

Machine Learning for Soil and Crop Management
Professor. Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology, Kharagpur
Lecture 07
Basics of Multivariate Data Analytics (Contd.)

Welcome students to this second lecture of week 2 or in other words there are seventh lecture of this NPTEL online certification course of Machine Learning for Soil and Crop Management. And in this week, we are discussing the basics of multivariate data analytics. In our last lecture, we have discussed about the multivariate data, what is multivariate data, and what is data matrix, and what is, what are the different kinds of representation of the data like means multi-dimensional data and how we can represent the multi-dimensional data.

Also we have seen the associations between multiple variables, or features in terms of correlation. Also we have learned what is the correlation coefficient, what are their values what are their features. We have learned about positive correlation, negative correlation, what is covariance. Also we have seen the simple linear regression.

(Refer Slide Time: 1:55)

The image shows a presentation slide with a dark blue header containing the text "CONCEPTS COVERED". Below the header, there is a bulleted list of three items: "SLR" (in blue), "CI and PI" (in red), and "Diagnostic plots of residuals" (in green). The slide is part of a video lecture, as evidenced by the video inset of Professor Somsubhra Chakraborty in the bottom right corner. The slide also features the IIT Kharagpur logo and the NPTEL logo at the bottom.

So, in this lecture, we are going to discuss this following concept. First of all we are going to see in details about different aspects of simple linear regression. And then we are going to learn what is confidence interval and prediction interval. And what is the difference between confidence

interval and prediction interval in case of SLR. And also we are going to see some of the diagnostic plots of residuals, based on the assumption of simple linear regression.

(Refer Slide Time: 2:30)

KEYWORDS

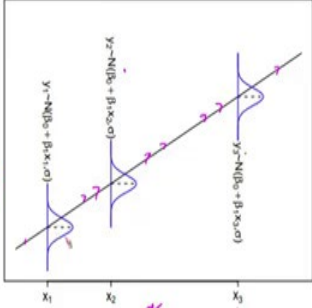
- SLR assumptions
- Confidence Interval
- Prediction Interval
- Least Square Estimate
- Slope and Intercept

So, these are the keywords, which we are going to discuss today. First of all the assumption of simple linear regression, also we are going to learn what is confidence interval, prediction interval. And then we are going to learn what is least square estimate? And then we are going to also discuss the slope and intercept of the simple linear regression.

The reason for discussing this, because unless we understand this features of SLR, we cannot understand the multiple linear regression and different types of pitfalls of multiple linear regression.

(Refer Slide Time: 3:16)

SIMPLE LINEAR REGRESSION



- ▶ Let Y be the response variable, and X be the explanatory variable
- ▶ The **simple linear regression** (SLR) model assumes that mean of Y given (a single) X , is a straight line:
$$\mu_{Y|X} = \beta_0 + \beta_1 X \quad \text{and} \quad Y = \mu_{Y|X} + \epsilon$$
- ▶ If we assume that the error ϵ is normally distributed with mean 0 and variance σ^2 then Y is also normally distributed
$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$
- ▶ Think of each value of X generating a different subpopulation.

So, let us start with the assumption of simple linear regression. We have already seen the discussion you know, we have already discussed in our previous lecture, what is the difference between a correlation and regression. Remember in case of correlation we try to see the you know linear relationship between two unlabeled variables and it does not depend and the correlation does not depend on the unit of measurement of the variables.

However, in case of regression the regression generally identifies one dependent variable, another independent variable and we have discussed this dependent and independent variable. So, let Y be the dependent variable or response variable, and X be the explanatory variable. So, if you can see this plot, this plot is showing the Y versus different levels of X .

So, the this is an example of simple linear regression, simple linear regression or model, where Y is the target variable and X is the predicted variable. And the relationship between the Y and X is linear or in other words the model assumes that the mean of Y of given X is a straight line, as you can see here, this is a straight line, and this straight line basically are the mean of Y given different levels of X .

So, this is basically represent by this mu of this mean that is mean of Y given values of X . So, we can represent its as $\beta_0 + \beta_1 X$, where β_0 is the Y intercept and β_1 is the slope, you know the equation of a straight line. So, it really resembles the equation of a straight line and

so this mean line can be represented by this equation, whereas the actual observation can be considered as the mean plus an error, or residual.

