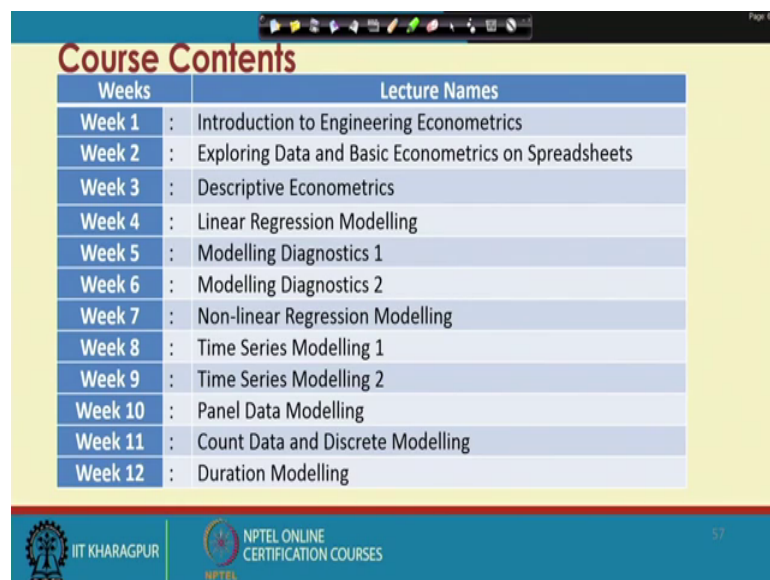


Engineering Econometrics
Prof. Rudra P. Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 22
Linear Regression Modelling (Contd.)

Hello everybody. This is Rudra Pradhan here. Welcome to Engineering Econometrics. Today, we will continue with Linear Regression Modelling that to the issue of multiple regression modelling.

(Refer Slide Time: 00:35)



Weeks	Lecture Names
Week 1	: Introduction to Engineering Econometrics
Week 2	: Exploring Data and Basic Econometrics on Spreadsheets
Week 3	: Descriptive Econometrics
Week 4	: Linear Regression Modelling
Week 5	: Modelling Diagnostics 1
Week 6	: Modelling Diagnostics 2
Week 7	: Non-linear Regression Modelling
Week 8	: Time Series Modelling 1
Week 9	: Time Series Modelling 2
Week 10	: Panel Data Modelling
Week 11	: Count Data and Discrete Modelling
Week 12	: Duration Modelling

So, in the last lecture we have discussed about this particular component. So, it is a issue here we have dependent variables and a many independent variables and the idea is to predict the dependent variable with at least two independent variables in the systems. So, how we have to actually address this problems and how to analyze we have briefly highlighted in the last lecture.

(Refer Slide Time: 01:05)

Real Estate Data

Observation	Market Price (\$1,000) Y	Square Feet X ₁	Age (Years) X ₂
1	63.0	1,605	35
2	65.1	2,489	45
3	69.9	1,553	20
4	76.8	2,404	32
5	73.9	1,884	25
6	77.9	1,558	14
7	74.9	1,748	8
8	78.0	3,105	10
9	79.0	1,682	28
10	63.4	2,470	30
11	79.5	1,820	2
12	83.9	2,143	6

Observation	Market Price (\$1,000) Y	Square Feet X ₁	Age (Years) X ₂
13	79.7	2,121	14
14	84.5	2,485	9
15	96.0	2,300	19
16	109.5	2,714	4
17	102.5	2,463	5
18	121.0	3,076	7
19	104.9	3,048	3
20	128.0	3,267	6
21	129.0	3,069	10
22	117.9	4,765	11
23	140.0	4,540	8

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

And, now what we can do; we can coordinate this particular technique with a kind of engineering problem and let us see how is the kinds of outcome and how we can address these outcomes to predict the dependent variables you know particularly with respect to two independent variables.

So, now we have here real state data here the objective is to predict housing price, subject to two independent variables, that is the square feet, the area of the house and the age years of a kind of housing. So, that means, technically when a particular house is built and over the time. So, there is a kind of depreciation. So, as a result the house price may get if actually affected.

So, actual housing price depends upon the how old is that particular building and against how is the size of that particular house. So, theoretical understanding is that if the square foot or size of the house is a high or more than by default housing price should high. If the square feet area is you know lower one then the housing price can be lower one. And similarly when the house is actually old then housing price will be lower and when the particular house is a new one then by default housing price can be high. That means, when it will you know over the years when the number of years will start increasing then there is a high chance that particular variable can affect the housing price negatively.

So, this is what actually the theoretical understanding before you apply the regression modelling that too trivariate modelling or multiple regression modelling with the two

independent variables case. So, we can actually connect this problem and then check how these variables can be affected or I mean say to study the impact of these two variables to dependent variable. That means, technically the issue is a means a how is the impact of square feet and age on housing price and what are the wage we can actually investigate the particular process and how to predict the housing price with respect to the square feet and you know age of the house.

So, compared to last problems which we have already connected in the case of part one where we have two variables that is connected to airline industry and where the objective is to you know predict airline cost with respect to airline passengers. Here the problem is different in the context of trivariate modelling in that too multiple regression modelling with the two independent variables.

