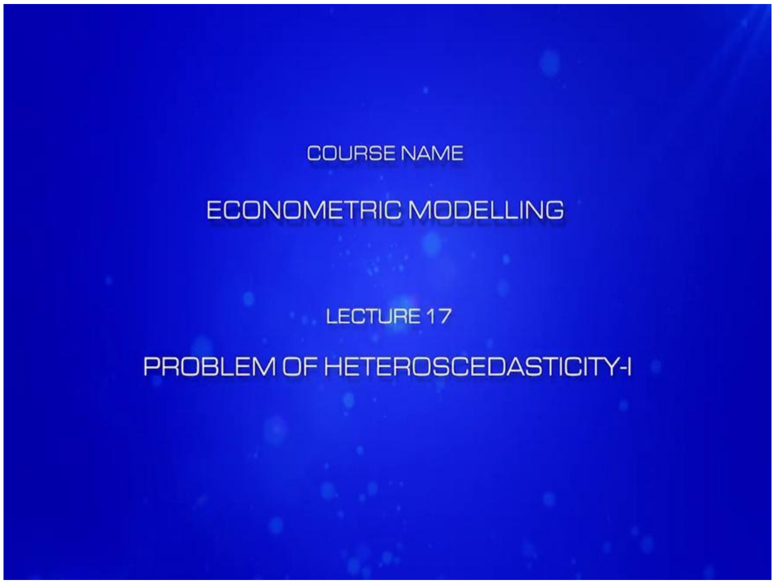
Econometric Modelling
Professor Sujata Kar
Department of Management Studies
Indian Institute of Technology, Roorkee
Lecture 17
Problem of Heteroscedasticity - I

(Refer Slide Time: 00:11)



COURSE NAME

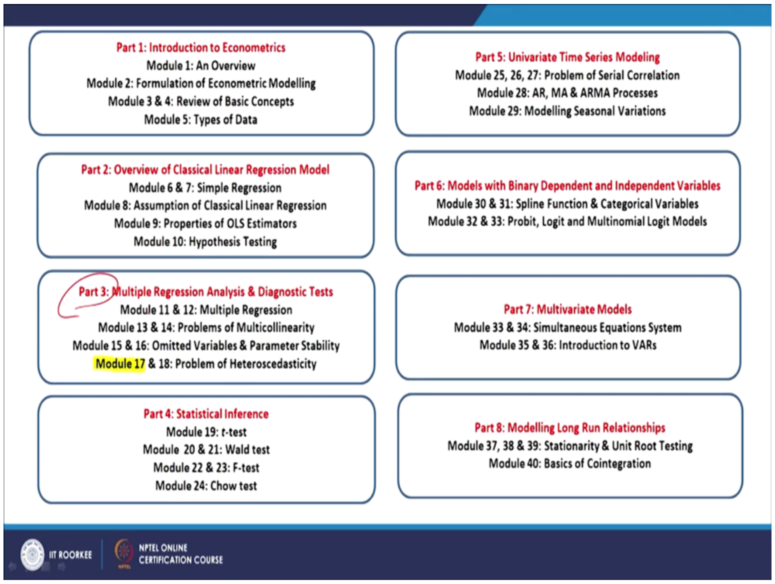
ECONOMETRIC MODELLING

LECTURE 17

PROBLEM OF HETEROSCEDASTICITY-I

Hello and welcome back to the course on Econometric Modelling, this is Module 17.

(Refer Slide Time: 00:33)



Part 1: Introduction to Econometrics
Module 1: An Overview
Module 2: Formulation of Econometric Modelling
Module 3 & 4: Review of Basic Concepts
Module 5: Types of Data

Part 2: Overview of Classical Linear Regression Model
Module 6 & 7: Simple Regression
Module 8: Assumption of Classical Linear Regression
Module 9: Properties of OLS Estimators
Module 10: Hypothesis Testing

Part 3: Multiple Regression Analysis & Diagnostic Tests
Module 11 & 12: Multiple Regression
Module 13 & 14: Problems of Multicollinearity
Module 15 & 16: Omitted Variables & Parameter Stability
Module 17 & 18: Problem of Heteroscedasticity

Part 4: Statistical Inference
Module 19: t-test
Module 20 & 21: Wald test
Module 22 & 23: F-test
Module 24: Chow test

Part 5: Univariate Time Series Modeling
Module 25, 26, 27: Problem of Serial Correlation
Module 28: AR, MA & ARMA Processes
Module 29: Modelling Seasonal Variations

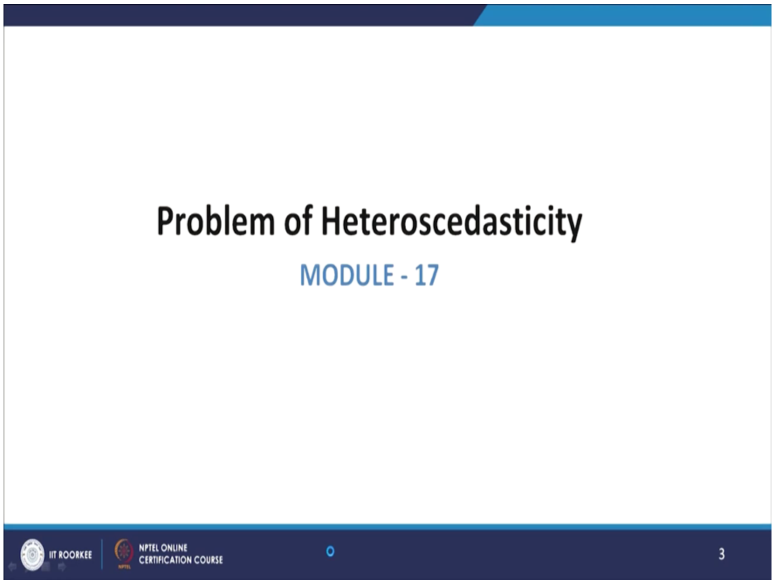
Part 6: Models with Binary Dependent and Independent Variables
Module 30 & 31: Spline Function & Categorical Variables
Module 32 & 33: Probit, Logit and Multinomial Logit Models

Part 7: Multivariate Models
Module 33 & 34: Simultaneous Equations System
Module 35 & 36: Introduction to VARs

Part 8: Modelling Long Run Relationships
Module 37, 38 & 39: Stationarity & Unit Root Testing
Module 40: Basics of Cointegration

Module 17 is part of Part 3 that deals with multiple regression analysis and several diagnostic tests. So, once we are through with the discussion of multiple regression analysis and after that, we discussed the problem of multicollinearity which is a problem outside the CLRM assumptions that it does not as such violate any CLRM assumptions. Then we discussed omitted variable and measurement error problems, which can lead to unbiased and inconsistent estimators.

(Refer Slide Time: 01:02)

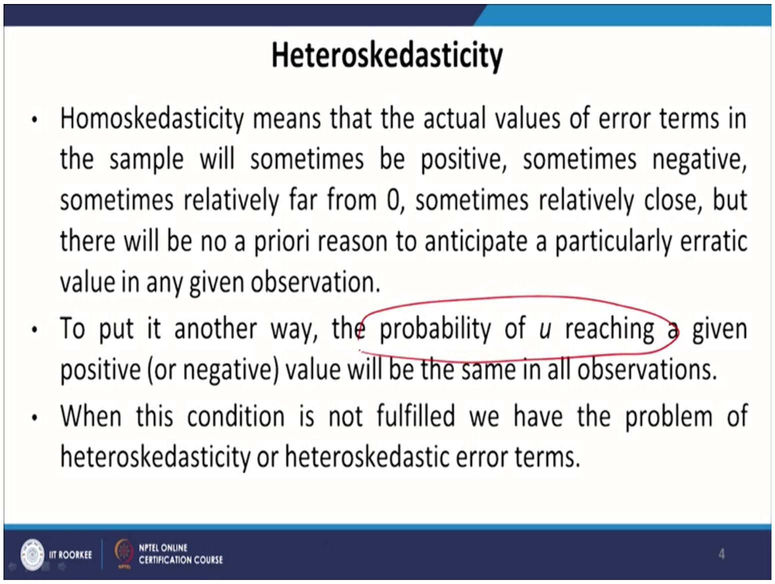


Problem of Heteroscedasticity

MODULE - 17

In the next two modules, we will be discussing the problem of Heteroscedasticity. So, module 17 first introduces the problem of Heteroscedasticity, how we define it and what does its presence means, what kind of problem it actually leads to, and also, we talk about the detection of Heteroscedasticity that is how we can test for the presence of Heteroscedasticity. And in the next module, we will talk about how to deal with Heteroscedasticity.

