

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

Lecture – 29
Heteroskedasticity problem

Hello, everybody. This is Rudra Pradhan here. Welcome, to Engineering Econometrics. Today, we will be continuing with model diagnostics and that to the problem of Heteroskedasticity. Like autocorrelation heteroskedasticity is also a problem with respect to error terms in the econometric modeling we have dependent variables, we have independent variable and that too independent variables explained and independent variables unexplained. So, independent variables unexplained are nothing, but the error terms. Now, while using any kind of you know econometric modeling for the management decision or to look for the solution of a particular engineering problem.

So, the models should be free from all these errors including this heteroskedasticity and for that we have to first have the estimated model get the error term and then check the error terms whether it is behaving with respect to heteroskedasticity or it is in favor of homoscedasticity. So, corresponding to heteroskedasticity the counterpart is called as a homoscedasticity. Homoscedasticity is good for the modeling, but heteroskedasticity is against the modeling. So far as you know decent making process is concerned and to look for the solution for a particular you know engineering problem.

So, like autocorrelation, it is also step by step process just to you know connect with the theory develop model, then must have data then with the help of where the OLS technique and there with the help of inner software we can first get the estimated output and after getting the estimated output then use the estimated equation to have the error terms.

So, now once you have the error terms then we can check the heteroskedasticity issue. So, obviously, we are very much interested here to know what is exactly heteroskedasticity and how to detect this particular component what are the regions through which heteroskedasticity can be a problem in the modeling environment and a what are the consequence, if it is there in the system and whether there is a kind of you

know solution and how will you go about this particular you know a requirement to set the model which is free from all kinds of you know diagnostics.

So, we start with the simple structure of you know heteroskedasticity. So, we have you know models and then corresponding to models we have the error terms. So, the game is like this.

(Refer Slide Time: 03:23)

What is Heteroskedasticity : $\sigma_{u1} \neq \sigma_{u2} \neq \sigma_{u3} = \sigma_{u4}$

Hetero (different or unequal) is the opposite of Homo (same or equal)...

Skedastic means spread or scatter...

Homoskedasticity = equal spread
Heteroskedasticity = unequal spread

The diagram shows a 3x3 grid with columns labeled U_1, U_2, U_3 and rows labeled U_1, U_2, U_3 . The cells contain error terms u_{ij} . A diagonal line is drawn from the top-left to the bottom-right, and a horizontal line is drawn across the middle row. The bottom-right cell contains the equation $\sigma_{u1} = \sigma_{u2} = \sigma_{u3} = \sigma_{u4}$.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, let us see. So, first of all we start with like that you know what exactly heteroskedasticity is. So, hetero means this is you know different or something called as an unequal and its counterpart is called as you know homo and that is nothing, but called as you know same or equal. So, that means a different or something called as you know same unequal and equal. So, that is with respect to the error terms. So, that means, usually once you have the estimated models and from the software itself we can check the behavior of the error terms that itself gives the signal about the autocorrelation and it also gives the signal about the heteroskedasticity.

Now, we are in the track of you know heteroskedasticity only and even that plotting itself will give you some kind of you know exposure whether the error term behaves normally or some kind of you know as per the requirement of the modeling or it is going against the modeling. So, you know; obviously, with respect to error terms where we are looking for actually covariance among the error terms here in the heteroskedasticity we are looking for the variance of the error terms.

For instance let us start with the understanding like this let us we have three error terms with U_1 , U_2 and U_3 and then we can create a covariance matrix with these three error terms. Then, obviously, we can have 3 into 3 a 3 into 3 you know what we can called as what we can called as you know you know error terms U_1 , U_2 , U_3 . So, that means, we just represent column wise and row wise then we will have a matrix like this. So, obviously, since there are three error variables, so obviously, the variance covariance matrix with respect to all these three variables will be having 3 into 3 matrix.

And, then you will find a it will find a the diagonal elements are you know error variance that is what it is called as you know sigma square U_1 . Then, this is called as a sigma square U_2 because the game is with respect to U_2 and U_2 and then sigma square U_3 that is between U_3 and U_3 . Rest of the you know you know place we will have actually covariance; that means, technically with U_1 and U_2 then it is become sigma $U_1 2$ and this is sigma $U_1 3$ and this is this is sigma $U_2 1$ and this is sigma $U_2 3$, then sigma $U_3 1$ and sigma $U_3 2$ and this is what the variance covariance matrix with respect to three error term and that can be extended with respect to k number of you know error variables.

We have lots of you know interesting or you know additional models with respect to all these error terms like you know as (Refer Time: 06:50) model (Refer Time: 06:51) models all these things are there we will discuss in details in the later stage. But, in the means time we need to understand what is the a error component, how what should be the behavior of the error term and how to detect the heteroskedasticity ok, then we will look for the kind of you know solution as far as the particular you know you know requirement is concerned.

