

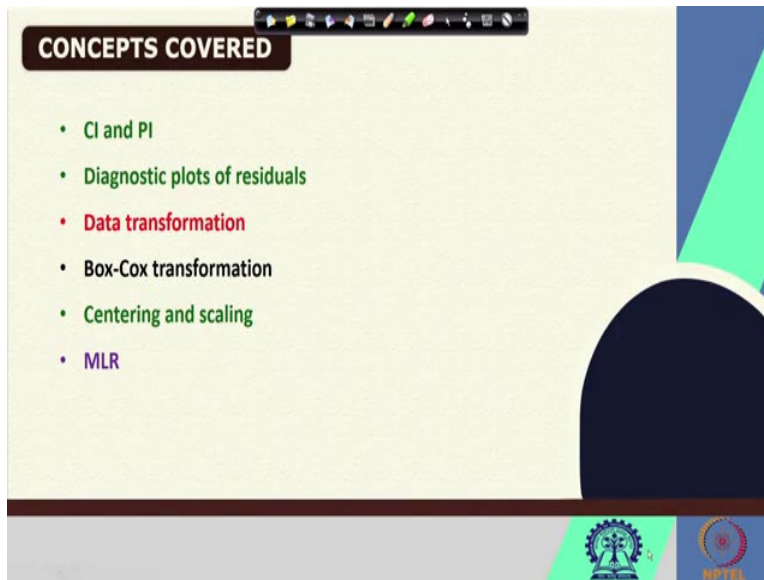
**Machine Learning for Soil and Crop Management**  
**Professor. Somsubhra Chakraborty**  
**Agricultural and Food Engineering Department**  
**Indian Institute of Technology, Kharagpur**  
**Lecture 08**  
**Basics of Multivariate Data Analytics (Contd.)**

Welcome friends to this eighth lecture of this NPTEL online certification course and we are in week 2, where we are discussing the basics of Multivariate Data Analytics. And in this week, we have already discussed the correlation, we have seen the basic structure of the spreadsheet, we have seen the what is the multivariate data and different types of association.

We have seen the correlation, features of different correlation, positive correlation, negative correlation, no correlation and how those plots look like, correlation features. And also we have seen the simple linear regression, in case of simple linear regression, we have seen the assumption of simple linear regression, four important assumption of simple linear regression if you remember, constancy of variance, independence of the observation, then normality of the observation and linearity of the mean. So, you can see that these are the four major assumption of linear regression.

And then we have seen the slope as well as the intercept for this simple linear regression of  $y$  versus  $x$ . And then we have seen how to calculate, how to visually represent the different sum squares part of the total regression scenario, how we can calculate the R square from the sum square error, sum square total and sum square regression. What is RMSE, how we can identify, how we can see and interpret the output of, the output of the slope and intercept. So, we have seen all these.

(Refer Slide Time: 2:23)



A slide titled "CONCEPTS COVERED" with a list of statistical concepts. The slide has a light green background with a dark blue and light green geometric design on the right side. At the bottom, there are logos for IIT Bombay and NPTEL.

- CI and PI
- Diagnostic plots of residuals
- **Data transformation**
- Box-Cox transformation
- Centering and scaling
- MLR

Now, in this lecture we are going to discuss these following concepts. First of all we are going to discuss the confidence interval and prediction interval and what are the differences between confidence interval and prediction interval. We are going to also discuss the diagnostic plots of residuals to identify whether all the assumptions of linear regressions are met or not. And then we are going to see if time permits then we are going to see that data transformation Box-Cox transformation, centering and scaling and multiple linear regression.

(Refer Slide Time: 2:54)

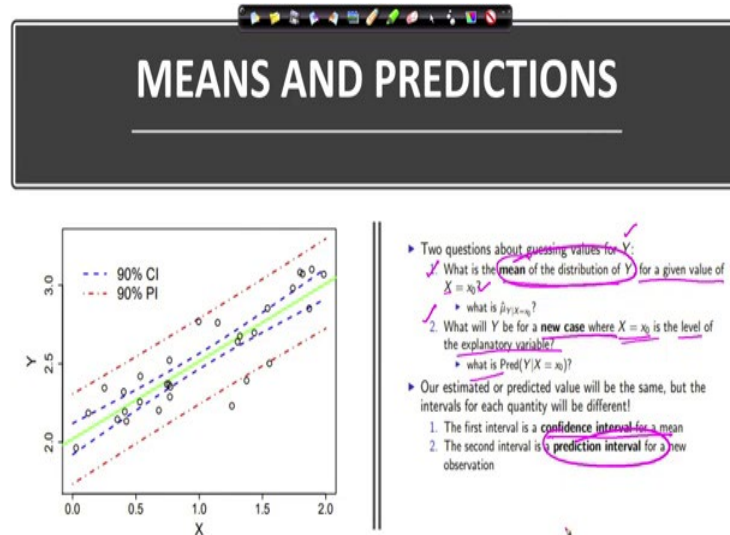


A slide titled "KEYWORDS" with a list of statistical concepts. The slide has a light green background with a dark blue and light green geometric design on the right side. At the bottom, there are logos for IIT Bombay and NPTEL. A video inset in the bottom right corner shows a man in a white shirt speaking.

- Centering and Scaling
- Box-Cox transformation
- Multivariate calibration
- Multivariate classification
- MLR

So, we will try to cover these keywords in this lecture. First of all, we will see the centering scaling, then a Box-Cox transformation, multivariate calibration, multivariate classification and multiple linear regression.

(Refer Slide Time: 3:08)



So, let us start with the means and prediction. So, you can see that in case of simple linear regression, if this is a simple linear regression plot, two questions always arise, first of all for guessing the values of  $Y$  always there are two questions. First of all, what is the mean of the distribution of  $Y$  for a given value of  $X$ ?

Suppose this given value of  $X$  is  $x_0$ , what should be the mean of the distribution of  $Y$  in that case? Or in other words, what is the  $\hat{\mu}_{Y|X}$  given the value of  $X$  equal to  $x_0$ . Now, what we, and second question is what will  $Y$  be for a new case where  $X$  equal to  $x_0$  is the level of the explanatory variable. And what is the, or in other words what is the predicted, prediction of  $Y$ , given the value of  $X$  is the  $x_0$ .

So, our predicted estimated or predicted value will be same, but the interval for these two, is these two questions will be different. So, each quantity for each quantity will be different. So, the first interval where we are interested for the mean of the distribution of  $Y$ , then we will call it a confidence interval for a mean and when we are more interested to know the  $Y$  for a new case where  $X$  equal to  $x_0$ , we will call it a prediction interval for a new observation. So, this is how these two are different from each other.