(Refer Slide Time: 01:32)



Heteroskedasticity

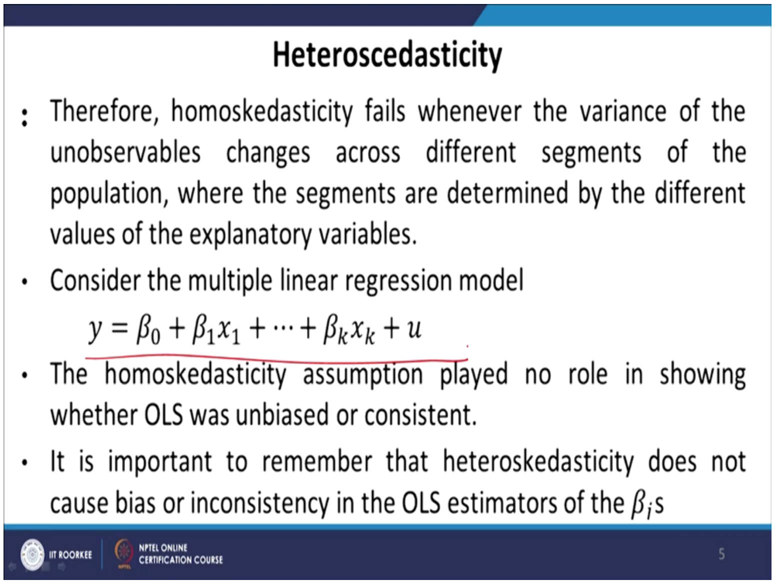
- Homoskedasticity means that the actual values of error terms in the sample will sometimes be positive, sometimes negative, sometimes relatively far from 0, sometimes relatively close, but there will be no a priori reason to anticipate a particularly erratic value in any given observation.
- To put it another way, the probability of *u* reaching a given positive (or negative) value will be the same in all observations.
- When this condition is not fulfilled we have the problem of heteroskedasticity or heteroskedastic error terms.

So, first, we have already introduced Homoscedasticity. So, homoscedasticity means that the actual values of the error terms in the sample will sometimes be positive, sometimes negative, sometimes relatively far from 0, sometimes relatively close, but there will be no a priori reason to anticipate a particular erratic value in any given observation. So, since, we are making assumptions about the error variance.

So, what does this error variance imply, that is, first of all, is explained here that the deviations from the error term from its mean value could always be there, but then if there is no systematic pattern in it, and we do not have any reason to anticipate a particularly erratic or we anticipate a particularly erratic value in any given observation and we do not anticipate any systematic pattern, then we have heteroscedastic errors.

To put it another way, the probability of u the population error reaching a given positive or negative value will be the same in all observations. When this condition is not fulfilled, we have the problem of Heteroscedasticity or heteroscedastic error terms. So, we are basically focused here on the probability of u reaching a positive or negative value that is u taking up a positive or negative value should be actually constant and there is no systematic pattern.

(Refer Slide Time: 03:06)



Heteroscedasticity

- Therefore, homoskedasticity fails whenever the variance of the unobservables changes across different segments of the population, where the segments are determined by the different values of the explanatory variables.
- Consider the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

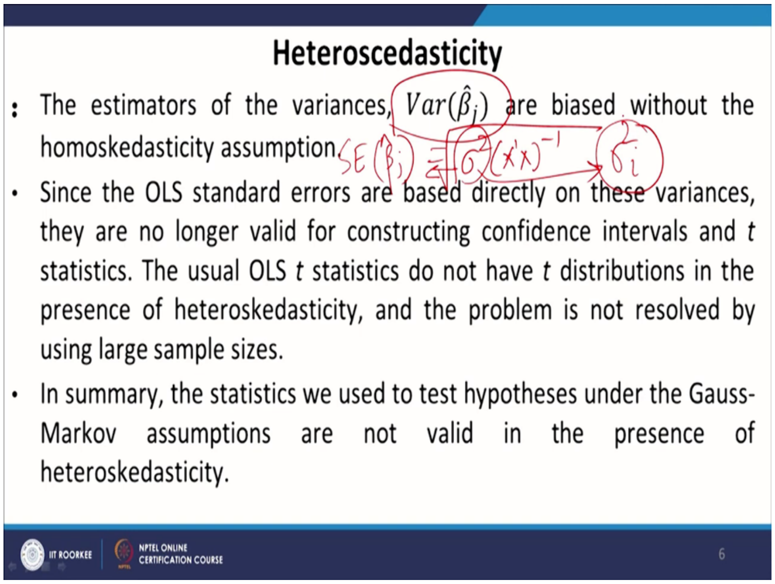
- The homoskedasticity assumption played no role in showing whether OLS was unbiased or consistent.
- It is important to remember that heteroskedasticity does not cause bias or inconsistency in the OLS estimators of the β_js

Therefore, homoscedasticity fails whenever the variance of the unobservables changes across different segments of the population, where the segments are determined by the different values of the explanatory variables. So, that is one possibility that we have different segments we can depending on the independent variable, we have different segments of the dependent variable and for different segments, we have different error variations.

So, then we can associate some non-constant probability in the dispersion of the population error from its mean value. Now, consider the multiple linear regression model the standard multiple linear regression model where we have k plus 1 variables, the homoskedasticity assumption played no role in showing whether OLS was unbiased or consistent.

It is important to remember that heteroscedasticity does not cause bias or inconsistency in the OLS estimators of the $\beta_j s$. So, unbiasedness and consistency property remains even when the errors are not homoscedastic or the errors are heteroscedastic.

(Refer Slide Time: 04:21)



Heteroscedasticity

- The estimators of the variances, $Var(\hat{\beta}_j)$ are biased without the homoskedasticity assumption. $SE(\hat{\beta}_j) = \sqrt{\sigma^2 (X'X)^{-1}} \rightarrow \sigma_i^2$
- Since the OLS standard errors are based directly on these variances, they are no longer valid for constructing confidence intervals and *t* statistics. The usual OLS *t* statistics do not have *t* distributions in the presence of heteroskedasticity, and the problem is not resolved by using large sample sizes.
- In summary, the statistics we used to test hypotheses under the Gauss-Markov assumptions are not valid in the presence of heteroskedasticity.