So, here the entire you know matrix can be divided into two parts, you know which we have already explained in the case of you know autocorrelation. So, now we have two parts one part is the diagonal elements the other part is the off diagonal elements. So, off diagonal elements so, we have sigma $U_1 2$ sigma $U_1 3$ and sigma $U_2 3$, just this side you know it is in nothing, but you know transpose of these elements that is nothing, but you know since they are symmetric so, that is why sigma $U_2 1$ sigma $U_3 1$ and sigma $U_3 2$ similarly sigma $U_3 3$ and sigma $U_2 3$ they are all almost all same.

So, now when we are targeting σ^2_{U1} , σ^2_{U2} , σ^2_{U3} then by default this problem will be going to oh autocorrelation. Now, we are looking for actually the variance of the error term that is over actually σ^2_{U1} , σ^2_{U2} and σ^2_{U3} and that with respect to U_1 , U_2 , U_3 . So, what is the exactly requirement for heteroskedasticity understanding is that, so, the error variance over the cross sectional unit or over the time should be equal. So, that means, technically so the requirement of the modeling as per the OLS technique is that σ^2_{U1} equal to σ^2_{U2} is equal to σ^2_{U3} and simply you can call as a σ^2_U , that is what the error variance and what this is what it is called as you know homoskedasticity.

So, we are looking for you know equal variance and the requirement of the OLS technique for using any kind of you know kind of you know decision making process is that error variance should be equal over the cross sectional unit and over the time. If they are different then that will automatically take you to the problem call as you know heteroskedasticity. So, if you say that you know what is homoscedasticity? Homoscedasticity is simply simply represents the equal error variance and a heteroskedasticity simply represents unequal error variance.

So, that means, technically if that is not the case then a heteroskedasticity the typical heteroskedasticity will be the case for these three error variable where σ^2_{U1} not equal to σ^2_{U2} not equal to σ^2_{U3} and by default it will take you the concept called as a σ^2_{Ui} . So, that means, if they are not same then they really create a kind of you know functions.

So, now we like to check whether the error variance or variances are equal over the time or they are you know unequals and if it is unequal and what is the level level of you know are you know unequal. So, that like you not the degree of autocorrelation. So, we have also the kind of an understanding the degree of you know heteroskedasticity. So, we like to check what extent we can actually identify the heteroskedasticity problem and what extent we would like to solve this heteroskedasticity problem.

So, that the other diagnostic will be in the truck and the model will not be getting affected negatively and that we can use this models for the management you know decent making a requirement. So, in the simple language almost capacity means equal

spread that is the equal error variance and heteroskedasticity represents unequal unequal spread that to unequal error variance. So, now, with respect to this understanding so, we will we will see a how is this particular you know case ok.

So, now we will move to this particular you know process like this, ok.

(Refer Slide Time: 11:18)

Regression Model

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

Homoskedasticity:

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

Or $E(U_i^2) = \sigma^2$

Heteroskedasticity:

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

Or $E(U_i^2) = \sigma_i^2$

$\sigma_i^2 = \sigma_{U_i}^2$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, to summarize the homoskedasticity represents the variance of the error terms where sigma square you know U and sigma square U 2 up to you know sigma square U k a they are all same and as a results we will simply write variance of error terms is equal to sigma squares or expected value of you know error square of the error term should be sigma square. That is what the requirement and that is what the term called as you know homoskedasticity.

This is good for the modeling as long as you know error variance are same over the time and over the a you know cross sectional unit, then we are in good track. When they are you know deviating from the you know various cross sectional units or deviating from different time = then by default that is what is called as you know homoskedasticity problem.

So, by default homoskedasticity definition in technically a like this; so, variance of error term should be sigma square I, so, it should not be sigma square U because the moment will say sigma square U means they are equals. So, when we are putting sigma square U

i. So, either you can call sigma square i or you can call you can call sigma square U i. So, that both are same just you have to understand you know accordingly.

So, or are you know expected value of square of the error term should not equal to sigma square U only it should be sigma square U i. So, in we write sigma square U i that that clear clearly indicates that you know all the error variance are not equal, so, they are unequal and when they are unequal then that is what the problem is all about and that is what the problem is called as you know heteroskedasticity and this one of the you know biggest virus in the modelling setup while looking for a kind of you know solution for any kind of you know an engineering problem and that with the help of engineering econometrics.

So, this is what the generalized concept of you know heteroskedasticity. So, I will give you the kind of you know clarity what is you know more about this particular component and how we can actually go about it.

(Refer Slide Time: 13:37)

What is Heteroskedasticity

Assumption of the CLRM states that the disturbances should have a constant (equal) variance independent of t:

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

Therefore, having an equal variance means that the disturbances are homoskedastic.

