

**Urban Transportation Planning**  
**Prof. Dr. V. Thamizh Arasan**  
**Department of Civil Engineering**  
**Indian Institute of Technology Madras**

**Module No. # 03**  
**Lecture No. # 10**  
**Trip Generation Analysis contd.**

This is lecture 10 on Urban Transportation Planning, in this lecture we will continue our discussion on Trip Generation Analysis. You may recall in the previous lecture, we started our discussion on trip generation analysis; we defined trip as one way movement from an origin to a destination, we also made a particular point very clear in connection with trip, as far as origins of trips are concerned, we consider zones centroids as trip origins, destination are also zone centroids.

So, it is a just movement between zone centroids, then we try to differentiate between trip production and trip attraction; trips generated at residential zones are termed as trip productions and trips generated at non residential zones are trip attractions, then we further classify the trips into home based and non home based trips.

Can anyone define home based trips?

One trip ends at household is home based.

**Yeah**, a trip having at least one end at household is a home based trip then; obviously, non home based trips will have neither of its ends at home, the ends will be connected to non home activity centres. Then, we classified the home based trips into different types, starting from work trip, educational trip, social recreation trip, shopping trips, personal business and work related business trips. Then in the process of a discussion on trip production modeling, we try to identify the set of variables or factors that might influence trip production.

What are the factors that might influence trip production? Of course, the factors are related to household characteristics is it not? Can anyone list the factors influencing trip production? Yes, please any volunteers no, I think you have to recollect faster. We started identifying the factors starting from the size and composition of household then,

number of workers in a household, number of students in a household, household vehicle ownership, household income and so on.

Then, we were just discussing about analytical technique, which will be suitable for this kind of analysis, trip production analysis, which basically involves relating a dependent variable by the set of independent variables. And for this purposes regression analysis is found to be more suitable compared to other analytical techniques.

To understand the basics of regression analysis, we took up a simple example involving one independent variable and try to understand the basic analytical procedure in the regression analysis and then, we also listed some of the important results related to regression analysis and the results that we are seen are this.

(Refer Slide Time: 04:15)

The slide displays the following content:

$$Y_e = a + bX$$
$$b = \frac{\sum xy}{\sum x^2} \quad a = \bar{Y} - b\bar{X};$$

Where,

$$x = X - \bar{X}, \quad y = Y - \bar{Y}$$

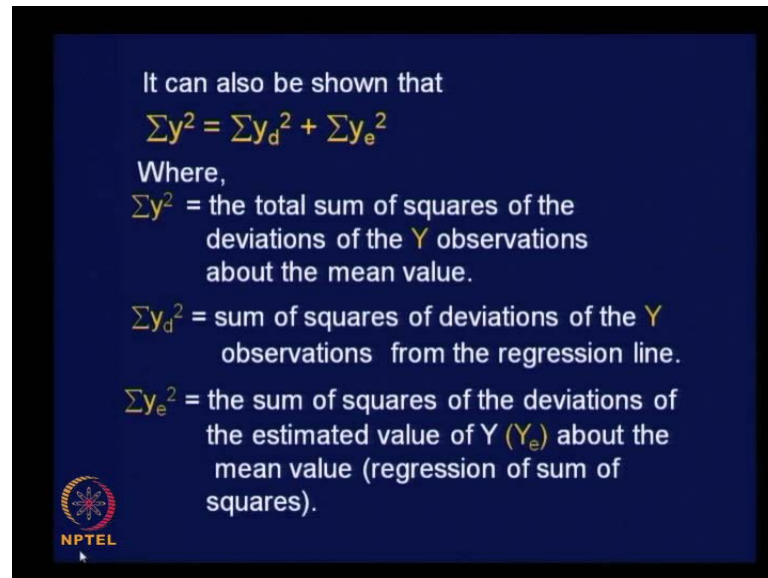
$\bar{X}, \bar{Y}$  = the means of the X and Y observations, respectively.

The NPTEL logo is visible in the bottom left corner of the slide.

We just assumed a very simple linear regression equation as,  $Y_e$  to be equal to  $a + bX$ . Now, we are clear that, we can estimate the regression coefficient  $b$  to be  $\frac{\sum xy}{\sum x^2}$  and the intercept constant  $a$  is given as  $\bar{Y} - b\bar{X}$  and of course, we define lower case  $x$  as  $X - \bar{X}$ ,  $y$  as  $Y - \bar{Y}$ , and  $\bar{X}$  and  $\bar{Y}$  are nothing but, means of  $X$  and  $Y$  observations, respectively observed values of  $X$ .

While please note here, we indicate the actual observed values with upper case letters and  $x$  and  $y$  used for your analysis are lower case letters, the difference between the observed value and a corresponding mean.

(Refer Slide Time: 05:24)



It can also be shown that


$$\sum y^2 = \sum y_d^2 + \sum y_e^2$$

Where,

$\sum y^2$  = the total sum of squares of the deviations of the  $Y$  observations about the mean value.

$\sum y_d^2$  = sum of squares of deviations of the  $Y$  observations from the regression line.

$\sum y_e^2$  = the sum of squares of the deviations of the estimated value of  $Y$  ( $Y_e$ ) about the mean value (regression sum of squares).

