disturbance term in any observation should be 0.

    - Sometimes it will be positive, sometimes negative, but it should not have a systematic tendency in either direction.

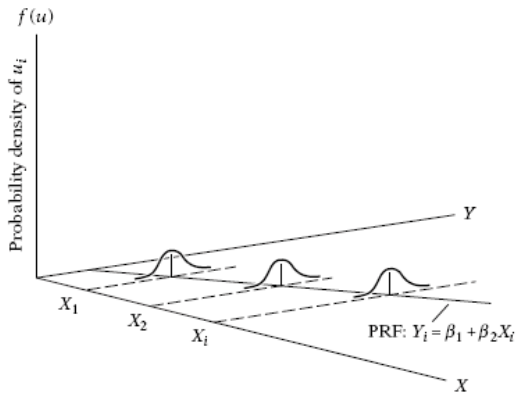
    - This assumption says that the factors not explicitly included in the model, and therefore subsumed in u_t, do not systematically affect the mean value of Y.

- **Assumption 4:** *Homoscedasticity* $[Var(u_t)=\sigma^2]$.

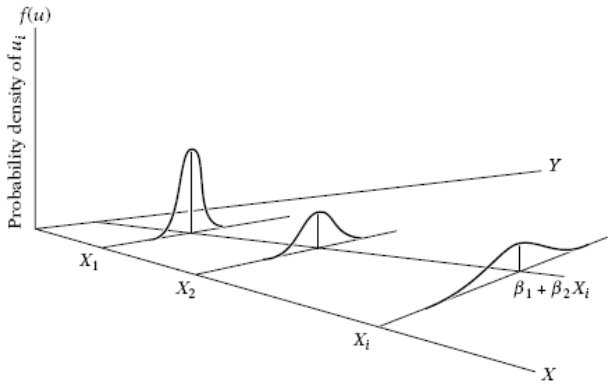
    - Population variance of the disturbance term (u_t) should be constant for all observations (Homoscedasticity).

    - If this condition is not satisfied, the OLS regression coefficients will be *inefficient*.

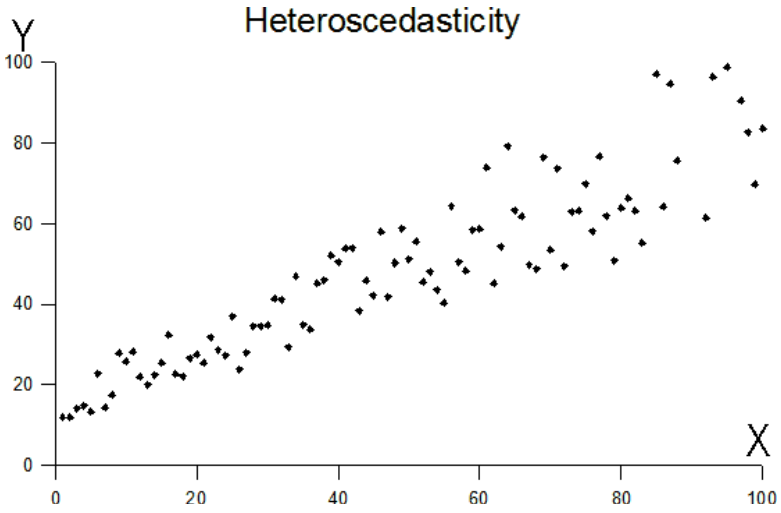
    - *In terms of our production example, this assumption implies that the variation in output is the same whether the quantity of labor is 20, 100, or any other number of units.*



Homoscedasticity.



Heteroscedasticity.



- **Assumption 5***:* No autocorrelation [correlation between any u_t and u_s $(t \neq s)$ is zero]

  - This condition states that there should be no systematic association between the values of the disturbance term in any two observations.
  - If this condition is not satisfied, OLS will again give inefficient estimates.
  - *In terms of our production example, this assumption implies that output is higher than expected today should not lead to higher (or lower) than expected output tomorrow.*
  - Recall that *correlation* between X and Y is given by:

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are standard deviations of X and Y, respectively. Therefore, if the correlation between X and Y is zero ($\rho = 0$), it implies that $Cov(X,Y) = 0$. As a result, the autocorrelation implies that $Cov(u_t, u_s) = 0$ where $t \neq s$.[2]

- Note that covariance between X and Y is given by

$$Cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}$$

Therefore, no autocorrelation implies that

$$Cov(u_t, u_s) = E\{[u_t - E(u_t)][u_s - E(u_s)]\} = 0$$

From *Assumption 3*, we know that $E(u_t) = 0$ and $E(u_s) = 0$. Thus, autocorrelation assumption implies that

$$Cov(u_t, u_s) = E\left\{\left[u_t - \underbrace{E(u_t)}_{0}\right]\left[u_s - \underbrace{E(u_s)}_{0}\right]\right\} = 0 \Rightarrow$$

$$E(u_t u_s) = 0$$

- **Assumption 6:** Zero covariance between u_t and X_t [$E(u_t X_t) = 0$]

- **Assumption 7:** The number of observations *n* must be greater than the number of parameters *k*.

- **Assumption 8:** Variability in X values [$Var(X) > 0$]

[2] Recall that a zero value of the covariance indicates <u>*no linear dependence*</u> between X and Y.

- **Assumption 9:** The regression model is correctly specified [no specification bias]

- **Assumption 10:** There is no perfect multicollinearity [there are no perfect linear relationships among explanatory variables]

V. Gauss Markov Theorem

Definition Given the assumptions of the CLRM[3], the OLS estimators, in the class of unbiased linear estimators, have minimum variance[4], that is they are *B*est[5] *L*inear *U*nbiased *E*stimators (*BLUE*).

VI. Normality Assumption

- In addition to the Gauss–Markov conditions, one usually assumes that the disturbance term is normally distributed.

- Reason is that if u is normally distributed, so will be the regression coefficients, and this will be useful to us later when we come to the business of performing tests of hypotheses and constructing confidence intervals for β_1 and β_2 using the regression results.

- Justification for the assumption depends on the Central Limit Theorem.

[3] See Handout 10 for the details of CLRM.
[4] In other words, they are efficient estimators.
[5] In other words, they are efficient or they have minimum variance.