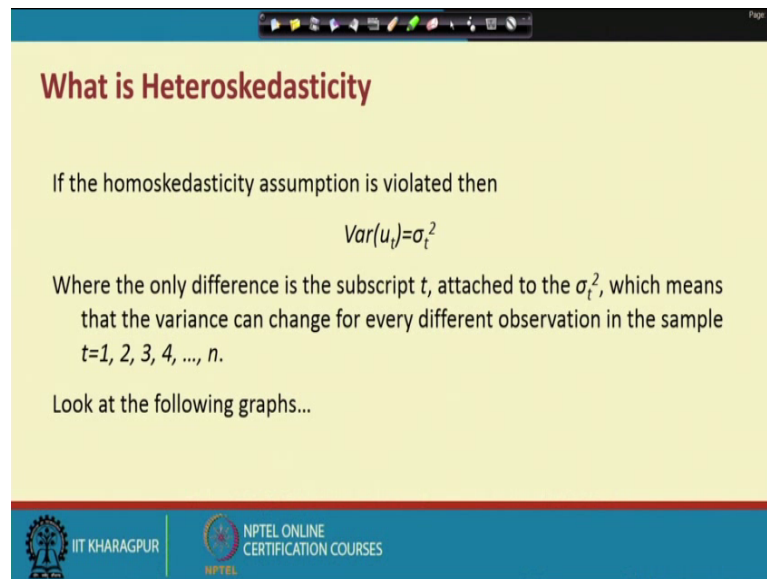
IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

So, against if you connect with you know OLS requirement that is what called as you know CLRM classical linear regression modeling and that with the help of you know OLS technique to estimate the parameters depending upon the problem formulation and the data availability. And, the simple understanding is that here variance of variance of u t is equal to sigma square. If that is not the case then it will be variance of u t is equal to simply sigma square I and that that clearly you know you know differentiate the

understanding of you know homoskedasticity and the understanding of the heteroskedasticity.

So, let us see how is this particular you know flow against to you know so far as you know detection criteria is a concerned.

(Refer Slide Time: 14:29)



The slide is titled "What is Heteroskedasticity" in a bold, dark red font. Below the title, it states: "If the homoskedasticity assumption is violated then". This is followed by the equation
$$Var(u_t) = \sigma_t^2$$
. The text then explains: "Where the only difference is the subscript t, attached to the σ_t^2 , which means that the variance can change for every different observation in the sample $t=1, 2, 3, 4, \dots, n$." It concludes with "Look at the following graphs...". The slide has a yellow background and is part of an NPTEL presentation, as indicated by the logos and text at the bottom.

What is Heteroskedasticity

If the homoskedasticity assumption is violated then

$$Var(u_t) = \sigma_t^2$$

Where the only difference is the subscript t, attached to the σ_t^2 , which means that the variance can change for every different observation in the sample $t=1, 2, 3, 4, \dots, n$.

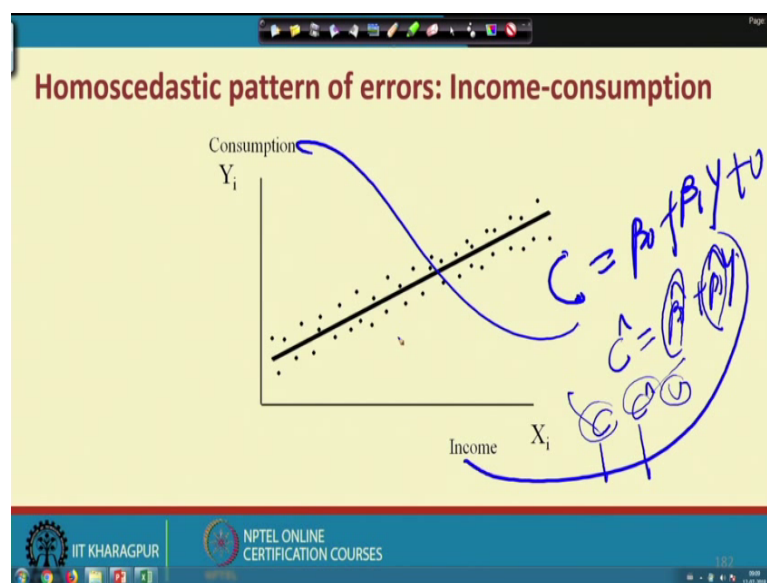
Look at the following graphs...

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Of course, we understand what is exactly or how to heteroskedasticity, but ultimately a you know we would like to know you know more something more about this particular component, how to detect and what are the ways you can actually check this particular you know component and what are the regions through which it will be it will be in the system. It may be it may be due to actually a wrong functional form or the kind of you know structuring of the model and with respect to the data structures it may be a you know due to you know anything.

But, we like to track how it happens and what is the level of this particular you know problem and once you identify the exact reason through which this problem is coming then we can also equally look for the solution. Until unless you know the typical reason behind this heteroskedasticity involvement so, you may not actually look you know come into a kind of you know perfect solution. So, that is that is why clear cut understanding must be requires.

