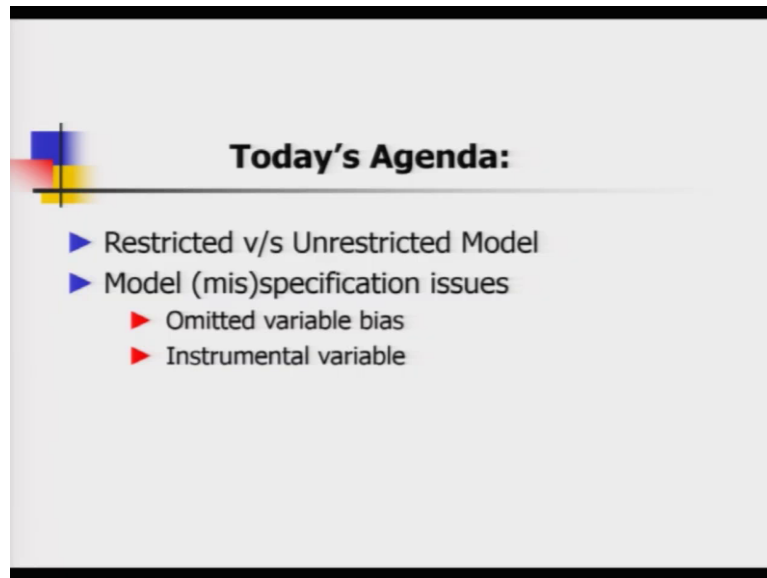


Applied Statistics and Econometrics
Professor. Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur
Lecture No. 29
More on Hypothesis Testing and Model Specification

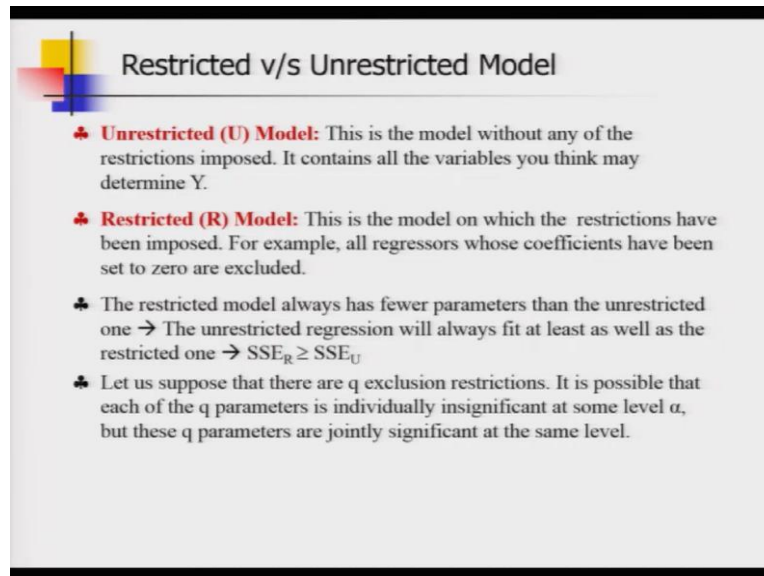
Hello students welcome back to the lecture series on Applied Statistics and Economic.

(Refer Slide Time: 0:20)



So, now let us have look at today's agenda items. So, I say we are going to first talk about the restricted versus unrestricted model and there we going to see an application of f test, our good all friend f test. And then we going to move on to an exciting area in econometrics and that is known as modal misspecification issues. Andthere are several types of modal misspecification issues in econometrics, but in today's lecture where we are going to talk about only one of them and that is called omitted variable bias. And finally we are going to propose a solution to this problem as well.

(Refer Slide Time: 1:00)



Restricted v/s Unrestricted Model

- ♣ **Unrestricted (U) Model:** This is the model without any of the restrictions imposed. It contains all the variables you think may determine Y.
- ♣ **Restricted (R) Model:** This is the model on which the restrictions have been imposed. For example, all regressors whose coefficients have been set to zero are excluded.
- ♣ The restricted model always has fewer parameters than the unrestricted one → The unrestricted regression will always fit at least as well as the restricted one → $SSE_R \geq SSE_U$
- ♣ Let us suppose that there are q exclusion restrictions. It is possible that each of the q parameters is individually insignificant at some level α , but these q parameters are jointly significant at the same level.

