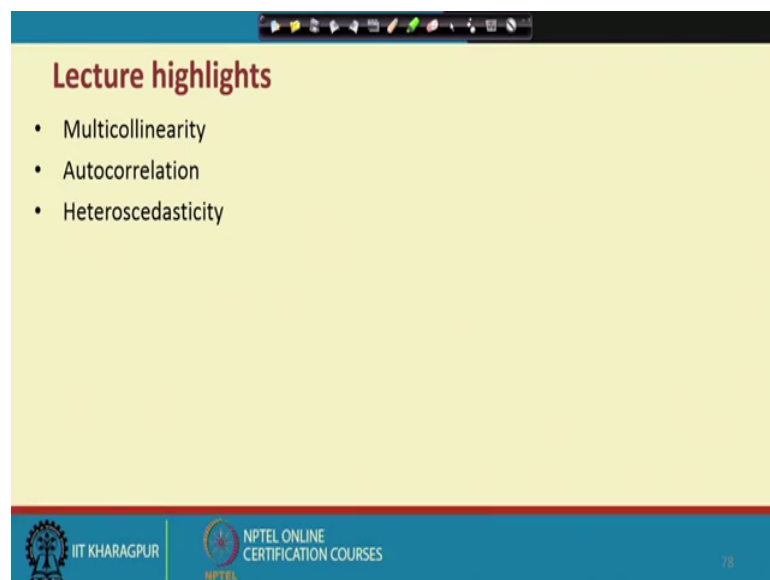


Business Analytics for Management Decision
Prof. Rudra P Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 30
Predictive Analytics (Diagnostics of Regression Modelling)

Hello everybody this is Rudra Pradhan here welcome you all to BMD lecture series. Today we will continue with the predictive analytics and that to the discussion is on diagnostic of regression modelling. So, or the typical coverage is on three topics; that is with respect to multicollinearity problem, autocorrelation problem and heteroscedasticity problems.

(Refer Slide Time: 00:39)



In fact, we have solved couple of problems earlier that to highlighting the case of Bivariate analysis and the case of multivariate analysis, and the objective behind this particular structure is to predict the dependent variable with respect to independent variables.

Now, while doing all these things, there are certain issues we like to address before you do the kind of prediction, and these issues are typically highlighted here. And in fact, suppose as diagnostic is concerned we have many things we are supposed to check, and here we specifically focus on three things; that is the multicollinearity issue, autocorrelation issue and heteroscedasticity issue, while multicollinearity is a

multivariate problem. In the case of autocorrelation and heteroscedasticity, it can be a bivariate problem and it can be also multivariate problem.

So, in the case of multicollinearity, it is the game with respect to independent variables and in the case of autocorrelation and heteroscedasticity, it is the game with respect to error terms; however, all these three components are treated as virus in the regression setup and that to in the predictive analytics. So, what is required actually? So, we have to first address all these problems, you have to check thoroughly and make ensure that the particular model is free from multicollinearity autocorrelation and heteroscedasticity before you go for the prediction.

So, what is exactly multicollinearity, autocorrelation and heteroscedasticity, how to detect and how to solve? So, these are the kind of discussions we can cover today. So, let us start with a simple problem here and here the problem is with respect to dependent variable and 5 independent variables.

(Refer Slide Time: 02:43)

Example: Predict Crude Oil Production

Y	X ₁	X ₂	X ₃	X ₄	X ₅
55.7	74.3	83.5	598.6	21.7	13.30
55.7	72.5	114.0	610.0	20.7	13.42
52.8	70.5	172.5	654.6	19.2	13.52
57.3	74.4	191.1	684.9	19.1	13.53
59.7	76.3	250.9	697.2	19.2	13.80
60.2	78.1	276.4	670.2	19.1	14.04
62.7	78.9	255.2	781.1	19.7	14.41
59.6	76.0	251.1	829.7	19.4	15.46
56.1	74.0	272.7	823.8	19.2	15.94
53.5	70.8	282.8	838.1	17.8	16.65
53.3	70.5	293.7	782.1	16.1	17.14
54.5	74.1	327.6	895.9	17.5	17.83
54.0	74.0	383.7	883.6	16.5	18.20
56.2	74.3	414.0	890.3	16.1	18.27
56.7	76.9	455.3	918.8	16.6	19.20
58.7	80.2	527.0	950.3	17.1	19.87
59.9	81.3	529.4	980.7	17.3	20.31
60.6	81.3	576.9	1029.1	17.8	21.02
60.2	81.1	612.6	996.0	17.7	21.69
60.2	82.1	618.8	997.5	17.8	21.68
60.6	83.9	610.3	945.4	18.2	21.04
60.9	85.6	640.4	1033.5	18.9	21.48

Y	World Crude Oil Production
X ₁	U.S. Energy Consumption
X ₂	U.S. Nuclear Generation
X ₃	U.S. Coal Production
X ₄	U.S. Dry Gas Production
X ₅	U.S. Fuel Rate for Autos

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, that means, technically the objective behind this problem is to predict Y with respect to 5 independent variables. Now when we will go for model building or the kind of model design, so, we have actually plenty of options.

(Refer Slide Time: 30:05)

Possible Regression Models with Five Independent Variables

Single Predictor	Two Predictors	Three Predictors	Four Predictors	Five Predictors
X_1	X_1, X_2	X_1, X_2, X_3	X_1, X_2, X_3, X_4	X_1, X_2, X_3, X_4, X_5
X_2	X_1, X_3	X_1, X_2, X_4	X_1, X_2, X_3, X_5	
X_3	X_1, X_4	X_1, X_2, X_5	X_1, X_2, X_4, X_5	
X_4	X_1, X_5	X_1, X_3, X_4	X_1, X_3, X_4, X_5	
X_5	X_2, X_3	X_1, X_3, X_5	X_2, X_3, X_4, X_5	
	X_2, X_4	X_1, X_4, X_5		
	X_2, X_5	X_2, X_3, X_4		
	X_3, X_4	X_2, X_3, X_5		
	X_3, X_5	X_2, X_4, X_5		
	X_4, X_5	X_3, X_4, X_5		

(Refer Slide Time: 03:12)

Multicollinearity

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_k X_{kt} + U_t$$

No Multicollinearity: $\text{Cor}(X_i, X_j) = 0$
 Note: $i \neq j$

Multicollinearity: $\text{Cor}(X_i, X_j) \neq 0$
 Note: $i \neq j$

$r < 1$ and $r = 1$

So, these options are; first case with respect to single predictor and with respect to two predictor and with respect to three predictor. So, now technically so, this is with respect to single predictor and this is with respect to two predictors. This is three predictors, 4 predictors and 5 predictors. When there is a question of multicollinearity, then in the case of single predictor this will not be the game, but it is the case for the two predictor case, three predator case and 4 predictors case and also 5 predictors case.

So, all together we have now plenty of models. So, the idea is which model will be considered as the best so as far as the prediction is concerned. So, now, typically our objective is here to predict the Y With respect to 5 independent variables. But technically after the diagnostic process, there may be a chance that Y can be predicted with only 1 variable or 2 variables or 3 variables or it can be 4 or it can be with 5 we have no idea, but over the time we can actually explore these issue.

