

Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras
Lecture 52

Relaxing the assumptions of CLRM-Autocorrelation and Heteroscedasticity Part - 3
(Refer Slide Time: 0:55)

Heteroscedasticity



What does it mean?

$\text{var}(u_i) = \sigma^2 \Rightarrow$ homoscedasticity / constant error variance

$\text{var}(u_i) = \sigma_i^2 \Rightarrow$ Heteroscedasticity

$y_i = \alpha + \beta x_i + u_i \Rightarrow \text{var}(u_i) = \sigma_i^2$ } what.

$y_j = \alpha + \beta x_j + u_j \Rightarrow \text{var}(u_j) = \sigma_j^2$ }



Welcome we will discuss today about another assumption of classical linear regression model that we made, that was the assumption of homoscedasticity and if that assumption is violated then we will say that the data set is basically suffering from Heteroscedasticity problem just the opposite of homoscedasticity.

So, we will first try to understand what is Heteroscedasticity. So, let us try to understand what does Heteroscedasticity mean. So, basically Heteroscedasticity mean that the error variance, variance of u_i if you recall, variance of u_i we assume that that is constant that was the assumption of classical linear regression model and if this assumption is satisfied then this is called homoscedasticity and homoscedasticity basically means a constant error variance the other name of this is constant error variance.

Now, why this is called constant variance? See here, I have mentioned variance of u_i but here there is no i subscript so that means vary the i th person's error variance, j th person's error variance, k th person's error variance, all the error variances are sigma square, since there is no

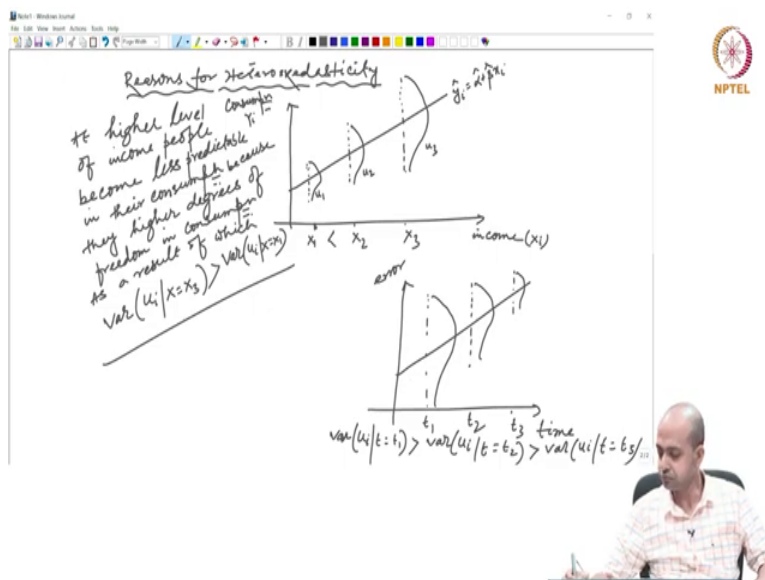
subscript like i, j and k after this σ that is why this is called homoscedasticity or constant error variance.

But if the error variance looks like this, variance of u_i equals to σ^2_i then this is called Heteroscedasticity that means I am saying that the error variance for the i th individual is σ^2_i and then error variance for the j th individual is σ^2_j . So, that means when I am saying this, this is Heteroscedasticity.

So, if you are writing your model like this y_i for the i th person equals to $\alpha + \beta x_i + u_i$ and then you are saying that variance of u_i equals to σ^2_i and then when you are writing the same equation for the j th person that would become $\alpha + \beta x_j + u_j$ and then you say that variance of u_j equals to σ^2_j and that is called actually Heteroscedasticity, Heteroscedasticity, that is the meaning.

Now, the question is why does Heteroscedasticity arise? Why does Heteroscedasticity arise in the context of a data that you are going to estimate in using your econometric model? So, we will need to understand the reasons behind Heteroscedasticity.