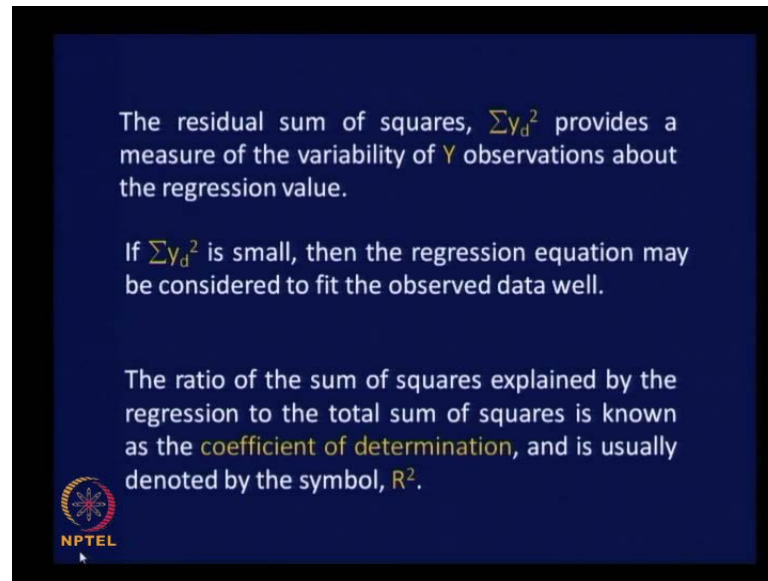


Then, we discussed about the sums of squares, we identified three different sums of squares and we also looked at this particular relationship as sigma  $y$  squared to be equal to sigma  $y_d$  square plus sigma  $y_e$  square, it is better to understand the description of each of these sums of squares.

Sigma  $y$  square is nothing but, the total sum of squares of the deviations of  $Y$  observations about the mean value; it is nothing but,  $y$  minus  $\bar{y}$  square summed over all the observations is it not, that is what is meant by this particular statement and sigma  $y_d$  square as we have seen with the help of a figure is sum of squares of deviations of the  $Y$  observations from the regression line.

We just measure the deviation up and down from the regression line of the observed values and get the value of  $y_d$  and then, sigma  $y_d$  square, then sigma  $y_e$  square is the sum of squares of the deviations of the estimated value of  $Y$ , which is  $Y_e$  about the mean value, which is also termed as regression sum of squares. We find the difference between  $Y_e$  and  $\bar{Y}$ , square it and then sum of to get the value of sigma  $y_e$  square.


(Refer Slide Time: 07:13)



The residual sum of squares,  $\sum y_d^2$  provides a measure of the variability of  $Y$  observations about the regression value.

If  $\sum y_d^2$  is small, then the regression equation may be considered to fit the observed data well.

The ratio of the sum of squares explained by the regression to the total sum of squares is known as the **coefficient of determination**, and is usually denoted by the symbol,  $R^2$ .




And to go further on this particular sum of squares, we need to understand that the residual sum of squares,  $\sum y_d^2$  provides a measure of the variability of  $Y$  observations about the regression value is it not?

And  $\sum y_d^2$ , if you find the value to be small, then the regression equation may be considered to fit the observed data well and we also discussed about this particular coefficient, the ratio of the sum of squares explained by the regression to the total sum squares is known as the coefficient of determination, which gives an idea about the level of utilization of the regression equation or the level of effectiveness of the regression equation and is usually denoted by the symbol,  $R^2$ .

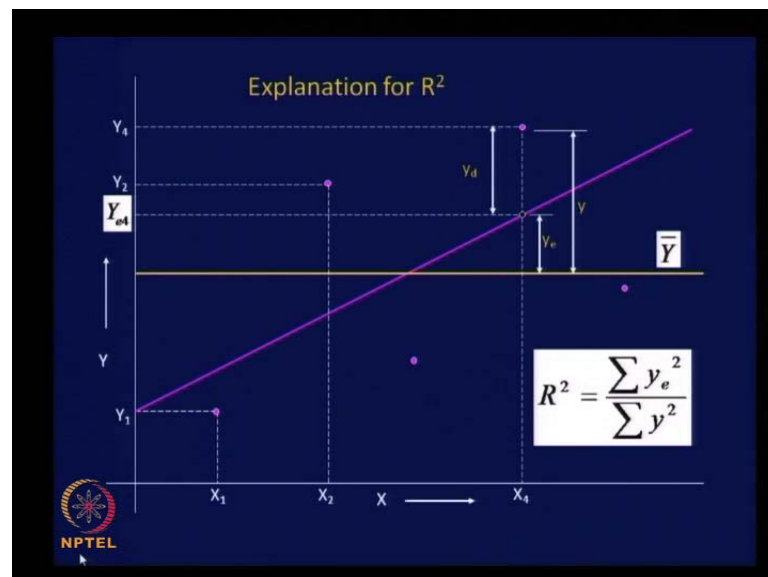
(Refer Slide Time: 08:25)

The value of  $R^2$  is given as,

$$R^2 = \frac{\sum y_e^2}{\sum y^2}$$


We can give the value of R squared to be sigma y e squared divided by sigma y square, you know now, what we really mean by sigma y e squared and sigma y square. Let me (()) little more light on the concept of R square, so that you are able to appreciate the significance and importance of this particular coefficient.