So, now, in this case obviously, the same sample point we cannot keep we must have actually more sample point. In the first problems we have used 12 samples case means sample size is 12 and here we are actually using a problems with you know 23 samplings. So, that means, this is what actually the problems. So, these are all sample observations. So, first 12 observations and again remaining observation up to 23; so, that means, technically so, we have n equal to 23, that is the sample size and then K equal to heres number of variables. So, in the case of bivariate setup K was you know 2. So, in the case of trivariar or multiple regression modelling with two independent variables then that becomes K equal to 3.

So, that means, it is the simple you know kind of calculation is the how many variables in the systems and by default how many parameters are in the systems including the intercept and since in the case of bivariate case we have two variables so, by default K equal to 2. In the case of multiple regression with two independent variables, so, including dependent variables so, we have a 3 variables as a result K equal to 3.

So, now, for multiple regression modelling with two independent variables so, the game is n equal means for this problem game is an equal to 23 K equal to 3 and as a result degree of freedom degree of freedoms n minus K which is actually 20. So, this will be 20. So, this is how the typical understanding. So, before you start the process as usual you can go through descriptive statistics and the correlation matrix. The correlation matrix by default will give you the indication about the X_1 relationship with Y , the kind

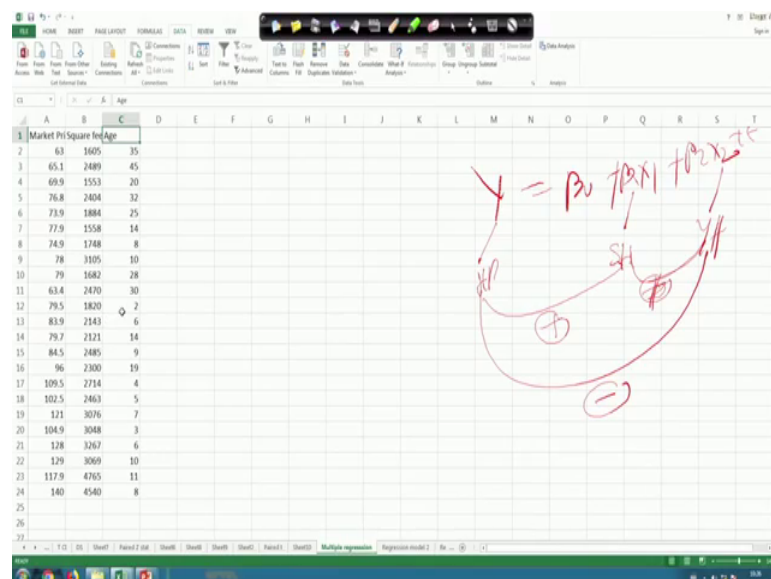
of nature of the relationship and whether there is a link or not having link then X_2 X_2 on Y and then finally, X_1 and X_2 .

For instance; so, we can create a link like this Y X_1 and X_2 similarly Y X_1 and X_2 . So, you can have a correlation matrix and then check; what is the kind of relationship; so here Y Y X_1 X_1 and X_1 X_2 . So, ultimately yx ones you must have here you must have here then Y X_2 you must have here and Y X_2 you must to here. So, this can be valid this can be valid, this can be valid, and in between. So, X_1 and X_2 so, X_1 and X_2 this place or X_2 and X_1 this place. So, this is what actually issue about a relationship between X_1 and X_2 .

So, that means, the problem will be like this you know Y with respect to X_1 and with respect to X_2 . So, X_1 and X_2 that correlation coefficient will give and Y X_1 is here and Y X_2 is Y X_2 is here. So obviously, we need more correlation with Y X ones and Y X_2 , but you know should not have any correlation or very low correlation between X_1 and X_2 . This what the theoretical overview and the descriptive statistic and correlation matrix can give and you should actually check before you go for the estimation of the regression you know models with respect 2 dependent variable and two independent variables.

So, this is how the problem. So, let us see how we can go for this particular problem and to solve these.

(Refer Slide Time: 08:15)



So, so, in that case what I can do I will take you to the softwares directly then I will show you how these variables can be used in through this softwares to get these parameters. So, I mean the first step of this processing to predict the housing price with respect to size of the house and the years of the house we need to fix the model like this. So, the technical of the model can be fixed it is like you know Y equal to β_0 plus $\beta_1 X_1$ plus $\beta_2 X_2$, and then in error terms. So, this is what actually a housing price and this is what actually size of the house and this is actually years of the house, ok.

So, this and this we expect positive relationship and this and this we expect negative relationship and then this and this we expect no relationships. So, there should not be no. So, this is how the a means theoretical kind of understanding and theoretical requirement before we start the processing, but the actual case is maybe something different, which we can actually get to know after using this data connect with this models and obtain the estimated model by using this software.