So you can see in this figure, this is Y versus X. And you can see these are the observation this blue line is showing the 90 percent confidence interval of mean. So, that shows that within this line the Y, the mean of the Y values will reside. And whereas, these dotted lines, these orange dotted line shows the two extremes of the prediction interval and it shows that Y will be given a value of X, the predicted value of Y will line here. So, one encompasses the mean of Y whereas, the other encompasses the prediction of Y. So, this is how they are different from each other.

(Refer Slide Time: 5:41)

## CI CALCULATION

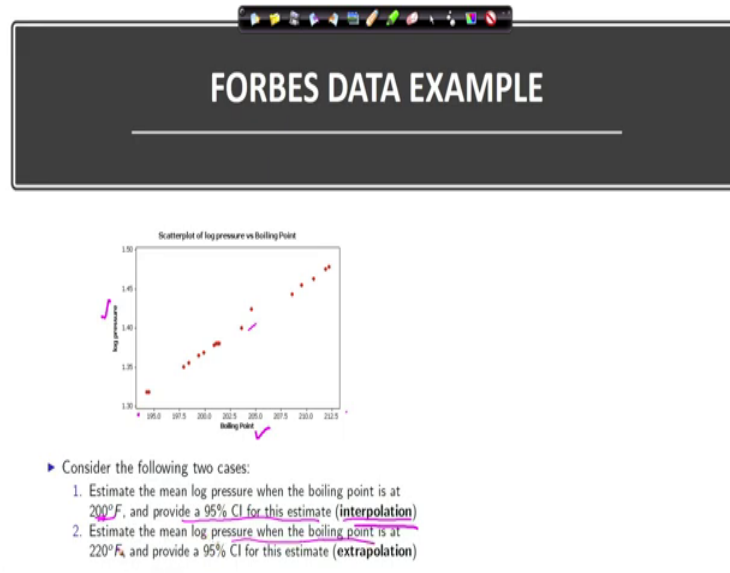
- ▶ The estimated mean of Y given  $X = x_0$  is
 
$$\hat{\mu}_{Y|X=x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$
- ▶ The standard error for this estimate is
 
$$S.E.(\hat{\mu}_{Y|X=x_0}) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{(n-1)S_X^2}}$$
- ▶ A  $100(1 - \alpha)\%$  CI of  $\mu_{Y|X=x_0}$  is
 
$$\hat{\mu}_{Y|X=x_0} \pm t_{(\alpha/2, n-2)} S.E.(\hat{\mu}_{Y|X=x_0})$$
- ▶  $S.E.(\hat{\mu}_{Y|X=x_0})$  depends on the value on  $x_0$ ,  $n$  and  $\hat{\sigma}$ .

So, if we how to calculate the confidence interval of mean so, if the estimated mean this is for given X equal to x0, we know that, we have already seen that now, the standard error for these estimates you can calculate by using this formula where this sigma hat root over of 1 by n plus x0 minus X bar whole square by n minus 1 SX square, where SX is the standard deviation of X.

So, for a given level of alpha, the confidence interval will be, this is the mean value which we know and which is equal, which is same for both calculating the confidence interval and prediction interval plus minus t alpha by 2 n minus 2 degree of freedom and then standard error of these which you have already expressed here.

So, if we calculate these that will give you that blue dotted line which we have seen in our previous slide, this blue dotted line which we have seen in our previous slide. So, this is the confidence interval of mean. Now, the standard error of these term depends on the value of x0, the number of samples are also the sigma hat or estimate of the standard deviation.

(Refer Slide Time: 7:06)




So, if we see the Forbes data example, if we consider the Forbes data example, the log of pressure versus boiling point and this is the scatter plot. Consider the following two cases, estimate the mean of log pressure when the boiling point is at 200 degree Fahrenheit and provide a 95 percent confidence estimate. So, in this case, since our data set encompasses these values, that means, it starts from somewhat around 190 and it goes up to 200, around 250 or something like that. So, for 200 which comes in between this range, this will be an example of interpolation.

However, if we want to estimate the mean log pressure when the boiling point is at 220 degrees Fahrenheit and provide a 95 percent confidence interval, that would create the problem, because in that case, there will be extrapolation. Now, why is it extrapolation? Because, this 220 does not come under this our data set data range. So, these 220 is somewhere here and that is why it will show the extrapolation.

(Refer Slide Time: 8:30)

## INTERPOLATION VS EXTRAPOLATION

- ▶ Estimating the mean of  $Y$  given  $X = x_0$  when  $x_0$  is in the range of the data is called **interpolation**
  - ▶ This is safe to do! ✨
- ▶ Estimating the mean of  $Y$  given  $X = x_0$  when  $x_0$  outside the range of the data is called **extrapolation**
  - ▶ This is dangerous!
- ▶ **Design** your experiments so that you always have data near to where you want to predict the mean!




Now, estimating the, so, what is the difference between interpolation and extrapolation you know. So, estimating the mean of  $Y$  given  $X$  equal to  $x_0$  is in the range of the data is called the interpolation, which we have seen, this is always safe to do, because you have already developed the model based on the data range, your calibration model has been developed based on the data range or estimating the mean of  $Y$  given  $X$  equal to  $x_0$  when  $x_0$  outside the range of the data is called the extrapolation.

Remember, extrapolation in case of linear model is a very, very dangerous issue. So, you should be very very careful, while extrapolating your value and you should interpret it very very cautiously. So, design your experiment so, that you always have data near to where you want to predict the mean. So, you should not extrapolate when you go for modeling any data or making a regression equation of  $Y$  versus  $X$ .

(Refer Slide Time: 9:34)

## SIMULTANEOUS CONFIDENCE INTERVALS

- ▶ The CIs for the mean that we have produced so far are valid only at one value
  - ▶ In the language of ANOVA, think "individual error rate"
- ▶ What about for "familywise error rate"?
  1. **Bonferroni** approach: to produce a simultaneous  $100(1 - \alpha)\%$  confidence band for the mean of  $Y$  valid at  $K$  different  $X$ -values  $(x_1, \dots, x_K)$ 
$$\hat{\mu}_{Y|X=x_k} \pm t_{(n-2, 1-\alpha/2K)} S.E.(\hat{\mu}_{Y|X=x_k})$$
  2. **Scheffé** approach: to produce a simultaneous  $100(1 - \alpha)\%$  confidence band for all values of the mean of  $Y$ , valid at all the observed  $X$ -values
$$\hat{\mu}_{Y|X=x_0} \pm \sqrt{2 \times F_{(2, n-2, 1-\alpha)}} S.E.(\hat{\mu}_{Y|X=x_0})$$



Now, we know that what is, we know that the confidence interval for the mean that we have produced so far are valued only at one value  $X$  equal to  $X_0$ , but in the language of ANOVA think individual error rate, if we consider the individual error rate. So, what about, so what about for familywise error rate? Here we are considering  $X$  equal to  $X_0$  but what about the familywise error rate?