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 6

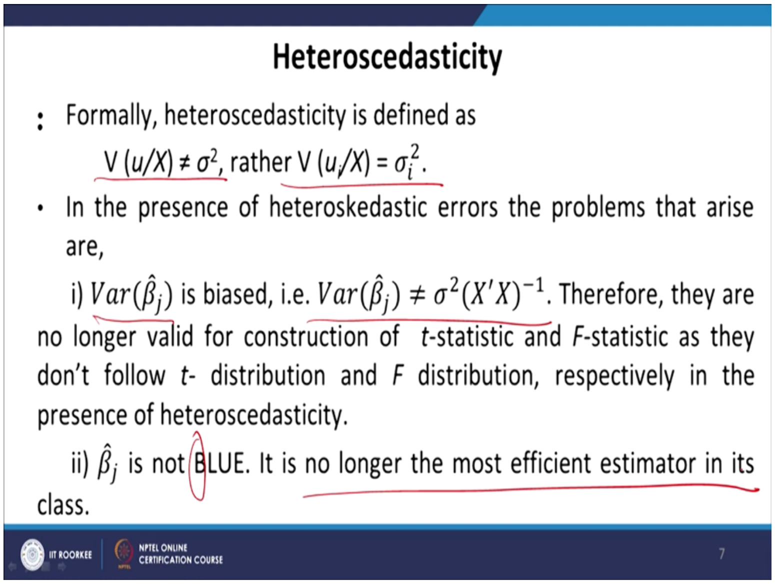
The estimators of the variances, variance of $\hat{\beta}_j$ are biased without the homoscedasticity assumption. So, the problem is that if we do not have homoscedasticity then the estimators of the variances that is the variance of $\hat{\beta}_j$, which we actually measure as (refer slide time:4:40) to recap that is these estimators are actually biased if the errors are heteroscedastic.

Since the OLS standard errors are based directly on these variances, they are no longer valid for the construction of confidence intervals and t statistics. So, this is the error variance. Now, when we have heteroscedasticity this error variance is no more sigma square, it is something else we can call it sigma $\sigma^2 i$, and it is varying from observations to observations. And the standard error of the estimated parameters that is the standard error of $\hat{\beta}_j$, is basically root over the variance of $\hat{\beta}_j$.

Now, what is happening is that, since we do not have homoscedastic errors or we have heteroscedastic errors. So, that is why the standard errors are no longer valid for constructing confidence intervals and t statistics, they are not valid for other statistics also. If we talk about f statistics or F- tests conducted in order to test for joint significance of several hypotheses, we have not talked about f test at length as of now, that is why I would not talk much about F-test right here.

In summary, the statistics we use to test hypothesis under the Gauss Markov assumptions are not valid in the presence of Heteroscedasticity. So, the Gauss Markov assumptions that we consider there were 5 assumptions, one of them was the assumption of homoscedasticity. So, in presence of Heteroscedasticity, that assumption is not fulfilled.

(Refer Slide Time: 06:27)



Heteroscedasticity

- Formally, heteroscedasticity is defined as

  $V(u/X) \neq \sigma^2$, rather $V(u_i/X) = \sigma_i^2$.

- In the presence of heteroskedastic errors the problems that arise are,

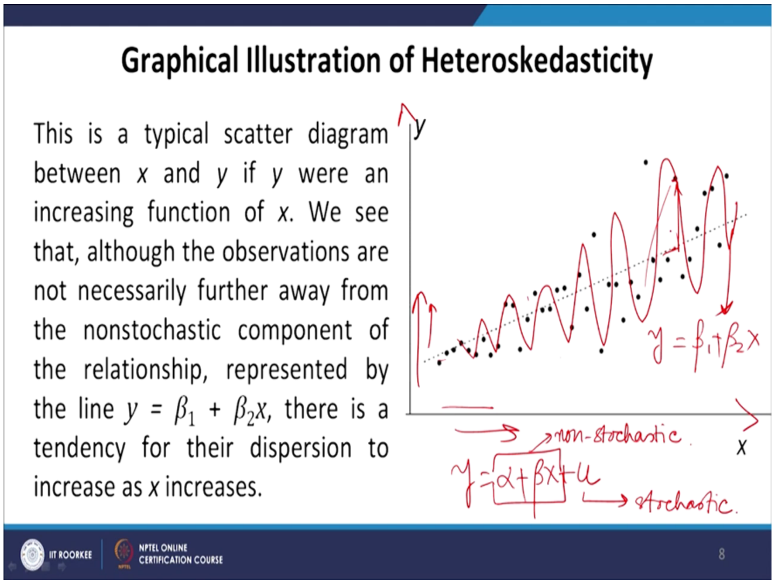
  i) $Var(\hat{\beta}_j)$ is biased, i.e. $Var(\hat{\beta}_j) \neq \sigma^2(X'X)^{-1}$. Therefore, they are no longer valid for construction of *t*-statistic and *F*-statistic as they don't follow *t*- distribution and *F* distribution, respectively in the presence of heteroscedasticity.

  ii) $\hat{\beta}_j$ is not BLUE. It is no longer the most efficient estimator in its class.

Formally, Heteroscedasticity is defined as (refer slide time: 6:30). In the presence of heteroscedastic errors, the problems that arise are first Var $\hat{\beta}_j$, is biased, which I have just mentioned, that is (refer slide time: 6:55). Therefore, they are no longer valid for the construction of t-statistic and F-statistic as they do not follow t distribution and F distribution anymore respectively in the presence of Heteroscedasticity.

And the second thing is that the estimated parameter that is $\hat{\beta}_j$, is not BLUE which is the best linear unbiased estimator. So, they could be still linear, they are still unbiased, but they are not best that is they are not the most efficient estimators anymore. So, it is no longer the most efficient estimator in its class. So, the property of OLS that is OLS estimators are BLUE is violated in the presence of Heteroscedasticity.

(Refer Slide Time: 07:43)



Graphical Illustration of Heteroskedasticity

This is a typical scatter diagram between x and y if y were an increasing function of x. We see that, although the observations are not necessarily further away from the nonstochastic component of the relationship, represented by the line $y = \beta_1 + \beta_2 x$, there is a tendency for their dispersion to increase as x increases.

y
x
$y = \beta_1 + \beta_2 x$
non-stochastic
$y = \boxed{\alpha + \beta x} + u$
stochastic
IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
8

Now, we graphically illustrate Heteroscedasticity. So, the presence of Heteroscedasticity can be also observed or detected through the graphical presentation, we present here a typical scatter diagram between x and y. So, we have y is measured on the vertical axis, x is measured on the horizontal axis, and if y were an increasing function of x. So, given the line plotted between x and y, it shows that y is an increasing function of x which implies that as x increases y also increases.

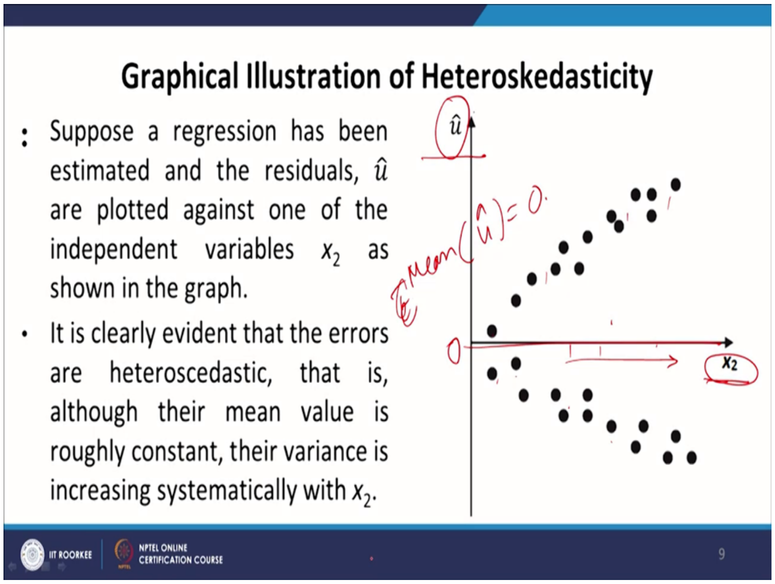
We see that although the observations are not necessarily farther away from the non-stochastic component of the relationship represented by the line (refer slide time: 8:28) there is a tendency for their dispersion to increase as x increases. So, this is my line, which is the non-stochastic component.

