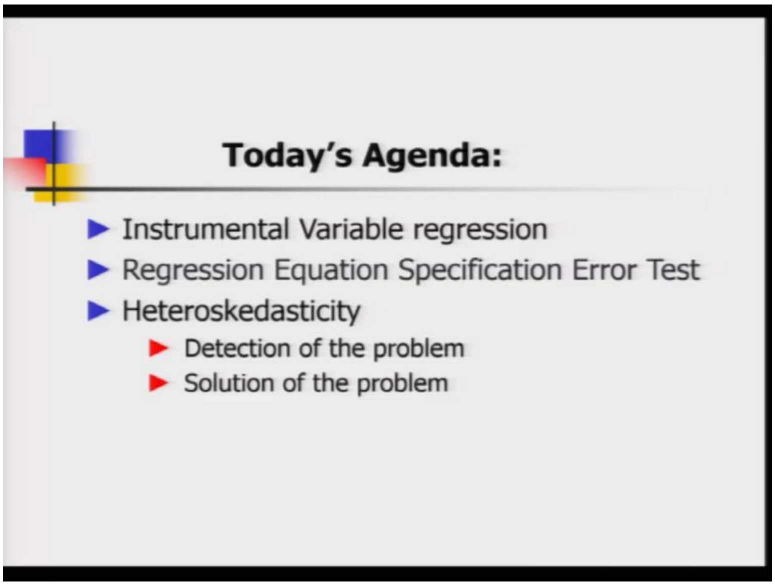
Applied Statistics and Econometrics
Professor Deep Muhkerjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur
Lecture 30
Violations of CLRM Assumptions (Heteroskedasticity)

Hello friends, welcome back to the lecture series on Applied Statistics and Econometrics. So, today we are going to continue our discussion on model misspecification related errors. And today I am going to talk about new types of model specification errors. So, before we start formal lecturing, let us have a look at today’s agenda items.

(Refer Slide Time: 0:39)



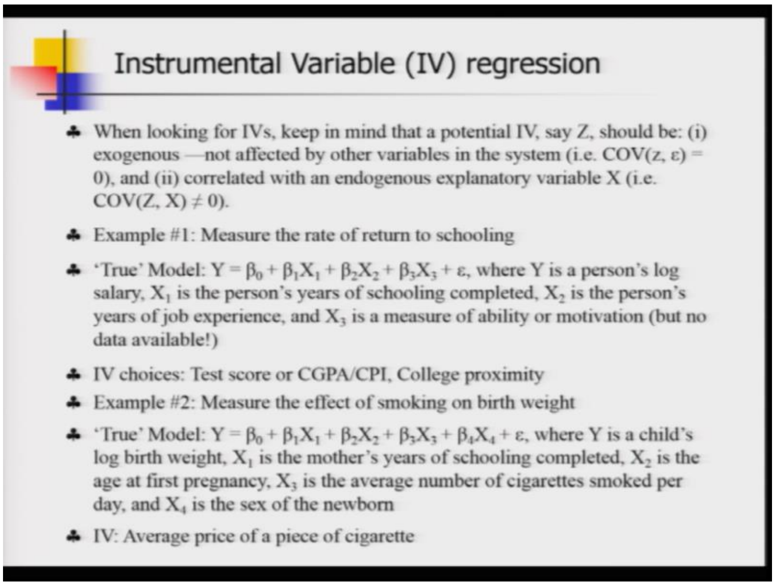
Today’s Agenda:

- Instrumental Variable regression
- Regression Equation Specification Error Test
- Heteroskedasticity
  - Detection of the problem
  - Solution of the problem

So, I will finish the discussion form last lecture and I am going to show you some examples of instrumental variables and how they are handy to tackle the case of omitted variable bias in linear regression analysis and then I am going to introduce two types of model misspecification problem. So, one is functional form misspecification and for that we are going to study something know as Ramsey’s reset test and then the third one in the least we are going to talk about heteroskedasticity being a model misspecification issue.

So, we do not have enough time left in the course to discuss this case of heteroskedasticity in great detail, it is a vast area so we are going to only mention some points which are the most important. So, in the last lecture I have formally introduced the concept of an instrumental variable. So, today I am going to start the discussion again by reminding you about the formal definition once again. It is not formal definition, but if you want to propose an instrumental variable to solve the omitted variable bias problem, then what things you should keep in mind, so that is what we are going to see first and then I am going to talk about two examples.