Because actual observation can occur anywhere and that can differ from this mean regression line and that is why the actual observation can be considered as a summation of both error as well as the mean of Y given different values of X . So, if we assume that the error, which is defined by this η is normally distributed with mean 0 and variance σ^2 , then Y is also normally distributed.

So, here you can see that differ at different levels of X , like X_1 , X_2 , and X_3 , you can see the corresponding values of Y_1 , Y_2 and Y_3 and so we are assuming that the error term is normally distributed and as a result, this whole Y term can also be considered as normally distributed with you know with a mean of $\beta_0 + \beta_1 X$, we have already know this line and also with a standard deviation of σ .

So, think of each values of X generating a difference of population. So, you can see that you can resemble this condition as each value of X is generating a difference of population. So, this is how we represent the simple linear regression between X and Y . Again here Y is the target variable and X is our predictor variable at different values of X , we are getting different values of Y . Let us assume that these are Y_1 , Y_2 and Y_3 .

So, how we can get this Y_1 , Y_2 , and Y_3 , to you know if we if we draw a model, linear model, this linear model assume that this line corresponds to the mean of Y given different values of X and then we just add, or subtract the error values, or residuals to get the actual observed Y . Now, this Y is normally distributed and since we are assuming that the this you know this error is also normally distributed. So, naturally the Y will be also capital Y will be also normally distributed.

(Refer Slide Time: 8:31)

SLR: ASSUMPTIONS

1. (Independent observations) The n responses Y_1, \dots, Y_n are independent, given the levels of the explanatory variable, X_1, \dots, X_n , respectively
2. (Linearity of the mean) The mean for the response, given the level of the explanatory variable, is linear; that is,
$$\mu_{Y|X} = \beta_0 + \beta_1 X$$
3. (Constant variation) The variance of the response, given the level of the explanatory variable, is σ^2 . This is true for all values of X
4. (Normality) The distribution of the response, given the level of the explanatory variable, is normal

So, let us move to the next slide, this slide shows the basic assumptions of simple linear regression, you can see you know there are independent observation like X_1, X_2, X_3 and the n responses are there suppose there are up to X_n . So, these n responses are Y_1, Y_2 up to Y_n are also independent. So, independent observations are this X_1, X_2 , and X_3 . However, Y_1, Y_2 , up to Y_n are dependent variables, given the levels of explanatory variables.

Now, linearity of the mean, that is. So, the first observation says that, the n responses these responses Y_1, Y_2, Y_3 up to Y_n are independent, given different levels of X_1, X_2, X_3 . Again, the first assumption is our independent you know the Y_1, Y_2 and Y_3 are independent given the different levels of explanatory variables. So, this Y_1 does not depend on Y_2 and Y_2 does not depend on Y_3 and so on so forth.

Second is the linearity of the mean, or in other words if we see this, that indicates that individual observations are independent to each other, or in other words if there are 10 observation in this data set, those 10 observations are independent to each other, they are not dependent to each other.

Now, second assumption is linearity of the mean, the mean for the responses we know that this line shows the mean for the responses, given the level of explanatory variable, these are X_1, X_2 , and X_3 , explanatory variable is linear that is $\mu_{Y|X}$, that means given X equal to $\beta_0 + \beta_1 X$

plus beta 1 X. So, this line is showing the linearity and so this line basically shows the mean of the responses given different values of explanatory variable.

The third assumption is constancy of variation. So, the variance of the response given the level of explanatory variable is sigma square and this is true for all the values of X. So, the variance is constant in case of responses the variance is constant. The fourth one is normality, the distribution of the responses given the level of explanatory variable is normal. So, we assume that this Y1, Y2, Y3, up to the Yn depend are they are normally distributed.

So, again what are the four assumption? First four assumption of simple linear regression is all the observations are independent to each other. Second assumption is the mean of Y given the different levels of explanatory variable is linear. Third is constancy of variance, that means the variance of the response given the different levels of explanatory variable is sigma square and this is true for all the values of X.

And finally, the normality of the normal distribution of the response given the level of explanatory variable. So, these are the four major assumptions of simple linear regression and why they are important we will see later on.

(Refer Slide Time: 12:24)

LEAST SQUARE ESTIMATE

- ▶ Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be estimates of β_0 and β_1
- ▶ The fitted value for case i is:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$
- ▶ The residual for case i is

$$\hat{e}_i = Y_i - \hat{Y}_i$$
- ▶ Least square criterion: find the $\hat{\beta}_1$ and $\hat{\beta}_0$ so as to minimize the sum of squared errors (SSE):

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{e}_i^2$$

The diagram on the right shows a scatter plot of data points with a fitted regression line. Vertical lines connect each data point to the regression line, representing the residuals. Handwritten labels include \hat{e}_i for residuals and \hat{Y}_i for fitted values.