(Refer Slide Time: 08:54)



Additional explanation about R square, let us try to understand the concept with the simple data set involving just five observed points of X and Y, that is what I have done here. I just made a scattered diagram pertaining to five points is it not and let us say, this

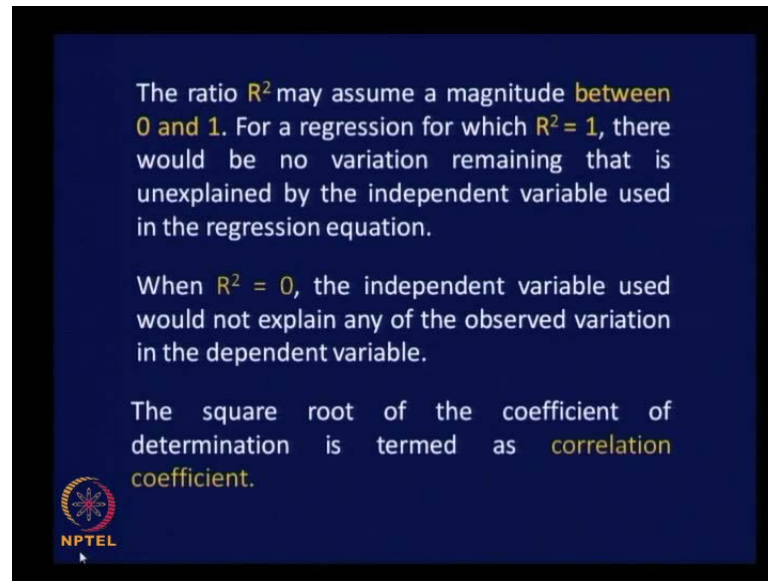
is the regression line, which fits the five points reasonably well, then let us say this is the mean value of the Y observations say  $\bar{Y}$ ; obviously, it is a constant, so the line is going to be a horizontal line and let us say the coordinates of point one or  $X_1$  and  $Y_1$  and point two,  $X_2$  and  $Y_2$  and point three and so on say point four,  $X_4$  and  $Y_4$ . So, we need to identify here the values of  $y$ ,  $y_e$  and  $y_d$  if we can pictorially perceive things it will remind in a memory better.

So, let us try to identify those values; obviously,  $Y_e$  is as shown here, this is the estimate of  $Y_4$ , estimate of  $Y_4$  is shown here as  $Y_e$  is it not. And this is the deviation from a mean and we name it as  $Y_e$  and this is how we should understand  $Y$ , deviation of observed values of  $Y$  from its mean and we are trying to calculate  $R^2$  using the values of  $Y_e$  and  $Y$  is it not (Refer Slide Time: 10:40).

So, we can write now  $R^2$  to be equal to  $\frac{\sum y_e^2}{\sum y^2}$  you can understand how we calculate the sum of squares of the deviations and then get the value of  $R^2$ . Now, what is the maximum possible value of  $R^2$ , you have the picture in front of you and you have the equation. Can anyone tell me, what is the maximum value for  $R^2$ ? Yes 1, perfect.

If there is no deviation at all between the observed and regressed values; obviously, the denominator and numerator **numerator** are going to be same is it not. So, the maximum possible value is 1. What is the least possible value, if the independent variable is not explaining the dependent variable at all, 0; can be 0 **right**. So, that is the range of values for  $R^2$  of course, for completion sake I will show the value of other deviation also, this is how we need to understand the deviation with respect to regression line as well as, with respect to the mean, is it clear.


(Refer Slide Time: 12:52)



The ratio  $R^2$  may assume a magnitude between 0 and 1. For a regression for which  $R^2 = 1$ , there would be no variation remaining that is unexplained by the independent variable used in the regression equation.

When  $R^2 = 0$ , the independent variable used would not explain any of the observed variation in the dependent variable.

The square root of the coefficient of determination is termed as correlation coefficient.



So, the ratio  $R$  squared may assume a magnitude between 0 and 1. For a regression for which  $R$  squared is equal to 1, there would be no variation remaining that is unexplained by the independent variable used in the regression equation; that means, your independent variable is so effective; There it is able to explain the whole of the variation of the dependent variable and will it happen in practice, it is very rare, it may not happen, it is only an ideal situation, which can be understood theoretically.

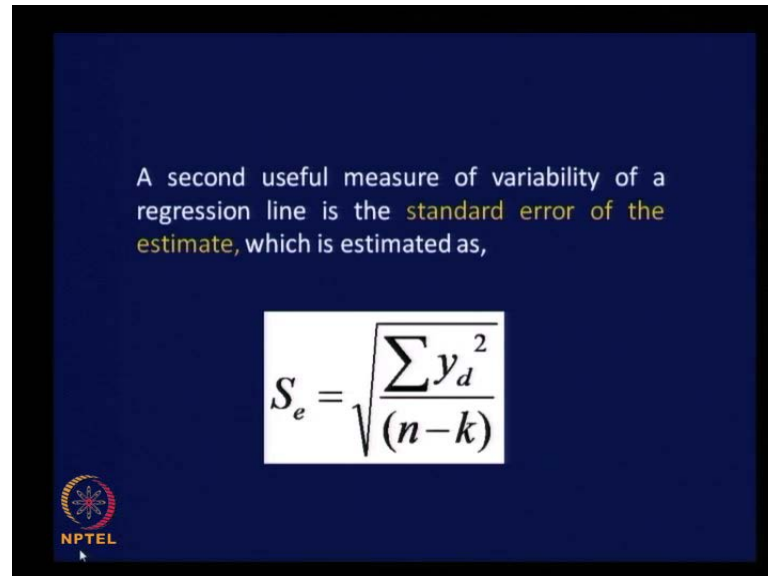
And when  $R$  squared is equal to 0, the independent variable used would not explain any of the observed variation in the dependent variable, there is no relationship at all between your independent variable and the dependent variable.