If you remember, initially I had mentioned that (refer slide time: 8:47) can be divided or decomposed into 2 components, this component is called the stochastic component that is the random term and this is a non-stochastic component, or essentially what we are trying to do is that through all these estimation processes, we are trying to estimate this non-stochastic component.

So, this is my non-stochastic component, which is (refer slide time 9:19), whatever you call it, and what we observe the presence of Heteroscedasticity, we can see that as x increases y also increases, but the dispersions are increasing. So, if I draw lines measuring the dispersions then you can see that the dispersions are increasing specifically, they are higher during this range, which may indicate that at least the variance is not constant.

The error variances or the differences between the plotted line and the actual value, so that is the error term, their differences or their deviations from the mean is actually not constant over time or may not be constant over time or the probability of the deviation from the mean value is not same for all observations. So, this is how we can observe Heteroscedasticity.

(Refer Slide Time: 10:19)



Graphical Illustration of Heteroskedasticity

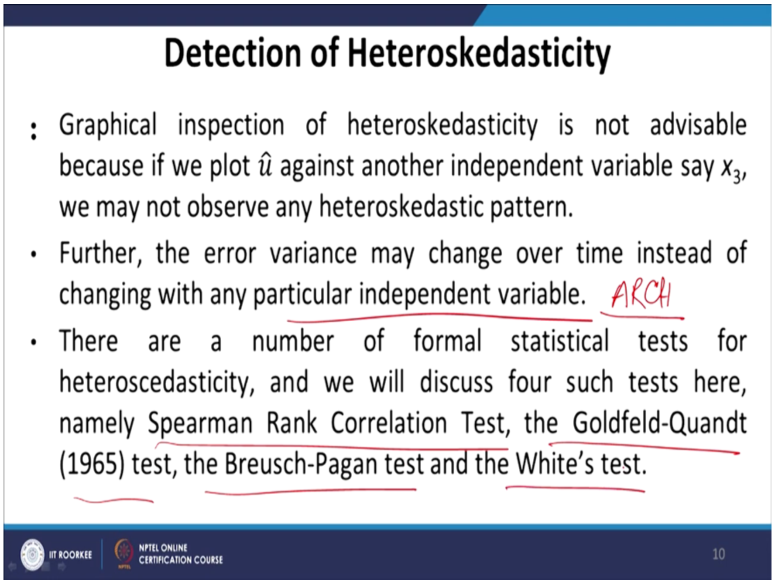
- Suppose a regression has been estimated and the residuals, $\hat{u}$ are plotted against one of the independent variables x_2 as shown in the graph.
- It is clearly evident that the errors are heteroscedastic, that is, although their mean value is roughly constant, their variance is increasing systematically with x_2.

Another illustration is like when we are plotting the estimated residuals that is the sample counterpart of the population error terms, we have residuals plotted against one of the independent variables say x_2. In case we have a multiple regression analysis and we have more than one independent variable. So, x_2 is one independent variable and $\hat{u}$ is plotted against them.

So, suppose a regression has been estimated and the residuals $\hat{u}$ plotted against one of the independent variables x_2 as shown in the graph, it is clearly evident that the errors are heteroscedastic that is, although their mean value is roughly constant, their variance is increasing systematically with x_2. Suppose this is the mean if this is 0, then you can understand that the positive value and the negative value roughly cancel out with each other, and because of which the mean of this residual term $\hat{u}$ is equal to 0.

So, if this is the mean have $\hat{u}$ the deviation of $\hat{u}$ from its mean value is systematically increasing as the value of x_2 increases. So, this is a very clear-cut pattern of the presence of Heteroscedasticity. But graphical inspection of Heteroscedasticity is not advisable.

(Refer Slide Time: 11:43)



Detection of Heteroskedasticity

- Graphical inspection of heteroskedasticity is not advisable because if we plot $\hat{u}$ against another independent variable say x_3, we may not observe any heteroskedastic pattern.
- Further, the error variance may change over time instead of changing with any particular independent variable. ARCH
- There are a number of formal statistical tests for heteroscedasticity, and we will discuss four such tests here, namely Spearman Rank Correlation Test, the Goldfeld-Quandt (1965) test, the Breusch-Pagan test and the White's test.

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 10

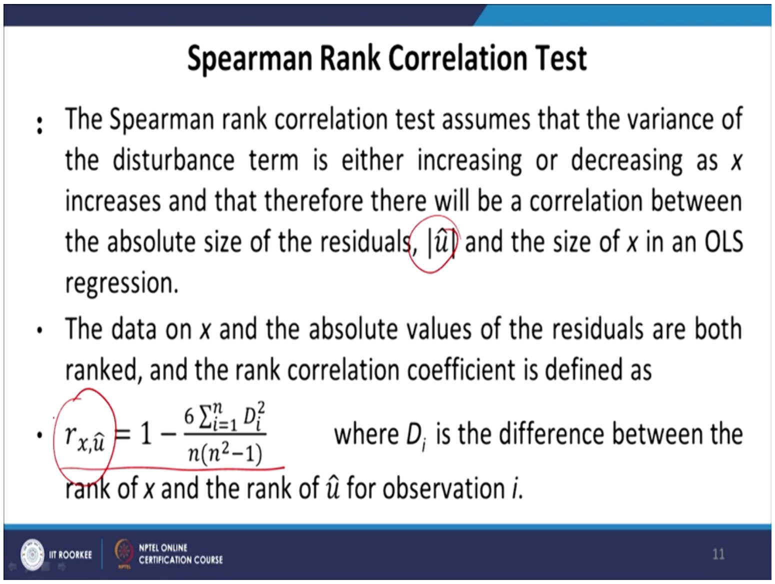
So, when we try to detect the presence of Heteroscedasticity graphical inspection is not advisable, because if we plot $\hat{u}$ against another independent variable, say x_3 we may not observe any heteroscedastic pattern.

So, we have to individually plot the residuals against each and every independent variable and then some of may show up some kind of a typical pattern in the movements or in the dispersion of the error term or the residuals terms from its mean. Further, the error variance may change over time, instead of changing with any particular independent variable.

So, with any independent variable, we may not observe any typical pattern, but the error variances are increasing over time. So, this is a special kind of case, which is often handled in time series analysis specifically, but also applicable in the context of cross-sectional data. And they are called autoregressive conditional Heteroscedasticity. And these kinds of models we will be dealing with in later modules.

So, there are several formal statistical tests for Heteroscedasticity. And we will discuss 4 such tests here, namely, the Spearman Rank Correlation Test, The Goldfeld-Quandt Test, which was suggested by Goldfeld in 1965, The Breusch-Pagan Test, and the Whites Test, these are the 4 alternative tests, which we are going to discuss here.

(Refer Slide Time: 13:09)



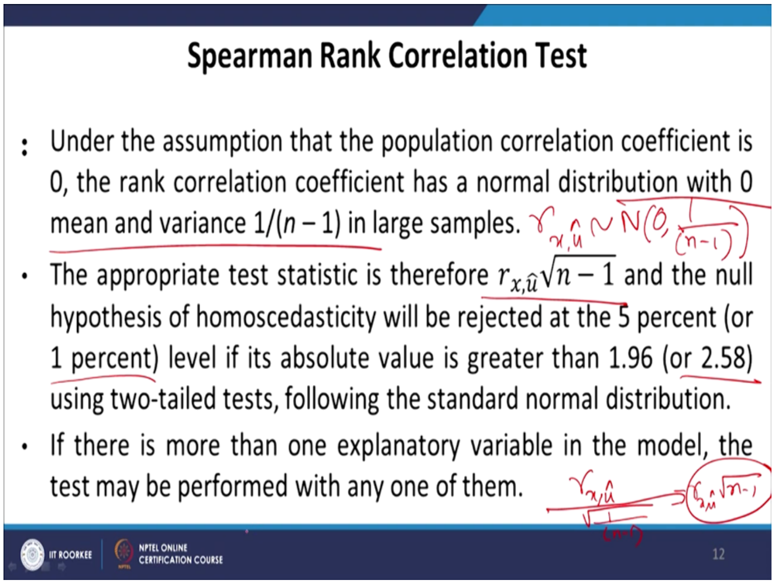
Spearman Rank Correlation Test

- The Spearman rank correlation test assumes that the variance of the disturbance term is either increasing or decreasing as x increases and that therefore there will be a correlation between the absolute size of the residuals, $|\hat{u}|$ and the size of x in an OLS regression.
- The data on x and the absolute values of the residuals are both ranked, and the rank correlation coefficient is defined as
- $r_{x,\hat{u}} = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$ where D_i is the difference between the rank of x and the rank of $\hat{u}$ for observation i.

