

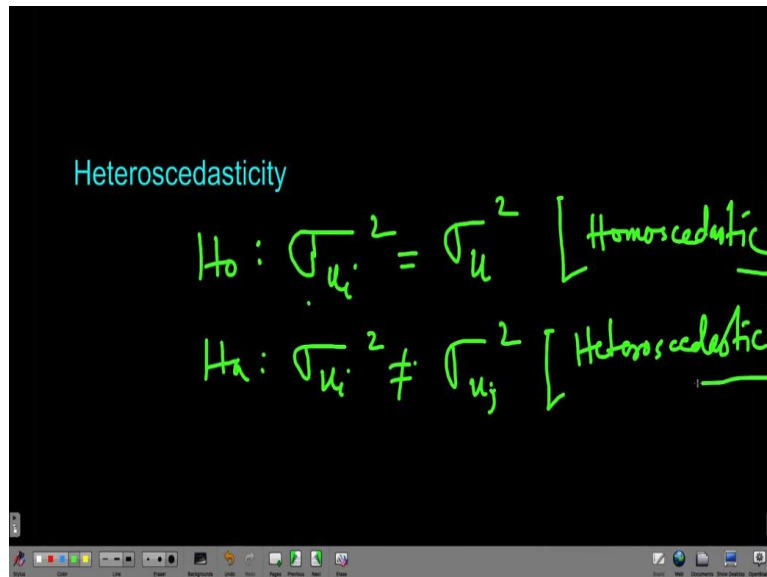
**Applied Econometrics**  
**Prof. Tutan Ahmed**  
**Vinod Gupta School of Management**  
**Indian Institute of Technology - Kharagpur**

**Module - 8**  
**Lecture - 64**  
**Heteroscedasticity**

Hello and welcome back to the lecture on Applied Econometrics. Today we are going to talk about heteroscedasticity. Now, when we talked about the Gauss-Markov assumptions, you remember that we wanted the error term to be homoscedastic. In other words, if the error term is heteroskedastic, so, we will say that the Gauss-Markov assumption is violated. Now, in this lecture, we are going to learn what exactly is heteroscedasticity.

Of course, we talked about it briefly, but we are going to elaborate it here, and we are going to see, what are the implications of heteroscedasticity? Wherefrom from it come? Or if you have heteroscedasticity, what are the symptoms? And how do you actually address the problem of heteroscedasticity? So, first let us talk about what exactly is heteroscedasticity.

**(Refer Slide Time: 01:13)**



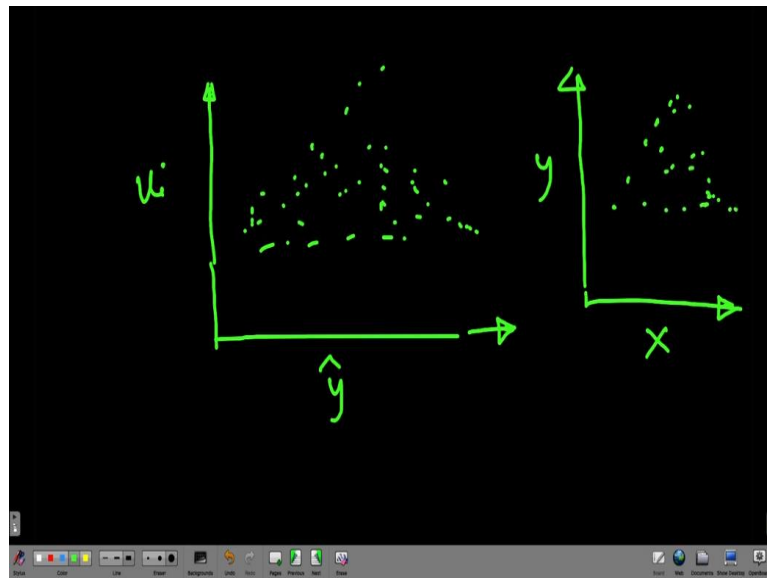
Heteroscedasticity

$$H_0: \sigma_{u_i}^2 = \sigma_u^2 \quad \text{[Homoscedastic]}$$
$$H_1: \sigma_{u_i}^2 \neq \sigma_{u_j}^2 \quad \text{[Heteroscedastic]}$$