So, first we are going to start with formal definition of unrestricted and restricted models. So, an unrestricted model is one model without any of the restriction imposed. So, it contains all the variables that you may think that they may determine Y, your dependent variable. And then here comes the concept of restricted model, this is a kind of model on which the restriction have been imposed. So, for example, you can say all regressors whose coefficients have been set to 0 they are called the excluded variables. And if you have some excluded variables in one regression model, then that leads to restricted model.

So, now note that restricted model works with fewer number of explanatory variables. So, no wonder that for the case of restricted model the sum of square from error residuals is going to be a bit higher than the unrestricted model. And we can also make similar observation for R square also. So, as this slide is showing you here, the unrestricted regression will always fit at least as well as the restricted one. So that means that unrestricted regression will have a higher R square value than the restricted regression model because the sum of square errors from the restricted model is always greater than or equal to sum of square of errors residuals from the unrestricted model.

So, here as you see, I am using the subscripts R and U for SSE symbol, that you already know. So, here the subscript R stands for the restricted model and the subscript U stands for the unrestricted model. Now the question is that how many restrictions one can put? So, if you are in a multiple regression setup and if you have k number of explanatory variables, maximum you can impose k number of restriction. So, here we are going to define a new

notation and we are going to denote the number of restrictions imposed on a linear regression by q .

Now it is very much possible that each of these q parameters that you are setting equal to 0, as you are putting some kind of restriction on the mother regression or the unrestricted model, they may be individually insignificant at some level of α . But it may be the case that these q parameters are jointly significant at the same level of significance. So, we need a statistical testing and for that f test is going to be used. And in the next slide we are going to see how.

(Refer Slide Time: 3:54)

Restricted v/s Unrestricted Model: F test

- ♣ Let us suppose that there are q exclusion restrictions to test. H_0 states that q ($< k$) of the variables have zero coefficients. H_1 states that H_0 is not true.
- ♣ If we obtain a large increase in SSE, this is evidence against H_0
- ♣ Test statistic F^{obs}

$$(i) \quad \frac{(R^2_U - R^2_R)/q}{(1 - R^2_U)/(N - k - 1)} \quad (ii) \quad \frac{(SSE_R - SSE_U)/q}{SSE_U/(N - k - 1)}$$
- ♣ F^{obs} will follow F distribution with q and $N - k - 1$ d.f.
- ♣ Decision:
 - ◆ Reject H_0 if $F^{obs} > \text{Critical value}$
 - ◆ Reject H_0 if $p\text{-value (Area in the } F\text{-distribution to the right of } F^{obs}) < \alpha$
 - ◆ A small $p\text{-value}$ is evidence against H_0 , while a large $p\text{-value}$ is not evidence against H_0 .

So, let us suppose that there are q restrictions imposed and these are exclusion restrictions so by that I mean that q parameter values are set equal to 0 and H_0 ; my null hypothesis states that q which is I am assuming here to be strictly less than k have 0 coefficients. And my alternative hypothesis says that H_0 is not true. So, that means that at least some of these exclusion restrictions are wrong.

Now, as I have shown you in the last slide that I expect a large increase in sum of square residuals or errors if I put restrictions on my mother regression or unrestricted model. And if I finally end up getting a very large increase in SSE score, then that is evidence against H_0 . So, the first step here is to propose a test statistic and that I say f observed is my notation for that. And as I have shown you in the previous lecture that an f test could be conducted using two different matrix or measures, of course, one can make use of the sum of square error from the ANOVA table and the other way to do an f test is to make use of R square.

So, here I am going to propose two formula as they are going to be related to two different approaches that we have seen in the last lecture. So, first we are going to talk about the R square measures. So, now I am showing you the formula here, so R^2_U denotes the R square obtained from the linear regression model which is unrestricted, so that means there we do not have any exclusion restriction.

And then R^2_R actually means the R square obtained from the linear regression or OLS regression of the restricted model where I have imposed q exclusion restrictions. And the difference needs to be divide by the degrees of freedom and that is basically q . And q is what here? q is basically the difference in the number of parameters between the unrestricted model and the restricted model.