(Refer Slide Time: 2:03)



Instrumental Variable (IV) regression

- When looking for IVs, keep in mind that a potential IV, say Z, should be: (i) exogenous —not affected by other variables in the system (i.e. $COV(z, \varepsilon) = 0$), and (ii) correlated with an endogenous explanatory variable X (i.e. $COV(Z, X) \neq 0$).
- Example #1: Measure the rate of return to schooling
- 'True' Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, where Y is a person's log salary, X_1 is the person's years of schooling completed, X_2 is the person's years of job experience, and X_3 is a measure of ability or motivation (but no data available!)
- IV choices: Test score or CGPA/CPI, College proximity
- Example #2: Measure the effect of smoking on birth weight
- 'True' Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$, where Y is a child's log birth weight, X_1 is the mother's years of schooling completed, X_2 is the age at first pregnancy, X_3 is the average number of cigarettes smoked per day, and X_4 is the sex of the newborn
- IV: Average price of a piece of cigarette

So, the first bullet point in this slide actually is talking about two conditions that you should keep in mind or the instrumental variable say Z should fulfil these conditions to be used in linear regression analysis. So, the first of all says that it must be exogenous so it should not be effected by any other variable in the regression equation system. So, in other way we can express these condition by statistical concept covariance. So, covariance between the proposed IV say a variable Z and the epsilon the error term, should be equal to 0.

And the second condition tells us that the proposed IV say Z again should be correlated with an endogenous explanatory variable X. So, in statistical terms covariance between Z and X should not be equal to 0. Now, I am going to talk about two different examples from real life and I am not going to show you statistical estimation or regression tables, I am just going to discuss these cases as case studies.

So, the first example I take from the field of education research or education economics and there are many scholars who are interested to measure the rate of return to schooling and note that it is not only an academic pursuit, there are policy makers and others stake holders in the system who also may be interested to know what is the return to schooling because if the return is high, then government may also decide to allocate more money in the annual budget to spread cheap but quality education to the mass.

And also think from an individual parent’s perspective or an individual’s student perspective. If the return to education is high, then actually the parents and the student of course, himself or herself will be much more interested to study for higher degrees and spend more years in schools or colleges. So, with this applied motivation, let us have a look at a very simple hypocritical case where we are going to discuss the omitted variable bias problem first and then we are going to propose some IV.

So, suppose there is a true model to figure out the rate of return to schooling, so the model has 3 explanatory variables and Y is basically a person’s log salary and X1 is the person’s years of schooling completed. So, you can say that the coefficient associated with these variables X1 which is beta 1 is the measure are rate of return to schoolings.

So, the marginal effect of variable X1 which is given by this coefficient beta 1 so the beta 1 coefficient which is associated with these explanatory variable X1 actually gives you a measure for return to schooling because it is a marginal effect. So, it tells you that if a student spends one more year in school, then how much it is going to increase its log salary. So, in a way it is a return to schooling measure.

Now, there are other two explanatory variables, say X2 measure a person's years of job experience because of course, salary of a person depends a lot on the job experience and finally we have another explanatory variable X3 and that is measuring the ability or motivation of the person in concern. Now, note that although this variable X3 is very interesting variable and of course, we know that salary may potentially depend on a person’s motivation and efforts that he or she puts in the job place.

But how to get a measure for this variable X3? No data is available to start wit so although you know that from intuition there is a variable that may impact the salary of a person, but you do not have any data on that particular variable, so then what to do? You have to find out a proxy an instrumental variable to be inserted in place of these unobserved variable X3, but it has to follow certain conditions that I have already stated at the very beginning of this slide.

So, now we are going to talk about what are the potential choices for instrumental variable. So, in academic literature tests scores like CGPA or CPI are often used as a proxy or instrumental variable for these unobserved variable ability or motivation of a person because you see the obvious linkage here. So, if the person is motivated and hardworking and has higher intelligence or ability than the others, then that should be reflected in the CPGA or the CPI or the tests scores from Board Examination.

At some point of time this ability is going to be reflected through marks or performance in the exams. So, if you somehow can access the scores from some kind of an independent test for the student, then you can make use of that information in the regression equation as an instrumental variable to take care of the unobserved ability factor. Also, in education economics scholars have come up with another interesting instrumental variable in this context and that is proximity to the educational institute.

So, here one can say that if a student is deciding close to an educational institute or if the distance is not that far, then the student will attend classes more frequently he or she will visit the educational institute more frequently and that will help him or her to study well and talk with friends or consult teachers or library and that will jack up his or her internal ability. So, now we are going to talk about another example and this time we are going to bring the example from the field of medical science literature or health economics also you can say.

So, we all know that the birth weight of the new-born babies are very important because by looking at these numbers doctors can actually say whether there is a problem with the new born or not and if there is a problem like especially when the babies are born with very low weight, then they need special care. So, doctors are very interested to know what are the determining factors of low birth weight because low birth weight is not at all a very good thing. With this, let us have a look at how we can frame a regression equation to explain variation in birth weights of new-born babies.