So, there are two approaches through which we can address this familywise error rate. First approach is known as the Bonferroni approach, where we want to produce a simultaneous confidence band for the mean of  $Y$  valid at  $K$  different  $X$  values. So, there are certain definite number of  $X$  values which are  $X_1$  up to  $X_K$ . So, this is how you calculate based on the Bonferroni approach.


However, there is another approach that is called Scheffé approach. So, Scheffé approach is aimed to produce a simultaneous prediction, simultaneous confidence band for all values in the mean of  $Y$ . So, it considers all values of the mean of  $Y$ , so valid at all the observed  $X$  values. So, this is the difference between the two approaches, but at the, you should remember that when you design your model you should avoid the extrapolation because that may create some problem.

(Refer Slide Time: 11:06)

## PI CALCULATION

- ▶ The predicted value of Y for a new case  $X = x_0$  is
 
$$\text{Pred}(Y|X = X_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

This value is the same as the predicted mean of Y at  $X = x_0$
- ▶ The **prediction error** (not standard error) associated with this value is
 
$$\begin{aligned} P.E.(\text{Pred}(Y|X = X_0)) &= \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{(n-1)S_X^2}} \\ &= \sqrt{\hat{\sigma}^2 + [S.E.(\hat{\mu}_{Y|X=x_0})]^2} \end{aligned}$$
- ▶ A  $100(1 - \alpha)\%$  prediction interval for the value of Y given  $X = x_0$  is:
 
$$\text{Pred}(Y|X = X_0) \pm t_{\alpha/2, n-2} P.F. (P.E.(Y|X = X_0))$$



So, now, we have seen the how to calculate the confidence interval. Now, how to calculate the prediction interval or PI. So, the predicted value of Y for a new case X equal to  $x_0$  is basically you can calculate by  $\beta_0$  plus  $\beta_1$  hat  $x_0$ . So, these value is the same as the predicted mean of a Y that is X equal to  $x_0$ . So, these terms is same for both confidence interval and prediction interval.

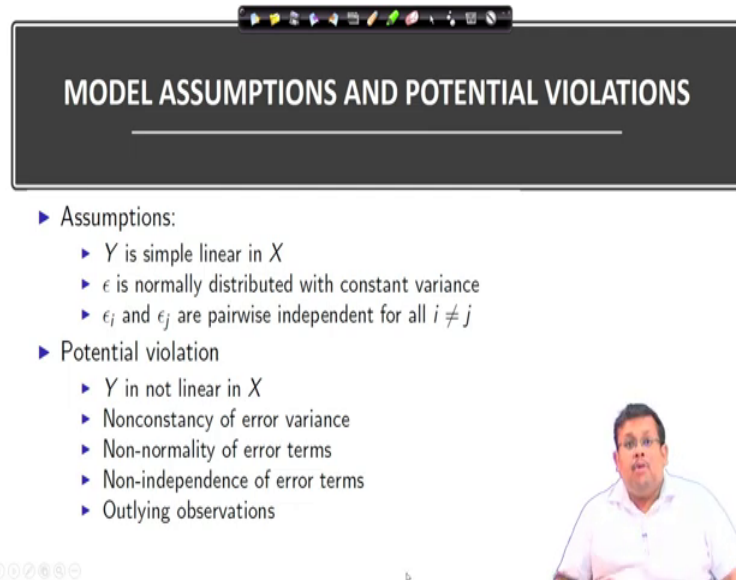
However, the only the difference is the in case of calculating the confidence interval we calculated the standard error, but here for calculating the prediction interval, we are going to calculate the prediction error. So, the prediction error is not equal to the standard error. In case of confidence interval of mean we can calculate the standard error, but in case a prediction interval calculation we have to calculate the prediction error.

So, prediction error is associated, prediction error we can calculate by using this formula where the sigma hat equal to  $1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{(n-1)S_X^2}$  square. So, basically the sigma hat square plus standard error of this term. So, if we want to develop a prediction interval for a given value of alpha, we can have these, we can have this is basically for a value of X equal to  $x_0$  this is basically the summation of this term.

Now, remember these term is equal in both the confidence interval as well as prediction interval. However, the second term is different in case of, this is the second term which is different in case of prediction interval calculation.




(Refer Slide Time: 13:32)



**MODEL ASSUMPTIONS AND POTENTIAL VIOLATIONS**

- ▶ Assumptions:
  - ▶  $Y$  is simple linear in  $X$
  - ▶  $\epsilon$  is normally distributed with constant variance
  - ▶  $\epsilon_i$  and  $\epsilon_j$  are pairwise independent for all  $i \neq j$
- ▶ Potential violation
  - ▶  $Y$  is not linear in  $X$
  - ▶ Nonconstancy of error variance
  - ▶ Non-normality of error terms
  - ▶ Non-independence of error terms
  - ▶ Outlying observations

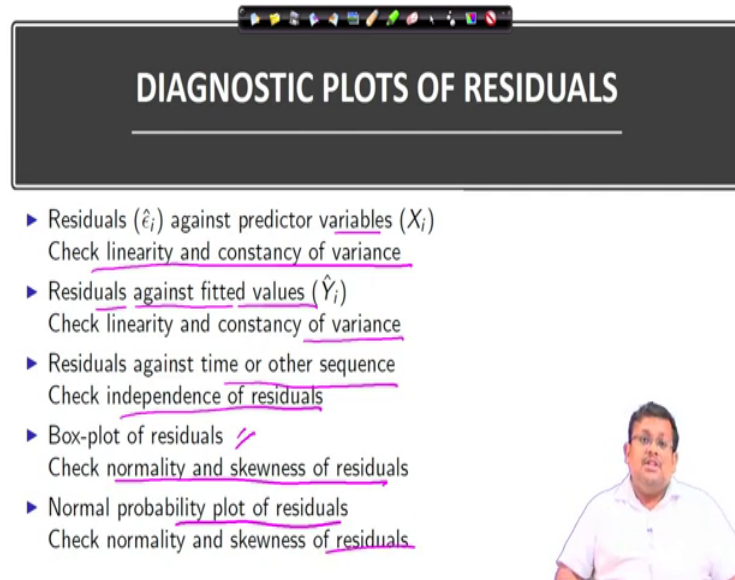


So, let us move to the, so, we have now seen the difference just. So, we have seen the difference between confidence interval and prediction interval and let us see the model assumption and the potential violation of the model assumption. So, we know that there are assumptions that is  $Y$  is a simple linear relationship with  $X$  and then the error term is normally distributed with a constant variance. And then, two error terms for two observations are pairwise independent for all  $i$  not equal to  $j$ .