(Refer Slide Time: 15:34)



And, as a result first we start with you know graphical inspections. Let us say there are two variables income and consumption and that is the (Refer Time: 15:44) mix problem. So, here consumption absolutely depends upon you know income. So, consumption absolutely depends upon you know income which which is a nothing, but here.

So, like you know c equal to C equal to β_0 plus $\beta_1 Y$, C stands for consumption and Y stands for let us say income and we have as usual the error term. So, we will go we will go about means the functional relationship between consumption and income is like this and with the help of the data we will estimate the consumption function and have the beta parameters like β_0 and β_1 .

And, as a result we will have actually estimated equation \hat{C} that is what is called as you know $\hat{\beta}_0$ plus $\hat{\beta}_1$ and in that to Y . So, by default error will be removed in the process this is the known parameter and this is the known parameters, now if you simplify we will have now C series and we will have now \hat{C} series that is what the actual consumption series and that is what the estimated consumption series now the difference between the two will give you the error terms.

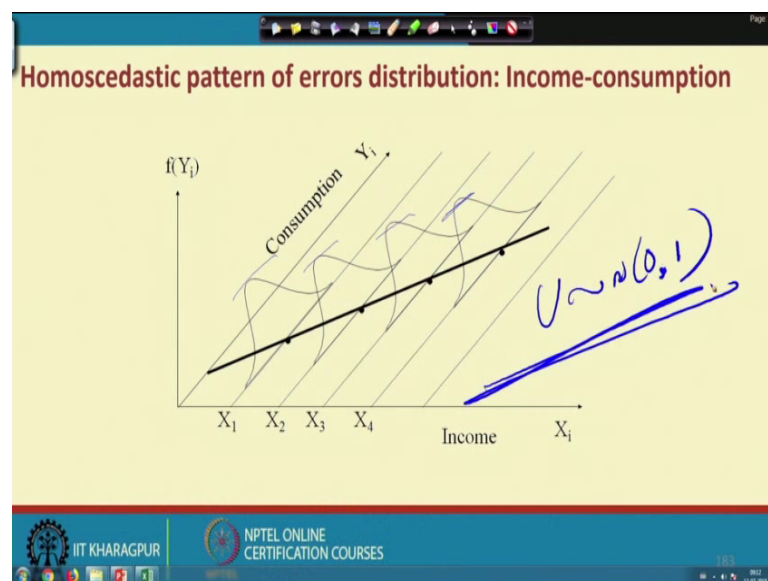
Now, what we are doing actually corresponding to income and consumption we will derive the error term and then plot the error terms and check whether actually it follows homoskedasticity pattern or it will it follows the heteroskedasticity patterns. So, now, if we will check here is in this in this graph and with respect to a particular you know data

set and the particular you know estimated a model and we will find the error behaviors you know little bit you know close each others.

So, when they are you know little bit close each other and within a particular you know small interval and as a result it gives indication that you know it is the issue of you know homoskedasticity and that is good for the modelling requirement or the modelling structure. If it is deviating you know maybe the spread is very high towards you know for you know towards down or in the left skewed or right skewed then by default this will give you the indication that you know it is the homoskedasticity involvement.

So, it means plotting the error term itself will give you the signal whether the model is correctly specified or model it is some kind of you know non diagnostics to correct the kind of you know requirement. So, that means obviously so, first requirement is to check the behavior of the error terms and then get to know what is the level of this particular you know flow.

(Refer Slide Time: 18:26)



So, the other way you can represent you know in a 3-dimensional framework. So, that is what you know Y and X and then the error behaviors you will find if you say equal spread and and a you know you plot this error terms you see here is the ad in a structure of this you know error plotting in that with you know a normal distribution, you will find the curve is actually more or less actually same. So, they are they are very much you know converging each others.

So, as a result it gives the signal that you know it follows the normal distribution pattern and when it follows the normal distribution then by default that will be appearance of the homoskedasticity. So, in short the error term should follow you should follow normally distributed with the mean 0 and the unit variance, that is what you know called as you know 0 one.