So, that is how you must understand the significance of this particular coefficient, it exactly gives you the extent of relationship between your independent variable and the dependent variable. Let us say in a particular case the value of  $R$  square is 0.85 means what? This implies what; I said that when  $R$  square value is 1, the independent variable is fully explaining the variation of  $Y$ , when  $R$  squared is 0.85 obviously, the independent variable is able to explained variation of  $Y$  to an extent of 85 percent, 85 percent of the variation of the dependent variable,  $e$  is explained by the independent variable chosen for regression analysis.

So, you can get a clear idea of the effectiveness of the set of independent variables that are using in the regression analysis from this particular coefficient. Of course, the square

root of the coefficient of determination is termed as correlation coefficient is one way or there are other methods also available to calculate correlation coefficient.

(Refer Slide Time: 15:27)



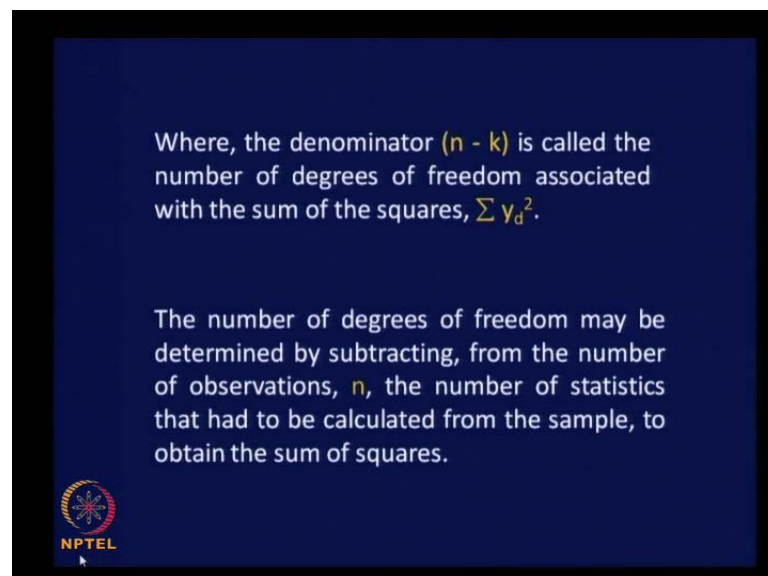
A second useful measure of variability of a regression line is the **standard error of the estimate**, which is estimated as,

$$S_e = \sqrt{\frac{\sum y_d^2}{(n-k)}}$$

NPTEL

And a second useful measure of variability of a regression line is standard error of estimate, which is estimated as,  $S_e$  to be equal to sigma  $y_d$  square by  $n$  minus  $k$ . I am directly giving you the result proof and other basic information you please go through relevant books and try to understand.  $S_e$  here is square root of sigma  $y_d$  square divided by  $n$  minus  $k$ , standard error of estimate.

(Refer Slide Time: 16:09)



Where, the denominator  $(n - k)$  is called the number of degrees of freedom associated with the sum of the squares,  $\sum y_d^2$ .

The number of degrees of freedom may be determined by subtracting, from the number of observations,  $n$ , the number of statistics that had to be calculated from the sample, to obtain the sum of squares.

NPTEL

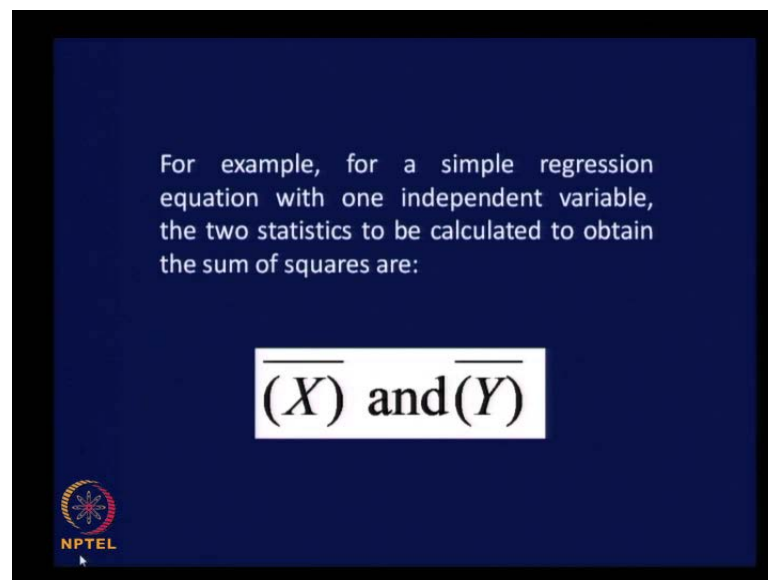


Where, the denominator  $n - k$  denominator of the equation for  $S_e$  is  $n - k$ , the denominator  $n - k$  is called the number of degrees of freedom associated with the sum of squares,  $\sum y_d^2$  number of degrees of freedom.