That means if these are two different observations, there are terms of those two different observations are independent to each other. In other words, the observations are independent to each other. So, what are the potential violations in these observe, in this, in these assumptions. First of all, you will see that  $Y$  is not linear to  $X$  that is the one potential violation. Second is non-constancy of error variance.


Now, constancy of error variance is also known as homoscedasticity. However, when there is a non-constancy of error variance that is called heteroscedasticity. So, heteroscedasticity violates the assumption of simple linear regression. And third is the non-independence of error terms is also another important violation. And finally, outlier observations are also potential violation.

(Refer Slide Time: 14:52)



## DIAGNOSTIC PLOTS OF RESIDUALS

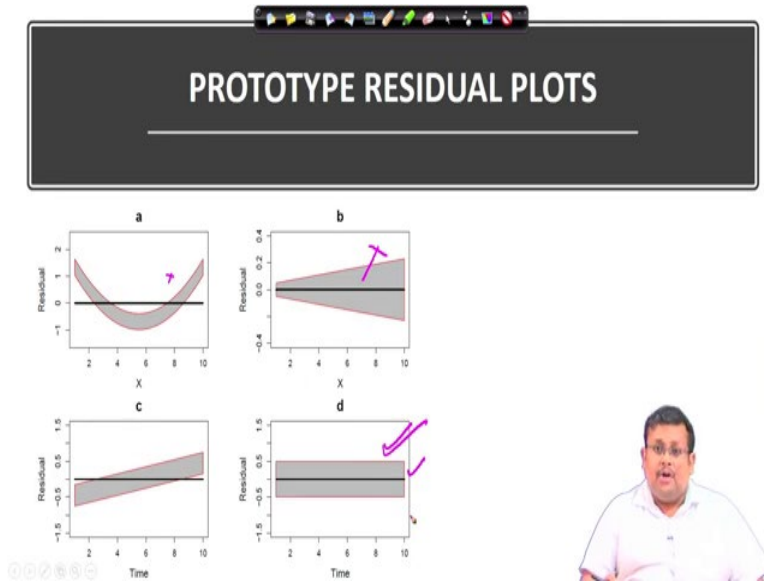
- ▶ Residuals ( $\hat{\epsilon}_i$ ) against predictor variables ( $X_i$ )  
Check linearity and constancy of variance
- ▶ Residuals against fitted values ( $\hat{Y}_i$ )  
Check linearity and constancy of variance
- ▶ Residuals against time or other sequence  
Check independence of residuals
- ▶ Box-plot of residuals   
Check normality and skewness of residuals
- ▶ Normal probability plot of residuals  
Check normality and skewness of residuals



So, diagnostic plot of residuals you will see that the residuals when the residuals that is  $E_i$  are against the predict, if you plot against the, if you plot the residuals against the predicted variables, you see, you should check the linearity and constancy of variant, you can also plot the residuals against the fitted values that is  $\hat{Y}$  and then check the linearity and constancy of variants.

Third, you want to plot the residuals against the time or other sequence then check the independency of the residuals. Then, you can do some box plots of the residuals and check normality and skewness of the residuals and then you can do some normal probability plot of residuals and check normality and skewness of the skewness of residuals. So, these are the some diagnostic plots you can check for seeing whether the assumptions has been maintained or they have been violated. We will see some example.

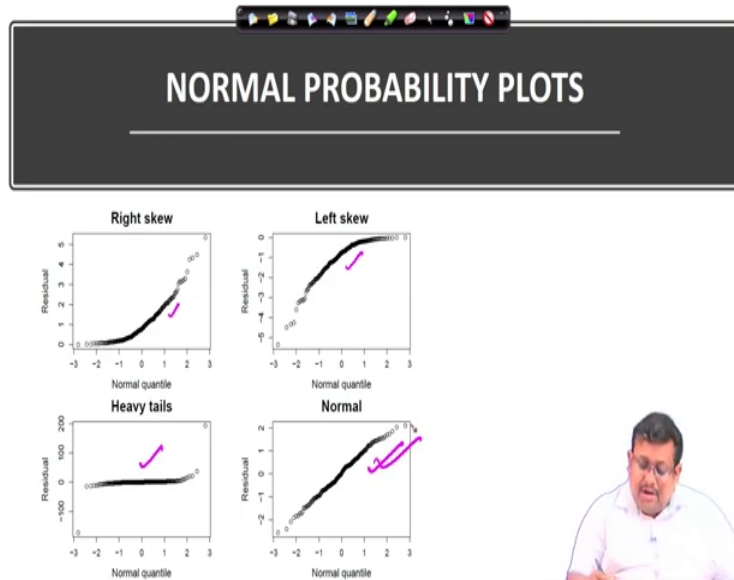
(Refer Slide Time: 15:54)



So, here you can see there are some prototype residual plot, here you can see if you are plotting the residual against the X, we will see that non-linearity of the distribution of the residuals. So, this is not okay, and this is here also you can see this is not constant error variance, there are these, in this way there are tapering, so, this is called heteroscedasticity. So, this heteroscedasticity is also not good.

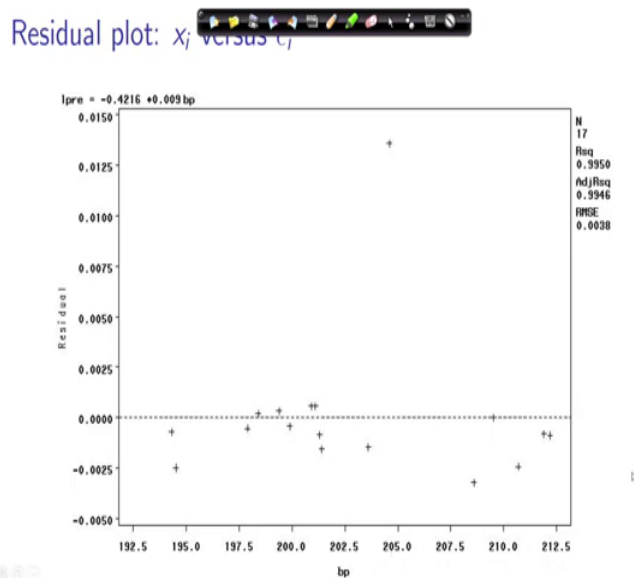
Here also you can see some trends, but here you can see the residuals versus time they are distributed evenly along the 0 line and that shows the accepted feature or accepted residual diagnostic plot. So, all the three cases we do not see any, they are violating our assumption, but here they are maintaining our or they are supporting the assumption where we are plotting the residual against the time, we can see both linearity as well as the constancy of the variance as well as we do not see any type of nonlinear trend. So, this is how you see the prototype residual plots.

