

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

Module #03
Lecture #12
Trip Generation Analysis Contd.

This is lecture 12 on Urban Transportation Planning. We will continue our discussion on trip generation analysis in this class too. Before we proceed further, let us try to recapitulate what was done in the previous class, you may recall, we started our discussion in the previous class on category analysis. Now, we know that in category analysis, households are taken as trip making centers where as in the case of regression analysis; we consider traffic zones as trip producing centers.

The advantage of category analysis is that, the wide variation in the house hold characteristic can be captured well in the case of category analysis where as in the case of regression analysis, since we consider zonal averages, these variations are not captured very well. Does it mean that regression analysis is totally useless? It is not so. Suppose you have a situation, in a particular urban area, there is not much of variation in socio economic characteristic of households, a city which has grown and matured almost come to a saturated condition.

In such cases, when there is no variation at all then you can go in for regression analysis to develop trip production equations, it is not that regression should not be used for trip production analysis. So, there are two possibilities, depending upon the conditions, you can choose any one of these two, take next to develop trip production equations. For Indian condition, which meteorology you feel is better, for trip production analysis, regression or category analysis?

For that, you must understand **are** the characteristic of house holds, and then the ranges over which characteristics are spread. Normally, under Indian conditions, you can understand that, we have very wide range of socio economic characteristics.

If you consider monthly income of persons living in urban areas, as we have seen in the previous class, it ranges between just 1000 rupees to more than 60000, 70000 rupees living out the extremes, where as in developed countries, the range is not that wide. So,

obviously, when income range is so wide, we can anticipate very wide range in other related house hold characteristics, under such conditions, it is better to go in for category analysis rather than regression analysis, when we come across such wide variation in socio economic characteristics of urban dwellers **right**.

So, after our discussion on category analysis, we started our discussion on trip attraction modeling. Now, we know that, we derive variables based on land used characteristic to model trip attraction, we also listed land used characteristics, which can be considered to formulate variables for developing trip attraction models.

(Refer Slide Time: 04:29)



The slide is titled "TRIP ATTRACTION ANALYSIS" and lists five causal variables. It includes the NPTEL logo in the bottom left corner.

TRIP ATTRACTION ANALYSIS

Causal Variables:

1. Retail trade floor area
2. Service and office floor area
3. Manufacturing and wholesale floor area
4. Number of employment opportunities in retail trade
5. Number of employment opportunities in service and office.


 NPTEL

Just to recollect the list of variables that, we listed in the previous class, the causal variables for trip attraction modelling could be like retail trade floor area, service and office floor area, manufacturing and whole sale floor area, number of employment opportunities in retail trade, the number of employment opportunities in service and office.

(Refer Slide Time: 05:56)

Causal Variables...

6. Number of employment opportunities in manufacturing and wholesale
7. School and college enrollment
8. Number of special activity centres like transport terminals, Sports stadium, Major recreational / cultural / religious places.




Also, we may have variables like number of employment opportunities in manufacturing and wholesale, school and college enrollment, number of special activity centers like transport terminals, sports stadium, major recreational, cultural and religious places. These are just factors listed to help us to derive variables for trip attraction modelling. Let us now, take very small or simple numerical example, and see how to develop trip attraction model for a given situation.

(Refer Slide Time: 05:39)

Numerical Example:
Develop a trip-attraction equation using the data given in the table below. Do the necessary statistical checks to assess the validity of the equation. The table value of t for this case @5% level of significance, is 1.77.

Zone NO.	No. of Employment Opportunities in zone		No. of daily work trips attracted
	Manufacturing	Service	
1	60	30	190
2	40	100	290
3	30	20	150
4	20	30	120
5	100	20	250



This is the problem, we need to develop a trip attraction equation, using the data given in the table that will be shown now. Do the necessary statistical checks to assess the validity of the equation, the table value of t for this case at 5 percent level of significance is 1.77. And this is the given data, we have five zones listed as 1, 2, 3, 4 and 5 and number of employment opportunities in the zones in manufacturing and service are given. Under manufacturing, we have the employment opportunities as 60, 40, 30, 20 and 100 and then under service, we have 30, 100, 20, 30 and 20. A number of daily work trips attracted are 190, 290, 150, 120 and 250. Which is the dependent variable in this case, dependent variable?

(O)

Obviously, the number of work trips attracted is a dependent variable and please note, we have two independent variables to deal with in this particular case; employment in manufacturing, and employment in service sector. Let us proceed further to develop the regression equation for modelling trip attraction and we need to recollect the regression equation for **two variables case**, two independent variables case.

(Refer Slide Time: 07:37)

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

This is the general form of the regression equation; Y is equal to a plus b1 X 1 plus b2 X 2, where b 1 is given as shown here, as we have seen earlier, b 1 is sigma x 2 square into sigma x 1 y minus sigma x 1 x 2 into sigma x 2 y whole divided by sigma x 1 square into sigma x 2 square minus sigma x 1 x 2 whole square.

On the same lines, we know the equation for b 2, similar to the equation for b 1 with some changes in respect of x 1 and x 2 as I indicated to you earlier in the previous class. And the intercept constant, a can be calculated as Y bar minus b1 X 1 bar minus b2 X 2 bar. And from the given data, we need to calculate the values of the symbols that, we are shown in the equations, starting from x 1 x 2, all lower case letters, please remember what we have as raw data or capital X 1, capital X 2 and capital Y and we need to calculate small x 1 small x 2 x 1 y x 2 y x 1 x 2 and square of all these values and so on.

(Refer Slide Time: 09:26)


Zone No.	X ₁	X ₂	Y	y	x ₁	x ₂	x ₁ x ₂	x ₁ y	x ₂ y
1	60	30	190	-10	10	-10	-100	-100	100
2	40	100	290	90	-10	60	-600	-900	5400
3	30	20	150	-50	-20	-20	400	1000	1000
4	20	30	120	-80	-30	-10	300	2400	800
5	100	20	250	50	50	-20	-1000	2500	-1000
Σ	250	200	1000				-1000	4900	6300

$\bar{Y} = \frac{\Sigma Y}{n} = \frac{1000}{5} = 200$	$\bar{X}_1 = \frac{\Sigma X_1}{n} = \frac{250}{5} = 50$	$\bar{X}_2 = \frac{\Sigma X_2}{n} = \frac{200}{5} = 40$
$y = Y - \bar{Y}$	$x_1 = X_1 - \bar{X}_1$	$x_2 = X_2 - \bar{X}_2$

Let us first, calculate the value of required parameters for the regression analysis, there are 5 traffic zones, the values of capital X 1 are as shown here, the values of capital X 2 for the 5 zones are as given here, and the values of capital Y are as shown. And once you know these values, you can calculate Y bar as sigma Y by n to be equal to 200 and X 1 bar is sigma X 1 by n which is 50 and X 2 bar is 40.