So, now, we remember that, we have actually written that the variance of the error term has to be constant; if we want the homoscedastic condition to be satisfied, then this property has to be there. That is, the variance of the error term is constant. Now, we can say that the null hypothesis is that the variance of the error term is constant. And this is simply, we can call it homoscedastic; spelling is always a little tricky.

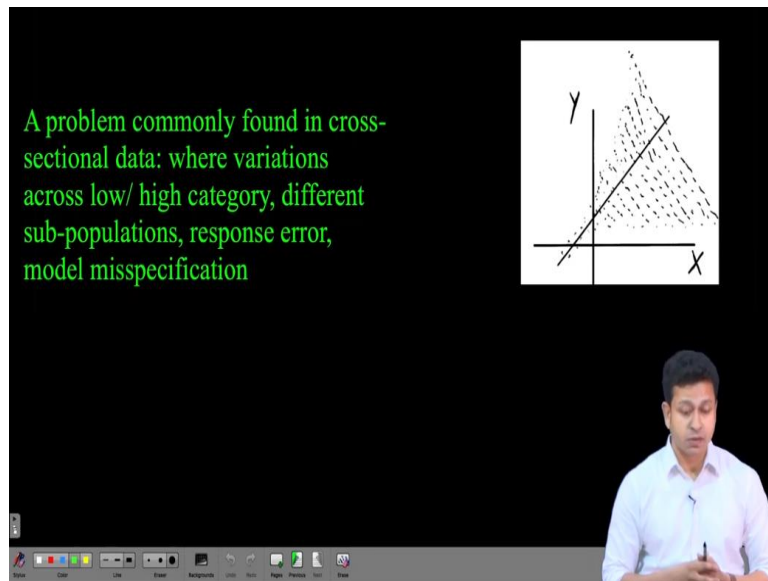
Here we have the error term; irrespective of whatever value of whatever different observation of the error term you take, the variance due to the error term is constant; whereas the alternative hypothesis is that the variance due to the error term is not constant. So, this is not equal to this. So, if you take different observations, you basically obtain the error terms. And if you calculate the variance of the error terms, so, what you will get is that, you will get different variance in case your error term is heteroskedastic. So, this is what is called heteroskedastic.

**(Refer Slide Time: 02:42)**



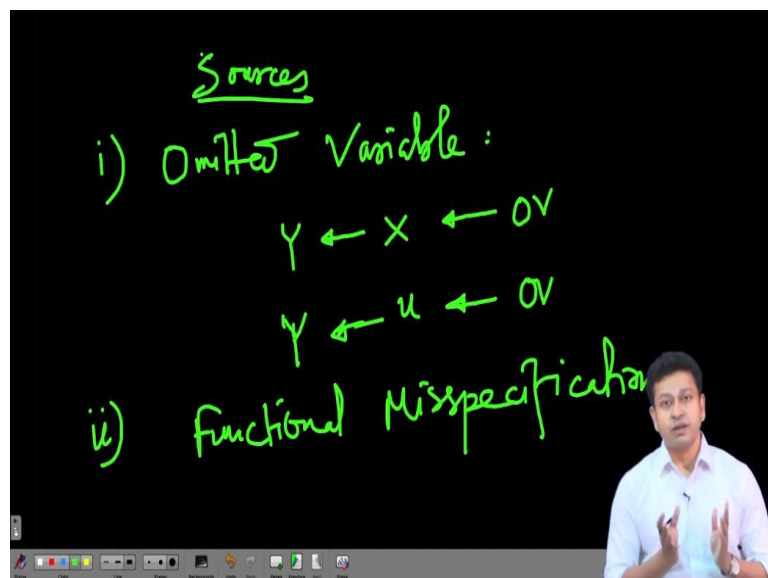
Now, if you want to pictorially show it, how it will look like is, we will have; you can plot, let us say the error term and let us say here you plot the fitted  $y$ . And you will see the error term is going to look like this, something like this. So, this is the condition of heteroscedasticity.