(Refer Slide Time: 17:15)



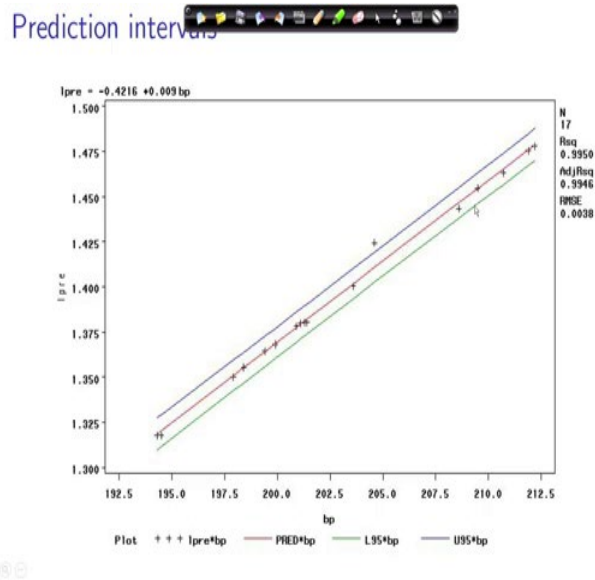
And now let us see some normal probability plots. Normal probability plots of the residual if you see the normal probability plot of the residuals if your residuals are right skewed, then you will see this kind of distribution, if they are left skewed, you will see this type of distribution. If they have heavy tails, you will see this type of normal probability plots, but the if they are normally distributed, you will see this type of observer, use this type of normal probability plot. So, these shows maintaining the assumption of linear regression. However, in this three condition, they are not maintaining the assumption of linear regression.

(Refer Slide Time: 18:01)



For Forbes example, you can see some residual plot here, predicted value versus residual, here you can see, also it is residual plot  $X_i$  versus  $E_i$  for each values of the independent variable you can see how these residual plots are varying.

(Refer Slide Time: 18:25)



And then you can also see the prediction interval and prediction interval and, so, this is the lower limit, this green line shows the lower limit of the prediction interval and this blue line is showing the upper limit of the prediction interval and this is basically 95 percent prediction interval and this red line is showing the mean line for prediction.

(Refer Slide Time: 18:51)

### MULTIVARIATE CALIBRATION

- Used to develop mathematical models that allow to predict a continuous  $y$  from variables  $x_1, \dots, x_m$ .
- Predicting soil organic carbon by near infrared spectra
- Predicting soil CEC by PXRF reported elements

Now, we have discussed the prediction interval and calibration and confidence interval. Now, we need to know what is multivariate calibration. So, we know now, what is simple linear regression and what are the features of simple linear regression, how to teach the assumption of simple linear equation. Now, we want to discuss, we can, we start discuss the multivariate calibration.

Now, multivariate calibration are used to develop mathematical models that allow us to predict a continuous Y from variable X values. So, the simple linear regression and multiple linear regression the only difference is in case of multi linear regression, there are more than 1x variable.

However, so, multivariate calibration depends on predicting a continuous value or continuous variable Y using the variable  $X_1$  to  $X_m$ . So, predicting soil organic carbon for example, predicting soil organic carbon by near infrared spectra or predicting cation extent capacity of the soil by using the portable XRF instrument reported elements. So, these are some examples of multivariate calibration. And in our coming weeks, we are going to discuss them in details with examples.

(Refer Slide Time: 20:13)

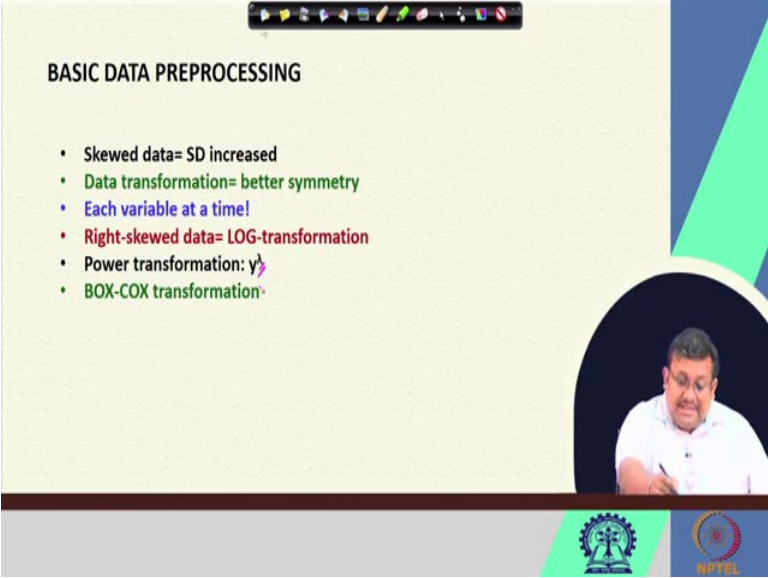
The slide is titled "MULTIVARIATE CLASSIFICATION" and contains a bullet point: "Classification/Clustering of samples into homogenized groups". The main content is a decision tree diagram with nodes and branches, and a bar chart below it. A small video inset shows a man speaking. The slide has a green and blue background with logos at the bottom.

Also multivariate classifications are there. When our target are several classes instead of continuous variable, then we can say it is a multivariate classification. Now, in case of, here you can see an example of classification and regression tree which is an important data mining or machine learning approach, where we try to classify, where we are trying to classify different

types of soil using the different elements, elemental values like, here you can see we are using zinc, potassium, zirconium, lead, manganese, copper and we are developing some rules, nonlinear rules to classify the soil samples into three groups, forest soil sample, converted soil samples and soil samples, which are coming from the agricultural fields.

So, these three are our final target clusters, and we want to classify the cell sample based on these input parameters, which are the different elemental content. So, this is an example of multivariate classification because, here more than one feature we are using and then we are using them to classify our target value. So, this is an example of multivariate classification.