Let us go further to get the value of small y as Y minus Y bar, all these values are already known to us and X 1 is X 1 minus X 1 bar and X 2 is X 2 minus X 2 bar. This is the value of lower case y, the value for small x values of x 2, x 1 x 2, x 1 y and x 2 y, is it not? So, this is how we get the required values, this is not the end of it, we need to calculate some more values to substitute in the regression equation.

(Refer Slide Time: 11:11)




Zone No.	X_1^2	X_2^2	y^2
1	100	100	100
2	100	3600	8100
3	400	400	2500
4	900	100	6400
5	2500	400	2500
Σ	4000	4600	19600

So, we will go further and get the other required values X_1 square, X_2 square y square and so on. Now, we are ready to calculate the values of b_1 , b_2 and a . We just substitute these values in the equation, you get b_1 to be 1.657, I am not shown the values substituted here, just I have written the equation given you the answer, you can substitute and check for the correctness of the result.

And b_2 works out to be 1.729, a is 47.99. Now, we are ready to write the regression equation, we can write now \hat{Y} to be 47.99 plus 1.657 X_1 plus 1.729 X_2 , and using this equation, we can get the value of \hat{Y} for all the five observations and then proceed to calculate the required statistics.

(Refer Slide Time: 12:43)




Y_e	$Y_e - \bar{Y}$ (y_e)	y_e^2	$Y - Y_e$ (y_d)	y_d^2
199.28	- 0.72	0.52	9.28	86.12
287.17	87.17	7598.60	-2.83	8.01
132.28	- 67.72	4585.99	- 17.72	314.00
133.00	-67.00	4489.00	13.00	169.00
248.27	48.27	2329.99	- 1.73	3.00
1000.00	101.00	19004.10		580.13

The value of Y_e is calculated as **Y minus** Y_e minus \bar{Y} and the values are shown here, and the other statistics are also calculated and the subsequent values are also calculated as shown here.


(Refer Slide Time: 13:13)

The coefficient of determination is given as,

$$R^2 = \frac{\sum y_e^2}{\sum y^2} = \frac{19004.10}{19600} = 0.9695$$


Now, let us determine the coefficient of determination, which is obtained as sigma y_e squared by sigma y square that is equal to 0.9695, implying that the set of these two independent variables are able to explain 96.95 percent of the variation of the dependent variable, namely work trip attraction.

(Refer Slide Time: 14:01)


$$S_e = \sqrt{\frac{\sum y_d^2}{(n-3)}} = 17.03 \quad S_d = \sqrt{\frac{\sum y^2}{(n-1)}} = 70.00$$

$S_e < S_d$, Hence O.K.

$$r_{12} = \frac{\sum x_1 x_2}{\sqrt{\sum x_1^2 \sum x_2^2}} = -0.2331$$

So, that is the inference of this particular result, then standard error of estimate calculated as σy_d squared by n minus 3 in this case, which is 17.03 and standard deviation is equal to σy squared by n minus 1, which is 70 and what is the inference here? We have standard error of estimate as well as standard deviation, any suggestion? Regression is ok or not ok? The standard deviation is more than standard error of estimate, the scatter of the observations with respect to the regression line is indicated to be less than the scatter of y observation with respect to its own mean, so obviously, regression is good and we can accept this regression equation, S_e is less than S_d , hence it is ok.


Then, we need to calculate the correlation coefficient for the involved variables or 1, 2 here is nothing that $\sigma x_1 x_2$ divided by square root of σx_1 squared into σx_2 squared, that works out to be minus 0.2331 is a negative value, does not matter there is a negative correlation this particular case.

(Refer Slide Time: 15:53)

The slide displays the following calculations:

$$S_{eb1} = \sqrt{\frac{S_e^2}{\sum x_1^2(1-r_{12}^2)}} = 0.276 \quad S_{eb2} = \sqrt{\frac{S_e^2}{\sum x_2^2(1-r_{12}^2)}} = 0.258$$
$$t_1 = \frac{b_1}{S_{eb1}} = \frac{1.657}{0.276} = 6.00 > 1.77$$
$$t_2 = \frac{b_2}{S_{eb2}} = \frac{1.729}{0.258} = 6.7 > 1.77$$

Since, t_1 & t_2 values are greater than the table value (i.e., 1.77), both b_1 & b_2 are significant.

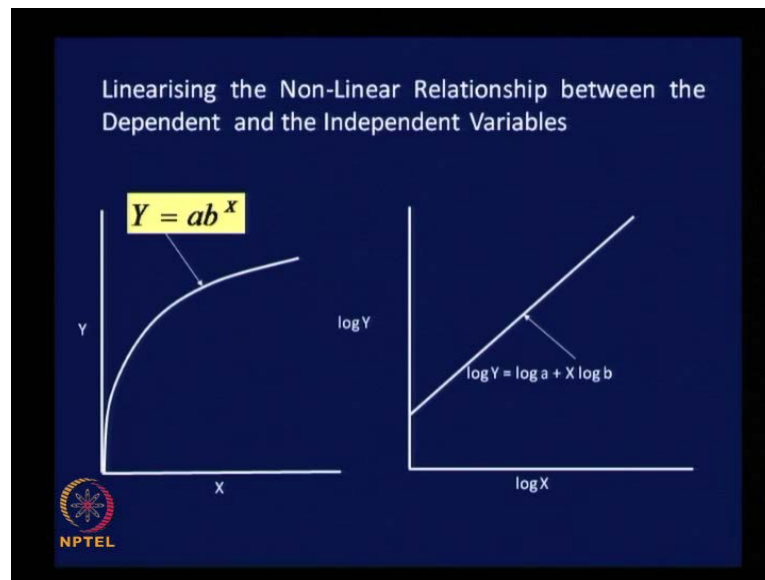


And the standard error of estimate for the regression coefficient b_1 is calculated here as shown in the equation as 0.276 and standard error of estimate for b_2 works out to be 0.258 and t value for the regression coefficient b_1 is 6.00 which is much higher than the table value of 1.77 at 5 percent level of significance.

Hence, the variable X_1 is significant statistically in explaining the variation of Y , that is the meaning of this particular result, t_2 b_2 by S_{eb2} is 1.729 divided by 0.258 that is equal to 6.7 which is again greater than 1.77 at 5 percent level of significance indicating that the variable X_2 is also significant statistically in explaining the variation of y .

Since t_1 and t_2 values are greater than the table, value both b_1 and b_2 is significant. So, that is how we check the statistical significance of the regression equation, to understand how to handle the situation when there is non-linear relationship between dependent and independent variables.

