Economics, Management and Entrepreneurship

Prof. Pratap K. J. Mohapatra
Department of Industrial Engineering & Management
Indian Institute of Technology – Kharagpur

Lecture – 34
Forecasting Revisited

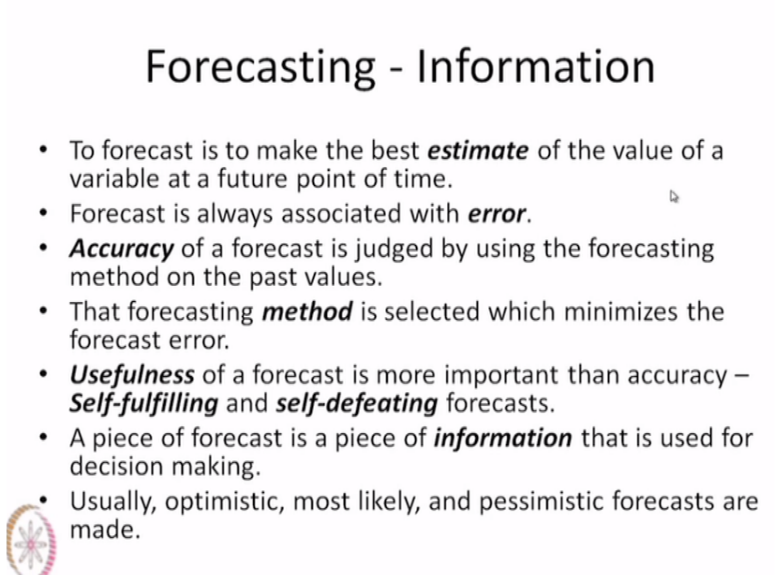
Good morning. Welcome to the 34th lecture on Economics, Management, and Entrepreneurship. In our last lecture, we discussed about product development. In particular, we discussed about the difference between products and services. We talked about different phases of product life cycle and also we discussed about features of product design and value engineering. Today, we shall discuss on forecasting.

If you recall in one of our earlier lectures, we had devoted a full session on demand forecasting. At that time, we had discussed at length on various qualitative methods and we had given only a glimpse of various quantitative methods. Today, we would like to revisit the forecasting methods and in particular, we shall discuss about 2 important quantitative methods: One for intermediate range forecast which is regression analysis are also called multiple regression analysis.

Which is a part of the broader econometric models and also we shall consider time series based forecasting, which is usually used for short term forecasting with the knowledge on these 2 important forecasting methods. It will be easy for us to use these methods to estimate forecasts of demand and then use these forecasts for different decisions such as plant location decisions, capacity requirement decision, machine requirement decisions, and of course production, planning, and inventory control decisions. So today's lecture is titled forecasting revisited.

Forecasting revisited is the title of today's lecture.

(Refer Slide Time: 03:23)



Forecasting - Information

- To forecast is to make the best ***estimate*** of the value of a variable at a future point of time.
- Forecast is always associated with ***error***.
- ***Accuracy*** of a forecast is judged by using the forecasting method on the past values.
- That forecasting ***method*** is selected which minimizes the forecast error.
- ***Usefulness*** of a forecast is more important than accuracy – ***Self-fulfilling*** and ***self-defeating*** forecasts.
- A piece of forecast is a piece of ***information*** that is used for decision making.
- Usually, optimistic, most likely, and pessimistic forecasts are made.

First of all, let us recall that to forecast is to make the best estimate of the value of a variable at a future point of time. Then naturally any estimate is always associated with error that means there is an error there will be always an error between the estimate or the forecast made at a particular time in the future and the actual value of the variable at that point of time. This difference is usually called the forecasting error.

The accuracy of a forecasting method is judged by the quantum of forecasting error not at one point of time, but at different points of time. Therefore, different criteria are used to judge the best forecasting method and normally the best forecasting method is chosen to minimize the mean absolute deviation or the mean absolute forecasting error or mean square error or similar such criteria, which is a function of forecasting errors at different points of time at future points of time of time.

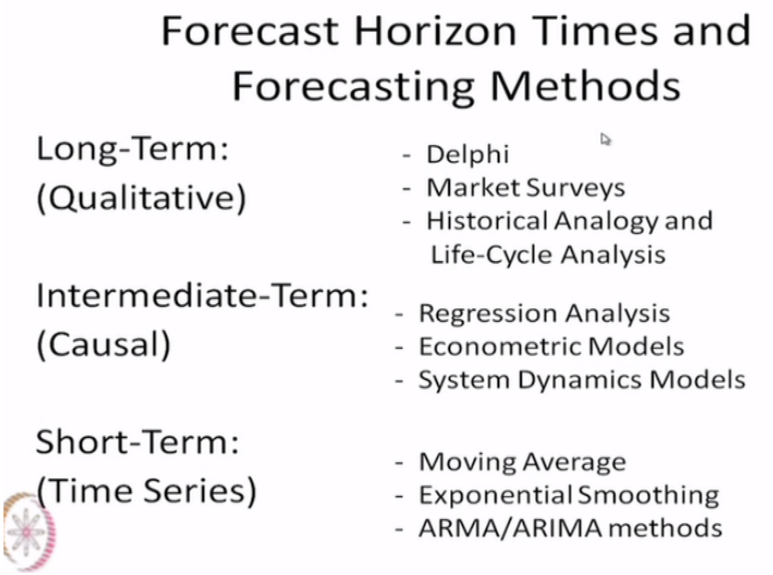
Now it must be also remembered at this point of time that usefulness of the forecast is more important than accuracy. The reason is this. That is if we know that a situation in the future is going to be very bad somebody makes a forecast we take preemptive actions to prevent the occurrence of such an event. So naturally in such a case, the forecast is not accurate, but it was highly useful. It is a self defeating type of a situation where a forecast is very useful, but is not accurate.

On the other hand, there are situations where we make a forecast at the future point of time and then needs a capacity decision, capacity requirement decisions are made on the basis of future projections of demand. If this projection is made, then the company tries its best to augment its capacity and also it produces and aligns it marketing forces such that it is able to utilize the capacity fully and sell the amount that was projected in the market.

Now this is a case of self full filling forecast. In any case, the forecast is judged not so much for its accuracy, but for its usefulness in taking decisions to improve the situation or to avert a bad future. Now so in any case a forecast is basically a piece of information and every information has got a value and a decision is sometimes defined as an information converter. A decision is made on the basis of information.

A forecast is a piece of information and that is an input to taking decisions. Since forecasts are associated with errors forecasting errors usually 3 types of forecasts are made: 1 is optimistic forecast. 2, the most likely forecast. 3, the pessimistic forecast.

(Refer Slide Time: 07:58)



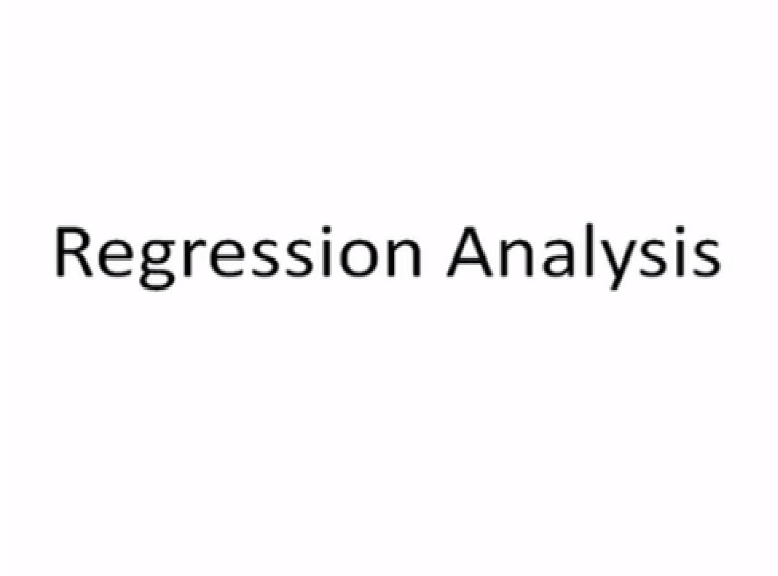
Forecast Horizon Times and Forecasting Methods

Long-Term: (Qualitative)	- Delphi - Market Surveys - Historical Analogy and Life-Cycle Analysis
Intermediate-Term: (Causal)	- Regression Analysis - Econometric Models - System Dynamics Models
Short-Term: (Time Series)	- Moving Average - Exponential Smoothing - ARMA/ARIMA methods

Forecast horizon times and forecasting methods are usually related. As we had discussed earlier for the long term forecast horizon time we use qualitative methods depending on the judgment of informed individuals and we already know of various methods such as Delphi methods, market survey methods, and we can also use historical analogy in which 1 product has gone through a product life cycle and another product of the same type is expected to go through similar life cycle variations or stages.