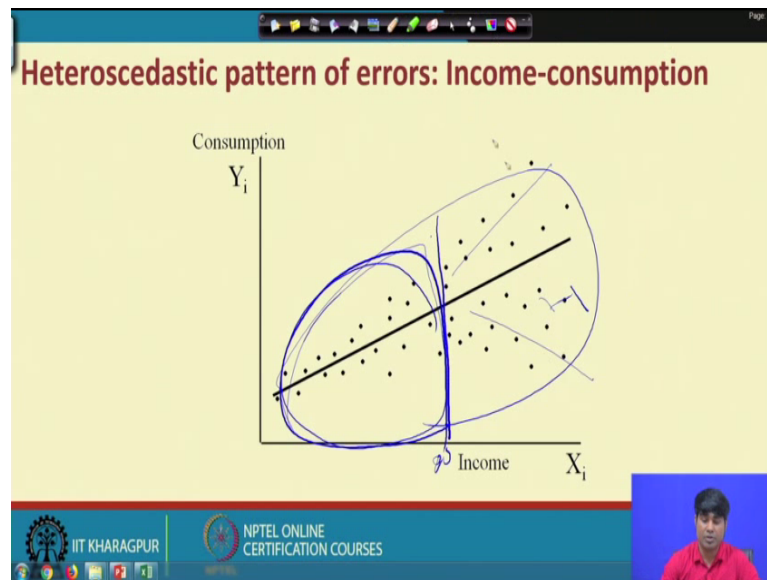
So, the behavior of the error term should be like this when we use OLS to help the estimation and use this estimated out for any kind of normal indecision and that is for the engineering requirement. So, you should you know follow you know zero mean and unit variance that is what the requirement if not then there is a problem and that too when there is a question of heteroskedasticity then the variance should not be unit or equal it will be unequal. It will it will give you some kind of you know different patterns or you know some kind of you know functionality.

Our job is a how to reduce this kind of you know functionality because the error terms and that to the you know explain independent variables, there in the independent clusters basket this will not form some kind of you know you know problematic like you know the relationship among them and then there should not be high volatile among them. So, if that is the case then you know what is happening it goes against the you know blue theorem that is what colors you know best linear unbiased estimators.

So, that means, the parameter which will have they have a you know simply a large variance large standard errors and as a result the significance of the particular parameters and the particular variables will be getting affected and as a result it will go in the model really go against the kind of you know specification test, goodness fit test and overall the reliability of the models will be getting affected and as a result we cannot use this model for the decision making process. This is it this is the simple kind of you know understanding in the first instance.

So, this is what the homoskedasticity you know look and if you look if you go for you know hetroskedasticity look, let us see here.

(Refer Slide Time: 21:19)



This is one way of you know plotting where you know, if you compare the plotting of this ones you see here the points are very close to you know very close to each other with the estimated line. And, as a results if you plot in a 3-dimensional and you know picture then the behavior of these error terms are a close you know more or less normally distributed and they look very similar. And, as a result it gives clear indication that you know there is the homoskedasticity.

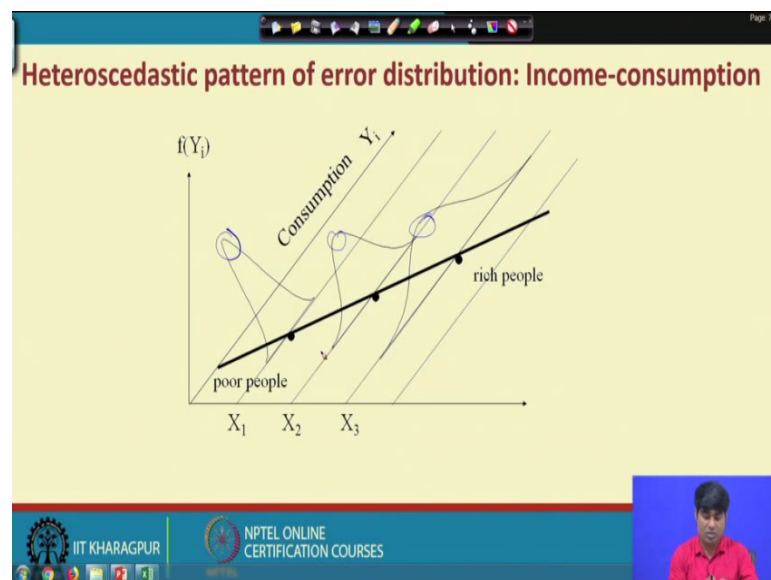
Now, if you look here the initially these you know points are you know a little bit you know converts each other or close to the estimated lines then after a particular point of time. So, like here if you see here is after this point times there is a spread actually. So, they are you know and I know diversifying here and there. So, as a result so, this gives the clarity that you know there is some you know heteroskedasticity problem and, that to up to a certain point it is, it looks like you know homoskedasticity and after that it gives like you know heteroskedasticity.

So, that means, clearly indicates that you know data itself the level of data itself you know bring you the issue of you know homoskedasticity and a heteroskedasticity. For instance, if you if you you know build a models up to this point this much of data let us say the entire data point is is say 100 and this will be up to 80 and the rest is a 20 percent. So, now if you build a model with respect to 80 percent of the data with respect to this plotting, then there is a high chance this will give you homoskedasticity picture.

Now, if you club the entire you know data then this will give you the heteroskedasticity pictures. So, in this case you know without any further discussion we the one ways to have the solution is either you will normalize the data or else you can you know divide the data into a two different groups and then you know estimate the model separately and then check what is the validation or the functional relationship between these two. So, this is another way of you know looking to this model to look you know to have the better solutions as per the particular you know requirement.

So, obviously, so, this is what the kind of you know requirement and then all right.