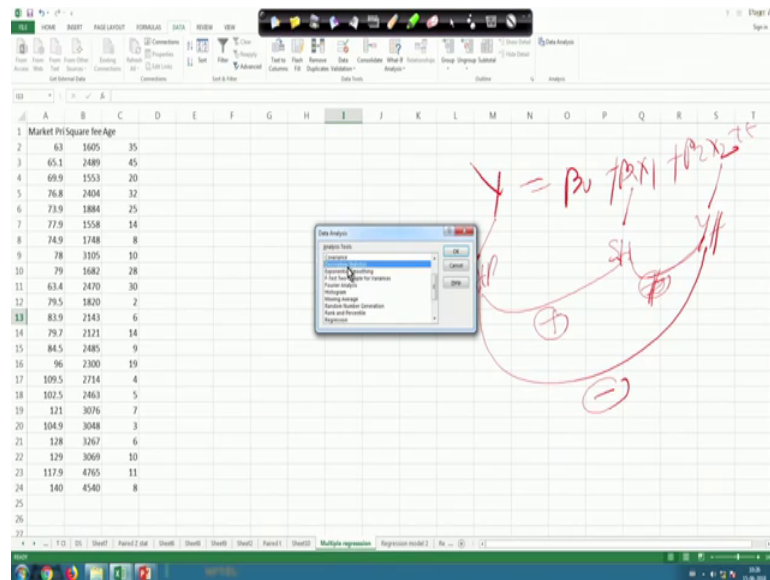
So, what will we do? Technically, you can start doing actually descriptive statistics, then correlation matrix and then after that we can get a regression output. So, let us see so, first of all with this is what actually what actually the kind of data. So, what I will do I will just little bit you know enlarge, ok. So, this what the three variable case this is house price and this is the square feet size of the house and this is the number of years of the house how old these particular house over the time.

So, as a result this is actually a completely cross sectional data. The different locations we have we know gather the data and that is what the kind of the requirement of regression modelling and that to use the random sampling to collect you know randomly data from different locations and then predict the kind of scenario that will be that will give you some kind of unbiased results and if it is time series data then in one particular locations then you can what the time you have to you have to collect the data. By default if it is a particular location and particular house, over the time then what will we do the same house and number of years of old; then that that will going on, then you will check how is the price housing price housing price is affected.

Then that is the time series modelling you know structure, but in this case we are using cross sectional modelling to know the particular scenario. That means how what should be the housing price if you know the number of years will increase or decrease and the

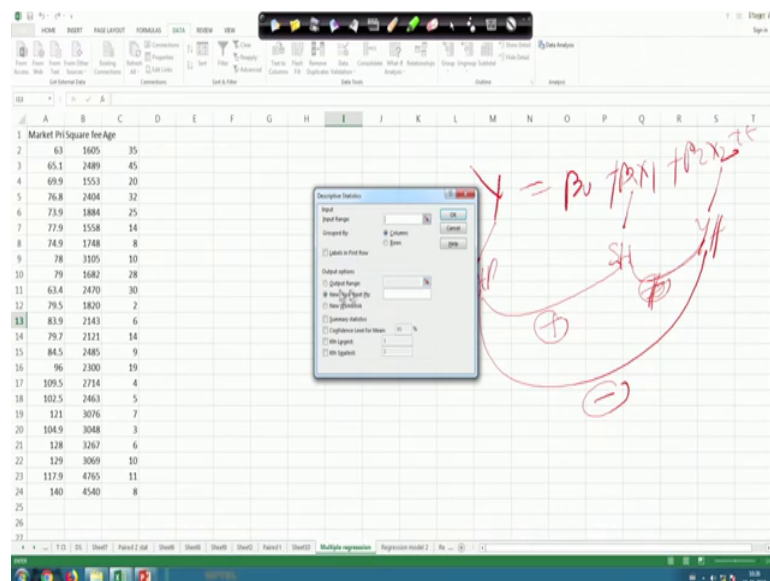
house size will be increase or decrease. So, that means, technically what should go you know housing price in the future kind of requirement that is with respect to size of the house and years of the house. So, this is how the kind of case.

(Refer Slide Time: 11:43)



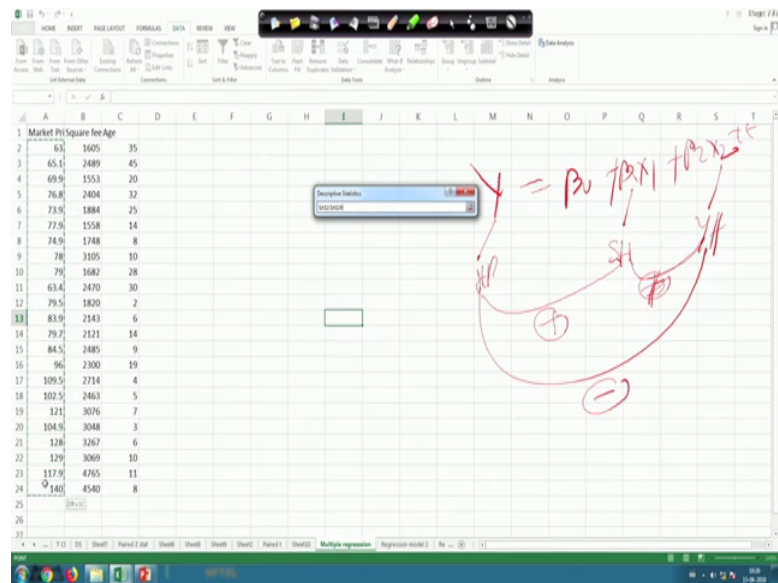
So, now what will we do? So, you just go to the data analysis package and click here and first of the requirement is descriptive statistics.