That is historical analogy and life-cycle analysis. Intermediate-term forecasting methods are usually cause-effect based and there are different types of methods. 1 single equation methods, they are normally known as regression methods and there can be multiple equations so econometric models are basically generalized models of regression analysis and there are methods that are useful in the short term and they constitute the various time series methods. Today we shall discuss mostly about the regression methods and the time series methods.

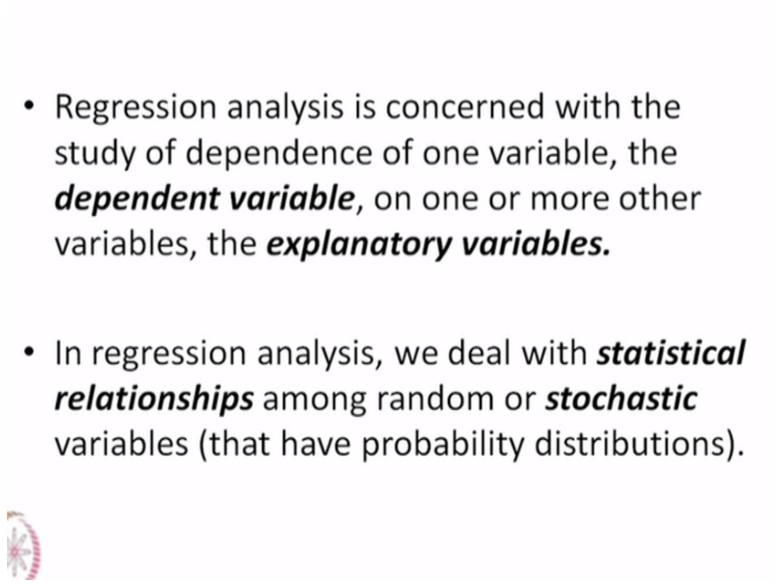
(Refer Slide Time: 09:40)



Regression Analysis

First, the regression methods of the analysis, that is made for regression methods. So what is regression analysis? It is concerned with the study of dependence of 1 variable, which is normally called the dependent variable on various independent or explanatory variables.

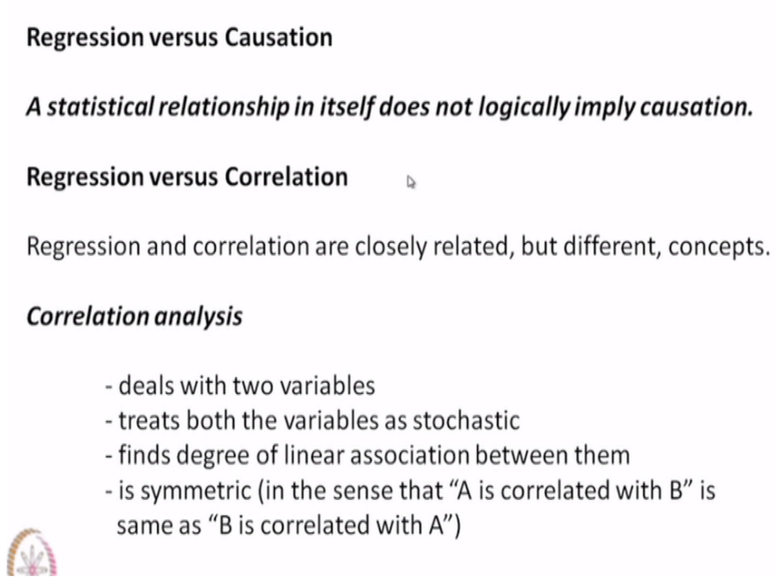
(Refer Slide Time: 10:18)



- Regression analysis is concerned with the study of dependence of one variable, the ***dependent variable***, on one or more other variables, the ***explanatory variables.***
- In regression analysis, we deal with ***statistical relationships*** among random or ***stochastic*** variables (that have probability distributions).

In regression analysis, we deal with statistical relationship among random or stochastic variables that have probability distributions. So we deal with statistical relationships.

(Refer Slide Time: 10:26)



Regression versus Causation

A statistical relationship in itself does not logically imply causation.

Regression versus Correlation

Regression and correlation are closely related, but different, concepts.

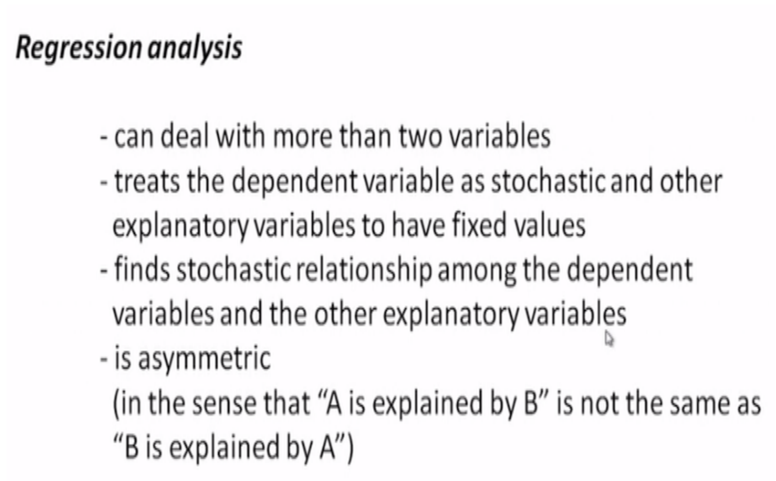
Correlation analysis

- deals with two variables
- treats both the variables as stochastic
- finds degree of linear association between them
- is symmetric (in the sense that "A is correlated with B" is same as "B is correlated with A")

Regression is not same as causation. A statistical relationship in itself does not logically imply causation. Regression is related to correlation, but they are different concepts. Correlation deals with 2 variables, treat both variables as stochastic, and finds linear relationship, or association between them. It is also symmetric in the sense that if A is correlated with B, then B is correlated with A, so in that sense it is symmetric.

Therefore, a correlation is always between 2 variables and it is symmetric and this always or usually discussed in the context of linear association between 2 variables.

(Refer Slide Time: 11:34)

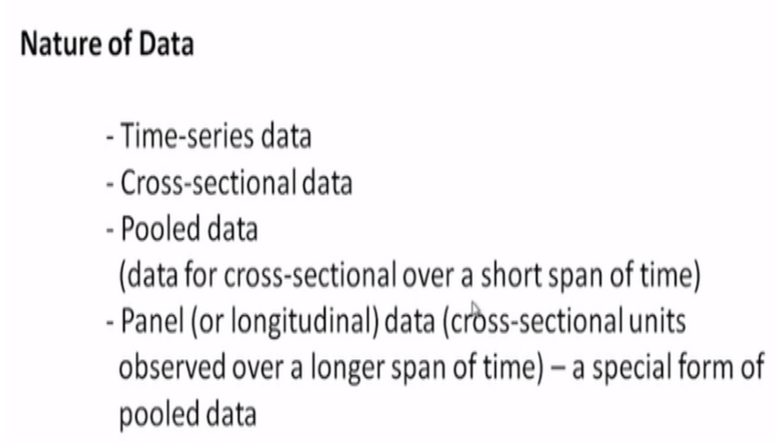


Regression analysis

- can deal with more than two variables
- treats the dependent variable as stochastic and other explanatory variables to have fixed values
- finds stochastic relationship among the dependent variables and the other explanatory variables
- is asymmetric (in the sense that "A is explained by B" is not the same as "B is explained by A")

Regression analysis on the other hand can deal with more than 2 variables. It uses the dependent variable as stochastic and explanatory variables as fixed and finds out a stochastic relationship between them. It is also not symmetric. It is asymmetric. In the sense that if A is explained by B, it is not same as B is explained by A. So these are the differences.