**(Refer Slide Time: 03:21)**



So, usually, we see heteroscedasticity when we deal with cross sectional data, when we have variations across different categories like low, high category, different subpopulations. It may also happen due to the response error, due to model misspecification, so forth. So, we will actually see where from this variance in the error term is coming. So, we said that the sigma square due to the error term is not constant if we have heteroscedasticity. So, now, we need to know why sigma square is not constant when I have heteroscedasticity. So, let us see the sources of heteroscedasticity.

**(Refer Slide Time: 04:01)**



So, there could be different sources of heteroscedasticity. The first one thing could be like the omitted variable bias. If you have omitted variable, suppose one variable is very important in the regression equation and that is influencing the Y variable; now, suppose you have

somehow missed to include that variable in your model; and what will happen, the variable, the influence is there.

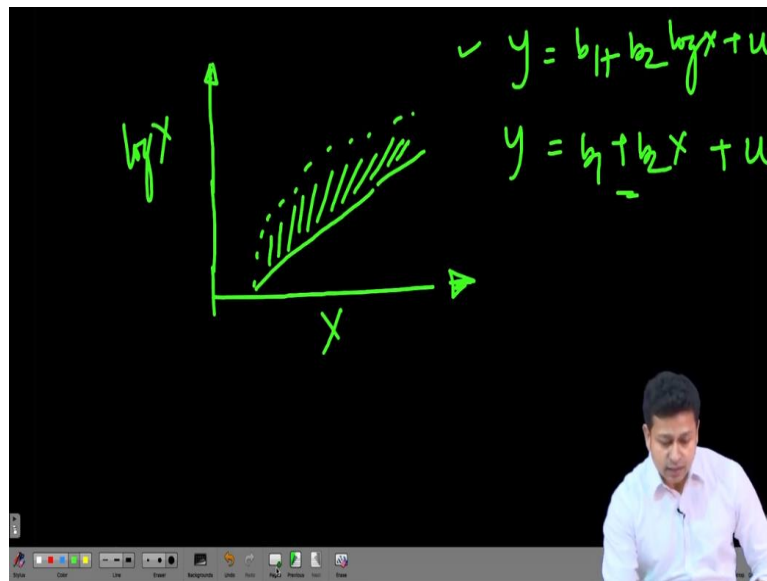
Now, the influence is there and that influence is captured by the model in one way or the other. Now, how it will be captured? So, one could be like, it could be like other independent variables may capture a part of the, impact that omitted variable. So, let us say this is the omitted variable OV. And if I see, OV is going to influence the X variables which are the part of the model. And that in turn is going to influence the Y variable.

So, partly, the impact of omitted variable could be captured by the X variables which are already present in the model. But it may also happen, if the X variables which are present in the model are not correlated with the omitted variable, the variable which we basically missed out; so, what will happen is that, the omitted variable will actually end up influencing the Y variable via the error term. So, it will influence the Y variable via the error term.

Now, this omitted variable might have some sort of trend of course, some sort of relationship with the Y variable it will have. And then, what will happen is that, since you have not captured the impact of the omitted variable in your model, it will show the trend of that omitted variable through the error term. So, basically, a pattern will be visible in the error term. So, that is precisely what you have seen. There is a pattern in the error term.

So, there is a pattern like, it is an increasing pattern in the error term. Now, the other thing, the other way that it can influence, we can actually end up having something called heteroskedasticity is say functional misspecification. And what do I mean by that? So, you remember that when we actually talked about different types of functional specification, so, you can actually have plot log linear, log log, like quadratic, polynomial; multiple ways, you can actually plot the relationship between Y and X. Now, when you do that, definitely, the way the curve will vary will change.