Now, here also you can see it is called the this simple linear regression is called also this type of linear regression is also known as the least square estimate. Here, the beta 0 hat and beta1 hat are considered as the estimates of original beta 0 and beta1, you know beta 0 is the intercept of Y

and then β_1 is the slope. So, the fitted value for case i , if we consider case i the fitted value for case i , which is denoted by \hat{y}_i can be considered as $\hat{\beta}_0 + \hat{\beta}_1 X_i$. So, this is a for a particular case and this is the predicted value. So, we are giving this hat.

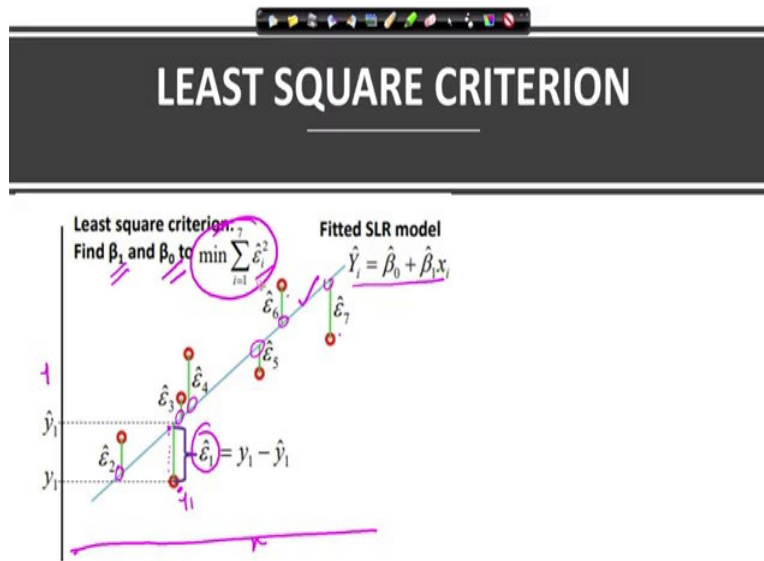
So, the residual for this case, so this is basically showing the mean value, this is showing the mean value, there is no residual here. But, if we subtract this value from the original observation, which is Y_i . So we will get the e_i hat, so or η_i hat, or residual. So, in this observation, if we subtract this, predicted values from the original observed value, then we will get the residual for this observation.

The least square criterion says that, we have to find that β_1 and β_0 , so as to minimize the sum square of error. So, we can see here, if sum square error varies from 1 to where i varies from 1 to n , the sum square error, that means the error term. So, the error term is basically the difference between the original observed values and their predicted values is and then if you take a square. So, this least square estimate gives the least you know least is you know least value for this type of condition.

So, you can see here, we can draw these X and Y relationship in different fashion. Now you can ask why we are sticking to a particular one? The reason for sticking to this particular one is this line is giving I mean for this line, if we draw this least square estimate, or linear regression line, for this linear regression line, we will get the least value if we take the summation of the all the error terms.

So, if you take the summation of the error terms in this case, suppose this is e_a squares, and then suppose this is e_b eta b squares and then you can take, and suppose this is i to n eta i square. So, in this case you can see this expression is giving the least value. So, that is why we are selecting this, we are selecting this line as a least square estimate. So, again guys, this is called the least square line, because this line gives the least square least estimate, or least value of the summation of the total error terms.

(Refer Slide Time: 16:11)

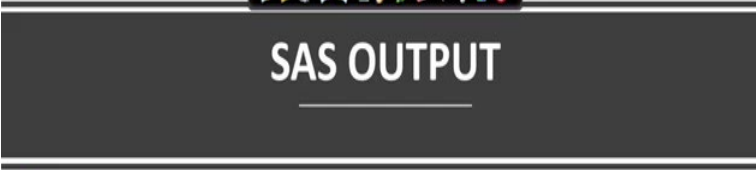


Now, if we move to see these in you know these terms visually, it will be much more clear. So, let us consider this is Y by X and we can see for different values of different values of X will get different values of Y like Y1, Y2, Y3, Y 4.