Then finally, we have to fix which one is the best for our prediction. So; that means, technically we have here plenty of options, and all these options cannot be simultaneously considered. So, only one option can be simultaneously considered, and how to find out the particular model is useful for these predictions we can go ahead with the kind of our discussions. So, what we are supposed to do here, that we have to find out which particular model is the best for this prediction. So, we have to go for the kind of search process, and one of these search process or mechanism we like to follow is the multicollinearity issue.

So, now the question is, what is exactly multicollinearity? as I have already pointed out multicollinearity is a game of multivariate analysis, and in the simple language, so multicollinearity represents the existence of linear relationship among the regressor. technically simply can calculate through correlation structure. So; that means, technically correlation among the regressor should not be equal to 0. if it is exactly equal to 0, then there is no multicollinearity. So; that means, if there is actually, they are not equal to 0. So, then there is your multicollinearity.

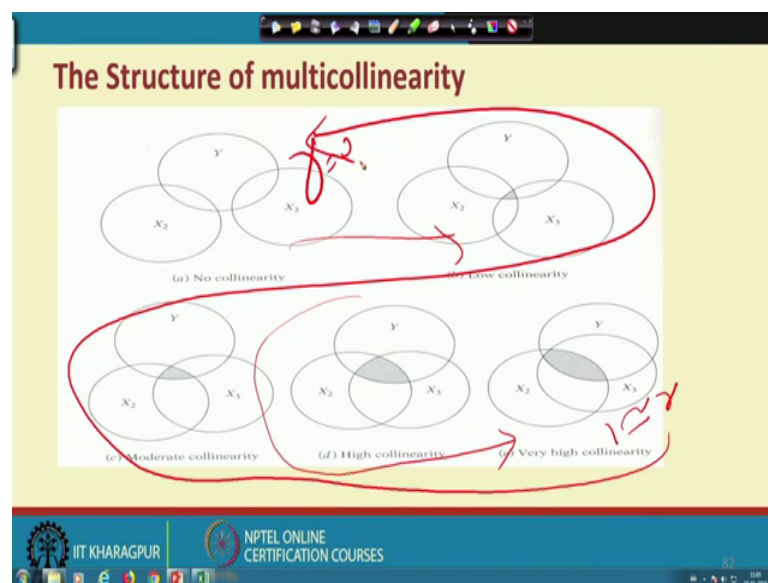
So, technically, so, we have a two options in the first case correlation among the regressors equal to 0. So, where the declaration is no multicollinearity, and correlation among the regressors not equal to 0, then there is a multicollinearity. Now in this case, so; that means, if the first case occurs. So; that means, the model is free from multicollinearity then model is good for the prediction, but in the case of the second case, where correlation coefficient between X_i and X_j not equal to 0. So, it may have two different options. In one options it may be equal to, correlation coefficient equal to 1, and in another case correlation coefficient may be less than 1.

So; that means. So, there is a third option which is equal to 0, here we have already declared that there is no multicollinearity. So, now, when there is r equal to 1; that means,

correlation among the regressor equal to 1, then it is called as, its a question of perfect multicollinearity. And in this case model cannot be used for any kind of prediction, but in the case of less than 1. So, it depends upon what is the degree of correlation; that is called as imperfect multicollinearity, but if the correlation coefficient close to 1 means. So, the degree of problem will be very high, when it will be close to 0 the degree of problem will be very low.

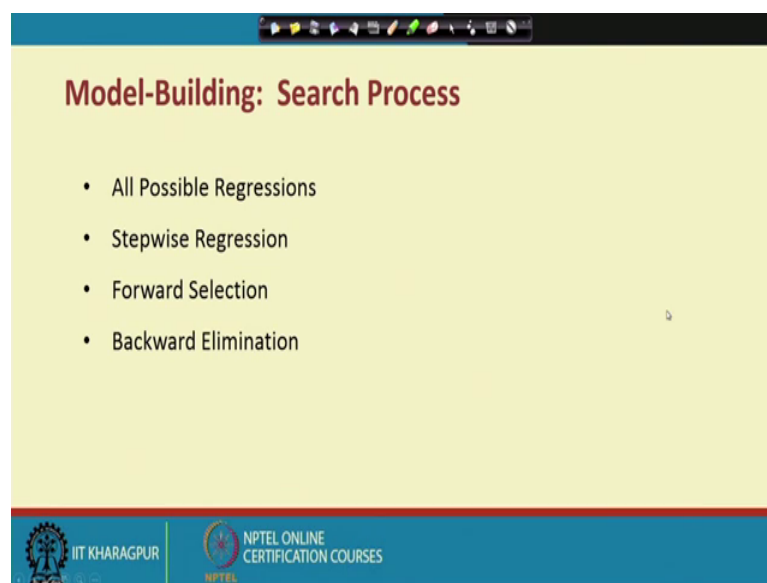
But whatever may be the case. So, we like to find out what is the exact nature or the degree of multicollinearity, then we will try to solve these problems as per the a model requirement. So, let us see, how is this particular structure. So, in the first instance the look will be like this. So, this is the case where r equal to 0, and here r will be in not exactly equal to 1, but it will be approximately close to ones.

(Refer Slide Time: 07:47)



So, it is the movement of no multicollinearity to low multicollinearity, then slightly increasing towards high multicollinearity. So; that means, it gives the signal that the problem is going to dangerous. So, we try to minimize the particular process; that means, technically our search process will be from this angle to this angle. So, this is how we are looking for the kind of search process. So, how you have to do that so we can actually get to know in this particular process.

(Refer Slide Time: 08:43).



The slide is titled "Model-Building: Search Process" in a bold, dark red font. Below the title, there is a bulleted list of four regression methods: "All Possible Regressions", "Stepwise Regression", "Forward Selection", and "Backward Elimination". The slide has a yellow background and a blue header bar. At the bottom, there is a blue footer bar containing the IIT Kharagpur logo and the text "NPTEL ONLINE CERTIFICATION COURSES".

- All Possible Regressions
- Stepwise Regression
- Forward Selection
- Backward Elimination

So, let us see the kind of game here, and there are many different ways actually you can check the multicollinearity. So, the first approach is simply called as correlation approach. So, you can prepare the correlation matrix, and try to see whether the correlation coefficients are very high, and if they are high and statistically significant, so how you have to sort out the particular problem. Besides there are many other issues or the tricks through which you can actually detect the multicollinearity. Like the conditioning index the BIF factors and through partial correlation coefficient and multiple correlation coefficient and through auxiliary regressions.

So, many ways we can actually check the multicollinearity, but by the way, if there is a multicollinearity, so you follow any methods, it will be showing the kind of multicollinearity problem. So, now, we try to find out what should be the particular structure or the degree, and then we like to sort out the solution and so far as the solution is concerned. So, these are the following mechanisms usually you follow. In fact, there are standard solutions are there, alternatives structure are there to increase the sample size, to drop the collinear variables, use different mechanisms. But here it is the standard structures without the changing sample size, without dropping any variables or something like that. So, we have to see the formal structure through which you have to find out a particular model which is free from multicollinearity.

(Refer Slide Time: 10:15)

Examples

We have annual data for

- ✓ Expenditure on clothing
- ✓ Disposable income
- ✓ Liquid assets
- ✓ Price index for clothing
- ✓ Producer Price Index

Handwritten equations:

$$Y = f(C, D, LA, P, PPI)$$

$$Y = f(D, LA, P, PPI)$$

Logos for IIT KHARAGPUR and NPTEL ONLINE CERTIFICATION COURSES are visible at the bottom.