**(Refer Slide Time: 07:07)**

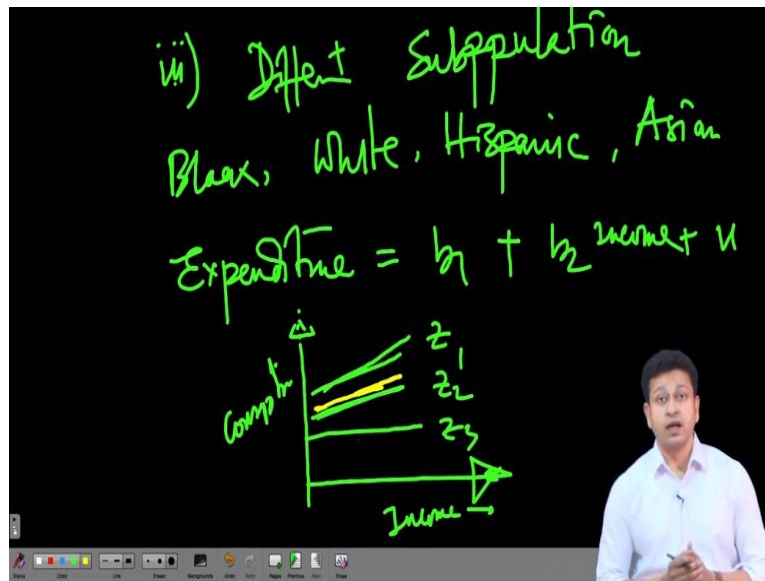


So, for example, let us say we want to plot a  $Y$  and  $X$  where my  $Y$ ; essentially, where is the  $\log X$ ; but what we have plotted instead is basically  $X$ . So, if I just plot  $\log X$  and  $X$ , so, I will get a graph, something like this. Now, you ended up plotting something like a linear equation where you have just; so, in the first case, let us say you have your  $Y$  is basically  $\beta_1$  plus  $\beta_2 \log X$  plus  $u$ ; whereas, in the second case, your  $Y$  is equal to  $\beta_1$  or  $\beta_1 + \beta_2 X$  plus  $u$ .

So, let us say the true equation is this one, whereas you have actually plotted this one. So, what will happen? This part; the remaining part will not be captured in your model. So, where will it go? So, essentially, the error term will basically show this remaining part. So, the moment error term actually captures this remaining part, what is going to happen is that, when you plot the error term, you are actually going to see this pattern which we actually left out because of your functional misspecification.

So, this is another reason why we can see the kind of heteroscedasticity. Now, the other reasons; there could be other reasons as well. And by the way, one point I should tell you here is that, when I say that you actually plot a  $u_i$  with a  $\hat{Y}$ ; so, in this kind of distribution, if you also plot, let us say  $Y$  and  $X$ , or the same sort of the data set, you will see, it will have the same type of dispersion. We will come back to this.

**(Refer Slide Time: 09:08)**



So, the third point I was talking about is that we can have different subpopulation. Now, different subpopulation might have different characteristics. And these different characteristics may actually be reflected when you actually take larger number of observations, when you take multiple types of subpopulations, and so forth. So, let us say we have different subpopulations like, say Black, White, Hispanic, Asian.

Now, I am interested to actually plot a regression where I have my expenditure as a dependent variable and my income is the independent variable; let us say  $\beta_1$  plus  $\beta_2$  income plus some error term. Now, in this case, what will happen is that; different group might have different tendency of expenditure. They have different propensity to consume if the income is changing.