(Refer Slide Time: 17:42)



We are dealing with so far very simple linear relationship between the independent and dependent variables, we dealt with equations of straight lines only; in practice, we may come across situations where the relationship is non-linear. We will just take few examples and see how to convert non-linear relationship into linear relationships. For example, a relationship between X and Y could be of this form with equation Y to be equal to a b power x.

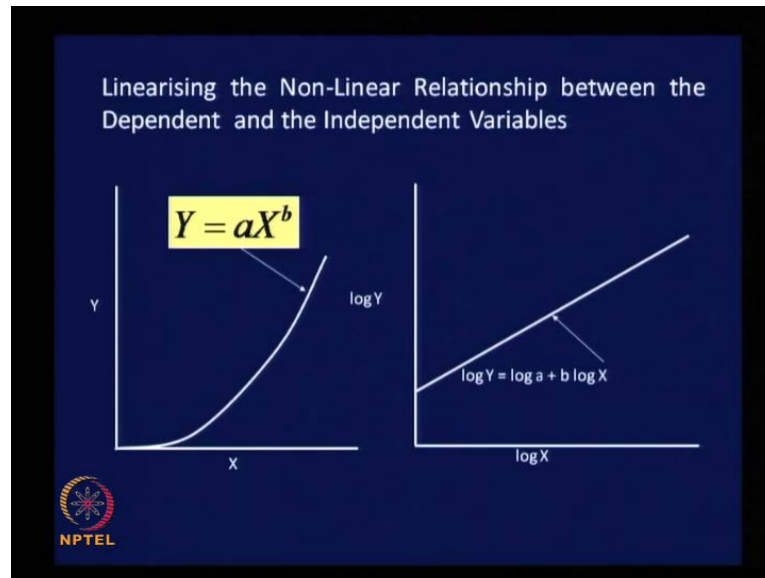
Any suggestion to modify this relationship into linear relationship we need to do simple manipulation to get linear relationship from this non-linear relationship?

Taking logarithm on both the sides, we can convert into a linear form

Good

You can take log on both sides and the equation will become log Y is equal to log a plus x log b, take x and log Y on X and Y axis respectively, and plot your data, you will get a simple linear relationship and this is not the only possibility that could be several possible non-linear relationships.

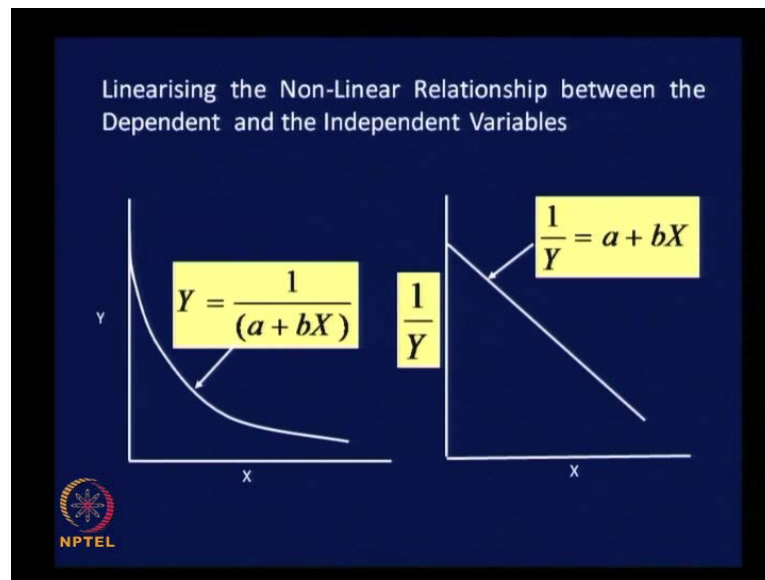
(Refer Slide Time: 19:42)



Let us take another case where the non-linearity could be slightly different from what we have seen now. Let us consider this kind of relationship, Y to be equal to a X power b, any suggestion to get linear relationship? You can follow the same technique and get a linear relationship, take log on both sides again, you will find that you are getting a linear equation $\log X$ $\log Y$ and equation becomes $\log Y$ to be equal to $\log a$ plus $b \log X$.

So, this is how we need to deal with non-linear situations, please remember you will be dealing with \log of X and \log of Y and finally to get the result, you must get the anti log of the values that you are dealing in your regression analysis to get the actual values, that is obvious.

(Refer Slide Time: 20:57)

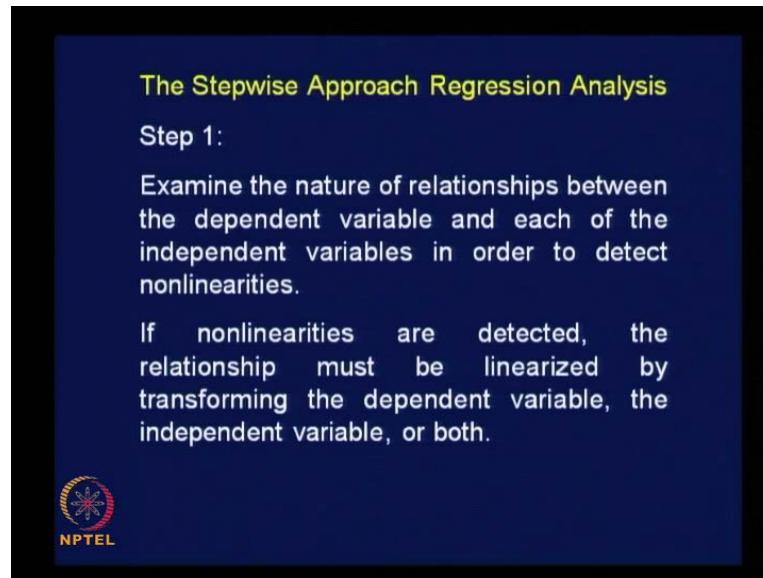


And let us consider one more situation of this kind; Y is equal to 1 by a plus bX , any suggestion for linearizing this relationship? Suppose, we just take inverse of the values on both sides, what will happen, become 1 by Y to be equal to a plus bX again it is a linear, is it not? So, you can just take inverse of the values on both sides and you can get this kind of linear relationship again, 1 by Y to be equal to a plus bX .

So, this is how we must deal with non-linear relationship between dependent and independent variables. Now, we are very clear that regression analysis is a very important analytical tool in developing both trip production as well as trip attraction models. Particularly, it is quite common to use regression analysis in trip attraction modelling, because we deal normally with aggregate data reflecting the land used characteristics, very rarely category analysis is used for trip attraction analysis.

So, it would be better for us to understand the important steps involved in a general regression analysis so that you know clearly given a set of data of independent and dependent variables, how to proceed with the analysis step by step. So, I will be just giving you an example illustrating the different steps involved in a general regression analysis and we will understand the steps with the aid of an example concerning trip attraction modelling.