(Refer Slide Time: 23:58)



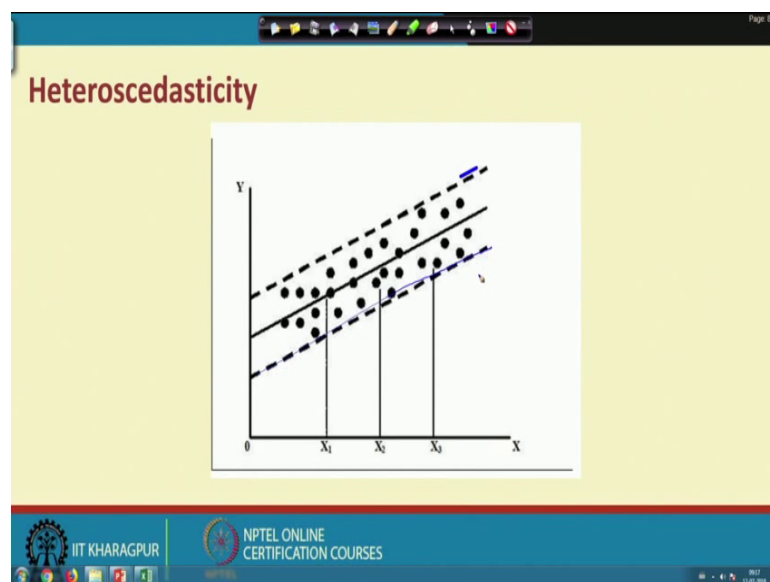
So, now, if you see here there is a variation, now the same you know instruction can be observed you know with this particular you know diagram this is the 3-dimensional plot and heres you will see. So, the plots are actually little bit you know are then down and then down like this so; that means, the look of these curves are not very much similar. So, this gives the signal that you know there is a heteroskedasticity for instance very simple here.

So, if you will simply you know compare this one's this ones like you know this called and then and these diagrams this ones you see here they are actually look like very similar, but if you look here is they are very you know you know what we can call it you know completely you know different we all to that, they are not at all similar. So, that means, they are you know deviating one you know differently from time to time or you

know sample to samples that itself gives the signal that you know this case there is heteroskedasticity and this case there is a homoskedasticity. That is the two different basket all together.

And, by the way if we are in the track of enormous capacity we are in you know good moment and when you are in the heteroskedasticity side then we are in the bad moment. So, as a result, we need to actually transfer the modeling structure or you know restructure the entire set up. So, that we should be in the track of you know homoskedasticity then you will to go ahead with the prediction for casting or something like that as per the particular requirement, ok. So, this is what the case.

(Refer Slide Time: 25:37)

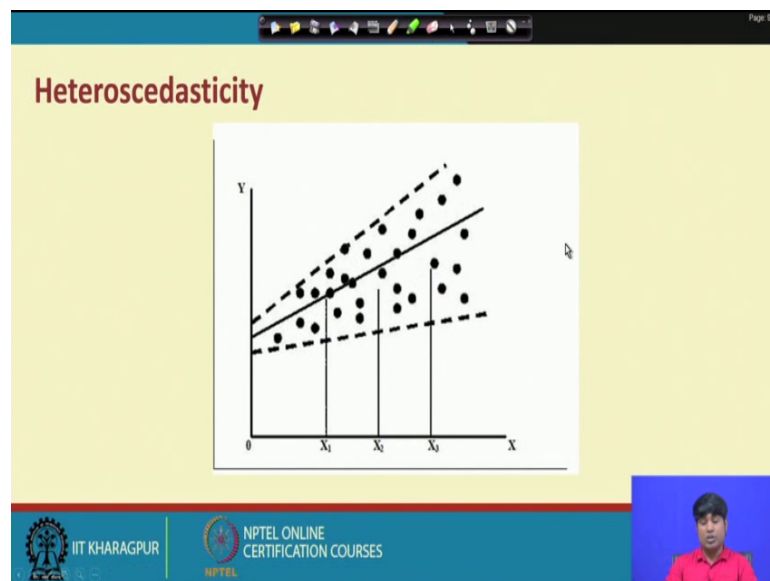


And, now this is another kind of you know heteroskedasticity loop. So, of course, they the points are little bit you know in so, you know these means a little bit deviation you know here with respect to error terms they are not very close to the you know estimated line, but what is the best part of this particular you know plotting is that you know they are in a kind of you know confidence intervals, right. So, if of course, any kind of you know plotting even if they are you know a big spread still we can have a confidence interval.