(Refer Slide Time: 12:12)



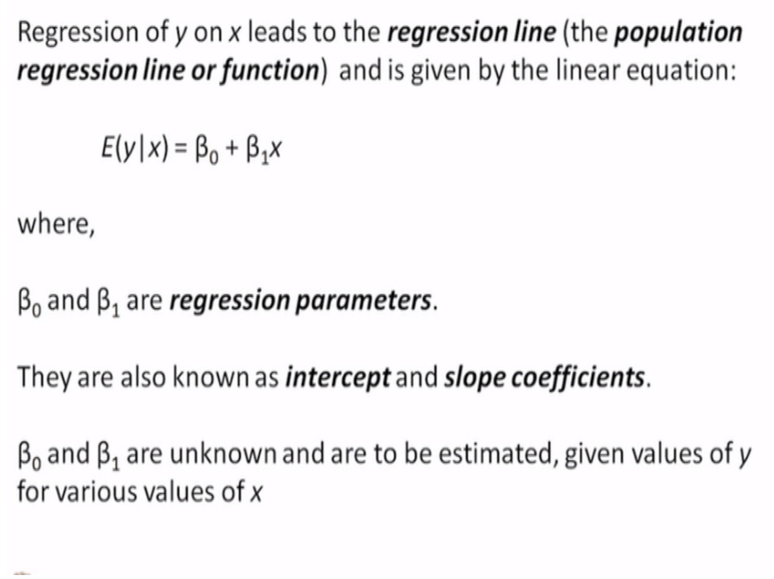
Nature of Data

- Time-series data
- Cross-sectional data
- Pooled data (data for cross-sectional over a short span of time)
- Panel (or longitudinal) data (cross-sectional units observed over a longer span of time) – a special form of pooled data

Now for both regression analysis and time-series analysis they depend on data. Regression can be done on time series data, on cross sectional data, on panel data, and pooled data, now what are they. Time series data is basically one variable or more than 1 variable at whose values are known at different points of time. This is time series data. Cross sectional data is at the same point of time many variables values are known for various subjects.

Pooled data is cross sectional data over a short span of time and panel data is a special form of pooled data, where cross sectional units are observed over a longer span of time. Regression analysis can be done on each type of data. We shall however discuss only the regression analysis done on cross sectional data, meaning that the values of the variable at a particular time are collected for different subjects.

(Refer Slide Time: 14:02)



Regression of *y* on *x* leads to the ***regression line*** (the ***population regression line or function***) and is given by the linear equation:

$$E(y|x) = \beta_0 + \beta_1 x$$

where,

β_0 and β_1 are ***regression parameters***.

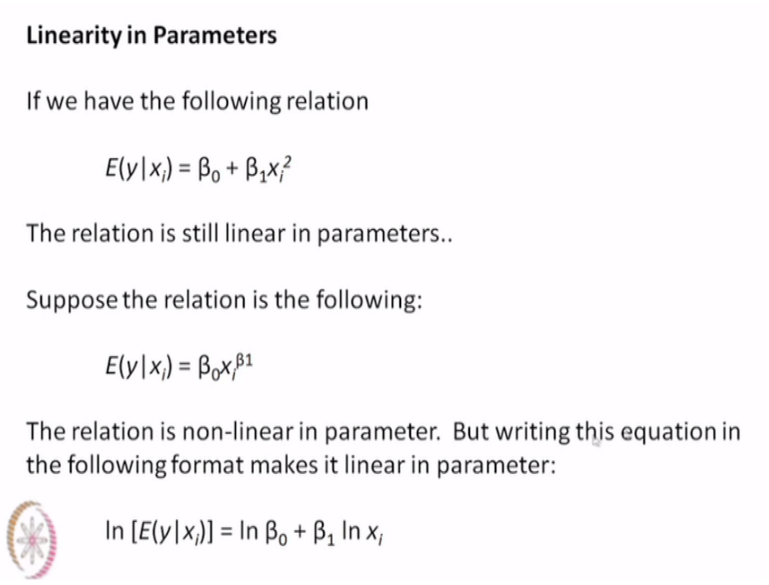
They are also known as ***intercept*** and ***slope coefficients***.

β_0 and β_1 are unknown and are to be estimated, given values of *y* for various values of *x*

Regression of a variable Y on x. Y is the dependent variable and x is the independent variable leads to what is known as the regression line and is given by the linear equation E y given x = beta 0 + beta1 x. So linear relationship x is the independent variable and Y it depends on x and y is a dependent variable and x is independent variable. Beta0 and beta 1 are regression. I am sorry this is regression parameters not coefficients, they are regression parameters.

Beta0 is the intercept and beta1 the slope coefficient. Usually beta0 and beta1 are unknown and are to be estimated given the different values of Y and X for different subjects at a particular point of time.

(Refer Slide Time: 15:33)



Linearity in Parameters

If we have the following relation

$$E(y|x_i) = \beta_0 + \beta_1 x_i^2$$

The relation is still linear in parameters..

Suppose the relation is the following:

$$E(y|x_i) = \beta_0 x_i^{\beta 1}$$

The relation is non-linear in parameter. But writing this equation in the following format makes it linear in parameter:

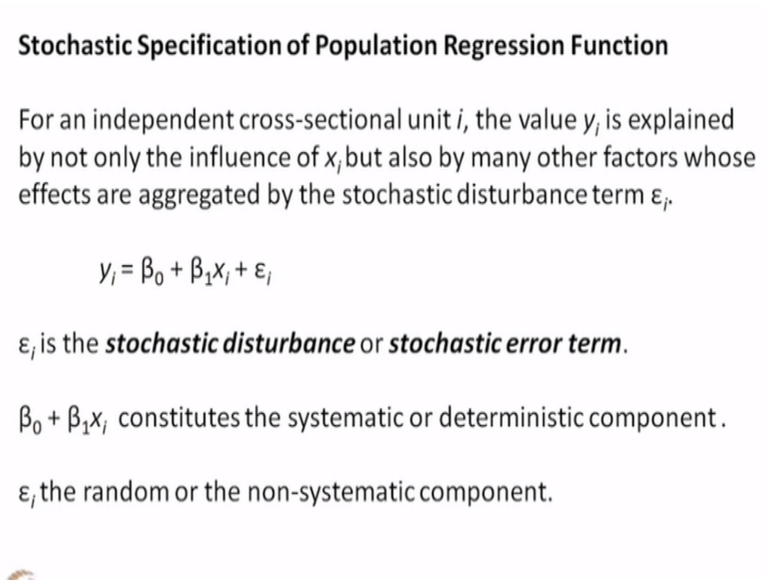
$$\ln [E(y|x_i)] = \ln \beta_0 + \beta_1 \ln x_i$$

Now we normally deal with equations that are linear in parameters. What does it mean? It means that our unknown variables are beta0 and beta1, unknown parameters whose values we would like to estimate. Therefore, if we have a relationship such as this although this contains xi square it is actually a numerical data. Y is also a numerical data. Therefore, the only unknown quantities are beta0 and beta1.

Thus, it says it is linear in parameter whereas if we have relationship such as this where beta 1 appears here as an exponent of the data remember that although we use x for an unknown quantity usually for variable. But actually it is a value of a variable in numerical quantity. It is a numerical quantity such as 5, 10, 15, 20 etc that raise to the power beta1 obviously the relationship is not linear, it is nonlinear.

You take log, and then it becomes ln beta0 + beta1 ln xi. This of course becomes ln. If you consider this as a variable and this is a one therefore this term becomes a linear relationship. This depicts a linear relationship. This is usually called a log linear relationship.

(Refer Slide Time: 17:31)



Stochastic Specification of Population Regression Function

For an independent cross-sectional unit i, the value y_i is explained by not only the influence of x_i but also by many other factors whose effects are aggregated by the stochastic disturbance term ε_i.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

ε_i is the ***stochastic disturbance*** or ***stochastic error term***.

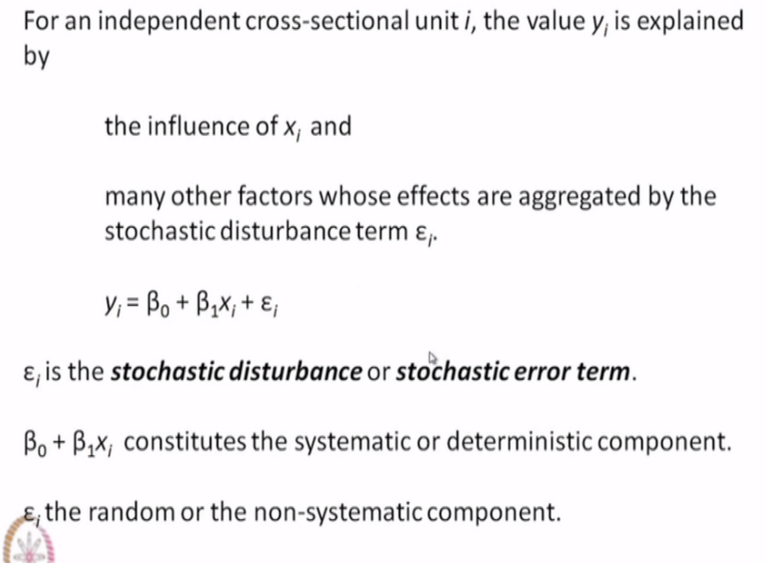
$\beta_0 + \beta_1 x_i$ constitutes the systematic or deterministic component .