(Refer Slide Time: 23:30)



The Stepwise Approach Regression Analysis

Step 1:

Examine the nature of relationships between the dependent variable and each of the independent variables in order to detect nonlinearities.

If nonlinearities are detected, the relationship must be linearized by transforming the dependent variable, the independent variable, or both.

NPTEL

The stepwise approach to regression analysis, step 1 is this; examine the nature of relationships between the dependent variable and each of the independent variables in order to detect nonlinearities that is what we are discussing just before few minutes.

We need to examine the relationship, nature of relationship between the dependent variable and each of the independent variables, let us see very important step. And if nonlinearities are detected, the relationship must be linearized by transforming the dependent variable, the independent variable or both, depending upon the requirement, linearized relationship that is the first important step. How to identify the natural relationship? How will you do that in practice, we will have a set of values of Y and set of values of X_1 , X_2 , and X_3 and so on.

Let's say thousands of observations of Y , X_1 , X_2 , X_3 , how will you understand the natural relationship between Y and X_1 ? What is the exercise to be gone through? Simply make a scatter diagram relating Y and X_1 , you will get dots as a plot, then try to fit a curve and straight line is also a curve with a extreme value of radius, that is it.

If you are able to fit a straight line with reasonable goodness of fit statistics, then your relationship is linear. And if you find that a curve fits better than a straight line, then the relationship is curvilinear and you can get the corresponding equation for the curve, get the equation for the curve, and then work out the methodology to linearize the relationship.

So, first important step is, make a scatter diagram involving the dependent variable and one independent variable at a time, do this exercise for all the independent variables and you can get the nature of relationship and if necessary, you can linearized the relationship **clear**, so this is step 1.

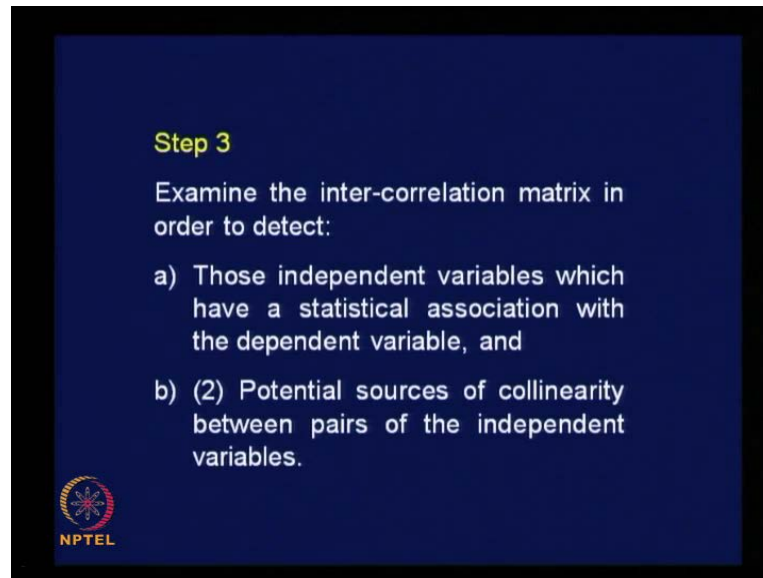
(Refer Slide Time: 26:28)



So, after linearizing, step 2 is this, develop an inter correlation matrix involving all the variables including both the dependent and the independent variables, get the correlation coefficient values for all the involved variables and put the values in a matrix form, that is what is indicated here as correlation or inter correlation matrix.

Get the value of correlation coefficient between Y and X 1, Y and X 2, Y and X 3, then between X 1 and X 2, X 1 and X 3, then X 2 and X 3 so that you get correlation between all the involved variables. Once you get this values, and put it in a tabular form, you will get automatically a matrix involving only variables. So, this is the very important second step, later on you will know why you are developing this matrix.


(Refer Slide Time: 27:40)



Step 3

Examine the inter-correlation matrix in order to detect:

- a) Those independent variables which have a statistical association with the dependent variable, and
- b) (2) Potential sources of collinearity between pairs of the independent variables.

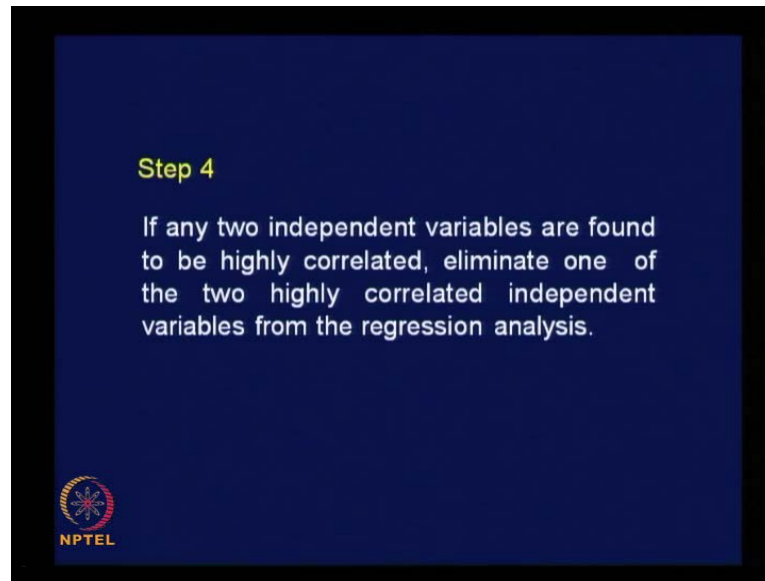
 NPTEL

Then, examine the inter correlation matrix in order to detect the following. Once you prepare the matrix, do the following exercise; detect first those independent variables which have a statistical association with the dependent variable. Find out whether all your independent variables are really statistically associated with your dependent variable, whether they have meaningful correlation with Y, look at the each of the independent variables and look at the correlation and check whether the correlation is significant in general and meaningful. So, that is what is implied by this particular step.

And then, potential sources of collinearity between pairs of independent variables are to be detected, what do you understand by collinearity between independent variables? It is nothing but the high level of correlation between independent variables or any two variables is termed as collinearity. When two variables will be highly correlated, when they are of more or less similar characteristics, their nature should be more or less same, and then they become highly correlated **right**. When you have to highly correlated independent variables, is it necessary to use both the variables in your regression analysis?