So, in order to justify this, so we start with a simple problem and then we check how is; the multicollinearity and how to sort out this particular problem. So, in this context, so the problem is a with respect to you here 5 variables, expenditure on clothing; that is the dependent variables and followed by independent variables are disposable income, liquid assets, price index for clothing and producer price index.

(Refer Slide Time: 10:48)

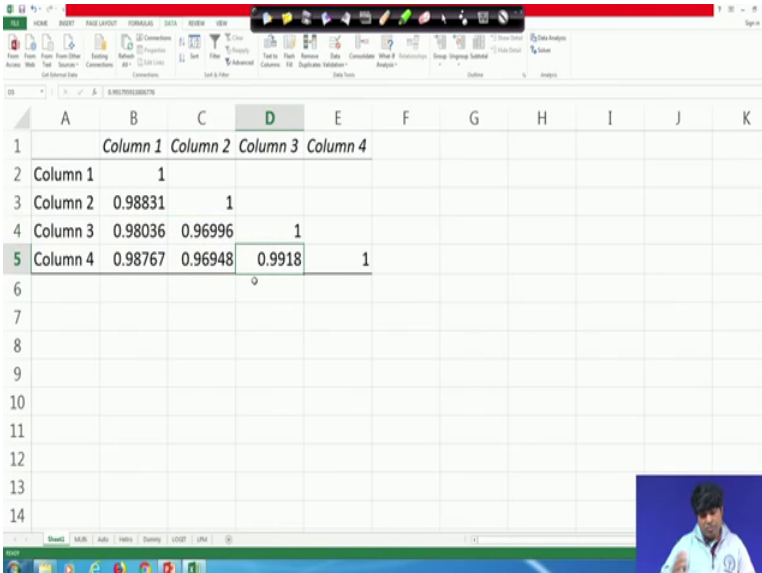
	A	B	C	D	E	F	G	H	I
1	Expenditure on clothing	Disposable income	Liquidity assets	Price index for clothing	General price index				
2	8.4	82.9	17.1	92	94				
3	9.6	88	21.3	93	96				
4	10.4	99.9	25.1	96	97				
5	11.4	105.3	29	94	97				
6	12.2	117.7	34	100	100				
7	14.2	131	40	101	101				
8	15.8	148.2	44	105	104				
9	17.9	161.8	49	112	109				
10	19.3	174.2	51	112	111				
11	20.8	184.7	53	112	111				
12									
13									
14									
15									
16									
17									
18									
19									

So, now, what we supposed to do? So, let us go to these spreadsheet and this is what the data all about, and what will you do here. So, we have actually following data set, and

we like to check whether there is a kind of multicollinearity or not. So, the first hand check is to find out the correlation matrix among the regressors. So; that means, the regressors are here.

These are the regressors and through which you have to check the kind of position. So, what will you do. So, go to the data analysis package, and then you find out correlation made structures, give the indication about the correlation, and then we will highlight the variable indications, all these independent variables; that is the 4 independent variables, and it will give you the kind of output matrix.

(Refer Slide Time: 11:33)



	Column 1	Column 2	Column 3	Column 4
Column 1	1			
Column 2	0.98831	1		
Column 3	0.98036	0.96996	1	
Column 4	0.98767	0.96948	0.9918	1

So, this is actually correlation matrix. Here is the correlation matrix gives X_1 to X_2 , X_1 to X_3 , X_1 to X_4 and X_2 to X_3 , X_2 to X_4 . Similarly X_3 and X_4 , but in all the cases we find there is a high correlation. so; that means, it is clear cut signal that there is a problem of multicollinearity. So, in the first instance before you start the kind of prediction. So, this gives some kind of negative signal. And again so what will it do that will be checked through regression analysis.

So, again you go to this kind of the data analysis package, because our basic objective of this problem is to predict the expenditure on clothing, subject to disposable income, liquidity assets, price index of clothing and general price index. So, then what will you do? We have to choose a regression package, and then we have to start working on this. So, the dependent variable is by default expenditure on clothing, and the

independent variables will start from disposable income to general price index. So, then we will find this is actually the complete regression output.

(Refer Slide Time: 12:46)

SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.999								
R Square	0.99801								
Adjusted R	0.99642								
Standard Error	0.25751								
Observations	10								
ANOVA									
	df	SS	MS	F	Significance F				
Regression	4	166.168	41.5421	626.463	0.000000				
Residual	5	0.33156	0.06631						
Total	9	166.5							
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95%	Upper 95%	
Intercept	-1.5335	7.51316	-1.80131	0.13154	-32.8467	5.77966	-32.8467	5.77966	
X Variable 1	0.09095	0.02649	3.66028	0.01459	0.02886	0.16504	0.02886	0.16504	
X Variable 2	0.01508	0.04939	0.30533	0.77242	-0.11189	0.14205	-0.11189	0.14205	
X Variable 3	-0.19913	0.09016	-2.20868	0.07823	-0.4309	0.03263	-0.4309	0.03263	
X Variable 4	0.34014	0.14975	2.27141	0.07231	-0.0448	0.72508	-0.0448	0.72508	

Here what is happening actually. So, these are all 4 variables and these are all T statistics. So, now, you will find some variables are statistically significant and some variables are not statistically significant. For instance, so this is actually X 2. So, it is statistically significant at 1 percent followed by X 3 and X 4, and X 3 is having low impact compared to X 4. So; that means, technically one variable is not statistically significant and in the contrary R square is very high, but what is required here. So, if there is a high R square close to ones, and most of the variables should be statistically significant, but here it is not the case. So; that means, it is the clear cut signal of multicollinearity so that means, the full model may not be actually work here.

So, out of 4 variables, at least 4 variables, it means few less than 4 variables can be the right solution to go for this prediction. So, which three or which two variables are right choice that we have to go for this particular process. So, there are three mechanisms. So, if you go to any software, there is a stepwise option. So, if you give the option to stepwise.

So, by default it will give you the exact variables combinations through which you can do the prediction and that to it is pre multicollinearity. Otherwise you can go by a forward search process and backward search process. in the forward search process the

most important variable should enter first. So; that means, technically, so we can start with like this, we can start with a first Y with the second variables; that is here or the first independent variables, and then followed by last variables and then a second last and then finally, this one.

So; that means, technically if you will go to this particular matrix here. So, what will he do the process. So, here, so the first model will be Y with respect to the first disposable income. So, this is a disposable income. So, then you have to check actually, let us say this is beta 1 coefficients and followed by R squares and F statistics and again in the second models. So, what will it do Y equal to function of disposable income with the liquid assets, then here the coefficient will be beta 1 and beta 2 then R squares and F like this.

So, now one after another variables you have to enter to the systems, and every time you have to check actually whether the variables are significant and the model fitness are as per the requirement. So, now, this is the first model, now you have to go for the second model, if the second model improves the first one then you can reject the first one and opt the second one. So, now, second one is the good fit for the prediction, again see you allow another variable and you check whether the variables are statistically significant R square is improving, F is improving, then you continuously go ahead with this kind of structure. Then you have to stop at a particular point of time where the new entry of any particular variables will not improve the significance level and will not improve the R square and adjusted R square.