ε_i the random or the non-systematic component.

Normally we add a random noise here a disturbance factor for every individual i, for every subject I or every cross section unit I it will be the population relationship that is beta 0 + beta 1xi + there will be a random error. This is the stochastic error term. Beta0 + beta1xi constitutes the systematic or the deterministic component and epsilon I is the random or the non-systematic component and I is for every cross section unit or the subject i.

Once again suppose that we have a relationship of this type this is still linear in parameters and if the relationship is this type by taking log, it is linear in parameter.

(Refer Slide Time: 18:51)



For an independent cross-sectional unit i, the value y_i is explained by

the influence of x_i and

many other factors whose effects are aggregated by the stochastic disturbance term ε_i.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

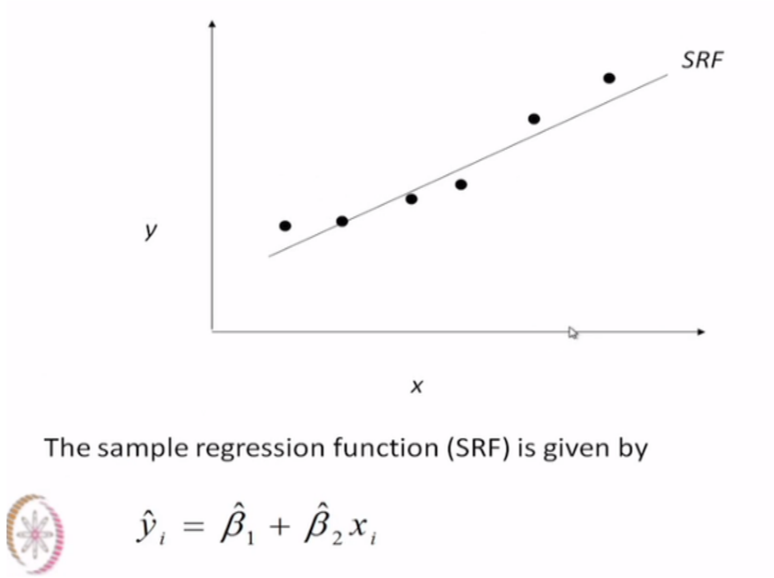
ε_i is the ***stochastic disturbance*** or ***stochastic error term***.

$\beta_0 + \beta_1 x_i$ constitutes the systematic or deterministic component.

ε_i the random or the non-systematic component.

And for an individual cross sectional unit it is this.

(Refer Slide Time: 18:52)



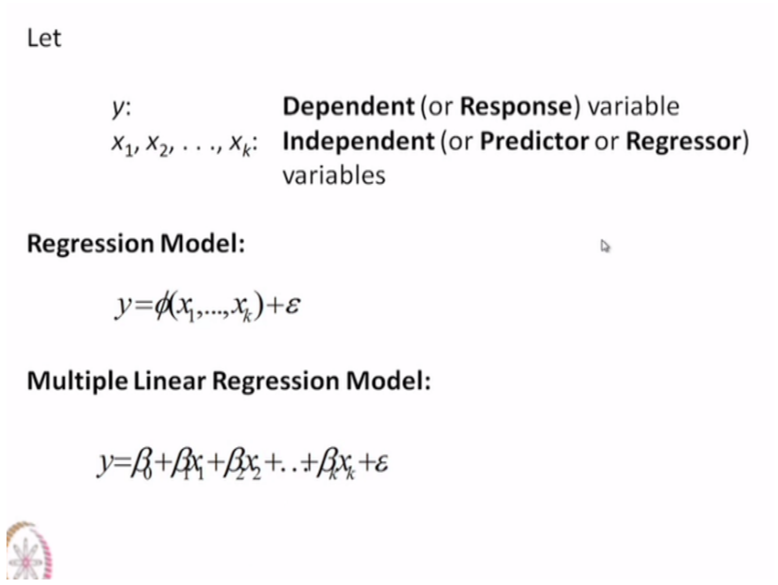
The sample regression function (SRF) is given by

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$

Now it can be shown in a graphical form. This is our sample regression function, the linear function, the systematic relationship between y and x of the deterministic relationship, but for every cross sectional unit or subject 1, 2, 3, 4, 5, and 6 the actual values are different. Therefore, this and this is due to the error, error that is attributable to this particular unit. This is the error for the third unit.

This is the error for the fourth unit, fifth unit, sixth unit, and if we can estimate the values of the parameters call them beta1 estimate and beta2 estimate then this relationship that does not contain the error term gives the value of yi that lies on the line SRF + this error term will give the actual value. Therefore, yi hat is basically a point that falls on SRF line and is just below the actual value line that means it is somewhere here. For the fourth one it is somewhere here, fifth it is here, sixth it is here.

(Refer Slide Time: 20:34)



Let

y: **Dependent** (or **Response**) variable
$x_1, x_2, \ldots, x_k$: **Independent** (or **Predictor** or **Regressor**) variables

Regression Model:

$$y = \phi(x_1, \ldots, x_k) + \varepsilon$$

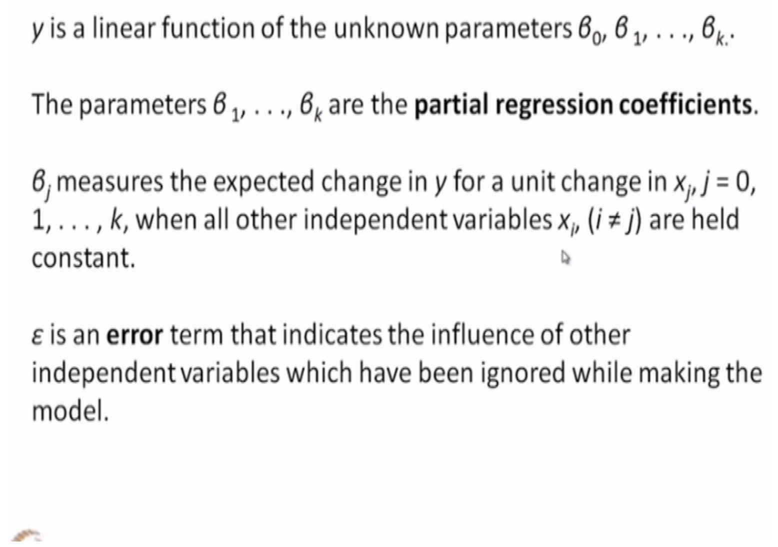
Multiple Linear Regression Model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

Now let us go for determining how the regression parameters are actually estimated. Let Y be the response variable or the dependent variable. X1, x2, xk and now considering more than 1 independent variable. I am considering k independent variables so, k independent variables and 1 dependent or response variable.

So the regression model is Y as a function of all this + an epsilon term and when written down in the expanded form it is written as y equals beta0 + beta1x1 etc beta k xk + epsilon, where beta0 is the intercept. Beta1, beta2, beta k are the regression coefficients and y is the value of the dependent variable.

(Refer Slide Time: 21:44)



y is a linear function of the unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$.

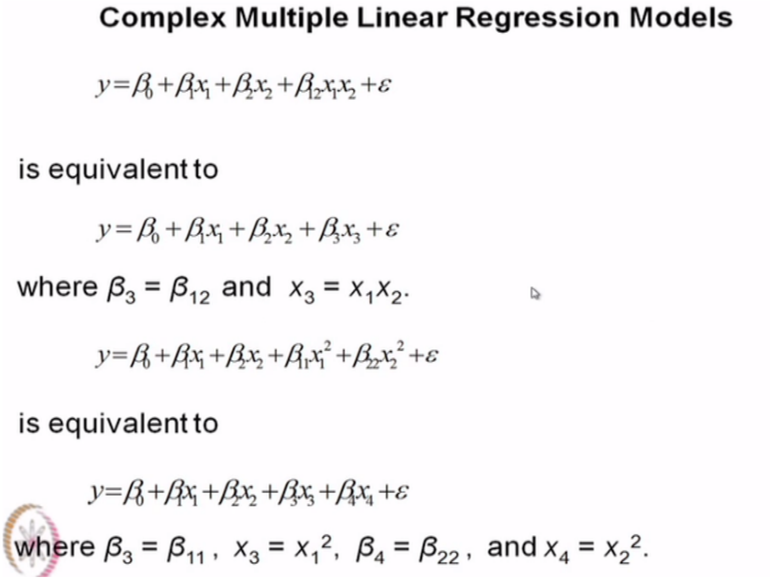
The parameters $\beta_1, \ldots, \beta_k$ are the **partial regression coefficients**.