So, we begin with the Spearman rank correlation test. The Spearman rank correlation test assumes that the variance of the disturbance term is either increasing or decreasing as x increases. So, there is a clear-cut pattern either increasing or decreasing between x and the residuals. And therefore, there will be a correlation between the absolute size of the residuals that is absolute value or mod value of $\hat{u}$, and the size of x in an OLS regression.

The data on x and the absolute values of the residuals are both ranked. And the rank correlation coefficient is calculated like this. So, this is denoted by (refer slide time: 13:51). I will explain the procedure using an example.

(Refer Slide Time: 14:09)



Spearman Rank Correlation Test

- Under the assumption that the population correlation coefficient is 0, the rank correlation coefficient has a normal distribution with 0 mean and variance 1/(*n* – 1) in large samples. $r_{x,\hat{u}} \sim N\left(0, \frac{1}{(n-1)}\right)$
- The appropriate test statistic is therefore $r_{x,\hat{u}}\sqrt{n-1}$ and the null hypothesis of homoscedasticity will be rejected at the 5 percent (or 1 percent) level if its absolute value is greater than 1.96 (or 2.58) using two-tailed tests, following the standard normal distribution.
- If there is more than one explanatory variable in the model, the test may be performed with any one of them. $\frac{r_{x,\hat{u}}}{\sqrt{\frac{1}{(n-1)}}} \rightarrow r_{x,\hat{u}}\sqrt{n-1}$

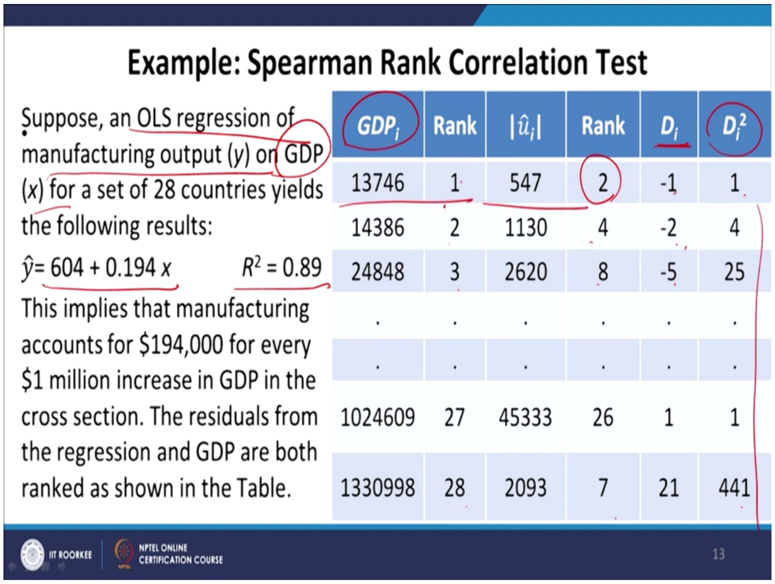
IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 12

But before that, let me tell you that under the assumption that the population correlation coefficient is 0, the rank correlation coefficient has a normal distribution with 0 mean and variance $1/(n-1)$ in large samples. The appropriate test statistic is, therefore (refer slide time: 14:28), follows a normal distribution with mean 0 and variance $1/(n-1)$, then what would be the standard normal distribution?

The standard normal distribution would be (refer slide time: 14:48- 15:04), and this is essentially the test statistic that we would go for. In case we are going for it T-test or if we are going to check it against in a standard normal distribution in both cases this is my test statistic and the null hypothesis of homoscedasticity will be rejected at the 5 percent or 1 percent level if its absolute value is greater than 1.96 or 2.58, which corresponds to the 1 percent level using a two-tailed test following the standard normal distribution.

If there is more than one explanatory variable in the model the test may be performed with any one of them. So, there is certainly some randomness involved in it that which variables you are going to pick up or you may consider all the possible variables alternatively, that is one by one.

(Refer Slide Time: 15:52)



Example: Spearman Rank Correlation Test

Suppose, an OLS regression of manufacturing output (y) on GDP (x) for a set of 28 countries yields the following results:

$\hat{y}$= 604 + 0.194 x R^2 = 0.89

This implies that manufacturing accounts for $194,000 for every $1 million increase in GDP in the cross section. The residuals from the regression and GDP are both ranked as shown in the Table.

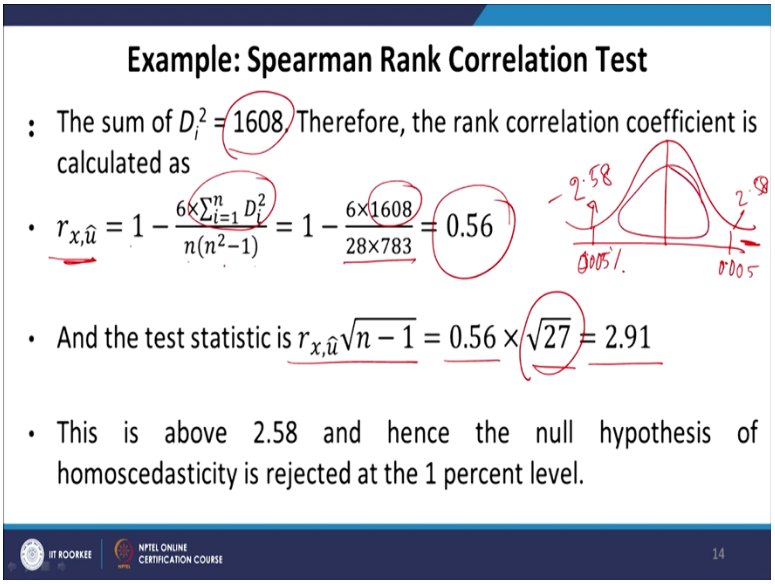
GDP_i	Rank	$\lvert\hat{u}_i\rvert$	Rank	D_i	D_i^2
13746	1	547	2	-1	1
14386	2	1130	4	-2	4
24848	3	2620	8	-5	25
.	.	.	.	.	.
.	.	.	.	.	.
1024609	27	45333	26	1	1
1330998	28	2093	7	21	441

Now, this is an example where we are considering GDP values or GDP figures of 28 countries. So, we are running an OLS regression of manufacturing output on GDP for a set of 28 countries, and this yields the following results. So, we have (refer slide time: 16:13). So, this is a cross-country analysis. This implies that manufacturing accounts for $194,000 for every $1 million increase in GDP in the cross-section.

The residuals from the regression and GDP are both ranked as shown in the table. So, GDP is the only independent variable we have considered from the analysis, it is first ranked, ranked from 1 to 27 and then the corresponding residuals are also ranked. So, the first country having a rank with respect to GDP as 1, in terms of residuals it has a rank of 2 and so on.