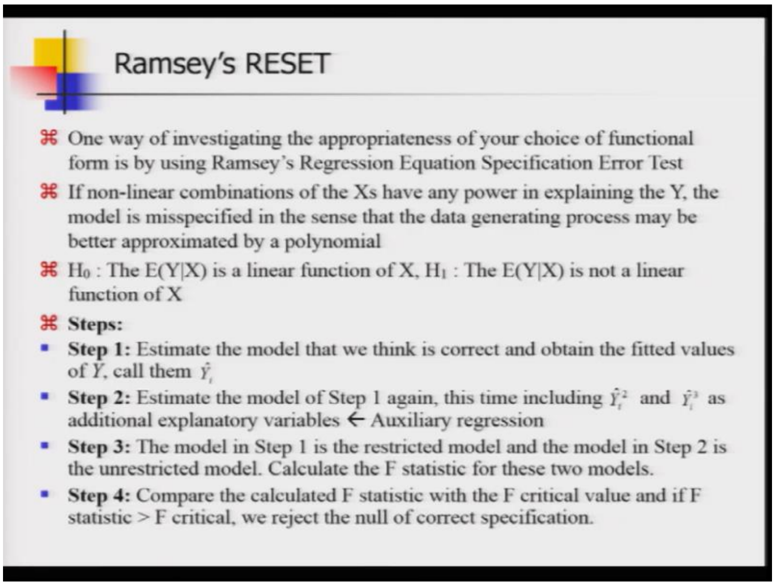
Suppose there is a true model where we are using 4 explanatory variables to explain the variation in Y which is a child’s log birth weight my explanatory variables are X1 which is mother’s years of schooling completed and X2 is basically the age at first pregnancy and X3 is average number of cigarettes smoked per day and X4 is the sex of the new-born. So, now we apparently you may think that I have got 4 explanatory variables to model the variation in Y and I know from medical science literature that all these 4 may have some impact on Y so they could be very good determining factors.

And then let us run the regression, but the story is not that simple. Wait a minute, are you missing some other variable? It may look that you have 4 variables in the regression equation, but there could be a few more which you are not taking care of. So, now let us look at this equation again. So, here there could be an omitted variable problem as well like if the mother was exposed to passive smoking. Now, how to control for that unobserved explanatory variable? So, here we can think about an IV that is used in literature and that is average price of cigarette, why?

Because see ultimately cigarette is a consumption good, you have to go to a shop and purchase it from the shopkeeper and of course, your demand or purchase will depend a lot on the unit price of the cigarette. So, if we are talking about the sample coming from a large country where you have many states, so every state government may have a different tax structure for the tobacco products and cigarette may be taxed differently from one state to the other. So, there could be a price differential expected and that could actually lead to higher or lower demand for cigarette purchase.

So, the passive smoking factor which you are not observing and do not have data, you can replace it by a proxy or instrumental variable say average price of per unit of a cigarette because you think that they could be related. So, this is it as of now on instrumental variable, now we are going to move on to the next case of misspecification error.

(Refer Slide Time: 12:52)



Ramsey's RESET

- ⌘ One way of investigating the appropriateness of your choice of functional form is by using Ramsey's Regression Equation Specification Error Test
- ⌘ If non-linear combinations of the Xs have any power in explaining the Y, the model is misspecified in the sense that the data generating process may be better approximated by a polynomial
- ⌘ H_0 : The E(Y|X) is a linear function of X, H_1 : The E(Y|X) is not a linear function of X
- ⌘ **Steps:**
- ▪ **Step 1:** Estimate the model that we think is correct and obtain the fitted values of Y, call them $\hat{Y}_i$
- ▪ **Step 2:** Estimate the model of Step 1 again, this time including $\hat{Y}_i^2$ and $\hat{Y}_i^3$ as additional explanatory variables ← Auxiliary regression
- ▪ **Step 3:** The model in Step 1 is the restricted model and the model in Step 2 is the unrestricted model. Calculate the F statistic for these two models.
- ▪ **Step 4:** Compare the calculated F statistic with the F critical value and if F statistic > F critical, we reject the null of correct specification.

So, now we are going to talk about the functional form misspecification problem and that we are going to solve by Ramsey's reset test. Now what do I mean by functional form misspecification? Suppose you are extremely lucky in the sense that you have got all the statistically relevant variables for your regression equation, so there should not be an omitted variable problem. Well, but there could be other types of problem and it is coming from the functional form that you are assuming for your regression equation.

So, suppose you are assuming that all these explanatory variables are linked to the dependent variable in linear manner. So, basically you are saying that the statistical dependence between the explanatory variable X and dependent variable Y is linear of nature. But who told you a priori that there is linear relationship? Real life is real life and data is data. So, without actually looking at a plot of Y versus X, you cannot even make this assumption that linear relationship exists between X and Y.

Because in the first part of this econometrics component of the course, we have discussed various functional forms, there could be sufficient non-linear relationship between two variables which is actually supported by the previous or past literature. So, if you oversimplify the model, suppose the relationship is highly non-linear and it requires a polynomial curve, but you are over simplifying by fitting a linear line, straight line, then there could be functional form misspecification and that will impact the distribution of the errors. Why? Because ultimately, if there is functional form misspecification problem like let us take the case of polynomial and linear because it is easy to understand.