(Refer Slide Time: 5:20)



So, reasons for Heteroscedasticity. We will try to understand the reasons for Heteroscedasticity using a simple diagram. Suppose, in the x axis I am measuring income and in the y axis I am measuring consumption and let us say this is the regression line that you have estimated.

Now, for any this is let us say your y_i and this is the x_i , now for given value of x_i let us say this is x_{1i} , what you will have this is let us say your \hat{y}_i , \hat{y}_i equals to $\hat{\alpha}$ plus $\hat{\beta}$ x_i this is the line you have fitted.

So, for any given value of x you will see some observations on y will lie for some individual that means using this line your regression line you are basically trying to predict the individual's consumption. So, for a given value of x_1 , given value of x at the level say x_1 what you will see that some individuals value will lie exactly on the line that means you could exactly predict their consumption while some individual will lie above the line, some individuals actual consumption will lie below the line, so that means you will have a range like this.

And we said what are these deviations? The deviations from your prediction is obviously error because your predicted line says the actual in predicted consumption should be like this, the model predicts the consumption here but some observed consumption is actually above the line, so this is basically the error and this is the spread of your error term.

Similarly, when you go to x_2 level of income where x_2 is actually greater than x_1 then what you see that error variance looks like this. Then again for x_3 level of income your error variance looks like this. That means what is happening let us say this is, they say this is u_1 , this is u_2 and this is u_3 , spread of u_1 , u_2 and u_3 . Now, what is the assumption we made?

The assumption what we made that at different level of x that means as income increases your error variance actually should not increase that is the assumption we made. But here what is happening as the individuals income increases from x_1 to x_2 , x_2 to x_3 then what is happening the error variance is actually increasing, higher the income higher is the error variance that is why this type of pattern explains actually Heteroscedasticity.

But the question is why this error variance is increasing instead of being constant at different level of income? The reason is very simple, see when the individual's income level at a very lower level then you can actually predict very well about the consumption pattern of the individual.

For example, let us say this x_1 is very basically the income of a day laborer whose income is let us say some 300 or 400 rupees per day and what that individual day laborer will do with the 300

and 500? The day laborer will probably buy some 1 kg of rice, some half a liter of oil, some 200 gram of dal, so on and so forth.

So, 400 rupees you can easily predict how the individual is going to spend. But as income increases let us say from 300 now this individual's income is 3000 per day. Now, with that 3000 it is becoming very difficult to predict what that individual will do using the 3000 rupees. Now, the question is why? Can you think of at higher level of income why people become more unpredictable that is why you are committing so much of mistake, error variance is having larger spread?

The reason is at a higher level of income people have higher degrees of freedom in their consumption pattern. So, that means with 3000 rupees or 90000 or 1 lakh income per month I can buy a music system, I can buy rice, I can buy dal, I can buy computer or probably I can buy a new vehicle so on and so forth.

So, at higher level of income the consumers enjoy a higher degree of freedom to select from different consumption bundle and as a result of which high income people become less predictive and that is why this much of mistake and this much of error variance. That is why when you are estimating this type of consumption function it would be highly unrealistic to assume that your error variance would be constant at different level of consumption, that is the reason.

So, that means here in this consumption income relationship we can say that at higher level of income, people become less predictable in their consumption because they enjoy higher degrees of freedom in their consumption. As a result of which variance of u_i given x equals to x^3 is much higher than variance of u_i given x equals to x^1 , that is the reason.

Alternatively, I can give another example to explain Heteroscedasticity, I can give another example, let us say this x axis I am measuring time and the y axis I am measuring error that you may commit in your typing. Suppose you are typing using a typewriter I am noting down how much errors you are committing while typing.

So, when you have just started your typing with very less amount of experience let us say at the first hour, you may commit so many mistakes, so that means your error variance would be like

this, but as time goes on this is t equals to t_1 , then as time increases what will happen? You will learn how to type better and your error variance actually will come down and when you have too much of experience then you will have very less error, so just opposite will happen in this case.