(Refer Slide Time: 11:48)



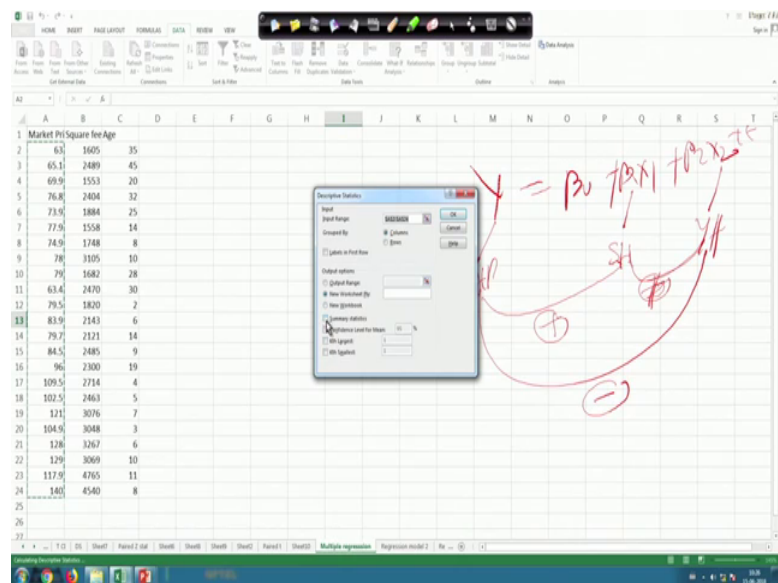
And, put the descriptive statistic here and you can go one by one.

(Refer Slide Time: 11:53)



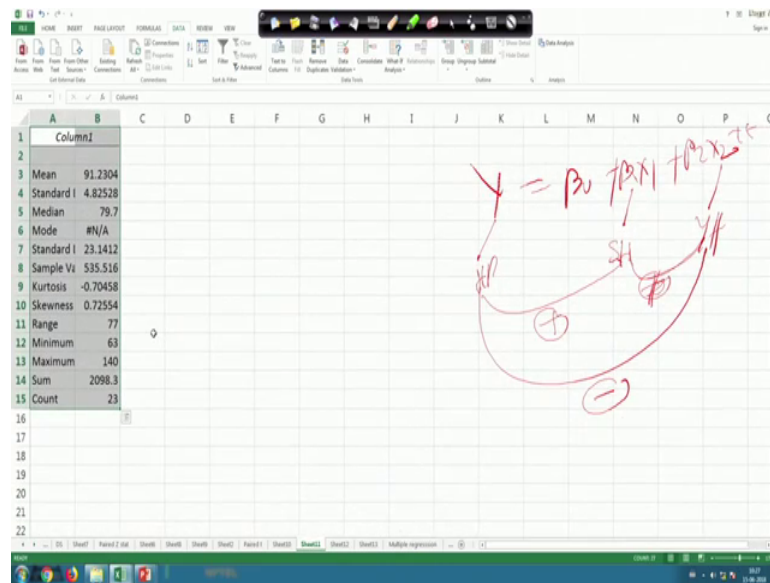
So, by keeping highlight this one.

(Refer Slide Time: 11:57)



Then, you can get this is what the case of descriptive statistic. You just put summary statistic here.

(Refer Slide Time: 12:02)



So, you will get summary statistics for X 1 that is a housing price. Again go to the data analysis put the descriptive statistic indicate a second variables and you will get second variable informations. Then again so, you go to the data analysis and again highlight descriptive statistic and give the indication about the third variable. So, you will get the descriptive statistic of third variables. In fact, so, you can indicate all the variables simultaneously, then you can get the descriptive statistic of the a three variables one at a time.

And, after knowing this so, this is descriptive statistic of one so, you can see the kind of in a things, just you have to check that you know mean standard deviation is minimum maximum and then count the particular range the range should not be 0 and range will usually 0 when all the all the informations are same. That means there is no difference between maximum and minimum.

So, before we start the processing of whether it is a bivariate structure or trivariate structure or any kind of multiple structure. So, the variable or data for a particular variable should not be uniform and that too you can dictate through the difference between minimum maximum and the range and standard deviations. In fact, if all the data points are you know same for a particular variable then standardization will be 0 by default. So, if standard deviation will be 0 and minimum maximum are same, the range

equal to 0; then by default this particular dataset for that particular variable should not be used for the empirical processing.

So, you mean that particular variable is very important, but if the data is not supporting so, you can simply drop the variable. If you drop the variable then software will not give you any kind of result because it is taking a lead you know technical defect case where you now same informations are there. If all the informations are same it cannot give any kind of impact to the dependent variable and if we the dependent variables data set is same so, there is no need to again predict the particular variables.