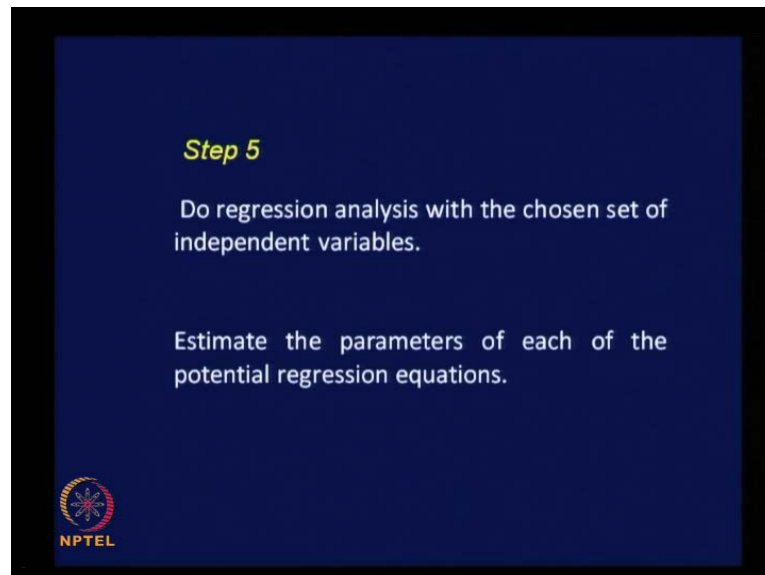
Because both will be explaining the same aspect of the dependent variables, there is no need to use both the independent variables, your regression will not be effective **right**. So, that is the purpose of doing this particular exercise, identify the level of collinearity between the involved variables.

(Refer Slide Time: 29:59)



Step 4, if any two independent variables are found to be highly correlated, eliminate one of the two highly correlated independent variables from the regression analysis for the obvious reason that we discussed just now. You have to eliminate one, later on we will see on what basis to eliminate one of the two highly correlated independent variables.

(Refer Slide Time: 30:34)



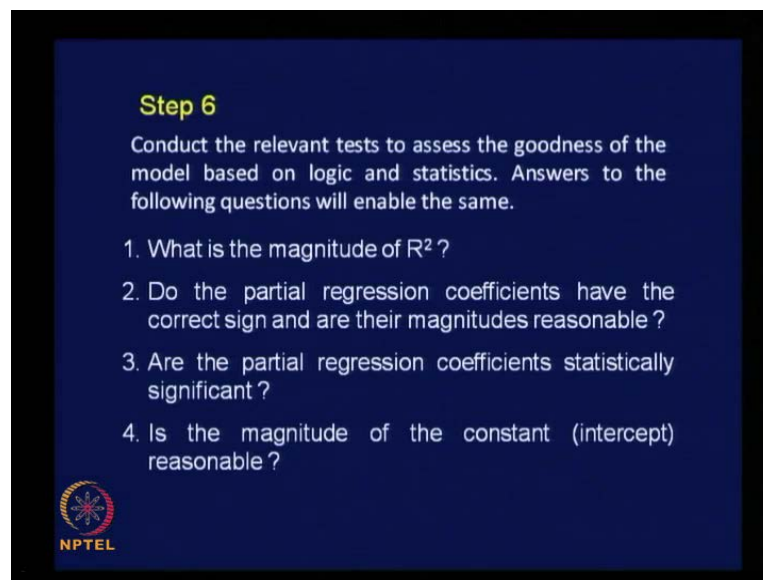
Then step 5, do regression analysis with the chosen set of independent variables after discarding some other variables if necessary, based on the collinearity aspect, then estimate the parameters of each of the potential regression equations, this implies that it

is possible to develop more than one regression equation given a set of independent variables and the dependent variable.

You can have several combinations, you can choose one independent variable and try to develop a regression equation relating the dependent variable and independent variable, then two independent variables and your dependent variable then three independent variables and the dependent variable.

And you can examine each of these regression models, and choose the one which is more suited to your requirement. So, that is what is implied by this particular statement, estimate the parameters of each of the potential regression equations.


(Refer Slide Time: 31:46)



Step 6

Conduct the relevant tests to assess the goodness of the model based on logic and statistics. Answers to the following questions will enable the same.

1. What is the magnitude of R^2 ?
2. Do the partial regression coefficients have the correct sign and are their magnitudes reasonable ?
3. Are the partial regression coefficients statistically significant ?
4. Is the magnitude of the constant (intercept) reasonable ?

 NPTEL

Step 6; conduct the relevant tests to assess the goodness of the model based on logic and statistics. We should not blindly calculate the values of coefficient of determination and comparison of standard error of estimate and standard deviation anti statistics and say that the regression model is fine.

There are certain logical aspects also to be looked into very carefully, a regression model should be both logically and statistically be valid, and answers to the following questions will enable the checking.

The first question is this, what is the magnitude of R squared, coefficient of determination, we need to answer this question, and then do the partial regression

coefficients have the correct sign and are their magnitudes reasonable? Why we should worry about the sign of the partial regression coefficients? Partial regression coefficients are nothing but b_1 , b_2 , and etcetera as we have seen in the regression equations.

These are partial regression coefficients, how the signs of these partial regression coefficients are important? In the previous example, we tried to relate work trips with the number of employment opportunities and manufacturing and then retail trade.

Logically, both the independent variables are to have positive relationship with the dependent variable, more the retail trade employment, more will be the work trip attraction, is it not? Suppose, when you develop a regression equation, you get a negative sign for the regression coefficient corresponding to retail trade employment.

Then, your equation is not really going to explain work trip attraction, is it not? It **it** gives a different meaning. So, that is where we need to be very careful, because we collect data pertaining to a set of independent variables and dependent variable, and there could be error in the data collection, or there could be some other source of erroneous entries in your data set which may result in illogical regression equations, you have to be very careful and check for the logical correctness of the equation first, before you go in for statistical checks, I would say **right**.

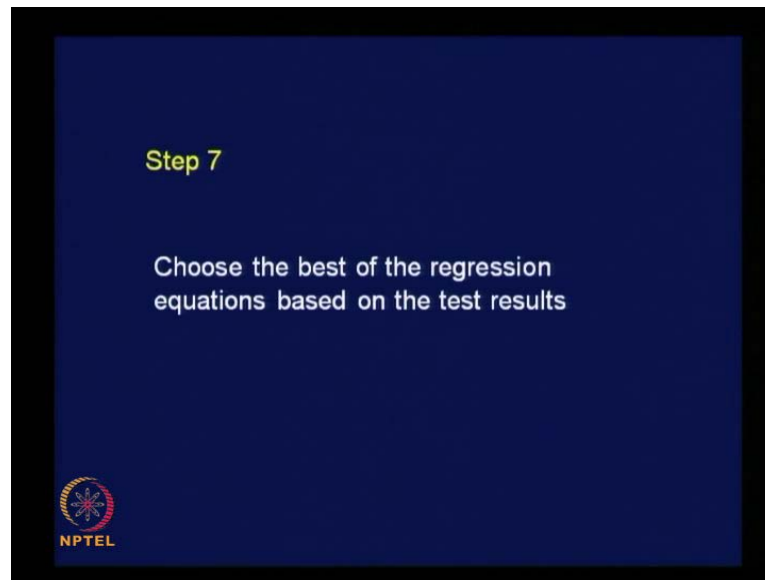
And their magnitudes should be reasonable, later on we will discuss about the significance of the magnitudes of the partial regression coefficients. So, their sign as well as magnitude are to be reasonable and it should reflect the overall correctness of distribution situation that we are dealing with.