So, suppose you actually require a cubic function, so in polynomial of degree 3 for your explanatory variable you require not only the X, but you require X square and X cube also as explanatory variable in the linear regression equation, but you have only X. So, now the impact of this non-linear components like X square and X cube, where will they go? So, they will go to the error component of your regression equation. So, you see, you are creating an omitted variable problem by choosing the wrong functional form or over simplifying your regression equation functional form. So, this is basically in a nutshell model misspecification error on functional form.

So, now we are going to conduct a regression equation specification error test as proposed by Ramsey and at this point, I cannot check myself to talk about little history stuff, so it is interesting to note that this Ramsey’s reset test was proposed by Ramsey when he was doing his PhD. So, sometimes very interesting innovations or findings or developments can happen in the PhD thesis itself. So, Ramsey’s reset test is one of the examples of that. So, here we start with defining my null hypothesis. So, I assume that the linear relationship actually is the true relationship, so I say that expected value of Y given X is a linear function of my explanatory variables X.

And the alternative could be that this assumption is false so you say that the expected value of Y given X is not a linear function of X. So, note that here when I am defining my null and alternative, I am talking about the expected value of Y given X. So, basically I am talking about the population regression function. So, that is very interesting point to be noted here. Now in this slide I am going to talk about 4 different steps that must be followed in order to conduct Ramsey's reset.

So, in step 1 you estimate the model that you think is correct and obtain the fitted values of the dependent variable and you can call them Y hat i and then in step 2 you estimate the model of step 1 again but this time you include the quadratic and the cubic terms of the fitted values of Y. So, you have extra two explanatory variables in the model namely, Y hat i square and Y hat i cube. And this concept where you add some artificial variables based on the fitted values in addition to the existing regressors in the regression equation is known as auxiliary regression. So, basically you have to now run an auxiliary regression but you have to apply it via the OLS method only.

In step 3 what we do? So, the model in step 1 is the restricted model of course, and the model 2 in step 2 is the unrestricted model so you see the Ramsey, it is a test can also been seen from a restricted versus unrestricted model problem. And now we know what to do, we have previously used f test, here also we are going to make use of our good old friend f test. So, from two regressions; the mother equation regression and the auxiliary regression, you calculate the f statistics values and then you come to the step 4 which is also the last step. So, there you compare the calculated value of f statistic with the f critical value and if f statistic is greater than f critical value, then we reject the null of the correct specification.

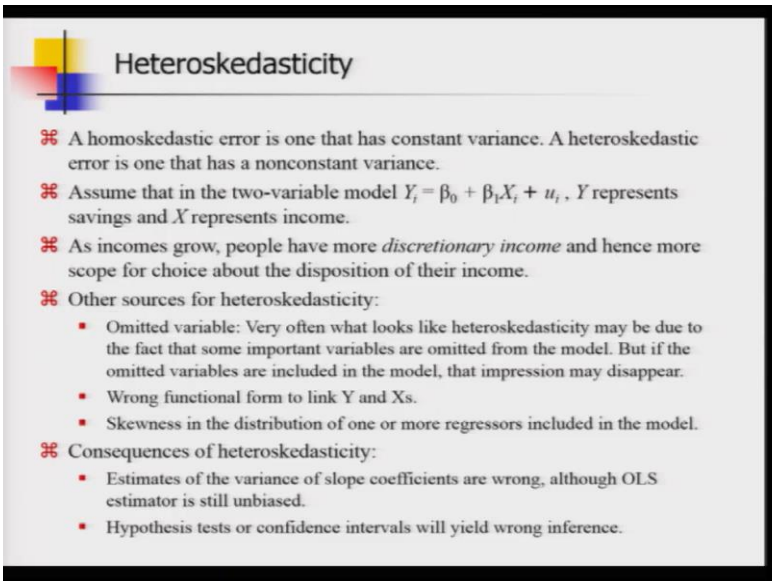
So, let me talk about the degrees of freedom issue here because of course, if you want to conduct the f test, you have to know the degrees of freedom and here as I do not have enough space left in this slide and I do not want to go to a new slide to finish my Ramsey's reset, I just want to talk about the degrees of freedom verbally. But please be attentive here. It is important because you cannot conduct reset without having a proper idea about the degrees of freedom for f.

So, now note that here I am actually introducing two new explanatory variables in the auxiliary regression and suppose the coefficients I am naming them as delta 1 and delta 2. So, delta 1 is associated with the square term of the fitted value and delta 2 is associated with the cubic value of the fitted value. Now you see that the null hypothesis, if it is correctly specified, then basically we can say that the null hypothesis can be restated in terms of delta 1 and delta 2.

So, for the auxiliary regression model, we can actually have these null hypothesis delta 1 is equal to 0 and delta 2 is also equal to 0. So, the distribution of the f statistic here is approximately following an f distribution with degrees of freedom 2 and n minus k minus 3. Now, we are going to move on to another type of model misspecification error problem and that is called heteroskedasticity.