β_j measures the expected change in y for a unit change in x_j, $j = 0, 1, \ldots, k$, when all other independent variables x_i, $(i \neq j)$ are held constant.

ε is an **error** term that indicates the influence of other independent variables which have been ignored while making the model.

Betas are called the partial regression coefficients. They measure the expected change in y for unit change in xj when all other variables are held constant and epsilon is an error term that indicates the influence of other independent variables that have been ignored while making the model.

(Refer Slide Time: 22:16)



Complex Multiple Linear Regression Models

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_{12}x_1x_2+\varepsilon$$

is equivalent to

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon$$

where $\beta_3 = \beta_{12}$ and $x_3 = x_1x_2$.

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_{11}x_1^2+\beta_{22}x_2^2+\varepsilon$$

is equivalent to

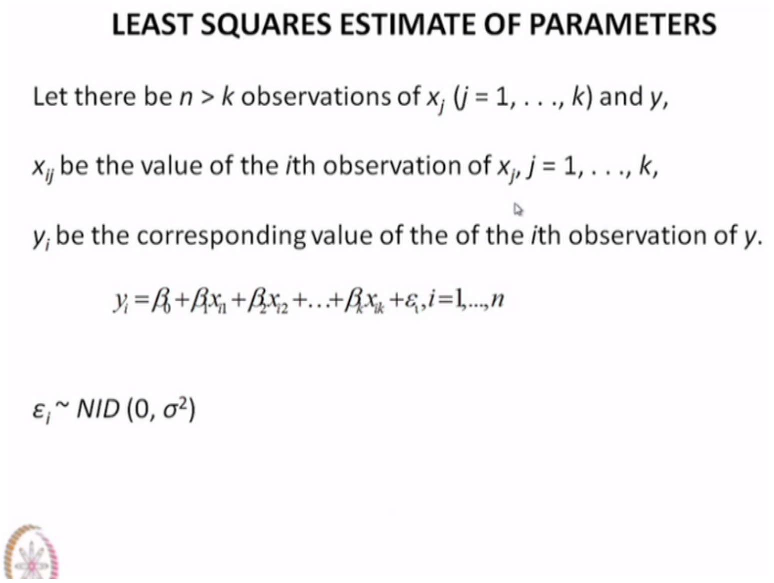
$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\varepsilon$$

where $\beta_3 = \beta_{11}$, $x_3 = x_1^2$, $\beta_4 = \beta_{22}$, and $x_4 = x_2^2$.

Now, we can have complex multiple linear regression models. Suppose we have an equation such as this beta0 + beta1x1 + beta2x2 + beta12x1x2. So even though it is written x1x2, but basically x1 and x2 are numerical values therefore the product is also a numerical value. So this can be written as x3, where x3 = x1x2 that can be found out given the values of x1 or x2. Therefore, this is equivalent to another equation where beta3 is nothing but beta12 and x3 = x1x2.

Similarly, if we have an equation such as this containing x1 square terms and x2 square terms we can write x1 square is = x3 and x2 square = x4. Then this becomes a linear form. Therefore, although we can have situations where we have in variables the relationship could be nonlinear it can be transformed into relationships or equations that are linear in parameter. So basically we are taking up cases that are linear in parameter.

(Refer Slide Time: 23:48)



LEAST SQUARES ESTIMATE OF PARAMETERS

Let there be $n > k$ observations of x_j ($j = 1, \ldots, k$) and y,

x_{ij} be the value of the ith observation of x_j, $j = 1, \ldots, k$,

y_i be the corresponding value of the of the ith observation of y.

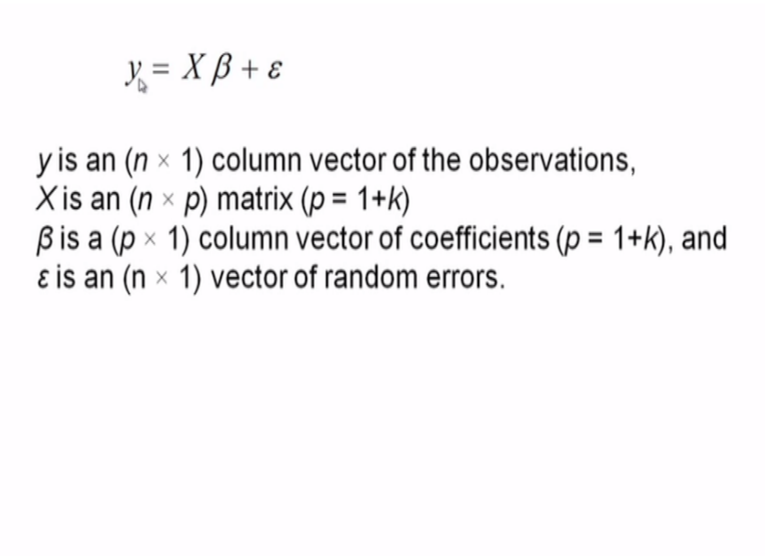
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, i = 1, \ldots, n$$

$\varepsilon_i \sim NID\ (0, \sigma^2)$

Now what we normally do is to estimate the values of the regression parameters by minimizing the least square errors. Will there be n observations if k is the number of independent variables then n should be larger than or higher than or more than the number of variables of xj and y. xij be the value of the ith observation of xj, j = 1 to k and yi be the corresponding value of the ith observation of y.

So we write yi = beta0 + beta1 xi1 + beta2 xi2 that means for the ith observation we have the values of x1, x2, xk etc and also y and normally epsilon the random error component is defined as normally distributed independent normally and independently distributed with 0 mean and constant variance sigma square. This is a symbolic notation for stochastic variation of the noise component epsilon normally and independently distributed random variable with 0 mean and variance sigma square.

(Refer Slide Time: 25:34)



$$y = X\beta + \varepsilon$$

y is an ($n \times 1$) column vector of the observations,
X is an ($n \times p$) matrix ($p = 1+k$)
β is a ($p \times 1$) column vector of coefficients ($p = 1+k$), and
ε is an ($n \times 1$) vector of random errors.

And since we have many observations it is possible to convert it into a vector matrix form. We can now write. We had y1, y2, y3 etc we can write y as a vector of all the observations of the dependent variable since there are n number of observations it will be n * 1 column vector, x can be a matrix n * p where p = 1 + k. beta is a column vector of coefficients with say dimensions' p = 1 + k and epsilon is a vector of random errors.

(Refer Slide Time: 26:22)