(Refer Slide Time: 21:39)



The slide is titled "BASIC DATA PREPROCESSING" and contains the following bullet points:

- Skewed data= SD increased
- Data transformation= better symmetry
- Each variable at a time!
- Right-skewed data= LOG-transformation
- Power transformation:  $y^p$
- BOX-COX transformation

The slide also features a video inset of a man in a white shirt speaking, and logos for IIT Bombay and NPTEL at the bottom.

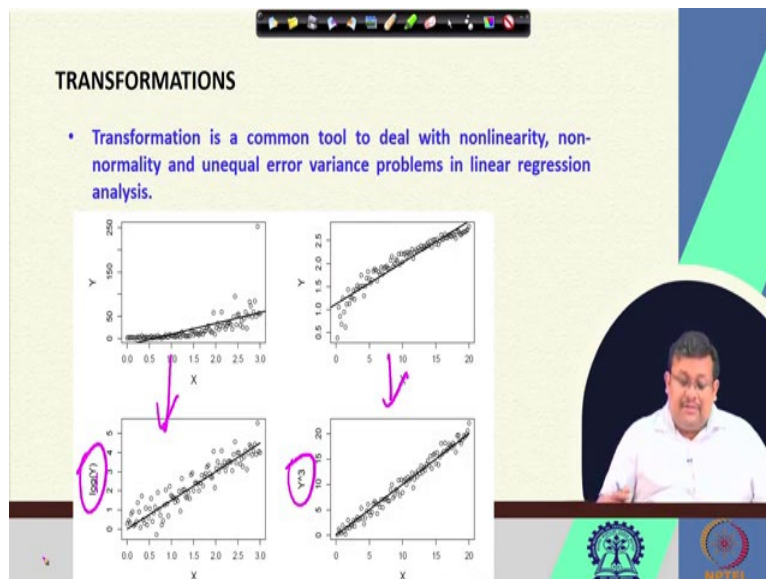
Now, let us see some basic data processing thing, which are generally basic data processing approaches which are required for handling the big data. So, generally when there is a Skewness in the data, I have already showed you when the residuals are highly skewed, that creates problem. So, when your data is skewed the in that case the standard deviation generally increased.

And so, in that case, we require data transformation for maintaining the better symmetry and further, and for standardization you cannot do the standardization of all the variables in a single time simultaneously. You have to do the standard deviation, you can, you have to do the standardization of each of the variable at a time and you can see, you have to see which one of

them are performing best. So, you have to try a couple of standardization technique and then you had to select the one which is giving the best result.

If there is a right skewed data you go with the log transformation and sometime there are some power transformation, log transformation is also power transformation. So, power transformation is denoted by  $Y$  power lamda, the power transformation is denoted by  $Y$  to the power lambda and the Box-Cox transformation is another important transformation which we are going to discuss. And Box-Cox transformation is also widely used in different data processing approaches.

(Refer Slide Time: 23:30)



Now, transformation, now transformation is a common tool to deal with nonlinearity, because and also non normality and unequal variance problem in linear regression analysis. So, here you can see that, here the data is kind of not very, this is an X versus Y plot, scatter plot and we can see that in this case, we cannot see very prominent linear relationship. However, when you take the log values of this Y, we can take that this linear relationship is more prominent. So, you can see we are making a transformation of the data here.

Similarly, also here you can see this is Y versus X is another data set and we can see some kind of nonlinear trend. However, when we are taking the Y cube value, then we can see there are linearly distributed. So, linear, there is a linear relationship between Y cube and X. So, this is also a data transformation.



So, you can see this is a power transformation, this is a log transformation and this type of transformations are helpful for dealing with some data where these linear relationship is not readily perceivable.

(Refer Slide Time: 25:00)

**BOX-COX TRANSFORMATIONS**

- ▶ Box-Cox procedure automatically identifies a power transformation on  $Y$ .
- ▶ Power transformation on  $Y$ :  $Y^* = Y^\lambda$  ( $\lambda$  is a parameter determined by data.) The power family includes:  $Y^2$ ,  $\sqrt{Y}$ ,  $1/Y$ , and  $1/\sqrt{Y}$ . Note that when  $\lambda = 0$ ,  $Y^*$  is defined to be  $\log(Y)$ .
- ▶ Box-Cox procedure
  - ▶ Given  $\lambda$ , "standardize" the  $Y_i$ 's to be  $W_i$

$$W_i = \begin{cases} K_1(Y_i^\lambda - 1) & \lambda \neq 0 \\ K_2 \log(Y_i) & \lambda = 0 \end{cases}; K_2 = \left( \prod_{i=1}^n Y_i \right)^{\frac{1}{n}}, K_1 = \frac{1}{\lambda K_2^{\lambda-1}}$$

- ▶ Apply simple linear regression on  $W_i$  (new response) and  $X_i$
- ▶ Find the "optimal"  $\lambda$  which minimizes  $SSE$

So, what is Box-Cox transformation? Box-Cox transformation is a Box-Cox procedure automatically identifies the power transformation of  $Y$ . So, power transformation of  $Y$  generally  $Y$  to the power  $\lambda$  where  $\lambda$  is a parameter determined by the data itself. So, the power family includes generally  $Y$  square, root over of  $Y$ , then  $1/Y$  and then  $1/\sqrt{Y}$ .

And remember that when the  $\lambda$  value is 0 that means, when the  $\lambda$  value is 0 this  $Y^*$  is defined to be  $\log$  of  $Y$ . So, this is a special case of power transformation, this  $\log$  transformation is a special case of power transformation where  $\lambda$  equal to 0.

Now, in case of Box-Cox procedure given  $\lambda$  standardized this  $Y$  and  $\lambda$  to be  $W_i$ , so,  $W_i$  takes this value  $K_1$  multiplied by  $Y_i$  to the power  $\lambda$  minus 1 where  $\lambda$  equal, not equal to 0, where  $\lambda$  equal to 0 you can just directly take  $K_2 \log$  of  $Y_i$ , where  $K_2$  is basically stands for  $1$  to  $n$   $Y_i$  to the power  $1/n$  and whereas,  $K_1$  stands for  $1$  by  $\lambda K_2^{\lambda-1}$ .

So, this is how you standardize your variable based on whether your  $\lambda$  is 0 or not and then you can use them, this is called the Box-Cox transformation, these is a generalized formula of Box-Cox transformation and so, you have to apply the simple linear regression on this  $W_i$ , this is

a new response or transform response and  $X_1$  which are the inputs and find the optimum lambda which minimizes the sum square error.