So, that means variance of u_i given t equals to t_1 is much larger than variance of u_i given t equals to t_2 and then variance of u_i equals to t_3 , whatever might be the case, whether it is increasing or decreasing does not matter, in this case we see an increasing pattern of the error variance, in this case we see a decreasing pattern of error variance, but whatever happens error variance is not constant that much we can say and these are the reasons for error variance not to be constant. That means there might be Heteroscedasticity problem.

(Refer Slide Time: 17:07)

The image shows a whiteboard with handwritten notes. The top section is titled "consequences of heteroskedasticity" and lists three points: (i) Unbiasedness, (ii) Consistency, and (iii) $\text{Var}(\hat{\beta})$ is no longer the min. The bottom section is titled "Detection of heteroskedasticity" and includes a sub-heading "graphical measure: \hat{u}_i ". It features three graphs: 1. "Homo.": A scatter plot of residuals \hat{u}_i versus \hat{y}_i showing a random, constant spread of points. 2. "Het": A graph showing a fan-shaped spread of residuals \hat{u}_i versus \hat{y}_i , indicating increasing variance. 3. "Het": A graph showing a downward-sloping spread of residuals \hat{u}_i versus \hat{y}_i , indicating decreasing variance. An NPTEL logo is in the top right corner of the whiteboard area.

Now, the next question is what are the consequences of the non-constant error variance? So, what will happen to unbiasedness? Unbiasedness is still maintained. What will happen to consistency? Consistency is still maintained.

But the problem happens even in the context of Heteroscedasticity also. Variance of β hat is no longer the minimum. The moment you change the assumption regarding the variance of u_i then what you observe that variance of β hat is no longer the minimum variance property is lost.

And as variance get disturbed again the standard error gets disturbed and again your t or any other statistic based on that standard error, the t statistic get disturbed and your hypothesis testing become problematic. That means you are getting a t statistic which is not the original t statistic rather that is disturbed because of the variance get disturbed in presence of Heteroscedasticity.

Now, the question is how will you detect Heteroscedasticity? So, there are different ways by which actually you can detect Heteroscedasticity. So, you can detect Heteroscedasticity by some graphical measures. The graphical measure says what you do after estimating your model you put your u_i hat square against your y_i hat and if you see some kind of pattern either this type or let us say this type, this type u_i hat square then this is y_i hat or let us say this type, some kind of down on, so both this condition says this is Heteroscedasticity.

But if you plot your u_i hat square like this then you see that your this type, so no pattern this is upward sloping pattern and but here there is no clear cut pattern that we can observe between u_i hat and y_i hat then we say that this is homoscedasticity. Now, the example you are thinking about income and consumption, we said that as income increases, error term increases.

That means we were, we were plotting the error variance against the explanatory variable, but here we are plotting error variance that means u_i hat square against y_i hat, what is the reason? Because this y_i hat basically captures the influence of all your explanatory variables, otherwise if you have multiple number of explanatory variables you have to plot this u_i hat square against each and every explanatory variable.

Suppose, you have 5 explanatory variable in the model and you are not sure which particular variable is showing Heteroscedasticity. So, what you have to do? You have to plot this u_i hat square against each and every explanatory variable, to solve that problem econometrician they suggest instead of plotting u_i hat square on individual explanatory variables. you plot u_i hat square against y_i hat.

Then you see whether you can observe some kind of pattern, either upward sloping or downward sloping, if there is some kind of pattern that is basically an evidence of Heteroscedasticity, if there is no such pattern emerges then we will say that the error variance is homoscedasticity.

But these are all simple graphical measure because sometimes from the pattern you may not even understand whether this is a pattern or not, so this is some kind of initial kind of impression that you can formulate out of your data but to detect Heteroscedasticity econometricians have developed different statistical measures as well and we need to apply some of those measures to detect Heteroscedasticity.