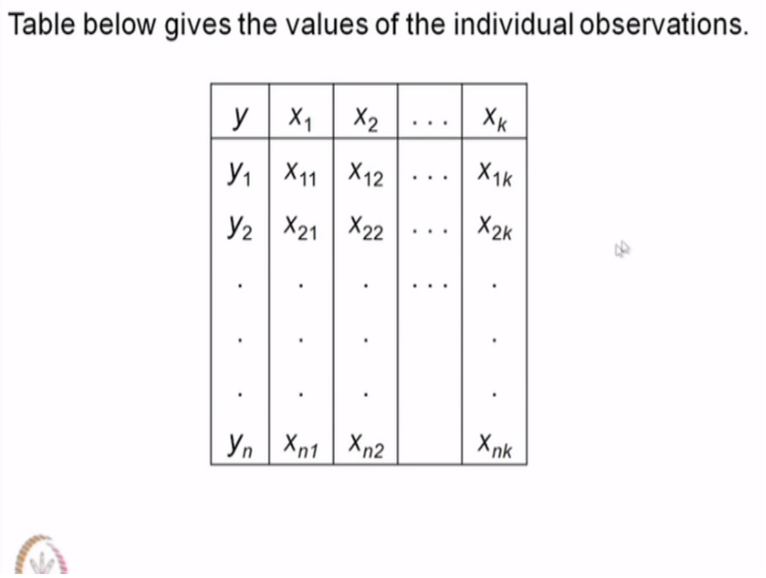
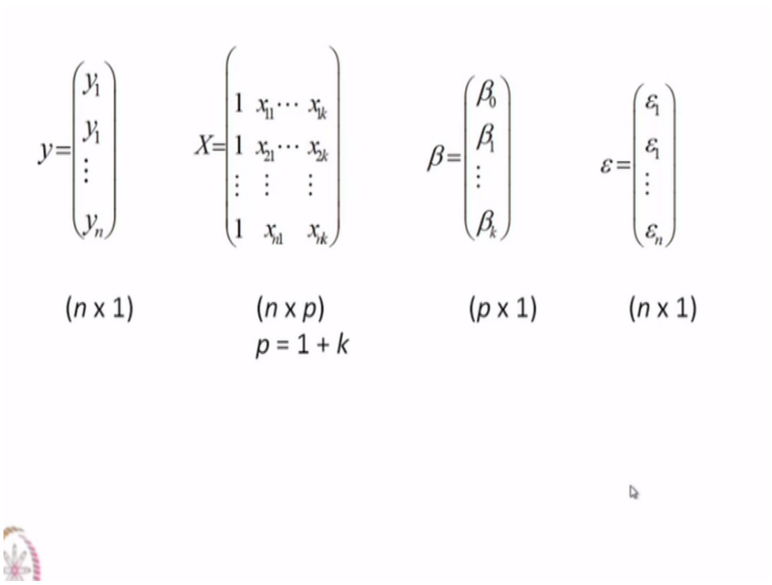


Table below gives the values of the individual observations.

y	x_1	x_2	. . .	x_k
y_1	x_{11}	x_{12}	. . .	x_{1k}
y_2	x_{21}	x_{22}	. . .	x_{2k}
.	.	.	. . .	.
.	.	.		.
.	.	.		.
y_n	x_{n1}	x_{n2}		x_{nk}

This will be clear from this table. These are the variables. Y, the dependent variable and x2, x2, xk are the independent variables. We have n different observations or n different cross sectional units or n subjects. For every subject the values are collected at a particular point of time. Let for the first cross sectional unit the values are this, for the second the values are this, and for the nth cross sectional unit the values are this. So what I am now saying is that these constitute a column vector y and this and a one vector we will constitute the matrix X.

(Refer Slide Time: 27:31)



$$y=\begin{pmatrix} y_1 \\ y_1 \\ \vdots \\ y_n \end{pmatrix} \quad X=\begin{pmatrix} 1 & x_{11}\cdots & x_{1k} \\ 1 & x_{21}\cdots & x_{2k} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{nk} \end{pmatrix} \quad \beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

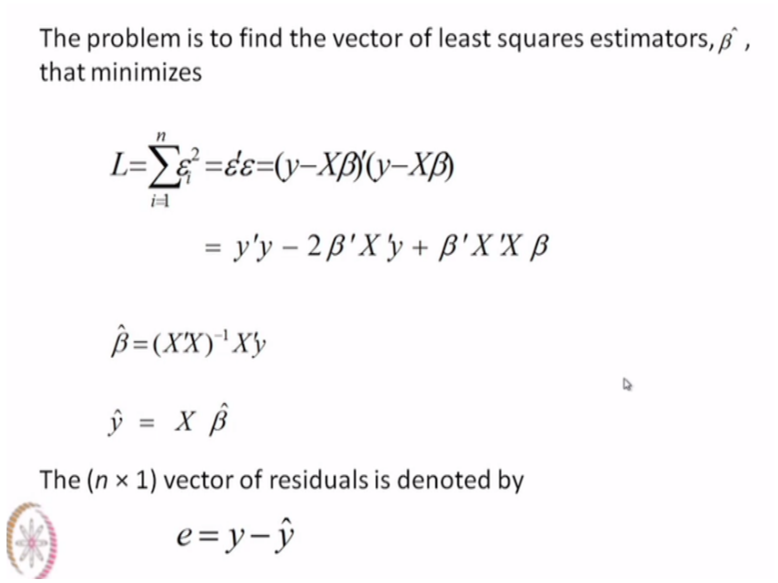
$(n \times 1)$ $(n \times p)$ $p = 1 + k$ $(p \times 1)$ $(n \times 1)$

Now this is shown here. Y is the column vector of the dependent variables. Beta is the column vector for all the regression parameters. As you know there are k number of variables and therefore associated number of parameters are k in number, but there is an intercept beta0 therefore it becomes k + 1. That k + 1 is written as p. So it is given a notation p * 1. So let us say p * 1 column vector and epsilon is also n * 1 there are n number of cross section units.

This 1 comes here because of beta0. If you multiply x with beta, then you will get back y1 = beta0 + beta1 x11 etc. There is one mistake here. This should have been xi k and not. So I basically said that we said that we collect data on the independent variables x1 through xk and on the dependent variable y for each of the n cross sectional units and then we have various equations y1 = beta01 + beta1 x11 + beta2 x12 etc.

Now, this I can now put in the matrix form such as this where this is a vector, this is a matrix, this is another vector is another vector with this dimensions.

(Refer Slide Time: 30:03)



The problem is to find the vector of least squares estimators, $\hat{\beta}$, that minimizes

$$L=\sum_{i=1}^{n}\varepsilon_i^2=\varepsilon'\varepsilon=(y-X\beta)'(y-X\beta)$$

$$= y'y-2\beta'X'y+\beta'X'X\beta$$

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

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

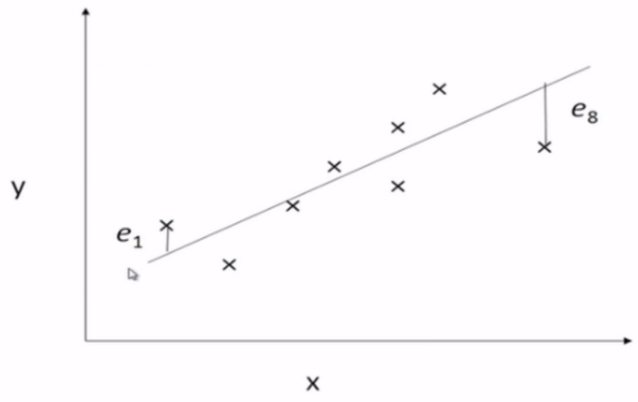
The (n × 1) vector of residuals is denoted by

$$e=y-\hat{y}$$

The problem is to find the vector of least squares estimators. This should come here. The problem is to find the vector of least square estimators called then beta hat that minimizes the some of the squares of the random errors epsilon I square I = 1 to n and that is nothing but epsilon transposed epsilon. What is this? This is y - x beta transposed. This will lead to by transposed y etc. From here without going through the derivation we can finally get the estimate for beta and the estimate for beta is it is x transposed x inverse x transposed y.

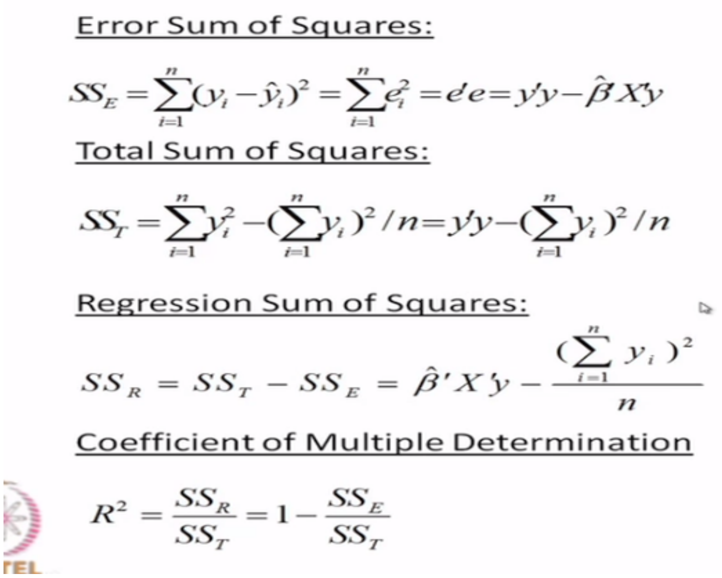
X is basically the if you recall x is this. This particular with 1's in its first column and the data that we had collected for every cross sectional unit’s independent variables are here. So this is the final expression for beta hat. X transpose x inverse x transposed y. Now that we know beta hat we can find out y hat, the estimate of y = x * beta hat. Now we know the actual value of y at the particular point of time and we have estimated the value of y and therefore we can find out the residual. We normally call this as residual e.