This is the explanation for number of degrees of freedom, the number of degrees of freedom may be determined by subtracting, from the number of observations,  $n$ , the number of statistics that had to be calculated from the sample, to obtain the sum of squares, total observations minus the number of statistics required to get the value of sum of squares. In this particular case,  $\sum y_d^2$ , how many statistics are to be calculated to get  $\sum y_d^2$ ;  $\sum y_d^2$  is nothing but, deviation of the observed value of  $Y$  with respect to the regression line.

So, you need to get the regression line first to get the regression line you need to have at least two statistics namely,  $\bar{X}$  and  $\bar{Y}$  you need to calculate, then only you able to go further with the values of small  $x$  and small  $y$  to calculate the regression coefficients and then, get the equation for the regression line, that is how that is what is meant by number of statistics required to estimate the value of  $\sum y_d^2$ .

(Refer Slide Time: 18:05)



For example, for a simple regression equation with one independent variable, the two statistics to be calculated to obtain the sum of squares are: as I said,  $\bar{X}$  and  $\bar{Y}$ .

(Refer Slide Time: 18:24)

The regression coefficient  $b$  is a statistical estimate and is therefore subject to error.


The **standard error** is the concept used to convey the magnitude of this error, and is estimated from the following expression:

$$S_{eb} = \sqrt{\frac{S_e^2}{\sum x^2}}$$

Where,

$S_e$  = as defined earlier

$\sum x^2$  = the sum of squares of the deviations of X observations about the mean.




So, the  $k$  value is 1 in this particular case and then, the regression coefficient  $b$  is a statistical estimate; obviously, we estimate the value of  $b$ , because there is no direct equation available to get accurate value of  $b$  and is therefore, subjected to error and it is better to check for the error associated with regression coefficient by some statistical measure, this standard error is a concept used to convey the magnitude of this error, and is estimated from the following expression: this is the standard error associated with the regression coefficient  $b$ ,  $S_{eb}$  is equal to square root of  $S_e$  square divided by  $\sum x^2$  and of course,  $S_e$  is nothing but, the standard error of estimate for the regression as a whole as explained earlier and  $\sum x^2$  is the sum of squares of deviations of X observations about the mean.

(Refer Slide Time: 19:44)

That is, 
$$\sum X^2 = \sum (X - \bar{X})^2$$

The 't' test may be used to determine whether an estimated regression coefficient is significant, by forming the following ratio:

t = the regression coefficient divided by standard error of the regression coefficient.



Or in other words, it is nothing but, X minus X bar whole square, summed over all the values. Now, the t test may be used to determine whether an estimated regression coefficient is significant, by forming the following ratio: the value of t can be estimated simply by dividing the regression coefficient by standard error of the regression coefficient, just a statistic divided by the standard error of that particular statistic, that is how we get the value of t.


(Refer Slide Time: 20:35)

**EXAMPLE:**

Develop a trip production equation and calculate all the relevant statistics to check the validity of the equation using the following data:

Average household size	:	2	3	4	5	6
Average total trips made per day	:	5	7	8	10	10

The value of t statistic for 3 degree of freedom at 5% level of significance is 2.353.



And with this understanding, let us consider a small numerical example of development of trip production equation involving again only one independent variable. The example is this, you have been asked to develop a trip production equation and calculate all the relevant statistics to check the validity of the equation using the following data: this is the data given, average household size 2 3 4 5 and 6 average total trips made by each of the households per day 5 7 8 10 10. Which is the dependent variable here, the first row values or second row values, dependent variable? Second row values are dependent variables and first row values are independent variables.

Trip production depends on the household characteristics, household size is the household characteristics and an average total trip made by the household per day is the trip production rate, is it not.

Let us try to work out the required statistics first and then, go for getting the regression equation of course, information available here is, the value of t statistic for 3 degrees of freedom at 5 percent level of significance is 2.353. This information is required, because finally to get this statistical validity of any regression coefficient will be comparing the table value of t with the actual calculated value. Here the table value given as 2.353 at 5 percent level of significance, why 5 percent, why not 10 percent or why not 1 percent; it can be anything it depends upon; however, expected level of accuracy.

Normally, in most statistical experiments, 5 percent level of significance is acceptable that is how here the 5 percent comes in; you can aim for 1 percent level of significance; that means, 99 percent accuracy that is also acceptable.

(Refer Slide Time: 23:33)

Y	X	$Y - \bar{Y}$ (y)	$X - \bar{X}$ (x)	$x^2$	$y^2$	xy	$Y_e$	$Y - Y_e$ ( $y_d$ )	$y_d^2$	$Y_e - \bar{Y}$ ( $y_e$ )	$y_e^2$
5	2	-3	-2	4	9	6	5.4	-0.4	0.16	-2.6	6.76
7	3	-1	-1	1	1	1	6.7	0.3	0.09	-1.3	1.69
8	4	0	0	0	0	0	8.0	0	0	0	0
10	5	2	1	1	4	2	9.3	0.7	0.49	1.3	1.69
10	6	2	2	4	4	4	10.6	-0.6	0.36	2.6	6.76
$\Sigma$	40	20		10	18	13	40		1.10		16.90

$\bar{Y} = \frac{40}{5} = 8$	$\bar{X} = \frac{20}{5} = 4$	$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{13}{10} = 1.3$
$a = \bar{Y} - b \bar{X}$		$Y_e = 2.8 + 1.3 X$
$= 8 - 1.3 \times 4 = 2.8$		

This is a given data of the dependent variable, Y values; on the sum of all the Y values is 40, we have five observations. And the five X values are shown here their sum is 20 and we are ready to get the values of Y bar and X bar, Y bar is simply 40 by 5, which is 8 and X bar is 20 by 5, which is 4 then, we can get the values of X minus X bar and Y minus Y bar or lower case x and lower case y. So, Y minus Y bar, which is small y, it is calculated and shown here as, minus 3, minus 1 0 2 and 2. Similarly, X minus X bar values are calculated as, minus 2 minus 1 0 1 and 2.