So, this is the search process through which you have to check the multicollinearity and to find out the solution for the prediction requirement. Now the second structure of this particular problem is with respect to, second structure is which with respect to autocorrelation problem and for that ok. So, the second problem is with respect to autocorrelation and this is actually with respect to error term.

(Refer Slide Time: 17:00)

Autocorrelation Problem

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + U_t$$

No autocorrelation: $\text{Cov}(U_i, U_j) = 0$
or $E(U_i, U_j) = 0$

Autocorrelation: $\text{Cov}(U_i, U_j) \neq 0$
or $E(U_i, U_j) \neq 0$

Note: $i \neq j$

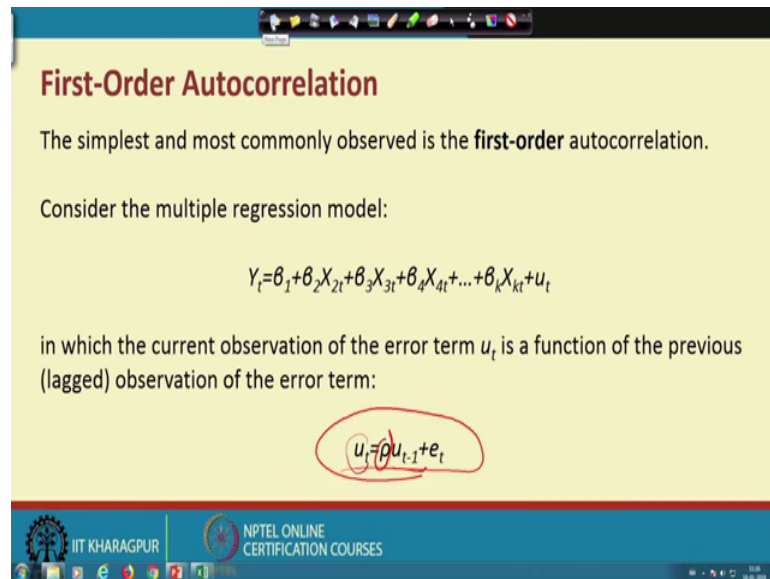
In general $E(U_t, U_{t-s}) \neq 0$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, like the previous one. So, here we like to try covariance between two error terms. If the covariance between two error term is coming 0, so then the declaration is, there is no auto correlations however, if the covariance between two error term is not coming equal to 0. So, then it is actually auto correlation; that means, technically like the multicollinearity case. So, here we are correlating X_1 upon X_2 or X_1 upon X_3 . So, here U_1 upon U_2 or E_1 upon U_2 like this.

So; that means, in this case we have no correlation, means autocorrelation, but in this case there is a autocorrelations. So, like multicollinearity it is also kind of virus that need to be actually detected and then finally, we look for the solution. So, what is the procedure here? first estimate the model, get the error term and then you connect with the one particular error term with another error term, and check whether there is a correlation among this error clusters. If there is a correlation, what is the degree of this correlation? If it is high then it is again problem, and that particular model cannot be used for prediction.

(Refer Slide Time: 18:39)



First-Order Autocorrelation

The simplest and most commonly observed is the **first-order** autocorrelation.

Consider the multiple regression model:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \dots + \beta_k X_{kt} + u_t$$

in which the current observation of the error term u_t is a function of the previous (lagged) observation of the error term:

$$u_t = \rho u_{t-1} + e_t$$

The slide is a screenshot of a presentation. At the top, there is a title bar with various icons. The main content area has a yellow background. The title 'First-Order Autocorrelation' is in bold red text. Below it, the text 'The simplest and most commonly observed is the first-order autocorrelation.' is in black. Then, 'Consider the multiple regression model:' is followed by a multiple regression equation. Below the equation, it says 'in which the current observation of the error term u_t is a function of the previous (lagged) observation of the error term:'. Finally, the equation $u_t = \rho u_{t-1} + e_t$ is shown, with the entire equation circled in red. At the bottom, there is a blue footer bar with the IIT Kharagpur logo and the text 'NPTEL ONLINE CERTIFICATION COURSES'.

So, like multicollinearity we have to find out the particular the degree of the autocorrelation and the kind of solutions. Let us see how we have to work out on this. And in this context, so the procedure actually, you can understand this is a simple multivariate model and in order to understand autocorrelation. So, you have to create a autocorrelation first order autocorrelation structures.

So, U_t upon ρU_{t-1} , and technically ρ is called as autocorrelation coefficient. So; that means, the technical term is here, error term depends upon the past error term, does the lag of 1 year, and we are assuming that today's error term depend upon yesterdays error terms. It may not there, but we have to check and find out whether there is a possibility or not. If it is possibility what is the kind of degree.

(Refer Slide Time: 19:27)


First-Order Autocorrelation


The coefficient ρ is called the first-order autocorrelation coefficient and takes values from -1 to +1.

It is obvious that the size of ρ will determine the strength of serial correlation.

We can have three different cases.

- (a) If ρ is zero, then we have **no autocorrelation**.
- (b) If ρ approaches unity, we have **positive autocorrelation**.
- (c) If ρ approaches -1, we have high degree of **negative autocorrelation**.





NPTEL ONLINE
CERTIFICATION COURSES

So, likewise. So, the kind of structure will be ρ is the autocorrelation coefficient, like correlation coefficient. So, the value will be either 0 or plus minus 1. So, if it is 0 then there is no autocorrelation. So, this is the case, and if there is a auto correlations then ρ equal to either positive value or ρ equal to maybe negative value. It may plus 1 it may be minus 1. If it is a plus 1 then positive autocorrelation, if it is a minus 1 it is the negative autocorrelation. So, now, what will you do here? So, we like to check; what is the actual structure through, which you have to check the particular item.

(Refer Slide Time: 19:59)

Higher-Order Autocorrelation

Second-order when:


$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + e_t$$


Third-order when

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + e_t$$


p-th order when:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \dots + \rho_p u_{t-p} + e_t$$





NPTEL ONLINE
CERTIFICATION COURSES



So, these are all second order, third order and a p th order of autocorrelation issues. So; that means, the error term of a particular year depends upon error term of previous years.

Previous years means 1 lag 2 lag 3 lag and so on. So, it is a kind of chain, its technically called as a auto regressive scheme. And we like to check, we like to see whether the particular error term is connected with the previous error terms. If there is a strong correlation, then that is actually very danger for the kind of prediction. So, we like to find out what is the degree of such relationship.