(Refer Slide Time: 32:24)



Now this is a graphic representation of what I was trying to say for a case when it is a single or a simple regression with one independent variable x, one dependent variable y. This is the SRF. Sample regression function and these are the errors or the residual e1 through e8 and by the (()) (32:57) a point lying here is the estimated value of y for this value of x. For this value of x the estimated value of e is this. So basically this line is y hat line and the actual value is here. The difference is the residual ei.

(Refer Slide Time: 33:25)



Error Sum of Squares:

$$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = e'e = y'y - \hat{\beta}X'y$$

Total Sum of Squares:

$$SS_T = \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2 / n = y'y - (\sum_{i=1}^{n} y_i)^2 / n$$

Regression Sum of Squares:

$$SS_R = SS_T - SS_E = \hat{\beta}'X'y - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

Coefficient of Multiple Determination

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

Now we can find out the error sum of squares, it is the actual value of y - the estimated value of y that is y hat y square them sum over I = 1 to n. This is nothing by ei square and this can be written as e transposed e if e is a vector and then we can final expression. The total sum of squares can be found out in this manner and 2 regressions the sum of squares will be total sum of squares - the error sum of squares.

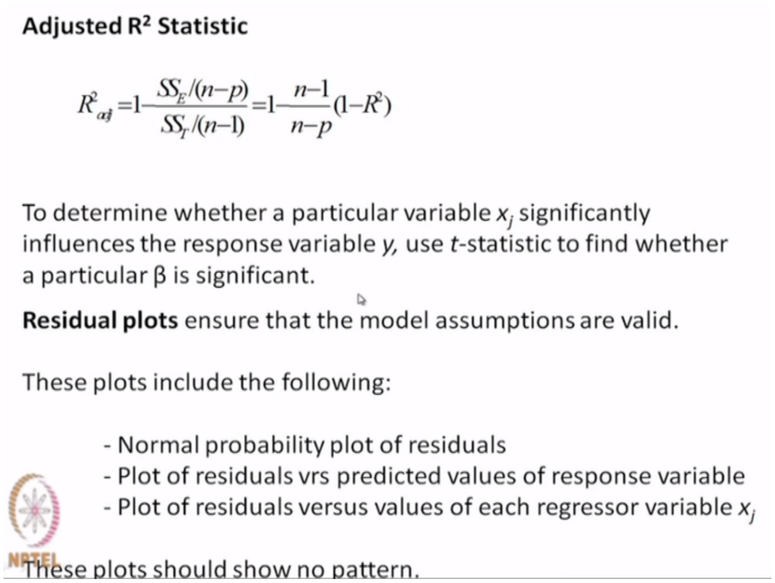
From here we define a coefficient of multiple determinations R square which is regression sum of squares/total sum of squares. Well regression modeling is mathematically quite complex. We did not want to go into the details of the mathematical details. The basically what I want to tell here is that per different cross section units or subjects we observe values of the independent variables and the dependent variables.

Dependent variable is y, independent variables are x1, x2, xk and then we find out the equation of the regression line so linear regression by estimating the best values of regression parameters beta0, beta1, up to beta k that minimizes the regression mean square errors or square error is called a least square estimation. It is given by in the vector matrix form we can write y vector = x beta + epsilon and we can find out beta as x transpose x inverse x transposed y.

That is the expression for beta. Once beta estimate is known we can find out, we can make the estimate of y at different values of x by using the relationship y hat = x beta hat. Then we can find the difference of the actual value of y and the estimated value y hat we call it error. Given the error we can find out 3 types of sum of squares: Error sum of squares, total sum of squares, and regression sum of squares.

The ratio of regression sum of squares to the total sum of squares tells us how much our regression equation explains the variation is called SSR/SST. The explained how much SSR is the error sum of squares defined by or explained by our regression model but the actual sum of squares is SST, the total sum of squares.

(Refer Slide Time: 37:39)



Adjusted R² Statistic

$$R^2_{adj} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} = 1 - \frac{n-1}{n-p}(1-R^2)$$

To determine whether a particular variable x_j significantly influences the response variable *y*, use *t*-statistic to find whether a particular β is significant.

Residual plots ensure that the model assumptions are valid.

These plots include the following:

- Normal probability plot of residuals
- Plot of residuals vrs predicted values of response variable
- Plot of residuals versus values of each regressor variable x_j

These plots should show no pattern.

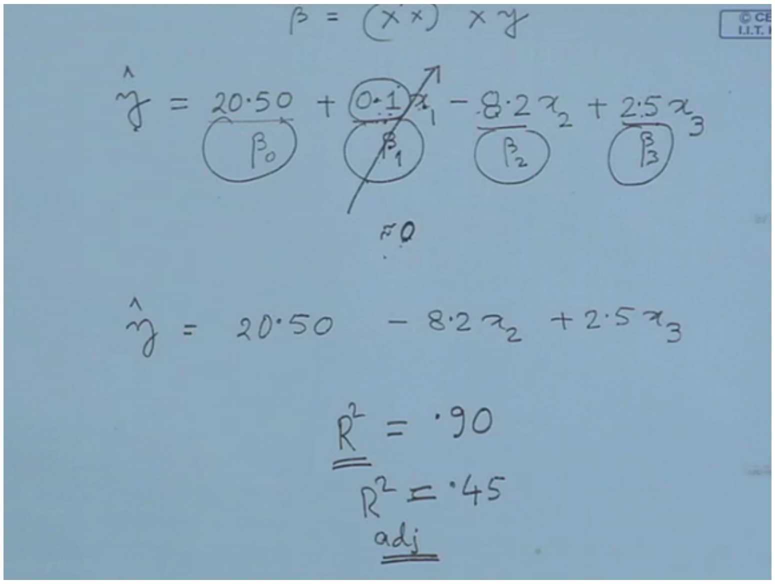
Normally, we go for an adjusted R square statistic because sometimes we are not very sure as to which variables influence the value of y which independent variable. The choice of independent variable is sometimes quite challenging. One may take large number of variables to explain y. It has been seen that quite often many of them are not important, many of them are related with each other and they can increase the explained sum of squares that means r square value.

So R square does not always indicate the adequacy of a regression model. It has to be adjusted for because of redundant or near redundant explanatory variables being included. That is why it is given by 1 - error sum of squares/n - p and/total sum of squares/ n - 1. Once again R square are adjusted as R square are good indicators of the extent to which the errors are explained by the regression line.

We also use t statistics to find whether a particular beta is significantly different from zero. This is beyond the scope to discuss how t statistics are calculated, but basically there are what I want to say is that there are quite a large number of software packages that deal with regression modeling, a regression analysis. If you can define the independent variables x1 through xk and independent variable y and give their values for different cross sectional units.

Then it will make the calculations for yourself the software package will make the computations for yourself and for you and then define the R square values find the R square value and find the t statistics and also do lot of other things. So we should know the meaning of these statistics rather than how to derive them.

(Refer Slide Time: 41:00)



$$\beta = (x x)\ x y$$

$$\hat{y} = \underset{\boxed{\beta_0}}{20.50} + \underset{\cancel{\boxed{\beta_1}}}{\cancel{\boxed{0.1}x_1}} - \underset{\boxed{\beta_2}}{8.2}x_2 + \underset{\boxed{\beta_3}}{2.5}x_3$$

$$\neq 0$$

$$\hat{y} = 20.50 - 8.2x_2 + 2.5x_3$$

$$R^2 = .90$$

$$R^2_{adj} = .45$$

So if for example what I want to say is that suppose by regression modeling I get a model such as this 20.50 + 0.1x1 - 8.2 x2 + 2.5 x3. Now these are the estimated values. This is estimated value of beta 0. This is estimated value of beta 1, estimated value of beta 2 and of beta 3. These are the standard values that came out of our equation which was beta 0, beta estimates = x transposed x inverse x transposed y. Now look at the values, 8, 2.5, 20.5, and 0.1.