Then third question is, are the partial regression coefficient statistically significant? How do you do that, statistical significance of partial regression coefficient? We have already done it, do the pre test that gives the, gives you the information about the significance of your partial regression coefficients. Then also check our answer, this question is a magnitude of the constant the intercept that you get is reasonable, you should not get huge values, **right** these intercept actually are the values which are not explained by the independent variables, these are residual values of the regression process.

So, we should see that the proportionate effect of this intercept is not harming or providing illogical results for extreme values of the independent variables, large values

of independent variables or small values of independent variables, there would be some significance of this particular intercept constant. So, we should be clear that the value of intercept constant is reasonable, now we are ready to look at an example and go through all the six steps.

(Refer Slide Time: 36:59)




Of course, the final steps is obvious, you have to chose the best of the regression equations based on the test results, you develop a series of regression equation when you have more independent variables and chose one which is the best among the set of equations.

(Refer Slide Time: 37:21)

Numerical Example: **Land-Use and Trip Data**

Zone No.	Employment				Peak – hour trips attracted
	Total (X_1)	Manufacturing (X_2)	Retail and Service (X_3)	Other (X_4)	
1	9,482	6,820	2,547	115	9,428
2	2,010	111	1,899	0	2,192
3	574	228	87	259	330
4	127	0	127	0	153
5	3,836	2,729	813	294	3,948
6	953	101	773	79	1,188
7	223	165	58	0	240
8	36	6	30	0	55




Let us now take an example, involving four independent variables and a dependent variable; the independent variables are total employment, employment in manufacturing, employment in retail and service, and other categories of employment. And in the last column, total of peak hour trips attracted, trips for all purposes is given; total of trips attracted during peak hour is given in the last column.

(Refer Slide Time: 38:17)

Land – Use and Trip Data

Zone No.	Employment				Peak – Hour Trip Attracted
	Total (X_1)	Manufacturing (X_2)	Retail and Service (X_3)	Other (X_4)	
9	2,223	1,550	499	174	2,064
10	272	0	166	106	280
11	50	0	48	2	52
12	209	36	173	0	230
13	410	140	7	263	420
14	11,023	10,932	63	28	9,654
15	527	188	325	14	450
16	183	123	59	1	130



Using this data, you need to develop a trip attraction equation, the information shown here pertains to 8 traffic zones and similar information is given for 8 more traffic zones,


totally we have 16 zones. Information of this kind is available to us for 16 traffic zones, total employment, employment in manufacturing, employment in retail service, other types of employment and peak hour trip attraction.

(Refer Slide Time: 38:56)

Simple Correlation Matrix

	X ₁	X ₂	X ₃	X ₄	Y
X ₁	1.000	0.978	0.486	0.110	0.996
X ₂		1.000	0.297	0.068	0.958
X ₃			1.000	0.073	0.552
X ₄				1.000	0.124
Y					1.000

X₁ = Total employment X₄ = Other employment
 X₂ = Manufacturing employment Y = Peak-hour trips Attracted
 X₃ = Retail and Service employment



Obviously, the dependent variable is peak hour trip attraction is Y, and we have X 1, X 2, X 3 and X 4 and we can assume that we have gone through the first step. At this stage, we can assume that we have checked for the nature of relationship between the dependent and each of the independent variables at this particular case and all are found to be linear.

And then, we are checking for collinearity between variables by developing inter correlation matrix, and this is the result of the calculation of correlation coefficient between a set of independent as well as dependent variables and the result is shown here. And you can see the correlation between X 1 and Y is how much? What is the value? 0.996, X 1 is highly correlated with the dependent variable Y, X 1 is what? Total employment opportunities, and total employment opportunities is able to explain very well the trip attraction during peak hour that is the meaning.

And if you look at the correlation between X 2 and Y, it is also high point 0.958 and the correlation between X 3 and Y is moderate 0.552 and the correlation between X 4 and Y is still less 0.124 only, **right**. So, we know now the nature of relationship between the set of independent variables and the dependent variable and we find that all the independent

variables are correlated with the dependent variable, and the correlation is very high in respect of X 1 and X 2. This implies that, it is possible for us to develop trip attraction model with only X 1 or X 2, because that itself will be explain more than 90 percent of the variations of the dependent variable.

What is the desired R squared value in regression analysis, any suggestion? In practice, you may get R squared values of wide range, suppose you get R squared value of 0.4 in one regression analysis, 0.6 in another case, 0.75 in third case, 0.8 in some other situation, 0.9, 0.95.

What is the limiting value to accept the model as a valid model? There is no hard and fast rule; it depends upon the problem that you are dealing with. If you are dealing with a every complex problem, representing complex socio economic characteristics which cannot be that easily quantified, and you are managing to pick some variables to explain variation of the dependent variable. In such a situation, you may say that a regression model with R squared value of 0.75 is fine, acceptable.

Also, it depends upon the level of accuracy that you expect **right**. So, in general you can understand that, when you are dealing with complex problems at macro level, you cannot anticipate very high R squared value. So, you should be realistic and should be able to decide about the correctness of the regression equation based on R squared value and you should substantiate why you are accepting a low R square valued for a particular situation. And if you are dealing with a micro level condition with well defined data set, then you can go in for, or you can anticipate, or you can fix a very high R squared value has a criterion to consider the equation to be valid, there is no hard and fast rule about the minimum value for R squared to declare equation to be good or valid, **right**.

So, accordingly in this case, you can easily say that with X 1 and X 2 alone, you will be able to reduce the reasonably good trip attraction model, we will find the correlation is very high. And since X 3 and X 4 are also correlated with Y, it should be possible for us to include these independent variables also to add little more clarity or effectiveness to your model **right**.