So, you can target for 0 that is okay. But for not equal to 0 you can try with many values and then you plot the values of  $W_i$  with against the sum square, the values of lambda against the sum square error and you can see which one is giving the minimum sum square error and that will be the optimum value of lambda. I will show you one example.

(Refer Slide Time: 27:38)

**POWER TRANSFORMATION**

**Normal Probability Plot**

The normal probability plot (Chambers et al., 1983) is a graphical technique for assessing whether or not a data set is approximately normally distributed. The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality.

The normal probability plot is a special case of the probability plot.

<https://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm>

NPTEL

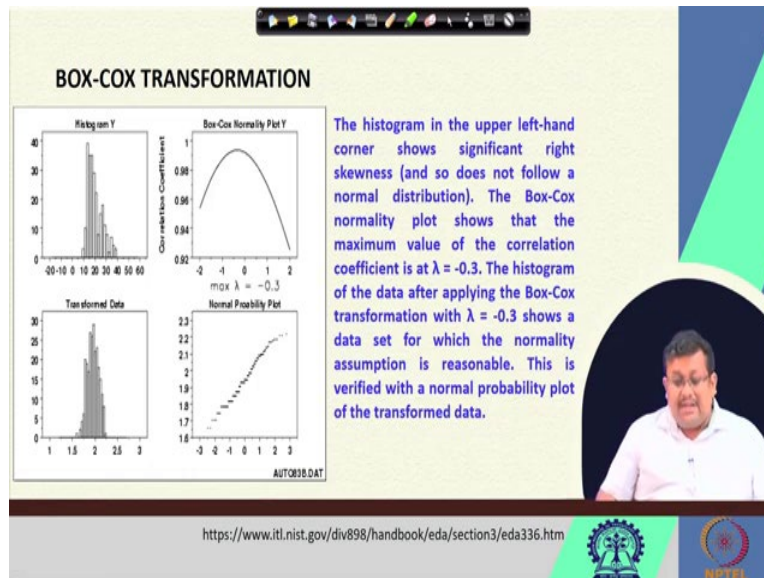
So, you can see that the normality assumptions are critical for many univariate interval and hypothesis tests. And it is important to test the normality assumption, so if the data are in fact clearly not normal. The Box-Cox normality plot can often be use to find the transformation that will approximately normalize the data.

So, you can see here this is a normal probability plot. So, generally in case of normal distribution, you should get this type of plot, but if you are not getting this type of plot in your data, you should try some kind of transformation. This normal probability plot given by this Chambers et al in 1983 is a graphical technique for assessing whether or not a data set is approximately normally distributed.

So, the data are plotted against a theoretical normal distribution in such a way that the point should follow, the point should form an approximate straight line, as you can see, they are

forming an approximate straight line and departures from the straight line indicate the departures from the normality. And the normal probability plot is a special case of probability plot.

(Refer Slide Time: 28:52)



So, you can see that the histogram in the upper left, our left hand corner shows the significant right skewness. So, you can see significant right skewed data here. So, and so, does not, it does not follow a normal distribution. So, the Box-Cox normality plot shows that the maximum value of the correlation coefficient is at lambda equal to minus 0.3. So, at maximum value of lambda you can get at the value of minus 0.3. So, this is another way of selecting the optimum lambda. So, you can select this optimum lambda based on the correlation coefficient.

So, the histogram. So, once you take the value of lambda equal to minus 0.3 then you could transform the data and now you can see they are following the normal distribution. So, we are getting in reasonable normal distribution. So, this is verified with a normal probability plot of the transformed data you can see here.

(Refer Slide Time: 29:57)

### CENTERING AND SCALING DATA

- Most straightforward data transformation
- It is always necessary to standardized data before processing
- Centers and scales a variable to mean 0 and SD 1

[https://subscription.packtpub.com/book/big\\_data\\_and\\_business\\_intelligence/9781785889622/3/ch03w1sec24/data-scaling-and-normalization](https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781785889622/3/ch03w1sec24/data-scaling-and-normalization)

Centering and scaling and centering of the data, it is the most straightforward data transformation procedure, and it is always necessary to standardize data before processing and center and scales, center and generally scaling a variable to mean 0 and standard deviation 1, it is basically very much needed for a scattered data. And you will see in case a principal component analysis what which we are going to discuss in our next week, we are going to use this centering and scaling extensively.

(Refer Slide Time: 30:28)

### BRAIN WEIGHT EXAMPLE

- ▶ The data set consists of brain weights (g) and body weights (kg) for 62 species of mammals. Three questions are of interest.
  - ▶ Any general pattern between brain and body weight across a number of species.
  - ▶ Is brain weight proportional to body weight? Are there any unusual species?
  - ▶ Do humans have unusually large brains given our body size?
- ▶ Sample data

Species (common name)	Body weight (kg)	Brain weight (g)
Cat	3.30	25.60
Pig	192.00	180.00
African Elephant	6654.00	5712.00
Kangaroo	35.00	56.00
Human	62.00	1320.00
Ground Squirrel	0.10	4.00

[https://subscription.packtpub.com/book/big\\_data\\_and\\_business\\_intelligence/9781785889622/3/ch03w1sec24/data-scaling-and-normalization](https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781785889622/3/ch03w1sec24/data-scaling-and-normalization)

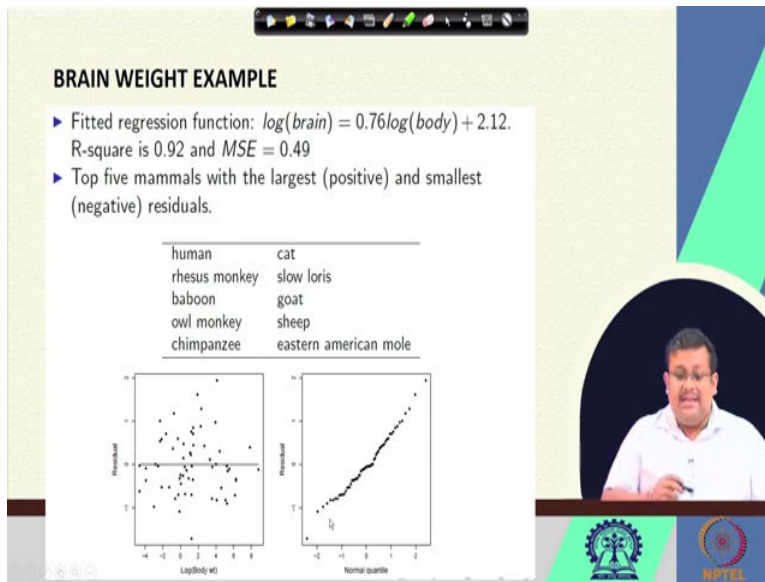
There are some brain weight example. So, the data set consists of brain weights and the body weights of 62 species of mammals. So, three questions are of interest any general pattern between brain body weight across a number of species or is brain weight proportional to the body weight? Are there any usable species? And do humans have usually large brain given our body size? So, these are some of the data, sample data.