Now let us concentrate on the denominator of the same formula. So, here $1 - R^2_U$ is divided by the degrees of freedom, and that is basically the degrees of freedom for the error term. If you remember the ANOVA table, the sources of variation will come from two angles and they are regression which is basically your regression model and the error.

And for regression you have k number of explanatory variables so you set k degrees of freedom and for the error source you set capital N minus capital K which is basically the difference between number of observations and number of parameters to be estimated from the regression model as the degrees of freedom. So, we are going to use the degrees of freedom from the error component, from the ANOVA table.

Now let us look at the second formula and here we are going to make use of the sum of square error directly. So, again, the subscript R stands for the restricted model and subscript U stands for the unrestricted model. And the difference between the sum of squares from restricted and unrestricted model needs to be divided by the degrees of freedom q and then this ratio needs to be further divided by sum of square error from the unrestricted model, and that too divided by the degrees of freedom.

So, now this f observed statistic, either way we have calculated, it does not matter which formula we have chosen for. Both formula will lead to this same theoretical result that f observed my test statistic is now going to follow f distribution with two degrees of freedom. And what are these two degrees of freedom? These are basically q , the number of exclusion restrictions you have imposed on the mother regression model and the capital N minus small k minus 1. So, the difference between the number of observations and the number of parameters to be estimated in the unrestricted model.

Then you set your alpha and then this is the way you are going to make your decision. It is not new to you, so we have done f test previous so you know you have to compare the value of the calculated test statistic which is f observed. And if it is greater than the critical value that you find from f table for the degrees of freedom and the level of significance, then you reject the null hypothesis. And if you want to follow the p value approach, then of course, you reject null hypothesis if p is less than alpha, the set level of significance.

And as I told you previously that a small p value is enough evidence against H_0 while a large p value is not an evidence against the H_0 . Now we are going to look at these restricted versus unrestricted model case and how f test is going to help us to resolve this fight between restricted and unrestricted model through an example.

(Refer Slide Time: 9:40)

Example: Restricted v/s Unrestricted Model

- ♣ Full or **Unrestricted (U) Model**: $\ln \text{butter purchases} = \beta_0 + \beta_1 \ln \text{price of butter} + \beta_2 \ln \text{price of margarine} + \beta_3 \ln \text{butter advertising} + \beta_4 \ln \text{margarine advertising} + \beta_5 \text{Time} + \epsilon$
- ♣ A company wants to know whether advertising expenses on butter and margarine impact demand for butter
- ♣ **Restricted (R) Model**: Butter and margarine advertising do not change demand $\rightarrow H_0: \beta_3 = \beta_4 = 0$ and H_1 : At least one of β_3 and β_4 is non-zero

Full Model (i.e. U) estimated on N = 51					Restricted Model (i.e. R) on N = 51				
Source	SS	df	MS	F	Source	SS	df	MS	F
Regression	0.3574	5	0.0715	11.7423	Regression	0.5233	3	0.1744	28.5506
Error	0.2740	45	0.0061		Error	0.2872	47	0.0061	

- ♣ F statistic = $\{(0.2872 - 0.2740)/2\} / \{0.2740 / (51 - 5 - 1)\} = 1.0838$
- ♣ F statistic value 1.08 < Critical F value (d.f. 2 & 45, $\alpha = 0.05$) 3.21
- ♣ Decision: We can't reject H_0

So, here in this slide I am going to talk about the restricted versus unrestricted model case from one demand function story. So, if you remember, a couple of lectures back I have introduced the case of butter and margarine as two related goods. I was actually trying to get a demand function for butter consumption and then I showed you that margarine also plays a role because margarine is a substitute good for butter.

Now if I want to go back to the same story of butter versus margarine and then how one can actually estimate the demand function for butter consumption. So, suppose I am interested to free the demand function for butter consumption. And I want to define a log log model for my demand function.

Now, we all know that demand or the purchase by consumers in the market is somewhat dependent on the advertisements made by the manufacturing company. And it is nowadays quite evident that people go by who is promoting that particular product and who is showing as the brand ambassador for a product. So, these things matter.