(Refer Slide Time: 20:41)

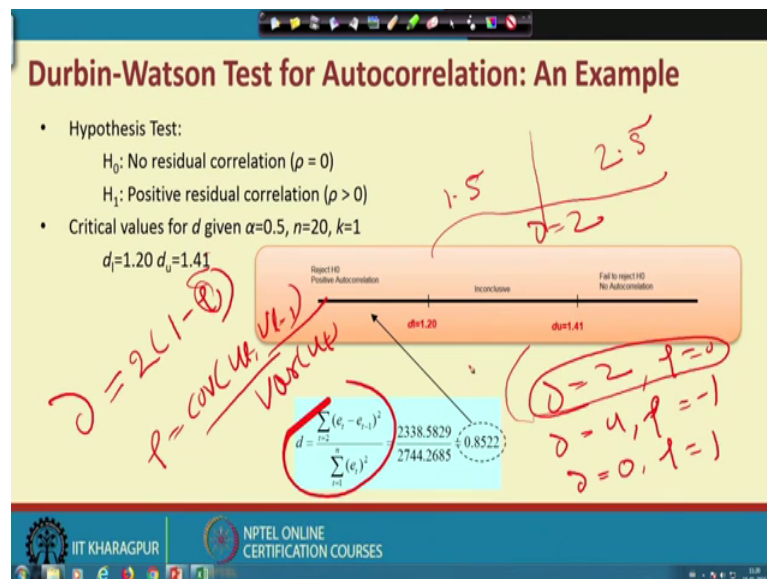
Detecting Autocorrelation

- Graphical method.
- Through **formal tests**, like the following ones:
 1. The Durbin Watson Test
 2. The Breusch-Godfrey Test
 3. The Durbin's h Test (for the presence of lagged dependent variables)
 4. The Engle's ARCH Test

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

So, technically there are many mechanisms through which you can check the process. So, either graphically you can plot the error term and get to know the kind of nature or else there are formal tests are there, through which you can check. So, one of the fantastic test is called as Durbin Watson and D statistics and B G test Durbin's h statistics and Engle's ARCH test, but most of the softwares gives the value of Durbin Watson D statistics, and through the value of Durbin Watson D statistic you can get to know what is the position of this autocorrelation; that means, technically what is the exact degree through which we go for the kind of prediction.

(Refer Slide Time: 21:34)



So, what will it do here? So, I will just highlight the Durbin Watson statistics and then let how is the particular structure Durbin Watson D statistics, the structure will be generally like this. So, it will depends upon the covariance between two error terms. So; that means, technically this one.

(Refer Slide Time: 21:49)

First-Order Autocorrelation

The simplest and most commonly observed is the **first-order** autocorrelation.

Consider the multiple regression model:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \dots + \beta_k X_{kt} + u_t$$

in which the current observation of the error term u_t is a function of the previous (lagged) observation of the error term:

$$u_t = \rho u_{t-1} + e_t$$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

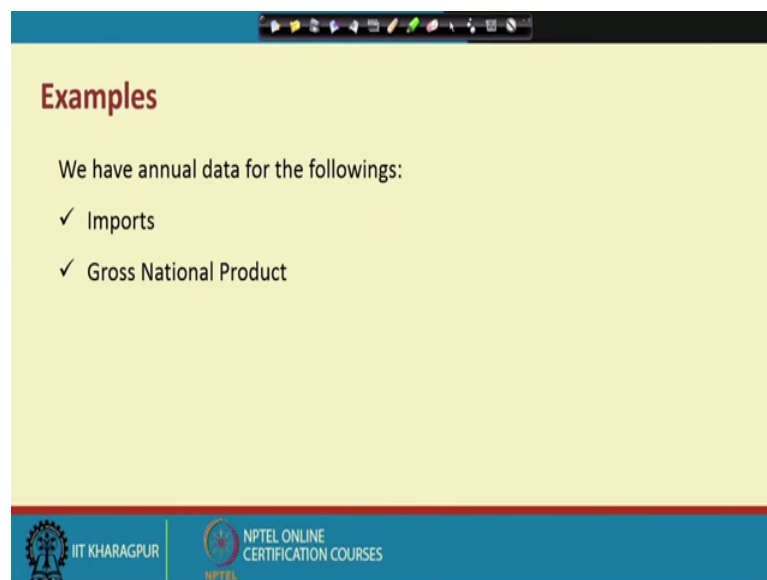
So, this is the rho coefficient that is autocorrelation. So; that means, technically Durbin Watson D statistic will be covariance between U_t U_{t-1} by variance of u_t . So, that is how we have actually reported here. So, this is how the Durbin Watson d statistic value. So, this gives the kind of procedure through which you have to calculate and then we have to check the degree of existence.

Generally in the Durbin Watson D statistic, d equal to 2 into 1 minus ρ and ρ is the autocorrelation coefficient. So, that is actually covariance of u_t of 1 U t minus 1 divided by variance of variance of u_t , and when ρ equal to 0 , then d equal to 2 ; that is where ρ equal to 0 , when d equal to 4 , then in that case ρ equal to minus 1 , and when d equal to, its equal to 0 where ρ equal to exactly. So, these are the four extremes.

And in the case of d equal to 2 where ρ equal to 0 ; that is actually free from autocorrelation. Otherwise if the range will be vary from 0 to 4 , but actually what is the optimum range through which you can do the prediction. So, usually we can see if the Durbin Watson d statistic will be ranging between 1.5 to 2.5 , then this is your tolerance through which you can actually go ahead with the prediction; otherwise perfect structure is if where d equal to 2 , but it is very rare to get this particular situation, but we try to find out what is the best possible way through which you have to manage the, a kind of predictions right.

So, what I will do. So, I will take you to this, I take you to a particular problem, and then take you how it can be actually connected with the autocorrelation.

(Refer Slide Time: 23:40)



The slide is titled "Examples" in a bold, dark red font. Below the title, it states "We have annual data for the followings:" followed by a list of two items, each preceded by a checkmark: "Imports" and "Gross National Product". The slide has a yellow background and a blue header bar. At the bottom, there is a blue footer bar containing the IIT Kharagpur logo and the text "NPTEL ONLINE CERTIFICATION COURSES".

(Refer Slide Time: 23:51)

Excel screenshot showing regression statistics and ANOVA table. The data is summarized in the following tables:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.99134					
R Square	0.98275					
Adjusted R	0.98179					
Standard Error	178.43					
Observations	20					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	3.3E+07	3.3E+07	1025.4	2.5E-17	
Residual	18	573069	31837.2			
Total	19	3.3E+07				
Coefficients						
	Coefficient	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-2461.38	250.007	-9.84522	1.1E-08	-2986.62	-1936.13
X Variable	0.27952	0.00873	32.0218	2.5E-17	0.26118	0.29786

(Refer Slide Time: 23:53)

Excel screenshot showing a dataset with columns for Imports, GNP, Imports Et, e-et-1, and Sqr e-et-1. The data is as follows:

	Imports	GNP	Imports Et	e-et-1	Sqr e-et-1
1	3748	21777	3636.18	111.82	12503.71
2	4010	22418	3815.66	194.34	6809.55
3	3711	22308	3784.86	-73.86	268.2
4	4004	23319	4067.94	-63.94	98.4064
5	4151	24180	4309.02	-158.02	94.08
6	4569	24893	4508.66	60.34	-218.36
7	4582	25310	4625.42	-43.42	103.76
8	4697	25799	4762.34	-65.34	21.92
9	4753	25886	4786.7	-33.7	-31.64
10	5062	26868	5061.66	0.34	-34.04
11	5669	28134	5416.14	252.86	-252.52
12	5628	29091	5684.1	-56.1	308.96
13	5736	29450	5784.62	-48.62	-7.48
14	5946	30705	6136.02	-190.02	141.4
15	6501	32372	6602.78	-101.78	7786.298
16	6549	33152	6821.18	-272.18	170.4
17	6705	33764	6992.54	-287.54	15.36
18	7104	34411	7173.7	-69.7	-217.84
19	7609	35429	7458.74	150.26	48382.4
20					

Let us have an examples, I have taken two variables here; imports data and GNP data and then you go to this spreadsheet and this is actually the kind of input the data structure. So, here, so inputs data and GNP is there. So, we are trying to estimate the inputs. So, what will you do technically? So, you go to the data analysis. First you go to the regressions, then highlight the input structure here. So, ranging to from this data to this data, then again you give the independent variable declarations; that is the GNP from this particular data point to this data points. So, you will get firsthand regression output that is actually from this all right.