So, when these are ranked, then I calculate D_i as the difference between rank 1, the rank of GDP minus rank of residuals. So, this is minus 1 here similarly 2 minus 4 is minus 2, 3 minus 8 is minus 5 and so on, and then I consider D_i^2, so minus 1 squared 1, minus 2 square 4 and so on.

(Refer Slide Time: 17:24)



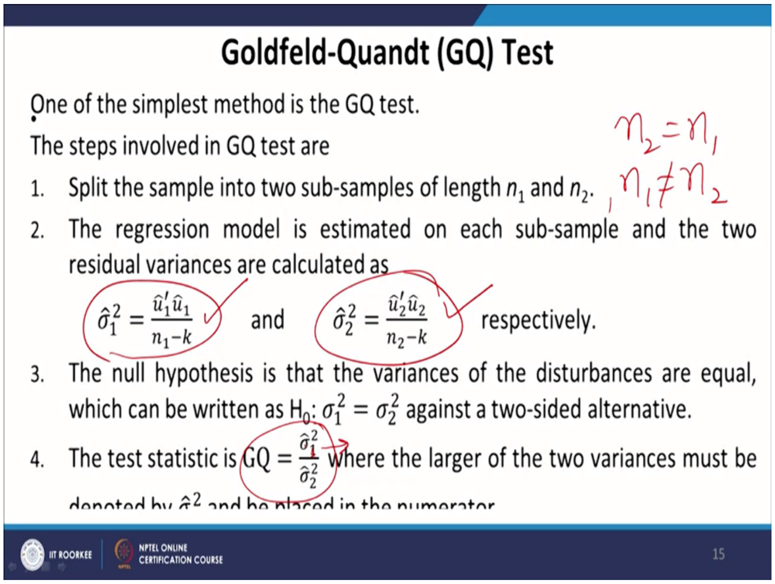
Example: Spearman Rank Correlation Test

- The sum of D_i^2 = 1608. Therefore, the rank correlation coefficient is calculated as
- $r_{x,\hat{u}} = 1 - \frac{6\times\sum_{i=1}^{n} D_i^2}{n(n^2-1)} = 1 - \frac{6\times 1608}{28\times 783} = 0.56$
- And the test statistic is $r_{x,\hat{u}}\sqrt{n-1} = 0.56 \times \sqrt{27} = 2.91$
- This is above 2.58 and hence the null hypothesis of homoscedasticity is rejected at the 1 percent level.

The sum of D_i^2 turns out to be 1608. Therefore, the rank correlation coefficient is calculated as (refer slide time: 17:32), this gives us a value of 0.56. So, the test statistic will be the rank correlation coefficient 0.56 multiplied by (refer slide time: 17:54).

And hence the null hypothesis of homoscedasticity is rejected at the 1 percent level and when it is rejected at the 1 percent level, it is obvious that it will be rejected also at the 5 percent level. So, the null hypothesis is rejected here. Essentially if you remember, this is our distribution and probably here I have 1 percent, so this is 0.05 percent. This is a two-tailed test. So, 0.005 percent, and this corresponds to 2.58 minus here it is 2.58. And 2.91 lies here because of which it does not fall into the acceptance region, it falls into the rejection region, we reject the null hypothesis.

(Refer Slide Time: 18:53)



Goldfeld-Quandt (GQ) Test

One of the simplest method is the GQ test.

The steps involved in GQ test are

$n_2 = n_1$

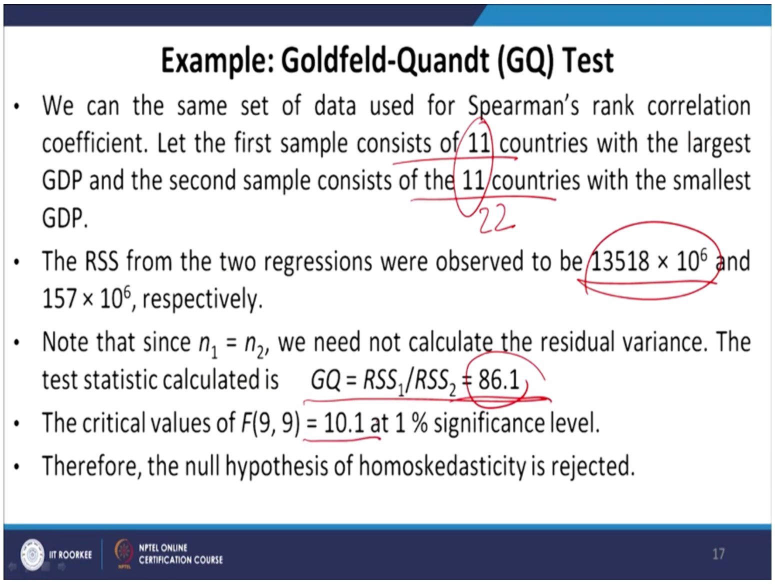
1. Split the sample into two sub-samples of length n_1 and n_2. $n_1 \neq n_2$
2. The regression model is estimated on each sub-sample and the two residual variances are calculated as $\hat{\sigma}_1^2 = \frac{\hat{u}_1'\hat{u}_1}{n_1-k}$ and $\hat{\sigma}_2^2 = \frac{\hat{u}_2'\hat{u}_2}{n_2-k}$ respectively.
3. The null hypothesis is that the variances of the disturbances are equal, which can be written as H_0: $\sigma_1^2 = \sigma_2^2$ against a two-sided alternative.
4. The test statistic is $GQ = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$ where the larger of the two variances must be denoted by $\hat{\sigma}_1^2$ and be placed in the numerator

Now we talk about Goldfeld-Quandt Test. One of the simplest methods is this Goldfeld-Quandt Test. The steps involved in the GQ test are first of all split the sample into two sub-samples of length n_1 and n_2, it is not necessarily that $n_1$has to be equal to n_2. So n_1 equal to n_2 is not necessary, we can have n_1 not equal to n_2, we also can have n_1 equals to n_2. The regression model is estimated on each sub-sample and the two residual variances are calculated as, σ_1^2 and σ_2^2.

So, you can see that (refer slide time: 19:32). The null hypothesis is that the variances of the disturbances are equal, which can be written as, $\sigma_1^2 = \sigma_2^2$ against a two-sided alternative. So, the null hypothesis talks about the two variances. If the two sub-samples have the same variance, then it means there is no Heteroscedasticity, the variances are the same across all the observations.

The test statistic is denoted by GQ following Goldfeld and Quandt is written as, $\hat{\sigma}_1^2 / \hat{\sigma}_2^2$, where the larger of the two variances must be denoted by $\hat{\sigma}_1^2$, or it should be in the numerator. Now, this may correspond to any one of the samples. Of course, here one refers to the first sample, but the point is that the larger variance must be in the numerator.

(Refer Slide Time: 20:38)



Example: Goldfeld-Quandt (GQ) Test

- We can the same set of data used for Spearman's rank correlation coefficient. Let the first sample consists of 11 countries with the largest GDP and the second sample consists of the 11 countries with the smallest GDP.

22

- The RSS from the two regressions were observed to be 13518 × 10^6 and 157 × 10^6, respectively.
- Note that since $n_1 = n_2$, we need not calculate the residual variance. The test statistic calculated is $GQ = RSS_1/RSS_2 = 86.1$
- The critical values of $F(9, 9) = 10.1$ at 1 % significance level.
- Therefore, the null hypothesis of homoskedasticity is rejected.

We can conduct an F-test to test for the significance of the test statistic and the null of constant variance is rejected if the test statistic exceeds the critical value. The GQ test is simple to construct, but its conclusion may be contingent upon a particular and probably arbitrary choice of where to split the sample.

Suppose that it is thought that the variance of the disturbance is related to some observable variable z which may or may not be one of the regressors. A better way to perform that test would be to order the sample according to the values of z and then to split the reorder sample into n_1 and n_2 observations. So, this just provides some guidelines about how to go for a split, but then there is no specific rule about how to split observations or the sample. So, that is one problem with the GQ test.