So, now suppose there is a company which wants to statistically test whether expenses made on these advertisement, be it any media, it can TV, it can be newspaper, it does not matter. The total advertisement expenditures made by a company, how it affects the demand for the commodity of interest, say butter. So, in our demand function not only we will have the price of the main commodity, which is butter; and margarine, the alternative or the substitute commodity from the point of the customer. We are also going to add two potentially relevant variables, and these are advertising expenses on butter and margarine.

So, now let us have a look at the final demand functional form that we are going to estimate. So, as I said earlier, we are going to estimate a log log form. So you see my main variables which are basically the dependent variable and the explanatory variables which are coming from economic theory, they are all logged, so logarithmic transformation has been used. And now note that I also have a fifth variable, that is time variable, and for that the regression coefficient is beta 5.

Now why do I need these extra variable which is not coming from microeconomic theory? It is because of the data. Suppose we have a data which is for several months or several quarters and then of course, with time population changes, taste and preferences changes. So, there could be some annual change in the overall demand for any particular consumption item. So, to capture the role of time, we are adding these time variable in the regression because we collect a data on all these variables over a substantially long period of time.

Now, we say that the model that I am showing you here under the first bullet point, that is my full model or that is my unrestricted model. Now as a statistician I have this task at hand, so the company, my client wants to know whether the advertising expenses on butter and margarine have impacted the demand for butter or not. So, if they have come to me with this simple question, how I am going to make use of this restricted versus unrestricted modelling framework.

To answer their question, so here I have to judiciously impose some exclusion criterions, some exclusion criterion or restrictions on the mother model or the full model. And what could and what could be the exclusion restrictions here? So, to develop the restricted model

we can start with these presumption that butter and margarine advertising expenditures do not change the demand. So, basically if you remember the previous slide when we were discussing the definition of restricted model, I said that some parameters in the original regression model has to be set equal to 0. And what does that mean?

That means that, that particular explanatory variables for which you have set the parameter value equal to 0, then basically you are excluding that variable from the mother regression model. So, if you do so, then that means that the variable for which you are imposing these parameter value 0 restriction, that is not linearly related with the dependent variable. So, if advertising expenditures made by the company does not impact the demand for butter, then basically I have to start with by setting the parameters for this advertising expense variables equal to 0. And that is what I am going to do here.

So, if you now concentrate again on the full or unrestricted model, you see the coefficients for the advertisements on butter and advertisement expenditure on margarine are β_3 and β_4 respectively. So, we are going to set null hypothesis which says that $\beta_3 = \beta_4 = 0$. So that is μ exclusion restriction. So, note that here I am imposing two exclusion restrictions. And what could be my alternative hypothesis? So, alternative says that at least one of these β_3 and β_4 must be non-zero.

Now I am going to show you the way how you can make use of ANOVA table or especially the information on sum of square error to conduct the f test. But you remember that I have also shown you the formula which involves R square formula. So, similar technique or the steps that I am going to show you, it can be followed for R square version of the formula or test as well.

So, now you concentrate on the ANOVA table. Suppose we have data on 51 time points, it could be 51 months, it could be 51 quarters, it does not matter. And so here in the main regression the time variable actually takes 1, 2, 3, to 51 values. So, that is the way I am going to measure my time variable in the regression equation. Now you look at this slice of the ANOVA table.

So, here, I am showing you two blocks of results, the first block talks about the full model or the unrestricted model, and the second block talks about the restricted model. And under both the models I am showing you the sources and the sum of squares for different components of total sum of squares, degrees of freedom and then mean square and then finally the f statistics.

So, if you look at the first block, then I am not showing the total sum of squares because if you add the sum of squares for the regression component and the error component, you are going to get the total sum of squares or SST. So, I am skipping that. Anyway, now let us look at the degrees of freedom column. So, here you note that look at the mother model or the full model, the unrestricted model. Here we have 5 explanatory variables excluding the intercept term, hence we have degrees of freedom equal to 5.

And for the error component we have 45 degrees of freedom, how come? Because if you remember we have 51 number of observations and including the parameter we have 6 parameters to be estimated, so 51 minus 6 you have 45. So, now let us be back on the ANOVA table computation, so you divide the sum of squares number for each source or component by respective degrees of freedom to get the MS values. And if you now divide the MS regression by MS error, so MSR divided by MSE, you get the F value. And this is the way you are going to get all these numbers that you are seeing here.