(Refer Slide Time: 24:24)

	A	B	C	D	E	F	G	H
5	4004	23319	4067.94	-63.94	-9.92	98.4064	4088.324	
6	4151	24180	4309.02	-158.02	94.08	8851.046	24970.32	
7	4569	24893	4508.66	60.34	-218.36	47681.00	3640.016	
8	4582	25310	4625.42	-43.42	103.76	10766	10766	
9	4697	25799	4762.34	-65.34	21.92	480.4864	4269.316	
10	4753	25886	4786.7	-33.7	-31.64	1001.09	1135.69	
11	5062	26868	5061.66	0.34	-34.04	1158.722	0.1156	
12	5669	28134	5416.14	252.86	-252.52	63766.35	63938.18	
13	5628	29091	5684.1	-56.1	308.96	95456.28	3147.21	
14	5736	29450	5784.62	-48.62	-7.48	55.9504	2363.904	
15	5946	30705	6136.02	-190.02	141.4	19993.96	36107.6	
16	6501	32372	6602.78	-101.78	-88.24	7786.298	10359.17	
17	6549	33152	6821.18	-272.18	170.4	29036.16	74081.95	
18	6705	33764	6992.54	-287.54	15.36	235.9296	82679.25	
19	7104	34411	7173.7	-69.7	-217.84	47454.27	4858.09	
20	7609	35429	7458.74	150.26	-219.96	48382.4	22578.07	
21	8100	36200	7674.62	425.38	-275.12	75691.01	180948.1	
22						536636.4	564274.9	
23							0.951019	

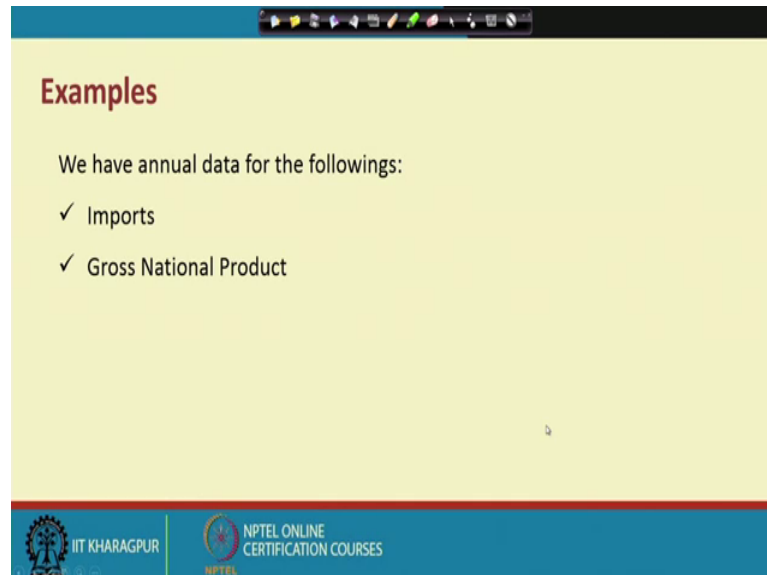
Now, this is what the regression output. So, so intercept is this much what about minus 2 4 6 1.4, and the beta coefficient is 0.28. By using this particular regression output, so what will it do? So, we first actually get the import estimated equation, just put alpha value and alpha value and beta value and then you scroll this structure. So, you will get the entire spreadsheet. So, that is actually estimated imports and this is the actual import, and the difference between actually important expected imports will give you the error component.

And now after getting the errors, so you will find out the e_t minus e_{t-1} ; that is the difference between two different situations error coverage that is actually like U_t and U_{t-1} . So, here this is e_t and then the previous data points will be e_{t-1} . So, the difference will be represent by e_t minus e_{t-1} . So, these are the difference between e_t minus e_{t-1} , and we need actually square of this variance of this particular structure. So, squaring this side, so we will have a the entire series like this. And then finally, we will get the some of this particular error variance. And then finally, we will get error sum squares that is square root of squaring the error sum squares. So, we will get the particular items.

So, now, so, as per the Durbin Watson d statistics. So, the covariance of e_t minus e_{t-1} by variance of e_t will give you this particular results. So, as a result, so this value by this

value will give you the Error component, means the plenty particular Durbin Watson d component.

(Refer Slide Time: 26:43)



The slide is titled "Examples" in a bold, dark red font. Below the title, it states "We have annual data for the followings:" followed by a bulleted list with two items: "✓ Imports" and "✓ Gross National Product". The slide has a yellow background and a blue header bar at the top. At the bottom, there is a blue footer bar containing the IIT Kharagpur logo and the text "NPTEL ONLINE CERTIFICATION COURSES".

So, as a result so, this particular value will give you the degree of multicollinearity, but in this context. So, the multicollinearity value is coming 0.95, which is actually not good for the model predictions. So, what will we do again. So, to solve these problems you have to increase the data points or you have to go for data transformations or you can add some variables or you can actually change the functional form.

So, somehow the Durbin Watson d statistics would improve. So, it should be coming in between the range of 1.2 or 2.5 so, that we can go ahead with these predictions. So, now; that means, technically we get to know, how Durbin Watson statistic can be calculated, and what should be the exact range through which actually prediction feasible or not.

So, taking typically if the curve or Durbin Watson statistic range is coming 1.2 2.5. So, then we will go ahead with the predictions, if it is not coming within that range. So, whether it is a lower range or upper range, then we need to do the kind of restructuring about the entire modelling process, either change the functional form, increase the sample size, go for the data transformations. In some extent then again re estimate the model, and find out the error term, again calculate the Durbin Watson d statistic and finally, check the value. Obviously, after doing all these process the value of the Durbin Watson statistic will improve and some of the software will directly give the kind of

value. So, what will you do? You have to just operate the process, then every time you rerun the model and then you will recheck the Durbin Watson statistic. So, you will continue to this particular process, till you get the model which is free from auto correlations.

(Refer Slide Time: 28:42)

Heteroscedasticity

$$Y_i = \beta_1 + \beta_2 X_i + U_i$$

Homoskedasticity:

$$\text{Var}(U_i) = \sigma^2$$

$$\text{Or } E(U_i^2) = \sigma^2$$

Heteroskedasticity:

$$\text{Var}(U_i) = \sigma_i^2$$

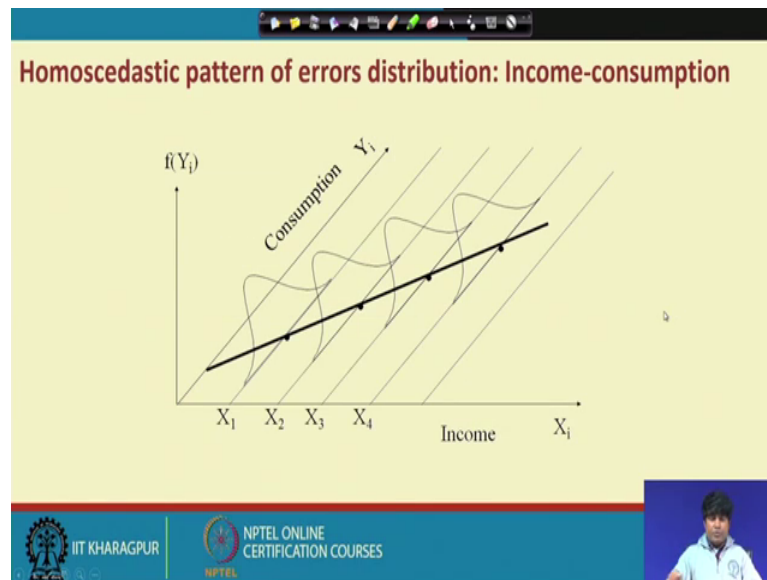
$$\text{Or } E(U_i^2) = \sigma_i^2$$

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

So, this is actually another virus through which you can move the particular model diagnostics, and this third one is the heteroscedasticity. Heteroscedasticity is a case where error variance is not actually homogeneous, but we have already a earlier in discussion that error variance should be same throughout the kind of models, but if that is in not the case, then this will give you the case called as heteroscedasticity; that means, technically variance of error term is not equal to constant that sigma square, then it is a question of heteroscedasticity. So, that is the case here heteroscedasticity. So; that means, if it is same then it is a homoscedasticity, if it is not same then it will be create a structure called as heteroscedasticity

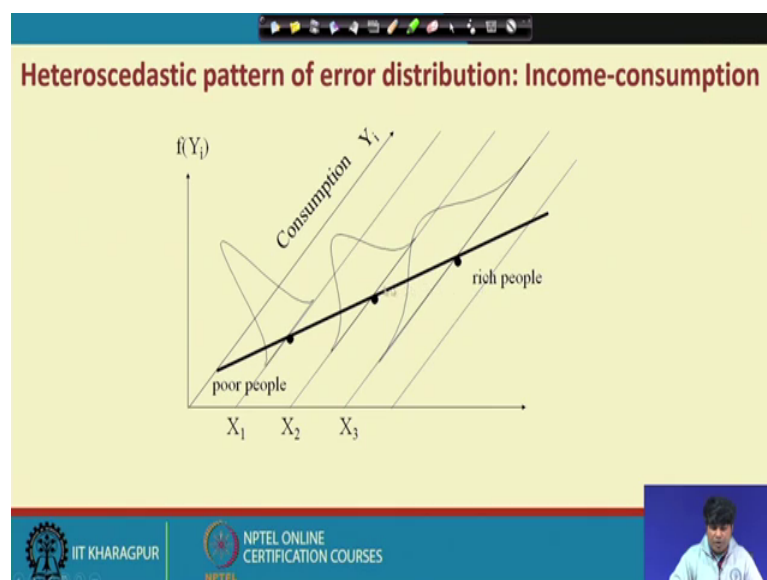
So, now again like multicollinearity and auto correlations here we have to check actually the structure of heteroscedasticity; and then we try to solve the problem of heteroscedasticity before you go for the predictions. So, what we will technically do here. So, again you have to go to check process and then find out whether there is a kind of a heteroscedasticity or not.

(Refer Slide Time: 29:48)

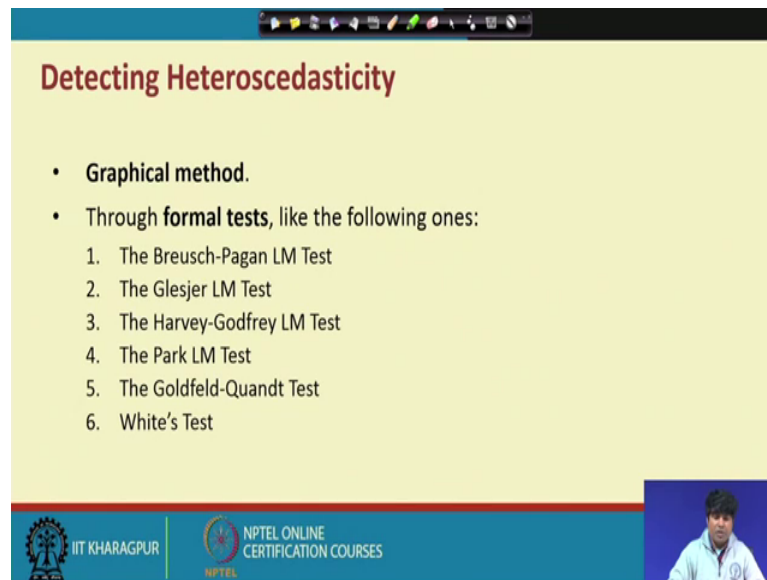


So, generally we have a here two options; one option is a homoscedasticity and heteroscedasticity. So, the graphically a homoscedasticity means you see the same structure, it is actually, coming actually similar kind of length and homogeneous variance altogether, but in other contrary, in the other case, here the degree of movement is changing. So, this is what signal of heteroscedasticity.

(Refer Slide Time: 30:10).



(Refer Slide Time: 30:22)



Detecting Heteroscedasticity

- Graphical method.
- Through **formal tests**, like the following ones:
 1. The Breusch-Pagan LM Test
 2. The Glesjer LM Test
 3. The Harvey-Godfrey LM Test
 4. The Park LM Test
 5. The Goldfeld-Quandt Test
 6. White's Test

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

This is a graphical inspection, but we have actually different mechanisms through which you can check the heteroscedasticity. So, means, like autocorrelation case. Here also we can graphically check, that is what we have already seen, and then the graphical check means, you just have the error term which is the difference between actual value and the estimated value, and then plot these error terms. If it will so a kind of trend, then this gives the signal that there is a heteroscedasticity problem.

In the mean times there are lots of formal tests through which you can check the heteroscedasticity. Like BPG test, Park test, Goldfeld test, White test, Harvey test. So, many tests are there through which you can actually check the heteroscedasticity, and find out whether there is a problem or not. If there is a problem you try to solve this problem before you do the prediction; otherwise your prediction will not be perfect as per the management requirement.

(Refer Slide Time: 31:24)

Examples

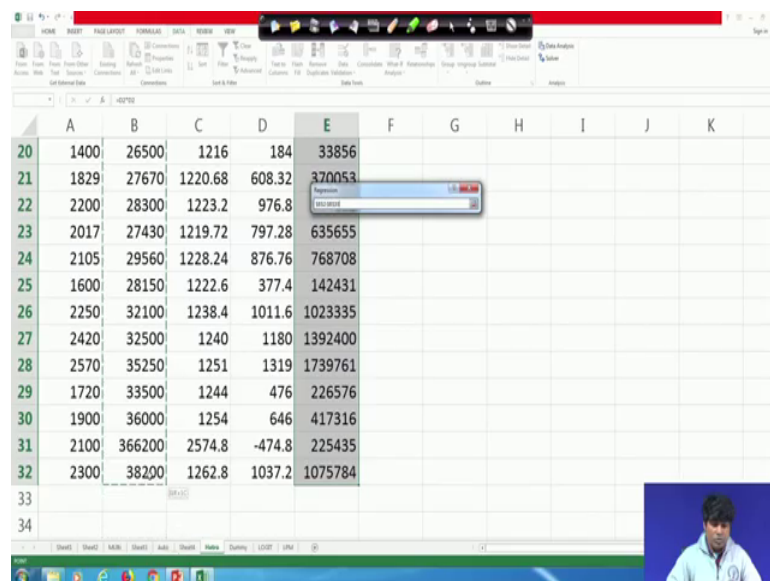
We have annual data for

- ✓ Savings
- ✓ Income

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, technically what I will do. I will take a problem again, then I will highlight how we can actually check and go for the kind of heteroscedasticity issue. So, what will you do? We can take another problem.