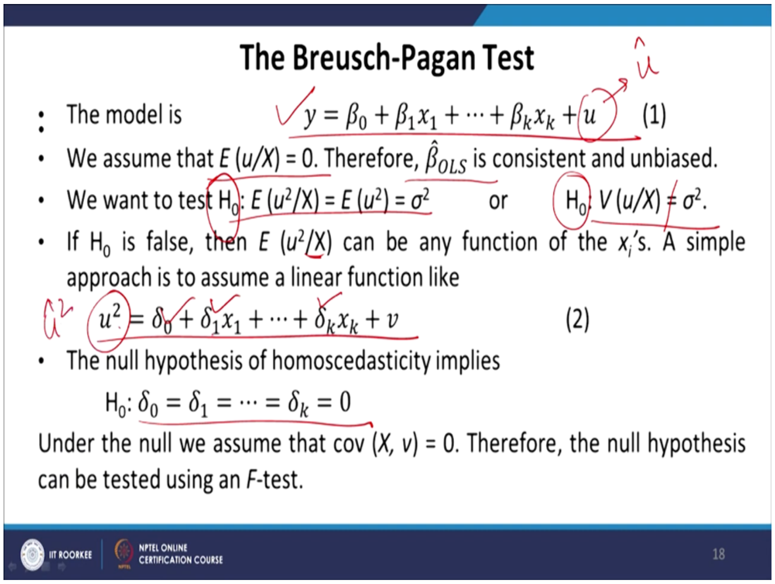
Now, if we continue with the previous example, then we see that using the same data used for a Spearman rank correlation coefficient. Let us take the first sample, we split the sample into two sub-samples, and we actually leave some of these observations from the middle part. So, that is also one solution to tackle Heteroscedasticity.

So, instead of 28 countries, let us take 11 plus 11, 22 countries, so we consider 11 countries with the largest GDP and 11 countries with the smallest GDP. We could have worked also with 14 countries with the largest GDP and 14 countries with the smallest GDP. The RSS from the two regressors is observed to be (refer slide time: 22:23).

Note that since $n_1 = n_2$, we need not calculate the residual variance because residual variances are simply (refer slide time: 22:41), because in both cases we have $n_1 - k$ and $n_2 - k$ in the denominators and since they are the same value so they cancel out and we observe a GQ value of 86.1.

The critical values of the F-statistic or the corresponding F-statistic with the degrees of freedom equal to 9 is 10.1 at a 1 percent significance level. Therefore, the null hypothesis of homoscedasticity is here also rejected because the critical value or this value of the statistic calculated is much higher than the critical value.

(Refer Slide Time: 23:28)



The Breusch-Pagan Test

- The model is $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \overset{\hat{u}}{u}$ (1)
- We assume that $E(u/X) = 0$. Therefore, $\hat{\beta}_{OLS}$ is consistent and unbiased.
- We want to test $H_0: E(u^2/X) = E(u^2) = \sigma^2$ or $H_0: V(u/X) = \sigma^2$.
- If H_0 is false, then $E(u^2/X)$ can be any function of the x_i's. A simple approach is to assume a linear function like

$$\hat{u}^2 \quad u^2 = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + v \qquad (2)$$

- The null hypothesis of homoscedasticity implies

$$H_0: \delta_0 = \delta_1 = \cdots = \delta_k = 0$$

Under the null we assume that cov $(X, v) = 0$. Therefore, the null hypothesis can be tested using an F-test.

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 18

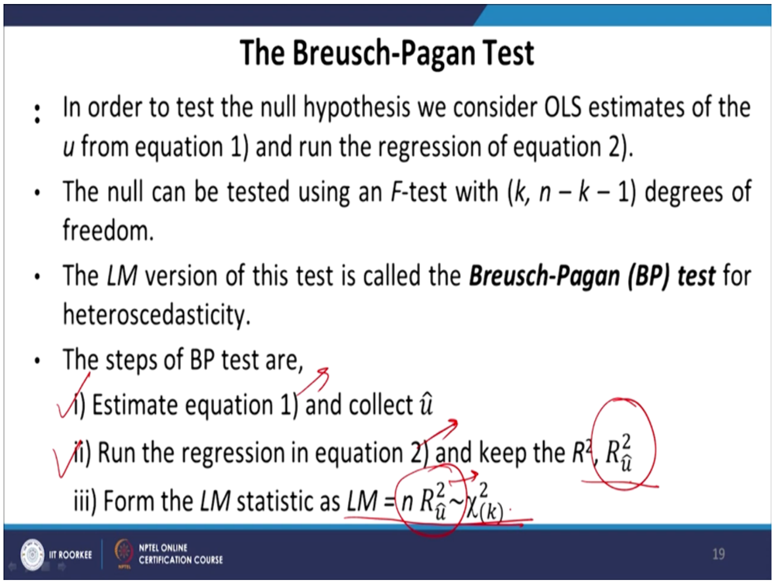
Now, we talk about the third test that is the Breusch-Pagan Test. Here we consider the model (refer slide time: 23:34) which is our original model or model of the world. We assume that the expected value of u given x is equal to 0. So, the mean value of the sample residuals is also expected to be 0. Therefore, $\hat{\beta}$ OLS is consistent and unbiased no problem.

And we want to test whether (refer slide time: 24:05). This is our null hypothesis against an alternative hypothesis that variance of u conditional upon x is not equal to sigma square. If H_0 is false, then the expected value of u^2 conditional upon x can be any function of the x_i's a simple approach to assume a linear function.

So, we assume a linear function between u^2 and the values of x. The null hypothesis of homoscedasticity implies that all these coefficients are going to be 0. So, if all these coefficients are equal to 0, then this implies that u^2 is not dependent on any of the independent variables at least linearly. Under the null, we assume that covariance x and v is equal to 0. Therefore, the null hypothesis can be tested using an F-test.

In order to test the null hypothesis, we consider OLS estimates of u from equation-1 and run the regression of equation-2. So, this is the first regression that we run, obtain the corresponding values as (refer slide time 25:23).

(Refer Slide Time: 25:32)



The Breusch-Pagan Test

- In order to test the null hypothesis we consider OLS estimates of the *u* from equation 1) and run the regression of equation 2).
- The null can be tested using an *F*-test with (k, $n - k - 1$) degrees of freedom.
- The *LM* version of this test is called the ***Breusch-Pagan (BP) test*** for heteroscedasticity.
- The steps of BP test are,

  i) Estimate equation 1) and collect $\hat{u}$

  ii) Run the regression in equation 2) and keep the R^2, $R_{\hat{u}}^2$

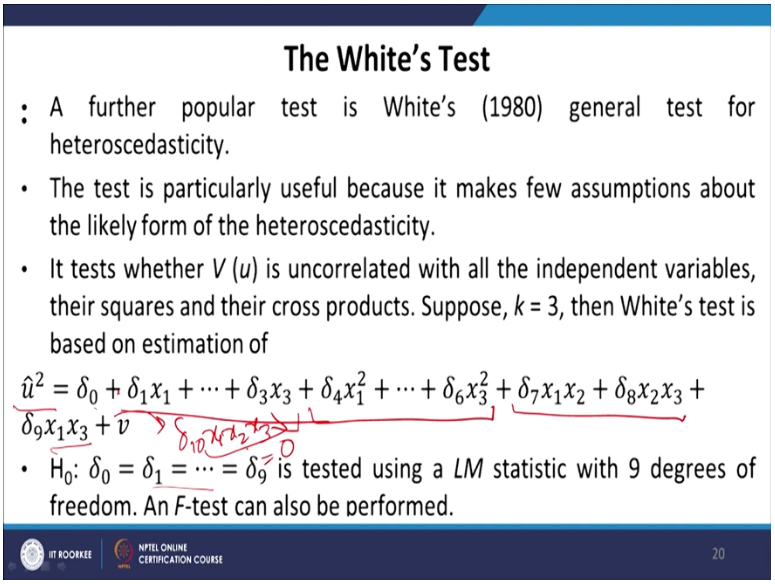
  iii) Form the *LM* statistic as $LM = n\, R_{\hat{u}}^2 \sim \chi_{(k)}^2$

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 19

The null can be tested using an F-test with (k, $n - k - 1$) degrees of freedom, the LM version of this test is called basically the Breusch-Pagan test or BP test for Heteroskedasticity. The steps of BP tests are; first of all estimate equation one and collect $\hat{u}$. Second, run the regression in equation-2 and keep the R^2, which we called $R^2_{\hat{u}}$, this is because this is not the R^2 generated from equation-1. Rather it is generated from equation-2 where the independent variable is (refer slide time: 26:07).