And now note that as I have conducted two different regression experiments, I have got two sets of coefficients, two different values of R square and two different ANOVA tables. Now, I am going to look at how I am going to combine the information from the restricted regression model and the full regression model into one formula which is my f observed test statistic and then conclude whether this restrictions are statistically significant or not.

So, now look at the fourth bullet point in the slide, just right at the bottom of the table. There I am going to show you how you calculate the f statistic value. So, here the formula I have shown you in the previous slide, you are just going to search for those particular entities in these two regression ANOVA tables. And then if you plug those numbers back in the formula, you are going to an f statistic value, that is going to be equal to be 1.0838.

So, once you get your calculated value of f statistic, what to do, you set alpha level of significance say 5 percent and then you have to now look at the degrees of freedom that you have used in the calculation of the f statistic and then you have to go to the f table to find out the critical value. So, what are the degrees of freedom here? So, for the numerator the degrees of freedom is q, which is the number of exclusion restrictions you have imposed on the mother model or the full model.

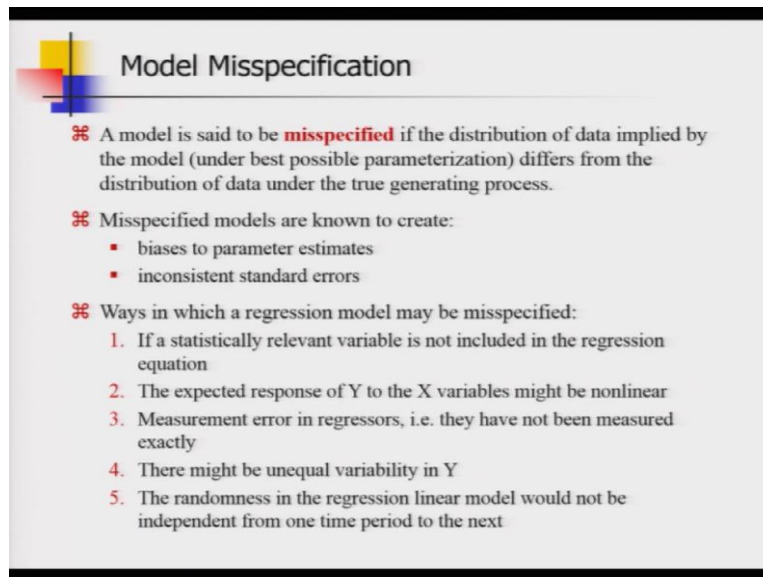
So, here I am putting two restrictions because I am setting 0 value for two population parameters, beta 3 and beta 4, so here the degrees of freedom is 2. And for the denominator, I have these degrees of freedom capital N minus capital K, and that is going to be 45. So, now

if you go to the f table, then the critical value for these two degrees of freedom and alpha value 0.05, that value is 3.21. And now note that that is much, much higher than the calculated f statistic value which is 1.08, so you take a decision that we cannot reject H_0 .

So, what do you mean by that you cannot reject H_0 . What is H_0 ? H_0 is basically talking about the exclusion restrictions, so it is setting 0 value for the population parameters β_3 and β_4 , and that actually tells you that jointly the advertising expenses on margarine and the butter commodities do not have any impact on the purchase or demand for butter in the market. And if you cannot reject null hypothesis, then you say that your hypothesis is true with which you have started.

So, then you can tell your client that from my statistical data analysis I have enough sample evidence to infer that the advertisement expenditures on margarine and butter they are not going to impact the demand for butter.

(Refer Slide Time: 22:22)



Model Misspecification

- ⌘ A model is said to be **misspecified** if the distribution of data implied by the model (under best possible parameterization) differs from the distribution of data under the true generating process.
- ⌘ Misspecified models are known to create:
 - biases to parameter estimates
 - inconsistent standard errors
- ⌘ Ways in which a regression model may be misspecified:
 1. If a statistically relevant variable is not included in the regression equation
 2. The expected response of Y to the X variables might be nonlinear
 3. Measurement error in regressors, i.e. they have not been measured exactly
 4. There might be unequal variability in Y
 5. The randomness in the regression linear model would not be independent from one time period to the next

Now, we are going to move on to a new topic and that is called model misspecification. Although, I am saying that it is a new topic, but it is not very different from the topic which we discussed last that is basically the case of hypothesis testing because here also you will see we are going to make use of this tool, hypothesis testing, to decide whether we have enough sample evidence so that we can infer something or not.