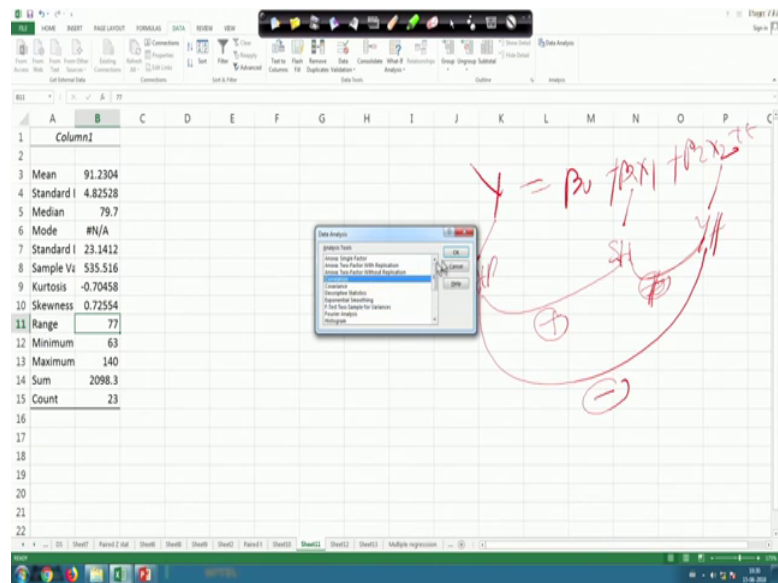
So, you need to predict the dependent variables, if there is a variation over the timer or over the cross sectional you need. If there is no such variations then you simply drop the idea to go for the kind of investigation and that too to predict the a independent variable in this case housing price with respect to the independent variables. So, these are the basic things you must have understanding before they start the process and do the needful.

So, that means, technically what I like to submit that you must have clear cut information about all these variables and that too with respect to data. So, theoretical understanding of these variables and then the particular data set for these variables. So, the theoretically these variables should be a correctly identify, they have logical kind of linkage or theoretical kind of linkage case and these data availability for these variables should be also consistent as per the process of investigation.

So, the investigation process requirement is that you know there should be some variations among these data points whether it is a dependent variable or independent variables or independent variables. If these variations are not there in that particular variable you mean it is important theoretically logically. And as per the engineering problems requirement still you should not use for any kind of regression modelling and that too predict and forecasting the engineering problem or engineering requirement.

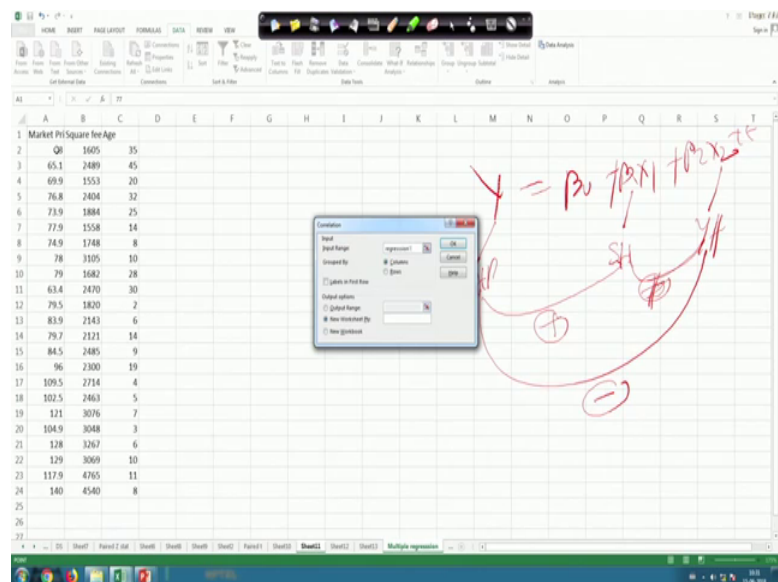
So, as a result so, first summation is that you know whether it is a bivariate structure trivariate structure or multiple structure. So, you are supposed to report the descriptive statistics of all variables, because it will give you the clue about the starting clue enter to the regression modelling.

(Refer Slide Time: 16:12)



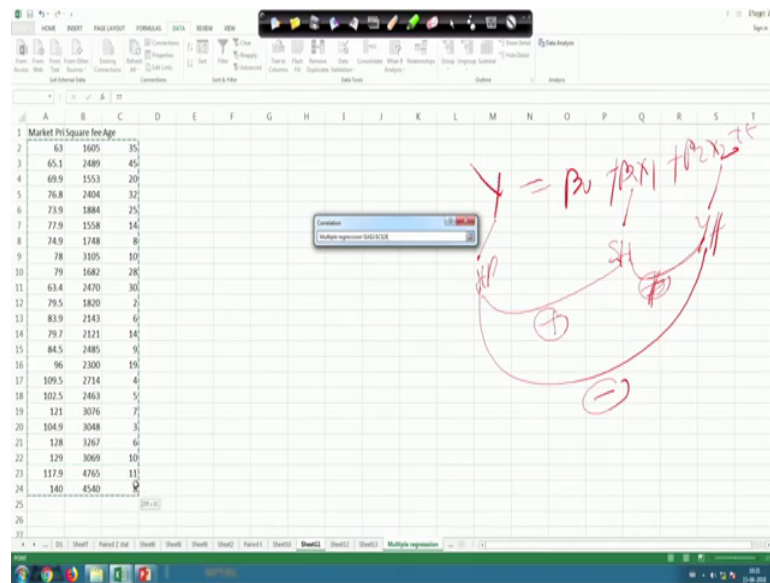
So, after knowing the descriptive statistic again you go to the data analysis package. Now, you give the integration about the correlation.