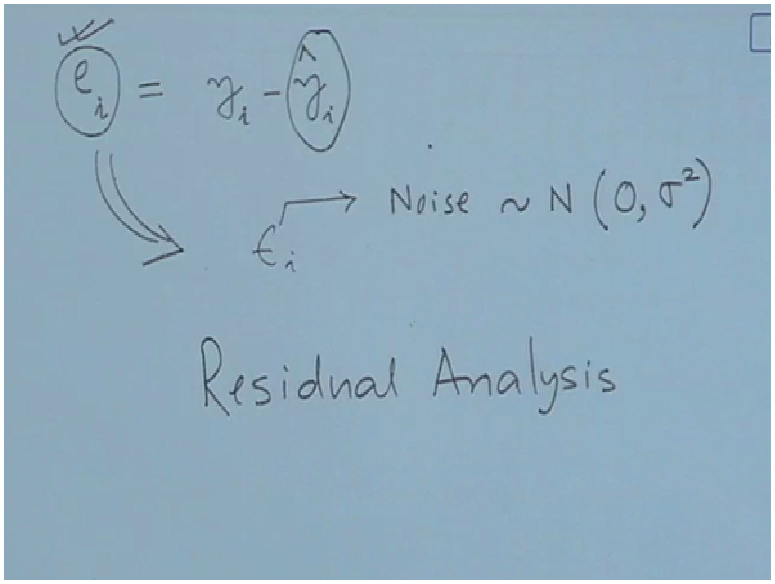
Now this 0.1 looks appears to be very small. The contribution of x1 to the change in y seems to be quite less. T statistics say whether this 0.1 is closed to 0 or whether this is closed to 0, this is closed to 0 and normally there are test of hypothesis with the method of which we can find out whether each of these coefficients are different from 0. Looking at it appears as though 0.1 is too small compared to these 2 and probably it can be considered that the effect of x1 is very negligible.

And therefore we can straight away neglect this we can say that the equation y hat it is = 20.5 - 8.2x1, x2 + 2.5 x3. So t statistics help to find out whether the coefficient is significantly different from 0 at what level of significance and then as I said R square values gives us R square = let say

0.9 tells us the 90% of the variation in y is explained by these 3 factors whereas if the R square value would have been only 0.45.

It will augment that only 45% of the variation in the value of y is explained by the regression equation. So this is the notation of these 2 statistics instead of writing R square as I told you we should normally go for some sort of an adjustment to indicate how far redundant variables have been added to our list of independent variables. Now after we do all this we also go for residual plots. If you remember

(Refer Slide Time: 45:01)



$e_i = y_i - \hat{y}_i$

$\epsilon_i \rightarrow$ Noise $\sim N(0, \sigma^2)$

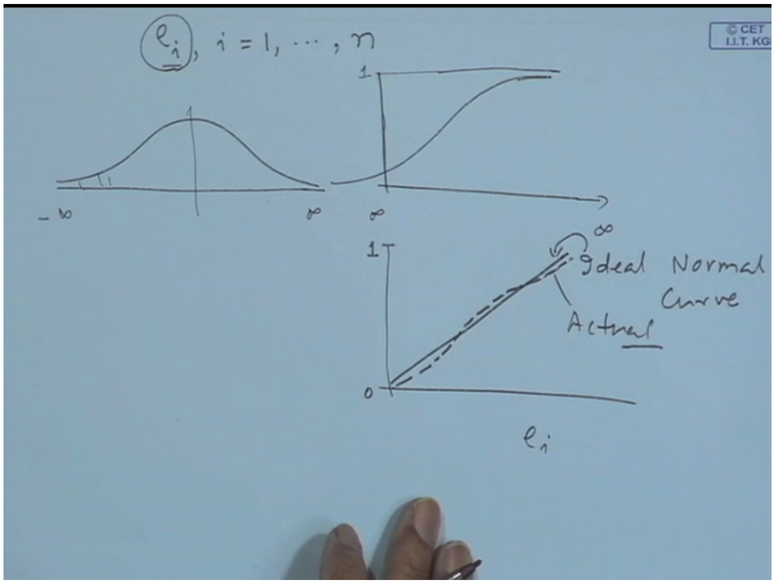
Residual Analysis

We calculate for each cross section unit I we calculate ei the residual yi - yi hat. This is the value of yi that we calculate from our regression equation and this is the actual value. Now that we have the residuals with us these residuals are actually surrogate measures of epsilon i. epsilon I is the noise which we assumed to be normally distributed with 0 mean, and constant variance sigma square. This was our original assumption when we derived the betas.

Therefore, now is the time to actually judge whether these assumptions are actually justified. This justification can be examined only by analyzing the residuals ei. So what we do there are different forms of analysis of the residuals. They are called residual analysis. Regression model is incomplete unless a residual analysis is also made. The residual analysis normally takes different forms; we have discussed here.

We have only indicated here different plots, residual plots. One is the normal probability plot of the residuals. Because we had assumed the errors to be normally distributed the residual should also be normally distributed. Then the residuals versus the predicted values of the response variable should not show a pattern and residuals versus values of each regressor variable xj should also not show any pattern let me explain this.

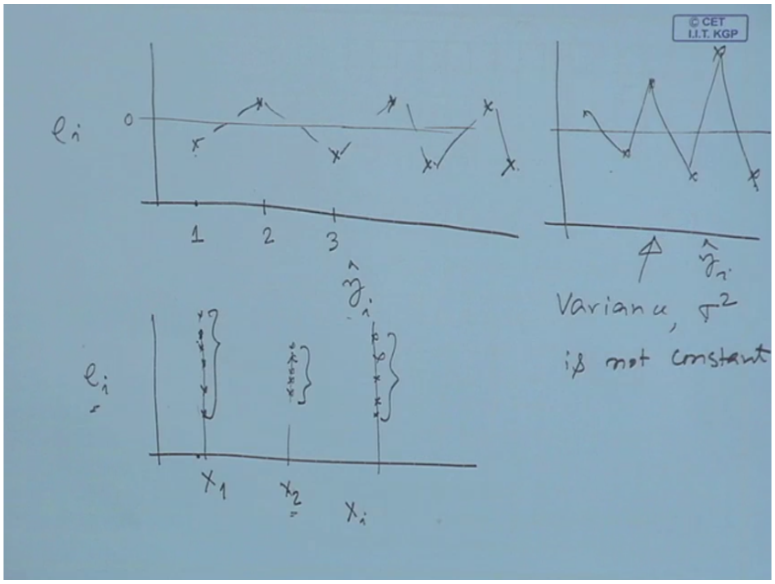
(Refer Slide Time: 47:39)



Suppose that we have residuals ei I = 1 through we have n number of observations. We can now plot ei as a histogram or as a cumulative distribution function. As you know a normal probability continuous normal function, normal density function looks like this and when we can draw a cumulative distribution function that goes up like this from 0 and become 1. This is from -infinity to+ infinity. This also goes to + infinity to – infinity, but the area is added up that is why this always rising. This is called a cumulative distribution function.

Now if this axis is probably scaled then it is possible that this curve looks like a straight line. If this axis is properly scaled, this raise from 0 to 1, but it is properly scaled. So we can say that this is the ideal normal curve, normal CDF, normal curve, and our actual ei when plotted in this manner may actually look like this. Now this extent of deviation from this is normal and this is actual indicates to what extent the normality assumption is deviated in the data. So this normal probability plot basically is done in this manner. The second type of plot that is required for residual analysis is plot of residuals

(Refer Slide Time: 50:10)



Versus the predicted values y bar for different i. So this is observation number 1, number 2, number 3, and like that 8 or 19 observations. The errors some will be positive; some will be negative. Let us say that the values are like this. Now you can see that there is a pattern. Or that you may normally if there is a relationship normally it will show a pattern such as this increasing pattern. There is also a pattern here.

There is a cyclic variation or a seasonal variation here and there is a diverging pattern of relationship between ei and yi hat. Both indicate that it is not random that there is a relationship and that variance sigma square iis not constant in this case. It is a function of yi hat. The third plot is the plot of residuals versus each regressor xi. Now here you will have for each regressor x1 and x2 and x3 you may see all the points lying here. There are so many observations.

For this it may be here. For the third you may see it here. So basically the deviation would indicate how far ei depends on xi. So these are the different ways by which the normality is judged. The independence is judged and the constant variance assumption is judged. So friends I discussed a full session on regression analysis because regression analysis is quite useful in making demand forecasting particularly for a new entrepreneur who has had not much of a pass data.

If pass data were available a time series forecasting method could have been used instead for a new entrepreneur, he can make assumption regarding demand of a product on the basis of various economic variables that are g and p population and so on and so forth. For such a situation the most adequate forecasting model is the regression model. In our next class we shall take an example and of regression model and then use or expose you to the time series forecasting method.