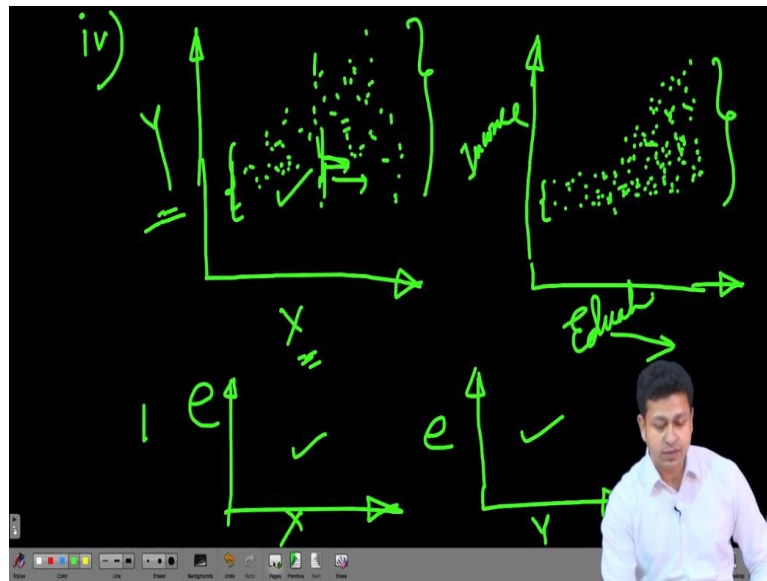
So, what we are going to see is that, for different groups, we are going to see different kind of regression lines. So, if we draw it, it will be something like; let us say this is for one category, this is for another category and this is for another category. So, for  $Z_3$  with change in income, it does not change in consumption as such; whereas the  $Z_2$  category actually has a slightly more consumption as compared to  $Z_3$ ; and  $Z_1$  has even more consumption than  $Z_2$ .

The graph could have been a little better; I could have actually made a little more steep. And this is the income. So, when I have this different regression lines for different categories, so, I essentially have heteroscedasticity present in my model; because, if I actually draw a regression line for all these 3 categories together, maybe I will get something like this. So, as the observation size is increasing, as I am trying to get more and more differences among

these different groups, we will see that the error term is also kind of capturing the differences among these different groups.

So, when we have these different subpopulations, it is likely that we will end up with this kind of heteroskedasticity. Here I am basically plotting X and Y; so, the moment my X and Y is actually diverging, I am going to see the presence of heteroskedasticity.

(Refer Slide Time: 12:43)



Now, the last point that I am going to talk about, the reason of heteroskedasticity is that, if Y and X has an increasing sort of relationship and what will happen is that, the error term that is left out after you fit the new model, it is going to be also quite high. So, when it is going to be quite high, what will happen is that, when you plot the error term, it is also going to give you that sort of pattern, like as with the increase with the value of the X variable, you are actually going to see an increasing value of the error term.

So, it is something like if my Y and X has this sort of relationship, increasing; or it could be decreasing also; I am just coming to that relationship; so, we are going to see the presence of heteroskedasticity. And here, my X could be, there could be many examples of it. A good example could be like, let us say the Y is my consumption decisions for travelling, so, like holidaying, and X is my income. Now, what will happen?

Lot of people, they might not have a lot of income. So, they might really not think about going abroad or even travelling within the country; but people, when they have certain amount of income, they cross this; let us say there is a threshold of it. When they cross this

certain level of income, what will happen is that, they will get this choice. So, they can choose whether to travel like neighbouring countries or weather to travel a different continent; and even in the different continent, how lavishly they want to travel and so forth.

So, as the income increases, what will happen is the dispersion or the different choices of consumption will actually also increase. And that is the reason the dispersion here is going to be pretty high as compared to the dispersion here. Similarly, you can think about the education and income. So, note here, I am basically plotting Y and X here. So, instead of Y and X, if I plot, say error term in this axis and X in this axis, we are going to see something similar.