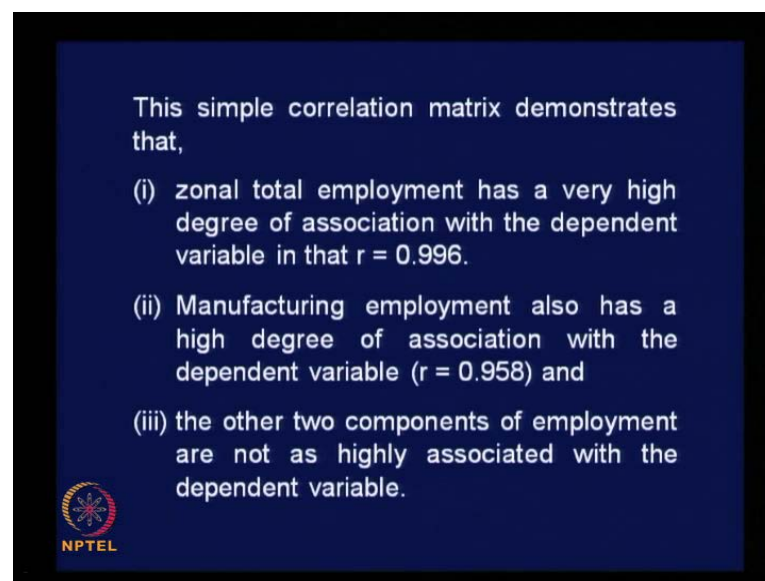
Another important point to be understood here is the collinearity between X 1 and X 2. What is the value of correlation coefficient of X 1 X 2 here? 0.978 is the value of correlation between X 1 and X 2, X 1 is total employment and X 2 is manufacturing

right. So, these two independent variables are highly correlated, this implies that there is no point in using X_1 and X_2 together to develop your regression equation, because they are already highly correlated, either you must do only with X_1 or only X_2 . Suppose, you have to eliminate 1, do with only 1, which one will be eliminated? What is the basis? To take a decision in this particular regard, look at the correlation of X_1 and X_2 with Y , X_1 is correlated to a greater extend compared to X_2 .

So, obviously, to explain the variation of Y , X_1 is going to be more useful to you than X_2 , and X_1 and X_2 are highly correlated. So, it is logical to draw X_2 and take X_1 for your analysis **right.** So, that is how you must decide about elimination of one of the two highly correlated independent variables, basis is where independent correlations with the dependent variable clear. So, that is how we must take the decision about dropping one of the highly correlated independent variables.


So, this is how an inter correlation matrix is a very important analytical step in any regression analysis, first for most important steps, look at the correlation matrix carefully, understand their effect of all the independent variables on the dependent variable as well as inter correlation between each of the among the independent variables before we proceed further.

(Refer Slide Time: 47:05)



This simple correlation matrix demonstrates that,

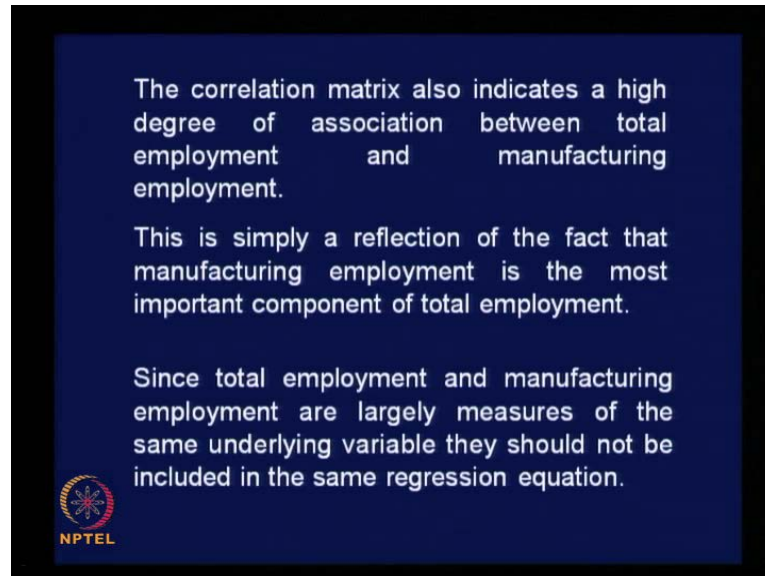
- (i) zonal total employment has a very high degree of association with the dependent variable in that $r = 0.996$.
- (ii) Manufacturing employment also has a high degree of association with the dependent variable ($r = 0.958$) and
- (iii) the other two components of employment are not as highly associated with the dependent variable.

 NPTEL

Now, let us go to the next step and this simple correlation matrix or inter correlation matrix demonstrate that zonal total employment has a very high degree of association

with the dependent variable r is 0.996, that is what we have seen and manufacturing employment has r value of 0.958.

(Refer Slide Time: 47:37)




The other two components of employment are not as highly associated with the dependent variable, that is what we have seen, the correlation matrix also indicates the high degree of association between total employment and manufacturing employment, this is simply a reflection of the fact that manufacturing employment is the most important component of total employment, that is why they are highly correlated, may be of the total of say 100, more than 80 are in manufacture, that is how they are highly correlated.

Since, total employment and manufacturing employment are largely measures of the same underlying variable; they should not be included in the same regression equation. If you like, you can develop separate regression equations, but not in the same regression equation.

(Refer Slide Time: 48:47)

In this problem, the analyst has the option of the following regression equations:

$$Y = a + b X_1 \quad (A)$$
$$Y = a + b X_2 \quad (B)$$
$$Y = a + b_1 X_2 + b_2 X_3 \quad (C)$$
$$Y = a + b_1 X_2 + b_2 X_3 + b_3 X_4 \quad (D)$$



These are the possible regression models with the four independent variables is it not? You can develop a simple model relating Y and X 1 with a linear form, Y to be equal to a plus b X 1, or you can develop another model of the form Y is equal to a plus b X 2, third model Y to be equal to a plus b 1 X 2 plus b 2 X 3, and fourth possibility is Y to be equal to a plus b 1 X 2 plus b 2 X 3 plus b 3 X 4. So, we can try all the four models, check each of these models for logic as well as statistical aspects, and then chose the one which is the best.

(Refer Slide Time: 49:27)

The four regression equations listed above [Eqs. (A)-(D)] when fitted to the given data, will give the following results:

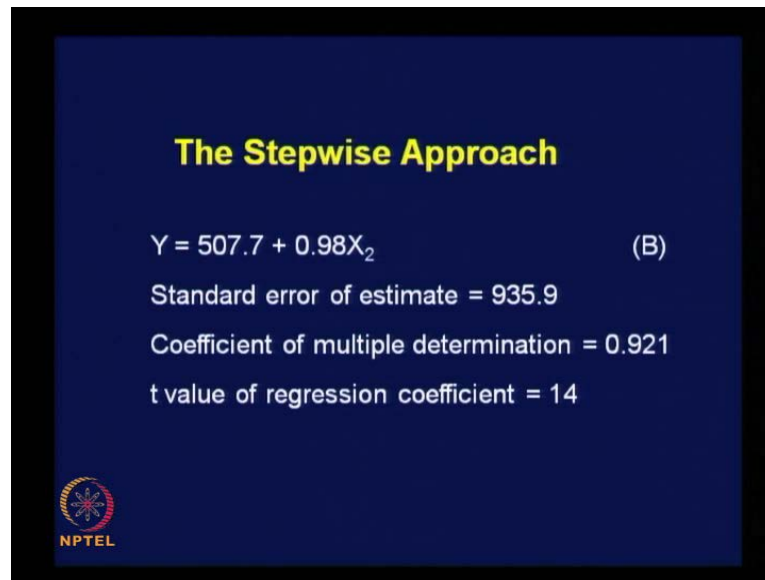
$$Y = 61.4 + 0.93X_1 \quad (A)$$

Standard error of estimate = 288.4
Coefficient of determination = 0.992
t value of regression coefficient = 42



And if you go for development of these models, I will just show you the results alone; this will be model A, the first one. You will get a regression equation as Y to be equal to 61.4 plus 0.93 X 1, and standard error of estimate is 288.4, coefficient of determination is 0.992 r squared and t value of regression coefficient 42.