But, now the confidence interval should not be very high while you know detecting in the issue of heteroscedasticity in compared to in comparison to you know how much capacity. So, there should be actually spread, but that spread should not be so, you know

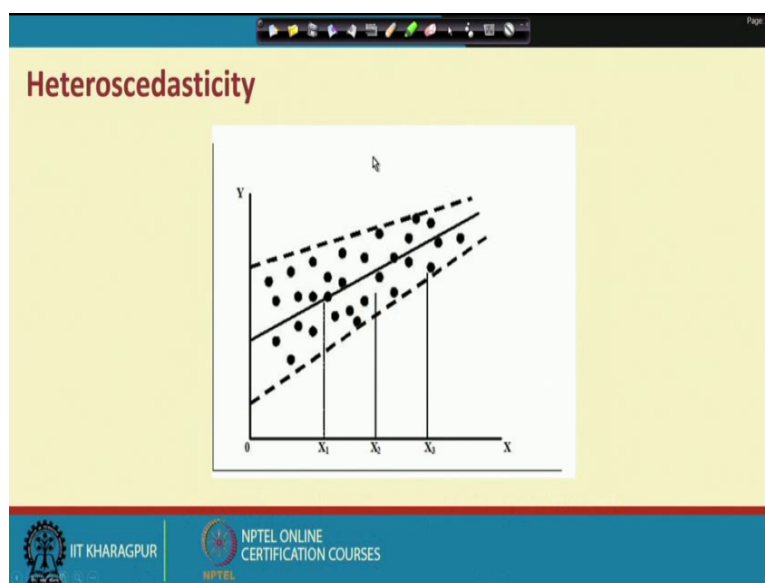
deviating in nature. So, it should be very close. So, that means, standard deviation should not equal to 0 standard deviation should not be very high that too with respect to error term and standard deviation should be you know very close to you know or a very low it that then you know this really give you the better judgment or in a better indication while using the kind of you know requirement, ok. So, that is what the case and so, we will see how is the further behavior.

(Refer Slide Time: 26:57)



So, this is another look actually heteroscedasticity, this is another look and; that means, the the first one is in a kind of you know you know in a moment that when a kind of you know linear angles.

(Refer Slide Time: 27:11)



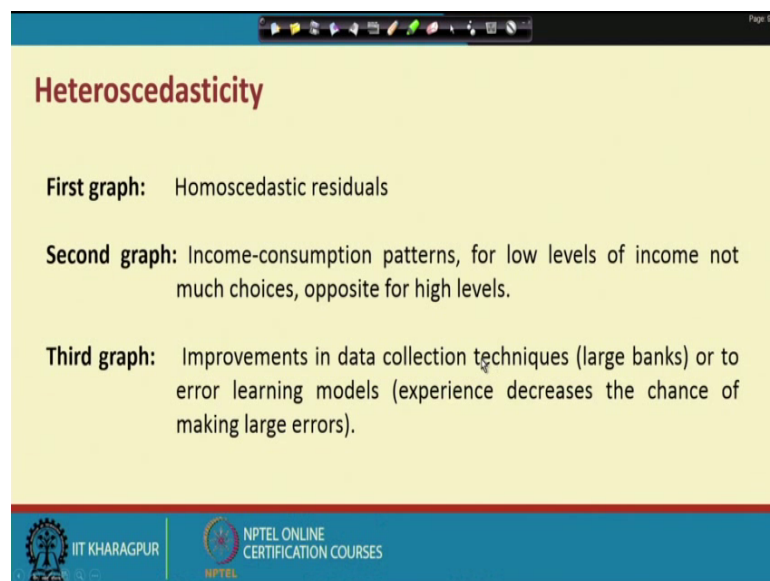
And, the previous one is a kind of an increasing kind of you know spread and; that means, it is a kind of you know positives spread and this is a kind of you know negative spread; that means, declining right kind up in a moment. So, initially the spread is very high and then over the time the spread is a slightly actually low and low.

So, that means, this is a good moment altogether, but this becomes you know bad moment because initially the the spread is very low over the time it is very high and by the way it is not in your control because when you are dealing with you know a multivariate problem. There are many variables you know initially they are very stables then over the time the there is a instability because of you know certain results or in external factors.

So, as a result so, the spread will be you know cannot be actually constant the spread will be you can say you know very much a you know different over the time over the you know over the kind of you know cross sectional units. So, now, this is actually the first one is the is increasing kind of you know spread over the time and this is the decreasing kind of you know spread over the time. But, by the way, but the both the cases and there is a heteroskedasticity and we should look for that you know and that means, technically we should you know quantify it and test it like you know multicollinearity and autocorrelation.

So we check whether the particular heteroskedasticity problem just artistically significant or not if that is the case then we cannot just go ahead. So, we have we we look for the kind of you know solution and then we need the declaration that this problem is free from you know heteroskedasticity, but you know it is very difficult to you know completely clean this virus. But, at least you know it should be in the tolerance level, like you know this issue you know multicolor you know and this you have to know out of correlation. So, that is the case actually.