(Refer Slide Time: 30:57)



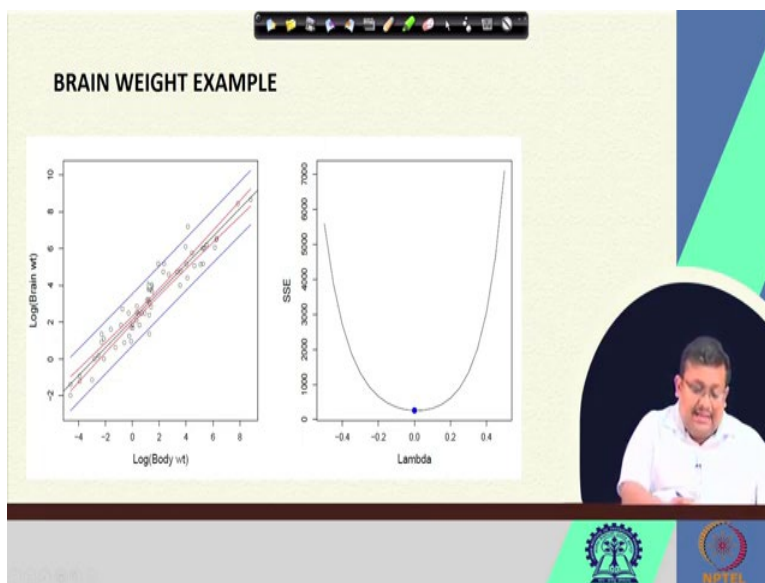
And we can see this is the brain weight versus body weight and then brain weight versus log of body weight. And you can see log of brain weight versus log of body weight which is where we are getting linear relationship. So, we should take this and then we are taking the residual versus fitted value they are meeting the assumption of linear regression.

(Refer Slide Time: 31:22)



And also the, so, the fitted regression line is log of brain equal to 0.76 log of body plus this intercept 2.1. R square value is 0.92 and MSE value is 0.49. And top five mammals with the largest and the smallest residuals we can see here. Now, this is the residual diagnostic plot and you can see the normal plot, normal quantile plot which is satisfying the assumption of linear regression.

(Refer Slide Time: 31:48)



And you can see that if we are taking different values of lambda at the values of 0 we are getting the least sum square error. That means, at the log transformation if you are doing the log

transformation, we are getting the linear regression relationship. And this is the prediction interval, this is the confidence interval of mean and this is the prediction interval we can see here.

(Refer Slide Time: 32:13)

**STANDARDIZED RESIDUALS**

- ▶ Properties of residuals
  - ▶ Mean:  $\sum \hat{e}_i = 0$
  - ▶ Variance:  $\sum \hat{e}_i^2 / (n - 2) = SSE / (n - 2) = MSE$
- ▶ Standardized or Semistudentized residuals:
$$\hat{e}_i^* = \hat{e}_i / \sqrt{MSE}$$
- ▶ In brain weight example, human has residual 1.944, while  $\sqrt{MSE} = 0.70$ . The standardized residual is 2.78, which is greater than 2. Hence humans do have unusually large brains adjusted by our body size.

Standardized residuals is the final thing, it is the properties of the residual. So, you can see that mean is summation of  $E_i$  equal to 0 and then variance it is, it can be calculated by this MSE, sum square error by  $n$  minus 2, MSE. So, standardized or semistudentized residuals can be calculated by this  $E_i$  hat by root over of MSE.

So, in brain example, human has a residual of 1.944, while root over of MSE is 0.70. So, we can see that the standardized residuals, if we calculate, this standardized residuals it will be greater than 2. So, humans have usually large brain adjusted to our body size.

(Refer Slide Time: 32:59)

**MLR**

- More than one predictor...

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon$$

where, for  $i = n$  observations:

- $y_i$  = dependent variable
- $x_i$  = explanatory variables
- $\beta_0$  = y-intercept (constant term)
- $\beta_p$  = slope coefficients for each explanatory variable
- $\epsilon$  = the model's error term (also known as the residuals)

The slide includes a video inset of a presenter and logos for IIT Bombay and NPTEL at the bottom.

Final slide, this MLR, so where there are more than one predictor you can see here. Instead of single predictor, so here you can see we are, it is the intercept and here  $x_{i1}$ ,  $x_{i2}$ , up to  $x_{ip}$  and then the error term, where for  $i$  equal to  $n$  number of observation,  $y_i$  is the dependent variable,  $x_i$  is the explanatory variable you know that,  $\beta_0$  is the  $y$  intercept,  $\beta_p$  is the slope coefficient for each explanatory variable, and  $\epsilon$  is the model's error term also known as the residuals.

(Refer Slide Time: 33:40)

**MLR**

Each regression coefficient is the amount of change in the outcome variable that would be expected per one-unit change of the predictor, if all other variables in the model were held constant.

The slide includes a video inset of a presenter and logos for IIT Bombay and NPTEL at the bottom.

So, this is the difference between simple linear regression and multiple linear regression. In case of multiple linear regression, each regression coefficient is the amount of change in the outcome



variable that would be expected per one unit change of the predictor, if all other variables in the model were held constant. We will discuss more about this in our next lecture.

(Refer Slide Time: 33:57)



**REFERENCES**

- [https://subscription.packtpub.com/book/big\\_data\\_and\\_business\\_intelligence/9781785889622/3/ch03lv1sec24/data-scaling-and-normalization](https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781785889622/3/ch03lv1sec24/data-scaling-and-normalization)
- <https://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm>

The slide features a navigation toolbar at the top with various icons. In the bottom right corner, there is a circular video feed showing a man in a white shirt speaking. At the bottom of the slide, there are two logos: the Indian Institute of Technology (IIT) logo on the left and the NPTEL logo on the right.

So, these are some of the references which are used and I hope that you have learned something new. Let us meet in our next lecture to discuss from here and see more diagnostic features of multivariate data analytics. Thank you guys, let us meet in our next lecture.