Then, we need these statistics also for our subsequent calculations, square of the deviations of the observed values from the respective means x square is required. So, we get this squared value of x as given here and the sum is 10. And y square are also given sum becomes 18 here, just squaring the values already given in the table and then, x y is also required product of x and y, so we get the values as 6 1 0 2 and 4 and sum is 13.

Now, we are ready to calculate the value of b, the regression coefficient given as sigma x y divided by sigma x square, the values already calculated sigma x y is 13 and sigma x squared is 10, 13 by 10 is just 1.3 that is the regression coefficient.

And we can also calculate the value of intercept constant a as, Y bar minus b X bar, because the values are known to us now. So, we can get value of a as 8 minus 1.3 into 4 to be equal to 2.8.

Now, we know the values of  $a$  as well as,  $b$ . So, we can even write the regression equation, this is the regression equation  $\hat{Y}$  is equal to  $2.8 + 1.3 X$ . So, once you have the regression equation you can get  $\hat{Y}$  values for all the five observations, what is the estimate of  $Y$  based on this regression equation that is your concern is it not, because we want to use this regression equation to estimate the value of  $Y$  for a similar situation; it is a very general equation. So, let's us try to get the value of  $Y$  based on regression, which is  $\hat{Y}$ .

So, if you substitute the value of  $X$  for each observation in the equation you get the value of  $\hat{Y}$  as 5.4 6.7 8.0 9.3 10.6 and 40 is it not, against the observed values of  $Y$  you just have a comparison; the observed value of  $Y$  is 5, the corresponding regress value is 5.4 observed value of second observed value of  $Y$  is 7, the corresponding regressed values is 6.7, then for 8, it is 8 exactly same value, for 10 it becomes 9.3, for 10 it is 10.6 of course, the sum is same, this is how we estimate the value of  $y$ , then deviation  $y_d$ , because finally, we would like to know the exact extent to which  $X$  is able to estimate the value of  $Y$  is it not, that is our objective.

So, we need to have the values of the deviations and then a squared sum. So,  $y_d$  value is this much and  $y_d$  square you calculate and then, sum up you get  $\sum y_d^2$  as 1.10, is it not.

Then get the value of  $y_e$ ,  $y_e$  is what,  $y_e$  minus  $\bar{Y}$ ,  $y_e$  values are known,  $\bar{Y}$  value is just 8. So, we find the difference and get the values as minus 2.6, minus 1.3, 0, 1.3 and 2.6. And  $y_e$  square is our interest,  $y_e$  square is simply the square of the previous column values and we get the sum as 16.90.

(Refer Slide Time: 29:41)

$$R^2 = \frac{\sum y_e^2}{\sum y^2} = \frac{16.9}{18} = 0.939$$
$$S_e = \sqrt{\frac{\sum y_d^2}{N-K}} = \sqrt{\frac{1.1}{5-2}} = 0.6055$$
$$S_d = \sqrt{\frac{\sum y^2}{N-(K-1)}} = \sqrt{\frac{18}{5-1}} = 2.12$$

$S_e < S_d$  Hence, O.K

And then, let us get the R squared value as, sigma y e squared by sigma y squared, which is equal to 16.9 divided by 18 that is equal to 0.939. And standard error of estimate is equal to square root of y d square by N minus K, sigma y d square is 1.1 as we have seen earlier and value of K, the number of degrees of freedom in this particular case is 2 as I said earlier, it was number of statistic to be calculated are X bar and Y bar. So, 5 is the value of N, the total number of observations, so 5 minus 2 give you the number of degrees of freedom. So, this is equal to 0.6055, what is the inference out of these statistics towards R square, the x is able to explain 93.9 percent of the variation of the dependent variable y. So, it is really an effective variable in this particular case.

Let us see how to go about using the value of S e and get some inference and standard deviation is related to the observed value of Y that is calculated using the normal formula again N minus, the number of degrees of freedom in this case, the number of statistics required to be calculated to get y squared is only one.

If you know the value of Y bar, you get the value of small y and y squared; there is no need to have two statistics, that is how a number of degrees of freedom are one more than the case, where we calculate the value of S e. So, that is how you need to understand the variation, the number of degrees of freedom, so this becomes 2.12. Will it be able to perceive physically, what we really mean by S e and S d, what is standard deviation in

general? This scatter the level of scatter of observed values with respect to the mean is it not, that is how we understand standard deviation.