And inference that I am trying to draw here is not about a particular value of the particular population parameter. Here, I am going to talk about drawing inferences about the

correctness or the validity of the assumptions that I made when I framed my linear regression model. So, in this lecture of course, we are not going to talk about the entire span of this problem, model misspecification, we are going to only talk about one type. In the next lecture we are going to continue about that discussion.

So, let us first begin with a formal definition for model misspecification. So, a model is said to be misspecified if the distribution of the data, here by data I mean the variable under focus which is basically your dependant variable Y . So, if the distribution of data implied by the regression model under best possible parameterization differs from the distribution of data under the true generating process.

So, what do I mean in a nutshell about this quite bookish definition? So, you remember I said that when you start the journey of building a classical linear regression model, you have made certain assumptions about the data generation process and what do I mean by data generation process? So, data generation process actually says that your Y is a random variable and there is a process of generating values of Y . And what is these process, from where it is coming?

So, when you write the classical linear regression model assumptions, you first of assume that your explanatory variables are all fixed in nature. So, you assume that your X s are not random. So, there is no process if there is no randomness in the X s. X s are all, exogenously given to you. So, in the regression equation what is random for which you can expect that there could be a stochastic process?

So, that is basically the epsilon term, the stochastic disturbance term or the random error term whatever you want to call it, that epsilon that is a random variable. And as its random variable Y will become a random variable too because your Y now becomes a function of realization of these random variable epsilon. So, there is a data generation process for Y as well as you see it for epsilon.

Now, you start model building by making some assumptions about these random error. So, remember the assumption of 0 conditional mean, then the assumption of homoskedasticity which is basically constant variance, then at last we have also said that one can assume normality as a distribution for the error term. So, now these are all assumptions and you may be interested to know that in reality how far true these assumptions are?

Note that a model is basically an abstraction from the reality. You with your dreams and tool sets developing models by making assumptions and using these statistical techniques, but real

life is real life, it works in its own way and the assumptions that you are going to make to develop your regression model may not hold in reality. Data is data, you do not know whether it is going to obey the assumptions you are making while you are sitting in an air conditioned room in front of a computer.

So, basically there has to be some mechanic to test whether the assumptions you are making to develop this classical linear regression models whether they are valid ones or not. And if somehow you figure out that the assumptions you made when you are developing your classical linear regression model; they are invalid, then actually they will lead to this model misspecification error.

So, why are we so much worried about misspecified models? What is the big deal if some of the assumptions that we made are not fulfilled by the data? The model misspecification errors can create havoc in your regression analysis and the inferences that you are going to draw out of your model can all be wrong, depending on the magnitude of the misspecification error. So, mostly they are going to create two types of nuisances, first they are going to create biases to the parameter estimates. And what do we mean by bias?

If you remember the discussions in part 1 of this course, I introduced this term called bias. So, bias is basically the difference between the true population parameter value and the mean of the estimator that you get from your OLS technique or whatever technique you are applying to draw some statistical inference from real life data. And secondly, the misspecified models are also going to create inconsistent standard errors.

So, inconsistent standard error means that if you get one sample, you get one value of standard error. And then if you move on to another sample, you are going to get another value of standard error. And if you are unlucky, if there is model misspecification issue, then they actually may be way apart from each other. So, the dispersion of the coefficients that you are going to get from different samples is going to be very high and that is not good because that is going to impact your hypothesis testing like t tests and all.

Now, there are many ways in which a regression model may be misspecified .And here I am going to list down 5 of them. So, first let us start with the case which is basically we are going to talk about in today's lecture. So, if a statistically and theoretically relevant variable is not included in the regression equation, then we will have the omitted variable problem. So, that is basically the number 1 source of model misspecification.