So, the fitted model is basically this \hat{Y}_1 equal to $\hat{\beta}_0$ plus $\hat{\beta}_1 X_i$. However, these are the error terms like e_1, e_2, e_3 , because if this is the original observation and we consider this Y_1 and its predicted value lies in this line. So, the vertical distance between these two points is the error term, which is denoted by \hat{e}_1 .

So, similarly for all other points, we are getting their corresponding error terms, that is \hat{e}_2 , \hat{e}_3 , \hat{e}_4 , \hat{e}_5 , \hat{e}_6 , \hat{e}_7 . So, if you take the sum of the square of each of this errors for this line will get the minimum value. So, that is why we are selecting this line not any other line. So, our idea is to find this $\hat{\beta}_1$ and $\hat{\beta}_0$. So, that we can get the minimum value of this term. So, for which for the line which gives the line which gives this minimum value of this term is considered as the fitted linear regression line.

(Refer Slide Time: 17:57)



```
proc reg data=forbes;
model lpre=bp;
run;
```

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-0.42164	0.03341	-12.62	<.0001
bp	1	0.00896	0.00016457	54.42	<.0001
Root MSE		0.00379	R-Square	0.9950	
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.04258	0.04258	2961.55	<.0001
Error	15	0.00021564	0.00001438		
Corrected Total	16	0.04279			

Now, if you see the SAS output for this type of regression problem, you can see here again we are using the pros reg and for the Forbes data set, we have already discussed the Forbes data set and our model is to predict the logarithm of pressure with the boiling point. And if we run it, we can see that some important matrices at this point of time I will just focus on 2 or 3.

So, here you can see the intersect values that is beta 0 you can get the value of 0.42 and the bp of you know the estimate value estimated value of bp or boiling point is 0 point the slope of bp, the slope of bp is basically 0.008. And root mean square error is 0.00379, whereas the r square values is 0.99.

So, that shows that this you know there is a very strong relationship between the pressure and barometric pressure logarithm of barometric pressure as well with the with the boiling point and you can see the t value and the probability of the t statistics for bp, which is less than 0.001, that means it is highly significant. So, this shows the (import), this shows the interpretation of the SAS output.

(Refer Slide Time: 19:33)

► Using calculus, the estimated slope of the line is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = r \left(\frac{s_y}{s_x} \right)$$

► Sign of $\hat{\beta}_1$ is the same as the sign of r .

► The estimate of the intercept is

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

► The fitted SLR model pass through the point (\bar{X}, \bar{Y})

Now, using calculus, we are not, we do not have time to discuss it, but using calculus the estimated slope of the line can be, can be calculate as beta1 hat equal to this is the term and ultimately if you simplify it, you will get r, which is the correlation coefficient multiplied by standard deviation of Y by standard deviation of X. So, this is how you get the estimated slope of the line.

So, remember that the sign of beta1 hat is the same as the sign of r. So, if we are having positive slope, positive correlation, then we will get the positive slope, if it is less than, if it is negative that means there is negative slope. So, the estimate for the intercept, we can calculate. Now, once we calculate the estimate for the slope, we can calculate the estimate for the intercept by simply subtracting this beta 1 X bar from the mean of Y. So, X bar is basically the mean of X values, whereas Y bar is the mean of Y values.

So the, remember that the fitted SLR model shall always pass to the point X bar and Y bar. So, these are, this is how you calculate the slope and intercept in case of simple linear regression.

(Refer Slide Time: 21:07)

PARTIAL SUM SQUARES IN SLR

$$\underbrace{y_i - \bar{y}}_{\text{Total}} = \underbrace{(y_i - \hat{y}_i)}_{\text{Error}} + \underbrace{(\hat{y}_i - \bar{y})}_{\text{Model}}$$

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$

$$\underbrace{SSTotal}_{n-1} = \underbrace{SSE}_{(n-2)} + \underbrace{SSR}_{1}$$

d.f. of SSTotal = d.f. of SSE + d.f. of SSR

- ▶ SSTotal: Sum of squared deviations from the mean \bar{y} (total variation of y without model adjustment).
- ▶ SSE: Sum of squared errors (or residuals) after adjusted by SLR model (part of variation of y cannot be explained by a SLR model).
- ▶ SSR: Reduction in the variation attributable to the SLR model (denoted by SSR).