And standard error of estimate is nothing but, again this scatter of the observations with respect to the regression line instead of mean line that is how we just understand y d deviation of the observation with respect to the regression line is it not. If the scatter of observed point with reference to a regression line is less than the actual scatter of the observations with respect to its own mean, what is the inference? Regression line is relatively closer to all the points compared to the mean of the observed points, is it not?


So, if  $S_e$  is less than  $S_d$  we can say that, it is good or bad, good.  $S_e$  in this case is less than  $S_d$  hence, it is O.K. There is another way of looking at the correctness of your regression analysis compare  $S_e$  and  $S_d$  value.

(Refer Slide Time: 34:01)

The slide contains two mathematical equations and a concluding statement. The first equation calculates the standard error of estimate ( $S_{cb}$ ) as the square root of the square of the standard error ( $S_e$ ) divided by the sum of squares of x ( $\sum x^2$ ). The second equation calculates the t-value as the regression coefficient ( $b$ ) divided by the standard error of estimate ( $S_{cb}$ ). The text below states that the calculated t-value is greater than the table value of 2.353 at a 5% level for 3 degrees of freedom.

$$S_{cb} = \sqrt{\frac{S_e^2}{\sum x^2}} = \sqrt{\frac{(0.6055)^2}{10}} = 0.1915$$
$$t = \frac{b}{S_{cb}} = \frac{1.3}{0.1915} = 6.789$$

't' is greater than the table value of 2.353 @ 5% level for 3 degrees of freedom.



And standard error of estimate of the regression coefficient itself is given by, this equation  $S_e$  squared by sigma x squared as I shown you earlier. So, substitute the corresponding values and we get the value to be 0.1915 and then, it should be possible for us now, to get the t value as regression coefficient divided by this standard error of estimate of that particular coefficient. So, t is nothing but, b by  $S_{e b}$  that is equal to 1.3 divided by 0.1915, which is 6.789 desirably, if the denominator is small we are going to get a higher value of t.



And for the regression coefficient to be significant do you expect the denominator to be large or small, again standard error is similar to standard deviation is it not, it is nothing but, statistic divided by standard deviation of that particular statistic or standard error of that particular statistic; the denominator is smaller the denominator, better is the effectiveness of the regression coefficient is it not.

So, that is what we do here,  $t$  here is greater than the table value of 2.353 at 5 percent level of significance for 3 degrees of freedom is it not. For estimation of standard error of estimate we had  $K$  value as 2. So, number of degrees of freedom is 5 minus 2 only 3 degrees of freedom. For that, the expected value  $t$  as per the expectation of 5 percent of significance is only 2.353, but our calculated value is much higher, so we can say that, the significance is more than 5 percent is it not.

Obviously, you can also work out the exact percentage by equating or by looking at the table and finding out the percentage, which matches your calculated value that is also possible. Now, the inference is the significance of  $b$  is much higher than the 5 percent level is it not, the value is very high and that is have few questions related to the analysis we have done.

We have regressed one independent variable against a dependent variable; the independent variable was nothing but, the number of trips made for any particular purpose or for all purposes, total number of trips made by household for all purposes together.

And the independent variable is just the household size, would it be a practical situation do we really developed production model relating trip made for all purposes and the household size, it is not going to help us in practice. In practice, we will be developing regression equations for trips made by households for different purposes; will have a set of five, six trip production equations; one trip production equation for work trips, another equation for educational trips, third one for shopping trips, fourth for personal business trips and so on, why? As I said earlier, the set of independent variables that will influence these categories of trips are going to be different or even if they are same, there extent up influence of different trips type, trip types will be different.

In case of educational trips for example, trip production for the purpose of education there likely independent variables are, number of students in the household definitely

that will be one of the variables then there could be other causal variables like household vehicle ownership or income very rarely, we just develop trip production equation with only one independent variable, more than one independent variables are common in any trip production equation.


If you take trip production equation for shopping trips, then household income might be an important variable plus could be level of vehicle ownership, is it not. So, that is how you must identify and pick the causal variables to develop trip production equation.

So, this example is not suited for real life situation; this example as was given to you just to understand or demonstrate the analytical principle involved in the regression analysis resulting in trip production equation, this is meaningless as far as, practical application is concerned. We need to understand this point very clearly. Then, the values of the variables we have given some household size as some numbers I will just quickly go back to the data.

(Refer Slide Time: 40:42)

**EXAMPLE:**  
Develop a trip production equation and calculate all the relevant statistics to check the validity of the equation using the following data:

Average household size	: 2	3	4	5	6
Average total trips made per day	: 5	7	8	10	10

  
NPTEL

This is the example, average household size is 2, 3, 4, 5 and 6 for five cases are they five independent households are they cannot be independent household, because it is clearly stated this average, average household size, average of how many households, how do we take this average, these are zonal averages we divide the entire urban area into traffic zones, so for trip production analysis we deal with zonal average values.

So, here the average household size means, if there are 10,000 households in a traffic zone; it is the average value of all the 10,000 observations and if you look at the other variable, average total trips made per day for that particular household the numbers given are whole numbers, but when you work out these values in practice you may end up with fractions also, you may end up with trips like 5.25, 7.101 and so on. Because, you are taking averages right, but you can work with this fractions even though, they are not realistic and then later on finally, you can round off the values.

Since, these are averages you have to put up with an unrealistic numbers and go ahead with you analysis and finally, round off the values to get the appropriate numbers for subsequent analysis, again these are zonal average values, this needs to be understood very clearly.