So, the sample observations or the number of observations in the sample multiplied by the R^2 obtained from regression 2, and this follows a chi-square distribution with K degrees of freedom. So, again, depending on the value of the statistic and the critical values given by the chi-square test, we may reject or not reject the null hypothesis.

(Refer Slide Time: 26:43)



The White's Test

- A further popular test is White's (1980) general test for heteroscedasticity.
- The test is particularly useful because it makes few assumptions about the likely form of the heteroscedasticity.
- It tests whether $V(u)$ is uncorrelated with all the independent variables, their squares and their cross products. Suppose, $k = 3$, then White's test is based on estimation of

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \cdots + \delta_3 x_3 + \delta_4 x_1^2 + \cdots + \delta_6 x_3^2 + \delta_7 x_1 x_2 + \delta_8 x_2 x_3 + \delta_9 x_1 x_3 + v$$

$\delta_{10} x_1 x_2 x_3$

- $H_0: \delta_0 = \delta_1 = \cdots = \delta_9$ $= 0$ is tested using a *LM* statistic with 9 degrees of freedom. An *F*-test can also be performed.

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 20

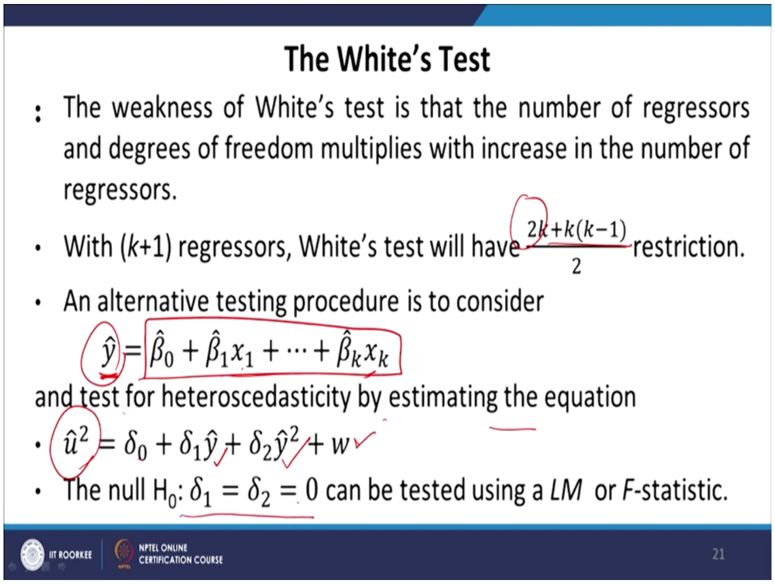
And the last test that we discussed is that of White's. A further popular test is White's test which was given by White in 1980, which is a general test for Heteroscedasticity. The test is particularly useful because it makes few assumptions about the likely form of Heteroscedasticity. So, unlike the Breusch-Pagan test, it does not consider only a linear combination of the independent variables, or $\hat{u}^2$ as a function of the linear combination of the independent variables.

It tests whether the variance of u is uncorrelated with all the independent variables, their squares, and their cross products. Suppose k is equal to 3 that is, there are only 3 independent variables

$x_{1,}x_{2,}x_3$ then White's test is based on estimation of $\hat{u}^2$ on $x_{1,}x_{2,}x_3$ and then x_1^2, x_2^2, x_3^2 and then cross multiplication of x_1x_2, x_2x_3, $x_{1,}x_3$.

We can also add terms like (refer slide time: 27:54). An F-test can also be performed in this context.

(Refer Slide Time: 28:16)



The White's Test

- The weakness of White's test is that the number of regressors and degrees of freedom multiplies with increase in the number of regressors.
- With (k+1) regressors, White's test will have $\frac{2k+k(k-1)}{2}$ restriction.
- An alternative testing procedure is to consider

$$\hat{y} = \boxed{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k}$$

and test for heteroscedasticity by estimating the equation

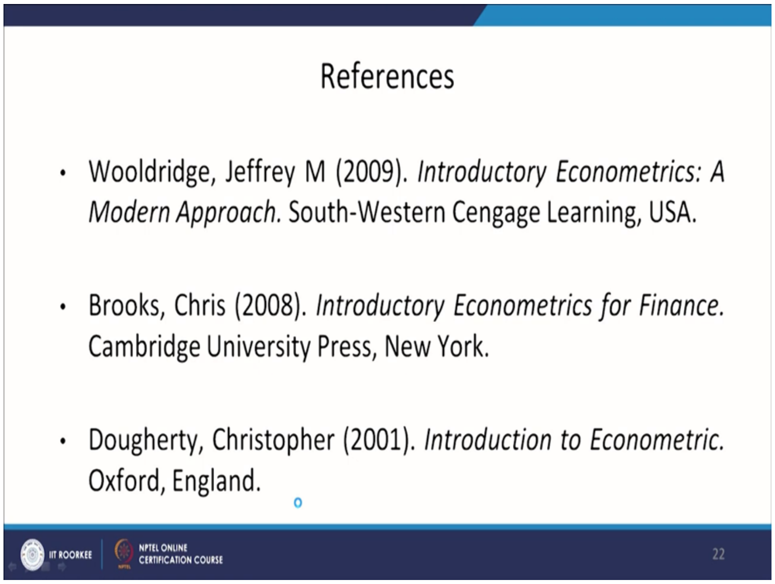
- $\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + w$ ✓
- The null H_0: $\delta_1 = \delta_2 = 0$ can be tested using a *LM* or *F*-statistic.

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 21

The weakness of White's test is that the number of regressors and degrees of freedom multiplies with an increase in the number of regressors. So, when we have $(k + 1)$ regressors White's test will (refer slide time: 28:34). So that is a large number of restrictions. An alternative testing procedure is to consider $\hat{y}$ this is very similar to the Hausman specification test, endogeneity test, that we discussed under omitted variable problem.

That we actually obtain (refer slide time: 28:55). And test for Heteroscedasticity by estimating the equation while (refer slide time: 29:04- 29:20). And this can also be tested using an LM or F-test statistic.

(Refer Slide Time: 29:25)



References

- Wooldridge, Jeffrey M (2009). *Introductory Econometrics: A Modern Approach.* South-Western Cengage Learning, USA.
- Brooks, Chris (2008). *Introductory Econometrics for Finance.* Cambridge University Press, New York.
- Dougherty, Christopher (2001). *Introduction to Econometric.* Oxford, England.

So, these are the references that I have considered in order to come up with a discussion on the definition of Heteroscedasticity and the detection of Heteroscedasticity. What kind of problems it causes to the OLS estimates and in the next module, that is module 18, I am going to deal with how to deal with Heteroscedasticity that is if there are heteroscedastic errors, then how we can still get further minimum variance estimates of the OLS regressors. Thank you.