(Refer Slide Time: 16:15)



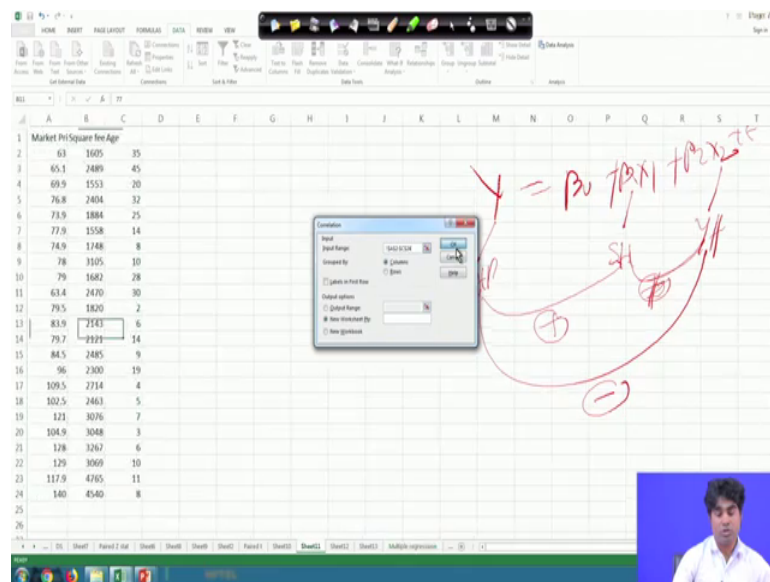
So, we need to know correlation and that too for this for these variables.

(Refer Slide Time: 16:25)



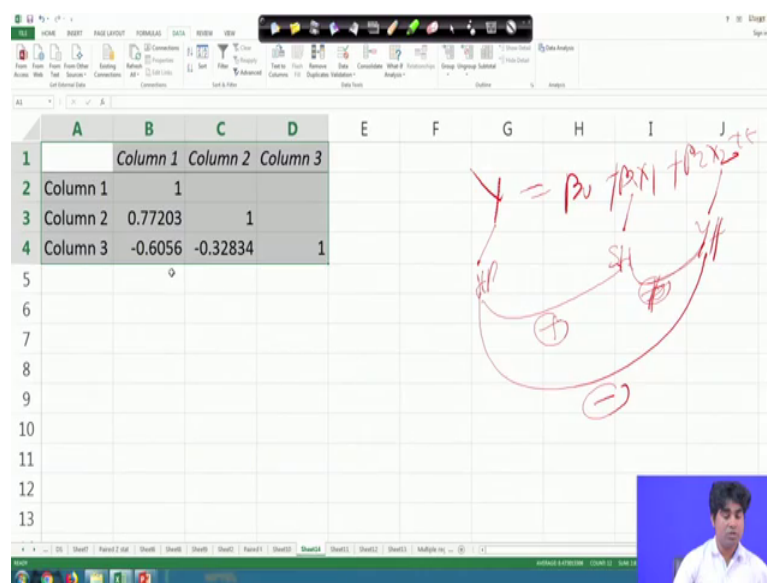
And, we like to check that you know the relationship between Y and X 1 should be high, Y X 2 should be high and no relationship between X 1 and X 2.

(Refer Slide Time: 16:31)



And for that you just give the link here.

(Refer Slide Time: 16:31)



And, this is what the correlation matrix and column 1 means it is Y and then column 2 is X 1 and column 2 X 3. For this problem this is housing price, this is the size of the house and this is years of the house that too how old that particular house.

We will find a housing price and then and the housing size, what theoretically you expect they there should be a positive association. Now, we in the correlation matrix we find the correlation coefficient is 0.77. So, that means it clearly indicates that you know if the size of the house is the high then housing price can be high if the size of the house is the low housing price and that means, there is a positive association between housing price and the size of the house. Since, the correlation coefficient is coming 0.77 and there is a high chance that that particular correlation or that particular link will be statistically significant.

Of course, we are we have not given the option about these testing of this particular correlation which is not required in the present contest because our aim is to predict the Y with respect to X 1 and X 2. And, then in the second items where you know we are the correlation between housing price and the years of the house and we expect theoretically there should be negative relationship because when the house becomes you know old then you can you cannot actually put high price.

So, obviously, there is a negative relationship between the two and the data will also data is also supporting here. So, that means, technically we find the kind of relationship

between years of the house and the housing price and negatively related, that means, technically the interpretation is that. So, the once the house is old, then the housing price will be low and if it is actually a low you know new one for or years wise it know very lower then you can put you know high price.