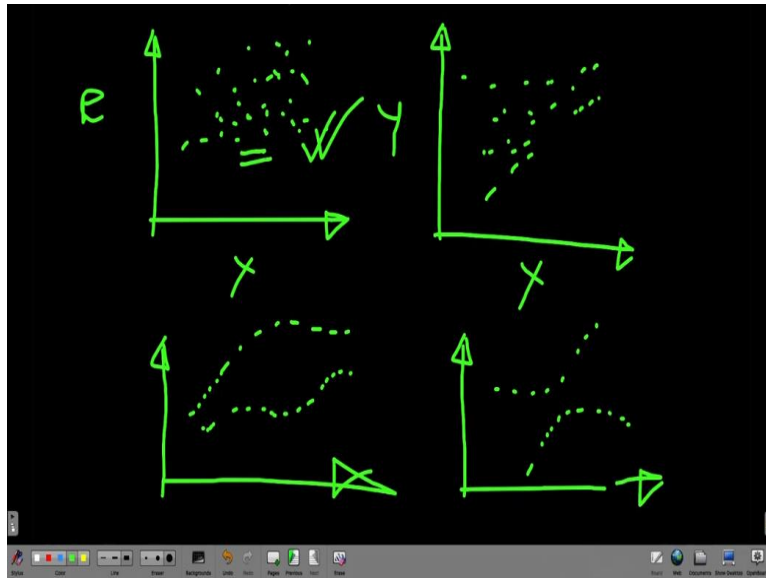
Or if we are going to plot error term here on the Y axis and fitted Y on the X axis, we are going to see the similar pattern. We just remember that, because the pattern is going to be represented in all these different diagrams that you use. Now, the other example I was giving is that, let us say income and education. So, if you see that the income level of people with lower education; normally, people, if they are below certain level of education, the income level is not going to change very much.

But as the education level increases, we will see, the income is going to be little, the dispersion is going to be higher. For example, take someone who has already done graduation. So, once someone has done graduation, he has a much higher scope of earning; or masters, he has much higher scope of earning; or once you are done with graduation or masters, you have the option to like do a MBA, to do different courses, vocational courses that will actually entitle you for a higher income; or a PhD; as you increase your education, your earning potential will also increase.

So, as you increase the education, the dispersion of your income is also going to increase. And that is why we will see the heteroskedasticity to be present when you have the value of X higher. Now, so far, we have shown the different reasons of heteroskedasticity, and we probably have got an idea that how heteroskedasticity looks like.

**(Refer Slide Time: 17:10)**





So, we probably have found that heteroscedasticity means; in all the examples, it is looking like something like this; it is a diverging nature of the error term, let us say. So, this is my error term, or I can plot error term or Y or whatever, depending on what you; let us say we just plot error term, so that we do not confuse it with X. Now, will we always have the heteroskedasticity looking like this? Not necessarily.

The point is, when I say it is heteroskedastic, we have the error term. So, that can also be something like this. So, it may actually converge, going forward. It is also a possibility that error terms will converge. So, that can also be error term or Y is converging with X. So, that is also a case of heteroskedasticity, because the variance is not constant. There could be many possibilities.

So, you can have something like this, let us say high error, and then again diverging, and then again converging. This is also a case of, of course, heteroskedasticity. Or it could be like this. So, there could be numerous cases of heteroscedasticity. So, you can just think about it. Now, to kind of sum it up, is that the heteroskedasticity essentially means the non-constant variance of the error term, and it can have different kind of shapes, but usually, we will see this kind of relationship, because, usually what happens is, Y and X has an increasing relationship; not necessarily all the time, but usually.

And when Y and X has an increasing relationship, we are also going to see this kind of distribution of the error term. And this is essentially what we call heteroskedasticity. So, in this lecture, we have understood what is heteroskedasticity and what are the different sources

of heteroskedasticity, where it comes from. And with this, we will end this lecture. And in the next lecture, we are actually going to talk about how to actually first understand that whether heteroscedasticity is there in the distribution or not.

And there, it will come like, we will actually see the importance of actually graph plot, we are actually going to plot the data to understand if there is heteroskedasticity in the model or not. So, with this, we will end the lecture here. Thank you.