(Refer Slide Time: 49:57)




The Stepwise Approach

$Y = 507.7 + 0.98X_2$ (B)

Standard error of estimate = 935.9

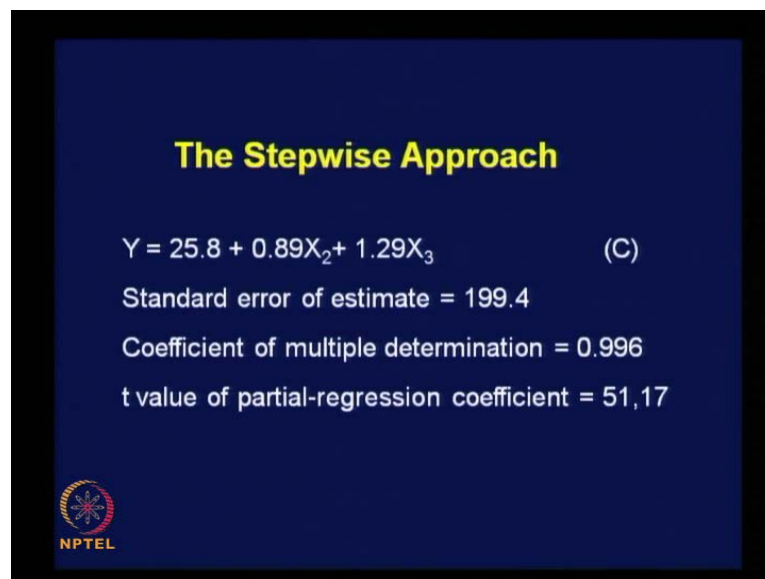
Coefficient of multiple determination = 0.921

t value of regression coefficient = 14

 NPTEL

And second model, the result is this, Y is equal to 507.7 plus 0.98 X 2, standard error of estimate is 935.9, coefficient of multiple determinations is 0.921 and t value of the regression coefficient is 14.

(Refer Slide Time: 50:22)




The Stepwise Approach

$Y = 25.8 + 0.89X_2 + 1.29X_3$ (C)

Standard error of estimate = 199.4

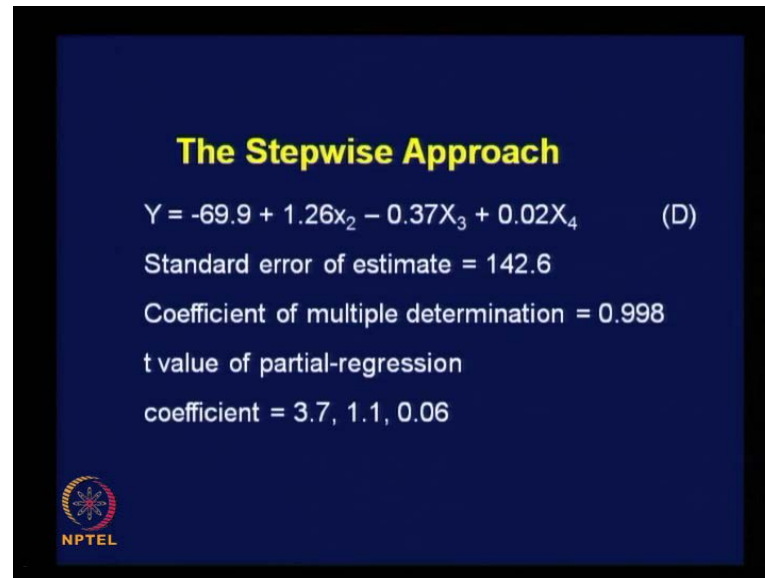
Coefficient of multiple determination = 0.996

t value of partial-regression coefficient = 51,17

 NPTEL

Third model, Y to be equal to 25.8 plus 0.89 X 2 plus 1.29 X 3 and standard error of estimate here is 199.4, coefficient of multiple determination 0.996, very high, and t value of partial regression coefficient are 51 and 17, because we have two variables here.

(Refer Slide Time: 50:53)




The Stepwise Approach

$$Y = -69.9 + 1.26X_2 - 0.37X_3 + 0.02X_4 \quad (D)$$

Standard error of estimate = 142.6

Coefficient of multiple determination = 0.998

t value of partial-regression
coefficient = 3.7, 1.1, 0.06

 NPTEL

And fourth model, Y is equal to minus 69.9 plus 1.26 X 2 minus 0.37 X 3 plus 0.02 X 4, standard error of estimate is equal to 142.6, coefficient of multiple determination r square is 0.998, t value of partial regression coefficients are 3.7, 1.1 and 0.06. Now, the question is which equation to choose finally, as your final regression model for this particular case, any suggestion? Or what are the aspects to be looked into to choose the best model? It is very simple; you must look at the logical aspects and statistical aspects, what are the logical aspect to be looked into?

There should be a logical relationship between your independent variable and dependent variable, each of the independent variable and dependent variable. In this case, we are using the independent variables to explain the peak hour total trip. Obviously, the relationship between your dependent variable and independent variable should be positive, because all these factors should add to the trip rate, where is employment in various sectors.

So, if you have negative sign in some case, probably there is some problem with their particular model that is how we need to check logical aspect. Then statistical aspects are coefficient of determination r squared, then comparison of standard error of estimate

with standard deviation and then checking for the individual statistical significance of each of independent variables based on their respective t values, is it not? This is how we need to check.

So, please do these exercise at home, and we will discuss on this particular aspect later when we meet in the next class. To summarize what we did today, we discussed mainly about trip attraction modelling, and we have seen one numerical example involving two independent variables and different steps in the regression model was discussed in detail.

And then, we tried to understand the step wise process of regression in general, starting from understanding the relationship or the nature of relationship between dependent and independent variables. And then, development of inter correlation matrix to study the level of relationship between each of the independent variables and the dependent variables as well as the possible multi collinearity between independent variables.

Then, we discussed about possibility of developing a set of trip attraction models for a given data and the process or the procedure to choose the best model among the set that we develop. So, with this, we will conclude for today and discuss the rest of it in the next class.