So, that means, technically so, here in the second case so, years of the house and a housing price are negatively related. So, higher the old lower the price housing price and lower the old lower the housing higher the housing price. So, that is how the negative relationship between housing price and age of the house. So, that means, technically more old means low housing price low lower order means number of that is the age of the house will have actually high housing price.

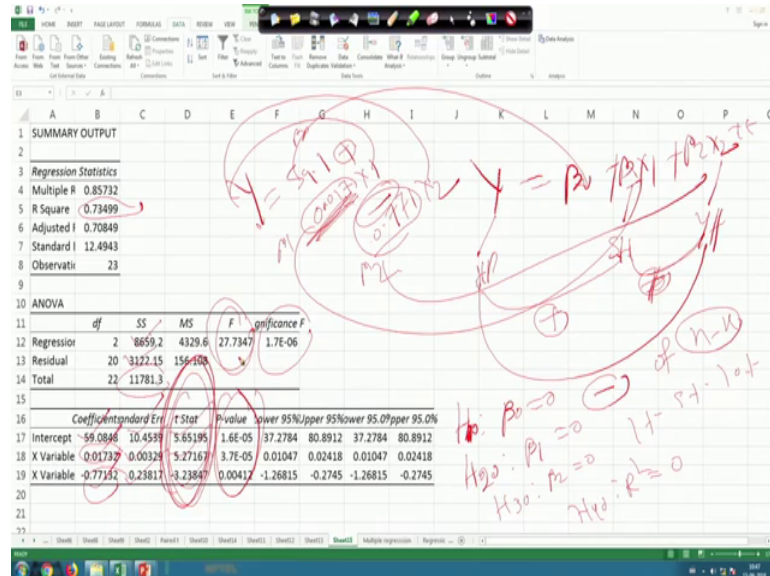
And, then finally the third link: so that is what actually between housing price sorry size of the house and a age of the house which is actually coming minus 0.33. So, that means, technically they are independent and they their correlation should be 0, but unfortunately in this case the data is behaving like that the there is a correlation and that correlation is coming 0 minus 0.33. That means, there is a negative relationship between these two that is with respect to size of the house and a kind of age of the house size of the house and age of the house.

So, they are negatively related to each other and there is a link and that link maybe statistical significant and I think minus 0.33 means there is a high chance this will also statistical significant, but we can actually minimize. But again since it is actually minus 0.32 and lesser to 0.5 means it is it indicates the low correlation coefficient and you can go ahead with the prediction. But this may affect the particular process and it may not be significantly affect the process, because the link between X_1 X_2 that is the size of the house and years of the house is very low correlation coefficient.

So, now it third case is the regression modelling and for that again you go to the data analysis package now instead of correlation you come to the regressions, then in the regressions so, our requirement is actually dependent variable with independent variables. So, what will we do? So, first is the input Y range. So, Y is here housing price that starts with you know 2 unit to here 24 unit because first one is the variables name and then again come to the input range. So, it should have actually same range otherwise

software not help you to process your output. So, now we have received the process you just put, ok.

(Refer Slide Time: 21:34)



So, will give you like previous case in the case of airline costing. So, here we get actually substantial you know regression output and as usual and we have discussed up to with this much till now and then this part. So that means, technically our model is Y equal to $\beta_0 + \beta_1 X_1 + \beta_2 X_2$. So now, here if you if I will be write like this then it becomes Y equal to $59.1 + 0.017 X_1 + \text{plus minus } 0.771 X_2$. So, this becomes then this becomes this becomes ok. So, what you can do you can directly put to minus here and then it becomes minus $0.771 X_2$, ok. So, the X_2 ; that means, technically X_1 on Y is the positive link and X_2 on Y is on negative link, right.

So, which we have also obtained through the correlation matrix where $Y X_1$ is a positive correlation and $Y X_2$ is negative correlation. So, that means, the evidence from the correlation and regressions are more or less same and what is what is more important correlation will not give you the important incidents case cause and effect, but here we can obtain the cause and effect. So, in that case actually here X_1 is positively influence, but the influence of external Y is the little bit lower compare to X_2 where the a impact is negative, but it is the high one.

So, the that means, technically the interpretation is that when house the age of the house is the actually declined drastically then house housing prices will be you know effect

very drastically and in the context of X_1 that is the size of the house. In fact, size of the house technically cannot actually increase for a particular house, but it may this model can be used for a new construction or something like that. But the thing is that if the size of the house is lower one then housing price will be lower size of the house is higher one then housing price will be higher one, ok.

So, this is how the coefficients this is for β_0 and this is for β_1 and this is for β_2 . So, like this. So, this is β_0 this is β_1 and this is for β_2 and as a results. So, we have β_0 coefficient β_1 coefficient β_2 coefficient n . So, now, this is what the estimated model now put X_1 value and X_2 value then predictive Y . And before that we should go for testing the reliability part of the model like the part a which we have already discussed in the case of airline costing that too airline cost with airline passengers.