(Refer Slide Time: 31:37)



	A	B	C	D	E	F	G	H	I	J	K
20	1400	26500	1216	184	33856						
21	1829	27670	1220.68	608.32	370053						
22	2200	28300	1223.2	976.8							
23	2017	27430	1219.72	797.28	635655						
24	2105	29560	1228.24	876.76	768708						
25	1600	28150	1222.6	377.4	142431						
26	2250	32100	1238.4	1011.6	1023335						
27	2420	32500	1240	1180	1392400						
28	2570	35250	1251	1319	1739761						
29	1720	33500	1244	476	226576						
30	1900	36000	1254	646	417316						
31	2100	366200	2574.8	-474.8	225435						
32	2300	38200	1262.8	1037.2	1075784						
33											
34											

So, this is actually problem with respect to saving and the income and again. So, the first check of the process is go to the data analysis and then choose the regression package. So, here is in this particular problem. So, we are allowing saving is a dependent variable, and it is typically depends upon income, and then accordingly here we will specify again income was the independent variables, and then what will it do. So, we will just allow them to run the model and after getting the estimated models we refine.

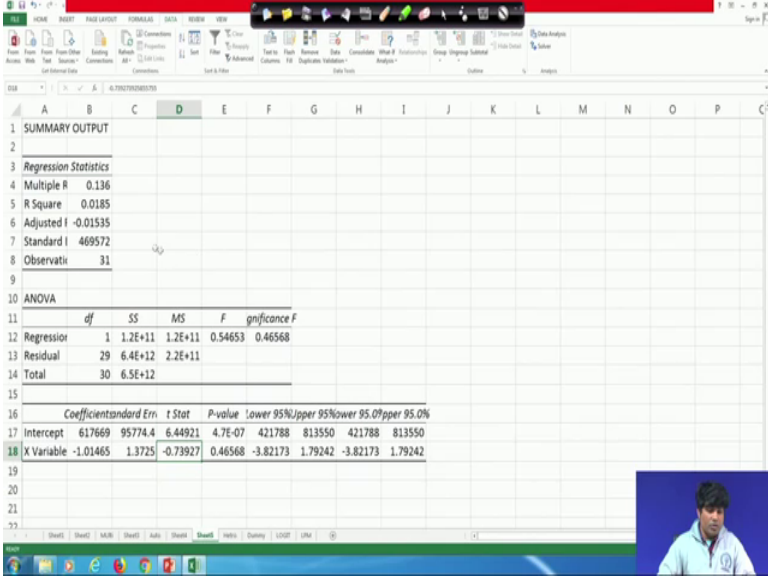
So, the saving and income are positively related to each other, and the alpha intercept is coming 1109 and beta intercept is coming 0.004. So, this gives the estimated structure. So, what will do? So, you go to the actually here. So, find out to saving estimates saving estimates that is equal to. Here the model outcome is 11 11 10 and 004. So, what will do. We will put here in 11 11 10 plus plus 0.004 into income. So, then you enter and then you scroll again. So, this will generate the saving estimates. So, this is you can get the saving estimates.

Then finally, what will you do, find out the error term. So, this is actually error observations. So, errors error observation will be the actual savings minus the estimated savings, and you will find the difference, and this is what the error observations. And now you can plot these error observations and check whether there is a kind of significant pattern or not, but actually in the heteroscedasticity process we like to find out the error variance. So, what will you do. You square the error terms, and then you connect this error terms with the any of independent variables. Here the independent variable is the income. So; that means, out of several methods which we have already highlighted here, one particular method can be applied to check the heteroscedasticity.

So; that means, technically. So, what will you do, after getting the error terms you find out the error variance, and then you check whether there is a heteroscedasticity or not. The simple structure of testing is after getting the error components. So, you can simply regress square of the error terms with any of X independent variables. here in the case of it is with respect to income. So, what will you do here again? So, you find out the square of the error term, then this is nothing, but actually this into this. So, you define the square of the error terms, and again what will it do. So, it can generate the particular series.

Now, what will it do again? So, you go to the data analysis package, and again choose the regression, and here the choice of the dependent variable is a square of the error terms, and that with respect to any one independent variable or a series of independent variables. So, you can connect with it this one, and this ones and then you put and you will find again a a regression output, and check whether its statistically significant or not. In fact, in this case your variable is not actually coming statistically significant.

(Refer Slide Time: 35:34).



SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.136					
R Square	0.0185					
Adjusted R Square	-0.01535					
Standard Error	469572					
Observations	31					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	1.2E+11	1.2E+11	0.54653	0.46568	
Residual	29	6.4E+12	2.2E+11			
Total	30	6.5E+12				
Coefficients						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	617669	95774.4	6.44921	4.7E-07	421788	813550
X Variable 1	-1.01465	1.3725	-0.73927	0.46568	-3.82173	1.79242

So; that means, it indicates that heteroscedasticity is not so serious problem, but still the model is not actually good fit, if you go to this particular original output, that is this, the original output this one and here R square is coming 32 percent. And so; that means, technically it may not have actually heteroscedasticity problem it may have a autocorrelation problem. So, technically what will it do? So, this is what actually check process.

So, every time you have to find out and check the particular process, and see whether there is a heteroscedasticity or not, but in this case it is not actually showing heteroscedasticity, but this is what the check process through which you have to check and finally, conclude whether there is a heteroscedasticity or not. So; that means, technically. So, the diagnostics issues are like this. So, a test stage we like to check multicollinearity problems, if there is more number of independent variables, and then you have to go through autocorrelation check, and then we have to go through heteroscedasticity check.

So; that means, technically these are all virus in the system, and until unless you check all these details and you cannot use this particular model for the kind of predictions and the kind of management requirement. So, every times you have to see the degree of multicollinearity in estimated model degree of auto correlations and the kind of heteroscedasticity.

So, finally, you have to use the model for the prediction, which is somehow free from multicollinearity issue. Even if it is not coming to 0, but to somehow it may have actually very low presence of low multicollinearity, and it may have the range of autocorrelation between 1.5 to 2.5, and it should also should not have actually so much heteroscedasticity issue. So; that means, error variance should be actually coming to the kind of homogeneous structure.

So, in the first instance when you will do this kind of modeling, you may not get the result as per the particular requirement, but over the time what we will do. So, you have to do the kind of continuous search process by changing the kind of modelling structures, by including one after another variables or dropping one after another variable, increase the sample size decrease the sample size, change the functional form, go for the data transformation structures, means there are various alternatives are there or options are there. So, what is actually the best requirement is, every times you have to get the first end output and that is the best model, and then you have to continuously check one after another items.

And then you have to declare that the model is free from all these obstacles, starting with multicollinearity, autocorrelation and heteroscedasticity. Once the model is actually free from all these errors, then you will finally, use this model for prediction, and the kind of management decision. Otherwise this particular model cannot give the better prediction and better management, managerial decision. So, the right choice is to go the kind of standard procedure, and then you find out the best models which is actually need for the a particular business predictions and the kind of management requirement. So, with this we will stop here.

Thank you very much have a nice day.