(Refer Slide Time: 29:14)



Heteroscedasticity

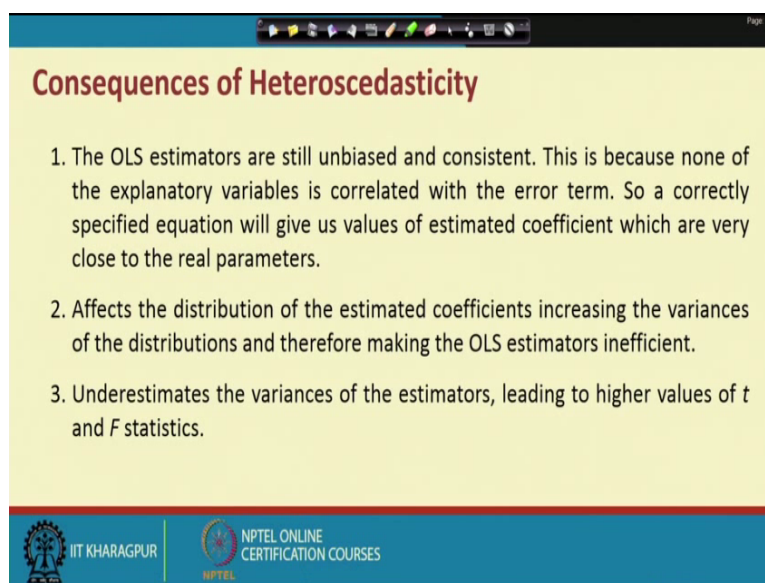
- First graph:** Homoscedastic residuals
- Second graph:** Income-consumption patterns, for low levels of income not much choices, opposite for high levels.
- Third graph:** Improvements in data collection techniques (large banks) or to error learning models (experience decreases the chance of making large errors).

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, now, if you ask me how it can go about this kind of not movement so, the first one is the graphical check and where you know you just start with the step by step process you know build the models, connect the data, use the software have the error terms. Then plot this you know error term that is what the residuals, then we will get to know how is the kinds of you know a moment with respect to error term and that to with this problem in terms of you know income concepts and relationships technically, ok.

So, these are all you know various requirement.

(Refer Slide Time: 29:55)



Consequences of Heteroscedasticity

1. The OLS estimators are still unbiased and consistent. This is because none of the explanatory variables is correlated with the error term. So a correctly specified equation will give us values of estimated coefficient which are very close to the real parameters.
2. Affects the distribution of the estimated coefficients increasing the variances of the distributions and therefore making the OLS estimators inefficient.
3. Underestimates the variances of the estimators, leading to higher values of t and F statistics.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, now means after knowing the kind of you know initial inspections. So, we can actually you must come to the kind of know conclusion about the consequence. If you say that it is a virus so, what is exactly the impact or the kind of you know incidence? The consequence of course, you know is not so good. It will largely effect the model reliability which I have already clearly high you know highlighted in the classical linear regression modeling.

If you connect with you know Gauss Markov theorem that is the BLUE theorem. So, we need actually the parameter should be completely unbiased this should have a very very minimum means minimum variance and then this is the parameter should be very consistent. So, unbiasedness, minimum variance and then the kind of you know consistency or sometimes we can use also sufficiency, but sufficiency criteria are not required here. So, we to the three things while you know checking this kind of you know diagnostics.

So, parameters should be completely unbiased and they should have a very minimum variance and they should you know very consistent of a different point of time or different cross sectional unit. Now what is happening in that case of you know or in the presence of you know how much heteroskedasticity. So, the parameters will not follow the BLUE theorem.

So, they may not actually you know means they may follows you know some items, but they may not follow the other items. For instance having actually heteroskedasticity the

parameters will be still unbiased and consistent, but it will not give you the minimum variance rather, it will increase the variance factors as a reserve standard error will be very high and what is happening the parameters will be not statistically significant even though they are unbiased and you know consistent. That is what is the actually game is all about you know heteroskedasticity.

So; that means, when these viruses are there like you know multi heteroskedasticity and autocorrelation then you know our parameters or the estimated models in you know you know setup is not to proceed for the forecasting or the decision making or as well you know you know kind of you know implication about any engineering problem.

So, now this is what the consequence simple understanding of you know heteroskedasticity and the behaviors and then we look for the kind of in solutions, means technically graphically we inspect and you know have the kind of you know solutions and a installation were about the homoscedasticity or you know heteroskedasticity.

And, then of course, so far as a consequence is concerned in the case of you know homoscedasticity so, they will be follow all the criteria like you know they are unbiased, they are consistent and they are also all of the pattern in a called as minimum variance. But, in the case of heteroskedasticity the parameters will be still unbiased and then consistent, but they cannot follow the minimum variance, that is what and that to the minimum variance is one of the most important requirement of the BLUE theorem. If that is not the case then this will largely affect the kind of you know systems. That is why we need actually you know, clean models which can give the better solutions as per the particular requirement. With this we will stop here and we will continue this lectures in the in the next class.

Thank you very much.