So, here also same things at the reliability part of the models we will pass through all these you know tests that is specification test, then the goodness fit test, then the diagnostic test and the out of sample prediction test since a since it is a cross sectional modellings out of sample prediction test is not required technically and then diagnostic test which we will discussed in the later part, but only important with this particular result is to check these specification test that too check whether the parameters are statistically significant or not.

So, that means, technically we have three hypothesis here. First knowledge β_0 equal to 0 then second $H_1: \beta_0 \neq 0$ you can put then H_2 where β_1 equal to 0 then $H_3: \beta_1 \neq 0$ where β_2 equal to 0 then accordingly you can applied t statistics and check whether β_0 is actually 0 or not 0 whether is statistical significant or not significant similarly for β_1 and similarly for β_2 and for that you can apply t statistics t is the provided level of significance at 1 percent, 5 percent or 10 percent.

Since we are using software, software is actually directly given the P-values here then you compare with the critical value and the calculated value. The calculated value will be a beta coefficient that is β_0 , β_1 , β_2 with respect to their standard error and standard error technically depends upon sum of the squares of the error and divide by degree of freedom which we have discussed in the part one case. And same case we will

be obtained here in the case of trivariate that is the dependent variable with two independent variables.

So, ultimately for all these parameter for testing purpose; we have a beta coefficients that too beta 0, beta 1, beta 2. So, what we are supposed to report before it start the testing. So, to find out the standard error of these you know parameters beta 0, beta 1, beta 2 and that is these square root of all the variance of these parameters. So, the standard errors for beta 0 is here 10.45 and standard error of beta 1 is 0.003 that is for X 1 variables and standard error for beta 2 is 0.24 that is for X 2 variables.

So, now, beta coefficients beta 0 divided by standard error beta 0 and beta 1 by standard error beta 1 and beta 2 by standard error beta 2 will give you t statistic which is having here and if you go through probability check for decision making process we are we have to compare the critical value these are all. So, these are all calculated value and when will be compare with the critical value that too t tables with a particular probability level. Let us say 1 percent, 5 percent and 10 percent and the degree of freedom here $n - K$ that too 23 minus 3 then you defined critical value for 1 percent, 5 percent and 10 percent then you compare with all these calculated value.

The rule is very same and that is very simple. So, if the critical value calculated value will overtake the critical value then the particular parameter then a particular null hypothesis will be rejected and justify that you know the particular parameter will be statistically significant. This is also for beta 1 and beta 2 and also for the interest rate beta 0.

And, in this particular problems all these parameters beta 0, beta 1, beta 2 are statistically significant and this is what the specification test about and that is clear and gives the green signal to go for the predictions of the housing price with respect to size of the house and the kind of age of the house. And, the second part of the testing is goodness of fit test that depends upon R square value which is nothing, but the ratio between explained sum of square by total sum of squares and where we have here and that is actually regression sum of square or expansion of square and this is actually unexpansion of squares that is error sum of square and this is total sum of square this by this with respect we will get in R square.

And, then the next null hypothesis that you know $X_4 = 0$ that R square naught equal to 0 and for that you use your statistics and F statistic is coming here 27 point 27.74 and that is also statistically significant. So, R square R is reported here and this is actually R square 0.74. So, 0.74 mean 70; 70 0.734 that is the case where are you know the influence is the 73 percent of independent variable to independent variable.

So, that means, a since R square is coming 0.73 73 percent 0.73 that is the 73 percent so, the impact of independent variable to dependent variable will influence you know out of hundred is 73. So, that means, if change the independent variable then the combined impact to the dependent variable will be 73. So, one unit of change of X_1 and X_2 will affect the dependent variables in you know with respect you know 73 percent. So, that is the kind of an interpretation you to bring here.

Then, you can also check the individual impact of X_1 and X_2 because R square will give you the combined impact and individual impact we can you can check through intercept estimator variable you know estimations that is actually through beta 1 and beta 2 and since R square is coming 0.73 which is close to high correlation. That means, the prediction is good. And, the impact of independent variable to dependent variable is also very high and at the same times R R square is coming statistically significant with the help of F statistic.

That means, technically this estimated models where our objective is to predict housing price with the a size of the house and age of the house is perfect to one with respect to specification test and goodness of fit test. Of course, the diagnostic part is there and it is left to again give the green signal before you start the prediction and forecasting and this diagnostic test includes the a relationship between also X_1 and X_2 . So, in the coming lectures we will be discuss the diagnostic test issue where it is not a kind of single one there are multiple kind of test are there in the diagnostic test baskets.

So, you will you will go through one by one and the and check the particular requirement and then finally, you give the green signal before you start a prediction and forecasting. And that means, the one of this which we have already highlighted is the relationship between independent variables and that can be also part of the diagnostic test, because it is not located in the specification test; and the kind of goodness of fit test which we can discuss in the next class, that too where you will be connect with the all these diagnostic

test, and including this particular that is the relationship between X_1 and X_2 . That is technically called as actually multi colinearity problem.

And with this, we will stop here and we will continue the same lecture in the next class.

Thank you very much. Have a nice day.