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$R^2 = 1 - \frac{SSE}{SST}$

Now, what is the partial sum square in SLR? So, there are different terms in the SLR. So, let us consider that this is a, this is a mean of Y. So, here if we consider this term Y_i minus \bar{Y} , so you can see here this is Y_i and the \bar{Y} lies here. So, the total variation, so total variation that is the this the deviation from the mean \bar{Y} without model adjustment is basically, this is the linear distance \bar{Y} minus Y_i . This is the total variation.

If there was no variation it should be in the mean line. However, since there is variation we can see Y_i is here. So, the total variation can be considered as a Y_i minus \bar{Y} . And this can be decomposed further into error term and model term. So, what is the error term? So, this is the model line.

So, the error, so the model, so the so the error term is of course Y_i minus \hat{Y}_i we know that, \hat{Y}_i is the predicted values which lies in this line and Y_i is the actual observation. So, the difference is Y_i minus \hat{Y}_i . So, this is the error term which is residual which cannot be explained by the model, but if we take \hat{Y}_i minus \bar{Y} .

So, this is the difference \hat{Y}_i minus \bar{Y} these difference these linear difference can be explained by the model. So, this variation can be explained by the model. So, if we decompose the total variation we can get this Y_i minus \hat{Y}_i plus \hat{Y}_i minus \bar{Y} . So, these two components will get error as well as model.

Now, if we take the sum square of total so this is the total variation, so if there is a sum square of total that shows the sum square of error plus sum square of model or regression. So, with the $n - 1$ degree freedom in case of sum square total and in case of sum square error we get the $n - 2$ degree of freedom, and in case of sum square regression will get 1 degree of freedom.

So, here you can see these terms, so sum square total is basically the sum square of deviation from the mean Y , so sum square of deviation from mean Y , so this deviation, this deviation. So, if you take this deviation for all the points for different values of X and then we will take a sum, we will get the sum square total, sum square error you already know, some square or some square residual we have discussed in our you know last slide.

So, some square error is basically the sum square of residual after adjusted by sum square, by the adjusted by the simple linear regression model which is part of the variation of Y which cannot be explained by the SLR model. So, this variation or residual cannot be explained by the residual model. And finally sum square regression is the reduction in the variation attributable to the SLR model which is denoted by SLR, SSR sum square regression.

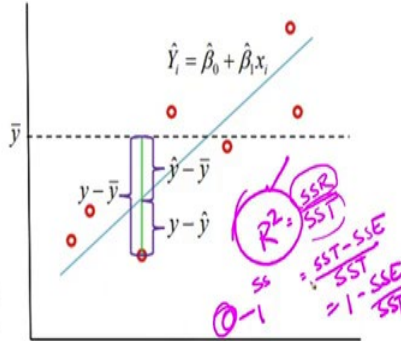
So, this variation which is from the \bar{Y} line to this regression line, this variation can be addressed by this regression, so sum square regression is basically the summation of square term of this difference. So, we can see that regression coefficient is basically $1 - \frac{\text{sum square error}}{\text{sum square total}}$, so this is how we calculate the or in other words actually what happens the regression coefficient basically shows how much variability you can explain through your regression model.

(Refer Slide Time: 25:49)

PARTIAL SUM SQUARES IN SLR

$$\begin{aligned}
 \underbrace{y_i - \bar{y}}_{\text{Total}} &= \underbrace{(y_i - \hat{y}_i)}_{\text{Error}} + \underbrace{(\hat{y}_i - \bar{y})}_{\text{Model}} \\
 \sum_i (y_i - \bar{y})^2 &= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 \\
 \underbrace{\text{SSTotal}}_{\substack{n-1 \\ \text{d.f. of SSTotal}}} &= \underbrace{\text{SSE}}_{\substack{(n-2) \\ \text{d.f. of SSE}}} + \underbrace{\text{SSR}}_{\substack{1 \\ \text{d.f. of SSR}}}
 \end{aligned}$$

- ▶ SSTotal: Sum of squared deviations from the mean \bar{y} (total variation of y without model adjustment).
- ▶ SSE: Sum of squared errors (or residuals) after adjusted by SLR model (part of variation of y cannot be explained by a SLR model).
- ▶ SSR: Reduction in the variation attributable to the SLR model (denoted by SSR).




So, in other words we can say that sum square or in other words we can say R square which is the indication of the variability which is explained by the model, so we can, since we can say that the sum square regression by sum square total. So, out of the total variation how much percentage is addressed by the regression model?