Now, heteroskedasticity is not a new concept to you, we have introduced these terms homoskedasticity and violation of homoskedasticity in special case. So, heteroskedasticity actually is linked with all these concepts that we all already have explored. So, it will not be a very difficult thing for you to appreciate what we mean by heteroskedasticity.

(Refer Slide Time: 21:25)



Heteroskedasticity

- A homoskedastic error is one that has constant variance. A heteroskedastic error is one that has a nonconstant variance.
- Assume that in the two-variable model $Y_i = \beta_0 + \beta_1 X_i + u_i$, Y represents savings and X represents income.
- As incomes grow, people have more *discretionary income* and hence more scope for choice about the disposition of their income.
- Other sources for heteroskedasticity:
  - Omitted variable: Very often what looks like heteroskedasticity may be due to the fact that some important variables are omitted from the model. But if the omitted variables are included in the model, that impression may disappear.
  - Wrong functional form to link Y and Xs.
  - Skewness in the distribution of one or more regressors included in the model.
- Consequences of heteroskedasticity:
  - Estimates of the variance of slope coefficients are wrong, although OLS estimator is still unbiased.
  - Hypothesis tests or confidence intervals will yield wrong inference.

So, a homoscedastic error is one that has constant variance and of course, heteroskedasticity is just opposite of that so a heteroskedastic error is one that has non-constant variance. Now, I am going to explain this concept of heteroskedasticity through a simple two variable equation model because it does not matter whether I go with k variable or not. The philosophy behind this concept or this topic is going to remain the same. So, why clatter our slides with unnecessary symbols and all, so let us stick to a simple two variable case.

So, let me start with the same story that we have discussed at the very beginning of part 2 of this course which is econometrics. So, do you remember how I started talking about OLS estimation in the context of econometrics? I actually started with a story that consumption of households is going to be a linear function of income. So, if there are two uses of income, one is of course, consumption and the other one is savings. So, of course, savings is income minus consumption so savings will also be a linear function of income, household income.

So, basically I now want to talk about savings because we have already dealt with the consumption, so I now want to tell the story in terms of savings. So, suppose we have this task at hand, we want to estimate a savings function and there I assume that household savings actually depends on the households total income. So, these are the only two variables to start with.

Now, here I am showing you this simple linear regression model where Y represents savings and X represents income and as incomes grow, people have more discretionary income and hence there is more scope for allocation or reallocation of their income in the savings and consumption baskets. So, basically what I want to say that as you have more income, then you have better flexibility, and you can decide more flexibly that how much you are going to save and how much you are going to spend because you have enough income.

So, you can expect a high degree of variability in the savings especially for middle income group or rich income group people. So, here we can say that this variance of Y which is basically savings, it depends on the level of X at which we are looking. So, for lower values of X, the error variance maybe smaller and for higher values of X, the error variance may go up.

Now, how it is going to go up, whether it is going to increase linearly or it is going to increase non-linearly that we do not know but what we are saying that sigma square i, the error variance for the ith individual is higher than sigma square say j when the jth observation or the household is a part of relatively lower income class.

There could be other sources for heteroskedasticity as well and we are going to talk about very briefly about three of such sources. So, one is our omitted variable case. So, quite often we get the heteroskedasticity problem because some important variables are omitted from the model and if these omitted variables are included back in the model, then this fear of heteroskedasticity will go.

And there could be another second source of heteroskedasticity, and it is the wrong functional form to link Y and Xs and third one is the case of skewness. So, if there is one or more explanatory variables which are highly skewed, then that will have an impact in the model and it will create heteroskedasticity.

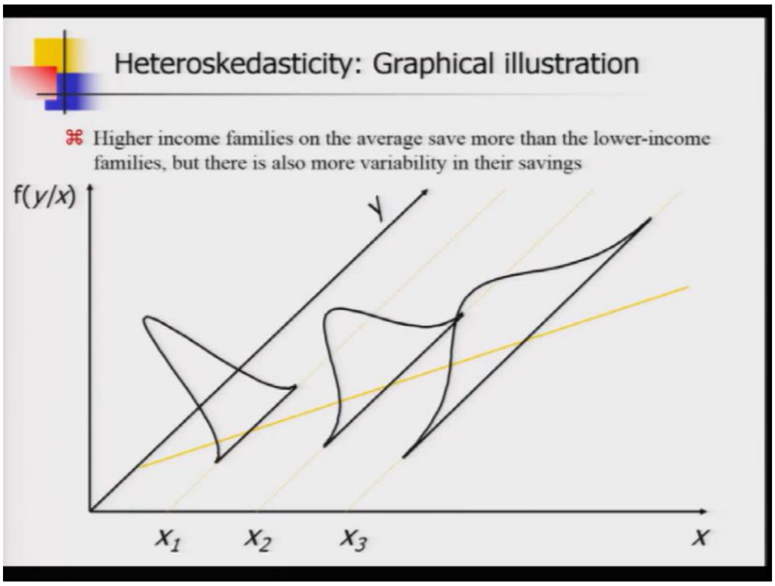
So, we have figured out the sources of heteroskedasticity, now let us look at the consequences of heteroskedasticity. What is a big deal if my data is showing presence of heteroskedasticity? Well, it is a big deal because the statistical inferences that you are drawing from the regression equation may all be faulty. Now, what do I mean by that? Now, statistical inference actually has two different components and we have spoken about them at length so you may be right if you are guessing that am I talking about estimation and hypothesis testing? Yes.