(Refer Slide Time: 42:58)

**Two Independent Variables**  
The general form of the regression equation for two-variables case, can be written as,


$$Y_e = a + b_1X_1 + b_2X_2$$

Where,

$$b_1 = \frac{(\sum x_2^2)(\sum x_1y) - (\sum x_1x_2)(\sum x_2y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2y) - (\sum x_1x_2)(\sum x_1y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

 NPTEL

And this is the case of single independent variable and as I said, in reality you may have to deal with more than one independent variable, what happens to the regression equation, if you have let say two independent variables I will just show the equation for two independent variables case as,  $Y_e$  to be equal to  $a + b_1 X_1 + b_2 X_2$ , there are two independent variables,  $X_1$  and  $X_2$ .

And if you look at the equations for calculating the regression coefficients, this is the equation for calculating the value of  $b_1$  you can perceive the way the complexity of the calculations increase; it is exponential growth of involved work in the form of

calculation of the values,  $b_1$  can be calculated as  $\frac{\sum x_2^2 \sum y - \sum x_2 \sum y x_1}{\sum x_2^2 - \frac{(\sum x_2)^2}{n}}$  and  $b_2$  can be calculated as  $\frac{\sum x_1 \sum y - \sum x_1 x_2 \sum y}{\sum x_1^2 - \frac{(\sum x_1)^2}{n}}$ .


Of course, some of you will be able to memorize these equations and still manage to do regression analysis involving two independent variables; it is easy to memorize you can just see the denominator same in both the cases, denominator is same in both the cases and if you change, interchange  $x_1$  and  $x_2$  for the numerator you will get the equation from the first equations, second equation can be obtained from the first equation you have to write  $x_1$  in place of  $x_2$  that is it.

So, it is possible to memorize and go ahead and you can calculate the intercept constant as,  $\bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$ , what will happen if there are more than two independent variables; equation is going to be highly complex, it may not be manually possible to manage such calculations and it will be highly time consuming, what to do? Nothing to worry, there are program packages available to do regression analysis; some of you might know that, even your excel of MS Office has one module for doing regression, simple regression analysis is possible and there are exclusive program packages for doing linear as well as, non-linear regression analysis involving any number of independent variables and give lot of statistical measures related to all the involved variables as well as, over all fitness of regression equation.

(Refer Slide Time: 47:01)

$$R^2 = \frac{\sum y_e^2}{\sum y^2}$$
$$S_e = \sqrt{\frac{\sum y_d^2}{(n - 3)}}$$
$$Sb_1 = \sqrt{\frac{S_e^2}{\sum x_1^2 (1 - r_{12}^2)}}$$
$$Sb_2 = \sqrt{\frac{S_e^2}{\sum x_2^2 (1 - r_{12}^2)}}$$

Where,  $r_{12}^2$  = the squared multiple correlation between variables 1 & 2.




And R square is calculated following the same from last we did in the case of one independent variable, R squared can be calculated as sigma y e squared by sigma y squared and standard error of estimate for the regression equation as we did earlier, only difference is number of degrees of freedom is different here, because we involve two independent variables, so it becomes three, sigma y d square divided by n minus 3.

And there are two regression coefficients. So, S e b 1, e is not shown here, but you can understand this as standard error of estimate for regression coefficient 1, S b 1 is square root of S e square divided by sigma x 1 square into 1 minus r 12 square and S b 2 is square root of S e square divided by sigma x 2 square into 1 minus r 1 2 whole square, what is r 1 2 here, any idea, r 1 2 is nothing but, correlation coefficient, where r 12 square is a squared multiple correlation between variables 1 and 2.

(Refer Slide Time: 48:43)

The coefficient of correlation between any two variables  $x$  and  $y$  can be calculated as:

$$r_{xy} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$


You may recall the formula for simple correlation coefficient, the coefficient of correlation between any two variables  $x$  and  $y$ , let's not confuse this  $x$  and  $y$  with our dependent and independent variables, this is just two variables,  $x$  and  $y$  to understand the concept of correlation coefficient.

We just calculate  $r_{xy}$  to be equal to  $\sum xy$  divided by square root of  $\sum x^2$  into  $\sum y^2$  on the same lines we calculate  $r_{12}$  in the previous case,  $x_1$  and  $x_2$  is calculated taking those two things as two variables is that clear.

Now, the points to be remembered in this lecture are that, the regression coefficient it is a very important indicator of the effectiveness of the regression analysis; we should not depend only on  $r^2$  value, it is better to check for the effectiveness or significance of the involved independent variable, effectiveness of the independent variable can be checked using  $t$  test for any desired level of significance.

We can also check for the overall effectiveness of the regression analysis by comparing standard error of estimate of the whole regression process and the standard deviation of the  $y$  observations is it not. And in practice, you may have to deal with more than one independent variable to develop trip production models.

So, when the number of independent variables increased the complexity of analysis increases exponentially and this complexity can be tackled by resorting to the help of

available program packages and you can do regression analysis involving any number of independent variables and develop trip production models. And finally, we need to understand that, the dependent and independent variables are zonal average values; they are not specific to any particular household.

So, we are dealing with zonal average values treating traffic zone as a trip production unit and this has to be understood, while developing trip production models with this we will conclude our discussion for today, we will continue the rest of it in the next class.