So, we know that sum square regression is basically sum square total minus sum square error, sum square total, so if we simplify it we will get sum square error by sum square total. So, this is how we calculate this R square. R square values generally varies from 0 to 1, I mean if there is no relationship between X and Y, we will get the value close to 0 and as there are much more strong relationship we will see that R square values goes towards 1.

(Refer Slide Time: 26:52)

SLR METRICS

- ▶ Root MSE is the estimate of σ :
$$\hat{\sigma} = \sqrt{\frac{SSE}{\text{d.f. of SSE}}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}}$$
- ▶ R^2 is the proportion of the total variance, accounted for by the model:
$$R^2 = \frac{SSR}{SSTotal} = 1 - \frac{SSE}{SSTotal}$$
- ▶ For the SLR model it equals the correlation, r , squared:
 - ▶ R^2 near to 1: X explains most of the variability in Y
 - ▶ R^2 near to 0: X explains little of the variability in Y



And what are the other matrix? The other matrix you can see root mean square error is the estimation of sigma, so is the estimate of sigma. So, you can see that sigma, so this sigma hat can be calculated by sum square error by degree of freedoms of SSE, so this is the sum square error which you know or sum square residual and degree of error, degree of freedom for some square error is n minus 2, so this is how you can calculate this the estimate of sigma.

R square is the proportion of the total variance accounted by the model, so we have already, I have already showed you sum square regression by sum square total that is 1 minus sum square error by sum square total. And for the SLR model it equals to the correlation. So, remember that the R for the for a simple linear regression model, R square basically denotes the square of the correlation coefficient. So, R square near to 1 where x explain most of the variability in Y and where R square near to 0 X explain little of the variability of Y.

(Refer Slide Time: 28:08)

CONFIDENCE INTERVAL OR TESTING β_1 (SLOPE)

- ▶ Given $\epsilon \sim N(0, \sigma)$, it can be shown that for SLR, $\hat{\beta}_1$ is a normal r.v. $\hat{\beta}_1 \sim N(\beta_1, \sigma_{\hat{\beta}_1})$, where $\sigma_{\hat{\beta}_1} = \sqrt{\frac{\sigma^2}{(n-1)s_x^2}}$
- ▶ We can estimate $\sigma_{\hat{\beta}_1}$ by $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{MSE}{(n-1)s_x^2}}$
- ▶ A $100(1 - \alpha)\%$ CI for β_1 is:
$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \sqrt{\frac{MSE}{(n-1)s_x^2}}$$
- ▶ For hypothesis testing on β_1 with $H_0: \beta_1 = \beta_1^*$, the test statistic is
$$T = \frac{\hat{\beta}_1 - \beta_1^*}{\hat{\sigma}_{\hat{\beta}_1}}$$

So, now confidence interval or testing beta 1 slope we can see that if given the error term is normally distributed with the mean of 0 and a standard deviation sigma, it can be shown that for SLR simple linear regression beta 1 hat or the estimate of the slope is also normal and which varies which normally distributed with a mean of beta 1 and with a standard deviation of sigma beta 1 hat, so where this sigma beta 1 hat is equal to this term that is square root of sigma square by n minus 1 standard deviation of X square.

So, we can estimate this sigma beta 1 hat by this sigma hat beta 1 and then we can calculate this by using this formula. And then, if we can take a confidence interval for a given level of alpha we will have, this is called the confidence interval, whereas this is beta 1 hat plus minus t alpha by 2 n minus 2 degree of freedom. So, root over of MSE n minus 1 then this standard deviation of X.

So, for hypothesis testing if we consider the confidence interval or testing of beta 1, if we want to test the confidence interval of beta 1 we can use this formula to get the confidence interval of the slope. Now, for hypothesis testing on beta 1 with H0 equal to B1, B1 star the test statistics is basically this one. So, this is how you can do the hypothesis testing for beta 1. The important take home message from this slide is using this formula you can easily calculate the confidence interval for the slope of any simple linear regression.

Friends so today let us wrap up our discussion here and I hope that you have learned, you have something new. And most of these things are already you have gone through it previously I assume. So, this will be really required when we discuss the multiple linear regression. So, please stay tuned and let us meet in our next lecture to discuss from here and we will discuss the different aspects of confidence interval, prediction interval and also the multiple linear regression. Thank you.