So, now let us discuss the consequences of heteroskedasticity problem if it exists on these two components of statistical inferences. So, first we are going to talk about the case of estimation. Now, it can be shown that even if there is problem of heteroskedasticity in my data, but still my OLS estimator are unbiased. So, I am not going to suffer a lot. Infact no suffering because I will get a set of unbiased estimates anyway, but the problem emerges when we are going to concentrate on the other component of statistical inference which is hypothesis testing.

So, when we are going to conduct hypothesis testing or when we are going to like make confidence intervals around the estimated regression coefficients, then actually we will have a problem and this problem is emerging because if there is this heteroskedasticity problem, then it will impact the standard deviation of the beta coefficient distribution or in other words, it will have an impact on the standard error of the regression coefficients. So, in a nutshell we can say that the hypothesis tests and confidence intervals may yield wrong inferences because the variance of the slope coefficients are wrong.

Now, let us have a look at a simple graph which is showing the presence of heteroskedasticity in the data. And I am going to only show you a simple case and this simple case is based on the savings and income story with which I started the discussion on heteroskedasticity.

(Refer Slide Time: 28:48)



Heteroskedasticity: Graphical illustration

⌘ Higher income families on the average save more than the lower-income families, but there is also more variability in their savings

So, here, I am measuring my income level as X variables and then I am also measuring savings as the Y variable and there is also one extra axis Z axis and on that I am going to measure the conditional probabilities of Y given X. Now, you see the gold colour or the orange colour straight line that upward stopping straight line that you are observing in this two dimensional plain X and Y that is basically my equation or population regression function, it does not have any stochastic component.

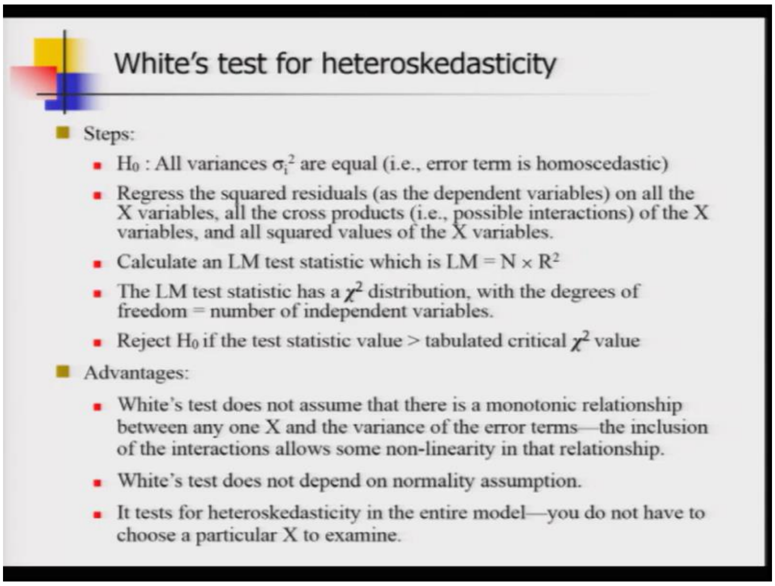
But note that this population regression equation or population regression function will pass through the mean point of the expected value of Y for a given value of X. So, here I am showing you 3 different possible values of X, so for each of these values, I have erected vertical lines in the X and Y plane. So, all my values of Y for a particular value of X variable will now lie on any one of these lines. So, if I say X equal to x1, then all my Ys for the household which is sharing its income value being x1 for that the savings Y values will lie on this vertical line which you here is drawn at the value of X equal to x1.

And of course, we have collected a sample, so there will be sampling distribution of the Y. So, there will be a distribution of the savings values. So, what you see here in terms of the shape of the distribution that it is becoming flatter and flatter. So, what does that mean? So, it means that the variance of the distribution is increasing that is why it is becoming flatter and flatter and there is more mass or density towards the tail. So, that tells you that the lower income households are actually reposting savings numbers for which you are observing lower variance but the high class or the rich group households they are basically reporting the savings numbers that are quite different from one each other.

So, the savings numbers reported by the households who belong to the rich class of the society, there is enough dispersion in the savings numbers and see that is why the variance is increasing. So, you see in this diagram that the variance of Y or the variance of epsilon is not constant for different values of X, in fact, it is increasing. So, there are 2 different approaches to detect heteroskedasticity after conducting the regression and one is graphical method and the other one is by conducting a formal hypothesis testing.

So, after plotting the squared residual values, you had i squared against the fitted values like Y hat i if you see that there is a pattern like it is showing a kind of a funnel shape or there is a positive correlation between the fitted values of Y and the squared residual values, then you can say that there is a problem a heteroskedasticity.

(Refer Slide Time: 32:34)



White's test for heteroskedasticity

- Steps:
  - H_0 : All variances σ_i^2 are equal (i.e., error term is homoscedastic)
  - Regress the squared residuals (as the dependent variables) on all the X variables, all the cross products (i.e., possible interactions) of the X variables, and all squared values of the X variables.
  - Calculate an LM test statistic which is $LM = N \times R^2$
  - The LM test statistic has a χ^2 distribution, with the degrees of freedom = number of independent variables.
  - Reject H_0 if the test statistic value > tabulated critical χ^2 value
- Advantages:
  - White's test does not assume that there is a monotonic relationship between any one X and the variance of the error terms—the inclusion of the interactions allows some non-linearity in that relationship.
  - White's test does not depend on normality assumption.
  - It tests for heteroskedasticity in the entire model—you do not have to choose a particular X to examine.

So, there could be other patterns of heteroskedasticity also. So, suppose you see a kind of parabolic pattern in the scatter diagram and so there could be a highly non-linear relationship between the fitted values of Y and the squares residual terms, then also you can suspect that there could be heteroskedasticity problem. Now, we are going to discuss what is called the White's test for heteroskedasticity. So, here I am going to talk about steps very quickly. So, you set a null hypothesis which is saying that the statusco is correct. So, the status co is homoscedastic.

So, the status co says that the error term is homoscedastic and here I just want to mention a very pity thing that in different textbooks you are going to see different spellings for heteroskedasticity, so someone may use letter ‘k’ someone may use letter ‘c' when they are trying to write this kcedastic part of the expression. So, do not get confused there. So, the alternative hypothesis says that all variances sigma square i are not equal for the observations like i equal 1 to N.

So, in first step now you must regress the squared residual terms as the dependent variable on all of the X variables and their cross products and the squared values of the X variables. So, basically it will now give you lot of new explanatory variables in the regression. So, to understand this auxiliary regression in the case of heteroskedasticity checking let us talk about a simple model where we have two explanatory variables.

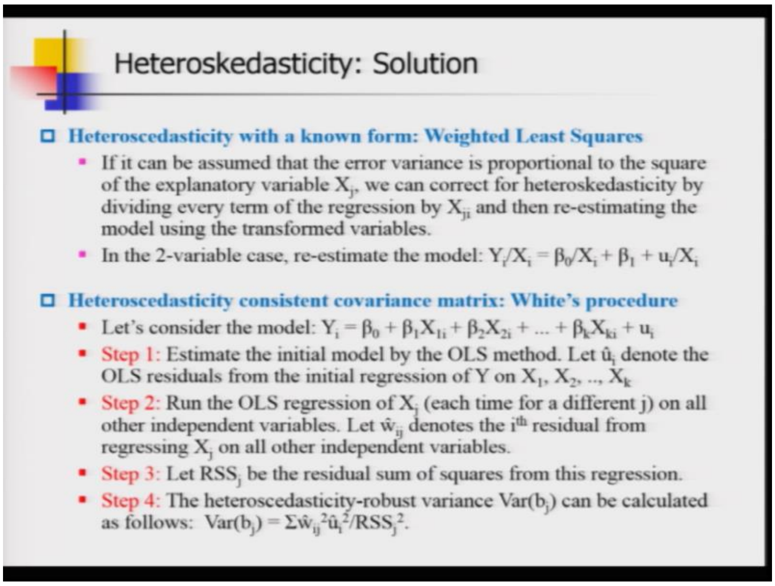
So, then you will have two square terms coming from these two explanatory variables and there will be one interaction term between these two explanatory variables. You are going to add 3 new explanatory variables in the auxiliary regression. And then you calculate the Lagrange multiplier test statistic which is given by the formula LM equals to N times R square and this R square will come from the auxiliary regression. Then what to do? Statisticians have shown that these LM test statistics follows the Chi square distribution with degrees of freedom equals to number of independent variables.

So, here you can say that if I am talking about the same model where we have only two explanatory variables, then basically we have this LM statistic following Chi square distribution with 5 degrees of freedom because there are 5 regressors in the auxiliary regression. 2 original variables, 2 square terms and 1 interaction. I hope this is clear to you now. And then what we need to do? We need to get the tabulated critical value for a specific alpha and then you reject your null hypothesis if the test statistic value is higher than the tabulated critical value.

So, now we are going to end our discussion on White's test by noting down some advantages of this testing procedure. So, White's test does not assume that there is a monotonic relationship between any one X and the variance of the error terms. So, the inclusion of the interactions allows some non-linearity in that relationship. Now, in my humble opinion the best part of White's test is that it does not require the normality assumption to check heteroskedasticity. So, that way it is very, pretty general even if the errors are non-normal, White test is going to give you correct result.

And finally, we can also say that White test actually is useful for checking heteroskedasticity in the entire model as you do not have to choose a particular X to examine. Now, we are going to end today’s discussion by talking very briefly about solution to the heteroskedasticity problem. Suppose you have conducted some kind of detecting exercise for heteroskedasticity, and you are now sure that you have the problem, so if you have conducted a White's test actually you have rejected the null hypothesis of homoskedasticity, then what to do? I am going to now talk about two different solutions proposed by statisticians and econometricians from two different perspectives.

(Refer Slide Time: 38:02)



Heteroskedasticity: Solution

□ Heteroscedasticity with a known form: Weighted Least Squares

- If it can be assumed that the error variance is proportional to the square of the explanatory variable X_j, we can correct for heteroskedasticity by dividing every term of the regression by X_{ji} and then re-estimating the model using the transformed variables.
- In the 2-variable case, re-estimate the model: $Y_i/X_i = \beta_0/X_i + \beta_1 + u_i/X_i$

□ Heteroscedasticity consistent covariance matrix: White's procedure

- Let's consider the model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$
- Step 1: Estimate the initial model by the OLS method. Let $\hat{u}_i$ denote the OLS residuals from the initial regression of Y on $X_1, X_2, \ldots, X_k$
- Step 2: Run the OLS regression of X_j (each time for a different j) on all other independent variables. Let $\hat{w}_{ij}$ denotes the i^{th} residual from regressing X_j on all other independent variables.
- Step 3: Let RSS_j be the residual sum of squares from this regression.
- Step 4: The heteroscedasticity-robust variance $Var(b_j)$ can be calculated as follows: $Var(b_j) = \Sigma \hat{w}_{ij}^2 \hat{u}_i^2 / RSS_j^2$.

So, in the first perspective, let us assume that we somehow know the form of the heteroskedasticity and the solution that is connected to this approach is called the weighted least squares approach. So, suppose if we can assume that this error variance sigma square i is proportional to the square of the explanatory variable Xj in a multiple regression setup, then we can correct for this heteroskedasticity by dividing every term of this regression equation by Xji. So, we are going to know re-estimate the mode using this transformed model.

Now, let us have a very simple illustration of this method. So, here I am going to go back to this two-variable case. So, suppose I have only one explanatory variable and now I know that my sigma square i is some constant proportion of X square i. So, if that is the case, then what I am going to do, I am going to divide my original or mother regression by Xi all observations will be now divided by values of X for that particular observation and now we have a new regression model to work with.

So, the dependent variable becomes Yi divided by Xi and the intercept term is also now going change to beta naught divided by Xi and note that the coefficient that was there associated with the explanatory variable X now it has become the intercept coefficient b1 or beta 1 in the revised model and then finally you see that the error term also has changed Ui divided by Xi and if you take the variance of this revised or redefined regression model, then concentrate on the newly defined error term.

So, variance of Ui is going to be sigma square i. Now suppose we define a new variable Vi which is defined as Ui divided by Xi. So, now we are interested to take the variance of that, so for that we have to take the expectation of Vi square. So, expectation of Vi square will be 1 over X square i times expected value of Ui square. Now we note from our assumption that expected value of Ui square is equal to sigma square times Xi because that is the assumption. So, now X square i will cancel out from numerator and denominator and we are left with sigma square.

So, if we follow the similar kind of adjustment, then we can also assume that error variance is proportional to X only and in that case now we have to divide each variables by square root of now the variable X. So, in that case the new variable, error variable Vi will be defined as Ui divided by square root of Xi, but note that there is a restriction, so your X should be positive here in this case because you are going to take square root of the values.

Now, we are going to briefly talk about the White's procedure and there we do not assume any specific form of heteroskedasticity so we will start with with a multiple linear regression model with k explanatory variables and then in step 1 we estimate the initial model by the OLS and let the U hat i denote the OLS residuals from the initial regression. And then you run OLS regression of Xj, so basically for each explanatory variable you have to run separate regressions on all other independent variables. And then basically let W hat ij denotes the ith residual form regressing the jth explanatory variable on the other independent explanatory variables.

So, in step 3, now you calculate the residuals sum of squares from this regression. So, for the jth explanatory variable the residual sum of squares is denoted is RSSj and finally in step 4 you have the heteroskedasticity robust variance, variance of bj and that can be calculated by the formula that I am showing at the bottom of the slide. So, we are done with our discussion on heteroskedasticity and next time when I will be back with more cases of model misspecification issues and in the next lecture I am going to start talking about autocorrelation. So, join me for the next class. Thank you.