Then we are going to talk about the second source of model misspecification and that has to do with the functional form of the regression equation. To start with, we have assumed that my X; the explanatory variable, is linearly related to my dependent variable Y, and that is why we have a linear-in parameter, linear-in variable models. But there could be non-linearity involved, and if you are simplifying a linear simplification maybe too much of fuzzy implication. And then, of course, it will have its own consequences on the model estimates and the model fit. And forcefully fit a linear regression equation, then there is a functional form misspecification and that will lead to the second type of model misspecification.

Now we move on to the third case and that is the measurement error in regressors. So, here what do we mean by measurement errors in regressors? So far we have assumed that Xs are given to you and Xs are perfectly measured, so there is no error in the value of Xs. But there could be error involved in the value of Xs if they are not properly measured. Suppose I give you an example of say temperature and rainfall, many, many times we use these variables in agricultural production function regression equations.

But in third world countries where the quality of tools and equipments are not very high and cutting edge, there may be some inconsistency and errors while the weather stations are measuring rainfall and temperature. So, in that case you may have some random noise embedded in the data on the temperature and rainfall.

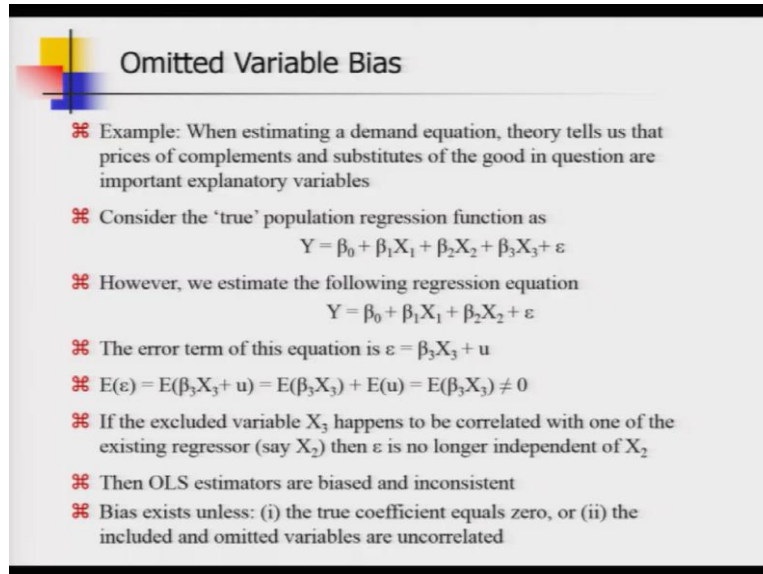
Now we are going to talk about the fourth source for model misspecification. So, there could be unequal variability in Y. So, what do we mean by that? So, remember when we started making assumptions of for classical linear regression model, I said that my errors are going to be homoscedastic in nature. So, I am talking about constant variance.

So, if epsilon is not homoscedastic, it implies that for ranges of data the variance for epsilon is now going to change, that will have the similar effect on Y as well because the distributions for Y and epsilon, they are linked to each other. So, now if the homoskedasticity for epsilon breaks down, then for Y also we cannot expect constant variance for the entire range of Y data. And homoskedasticity is very important assumption to make and if that breaks down, then we will have a problem. Of course, we are not going to discuss what sort of problem it can happen in this particular lecture, we save the discussion for the next lecture.

And finally we come to the last point and that is basically saying that randomness in the regression linear model would not be independent from one time period to the next. So, what do I mean by that? So, actually I have camouflaged one term that I have already introduced in

the course. Do you remember that I have introduced a term called serial correlation or autocorrelation? So, by that I mean that when two error terms are correlated with each other across cross sectional units or across time point units, then actually we say that the randomness or the independence between the consecutive values of dependent variable breaks down and that is not good for you. And that is what is specifically being told in point 5.

(Refer Slide Time: 33:52)



Omitted Variable Bias

- ⌘ Example: When estimating a demand equation, theory tells us that prices of complements and substitutes of the good in question are important explanatory variables
- ⌘ Consider the 'true' population regression function as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$
- ⌘ However, we estimate the following regression equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$
- ⌘ The error term of this equation is $\varepsilon = \beta_3 X_3 + u$
- ⌘ $E(\varepsilon) = E(\beta_3 X_3 + u) = E(\beta_3 X_3) + E(u) = E(\beta_3 X_3) \neq 0$
- ⌘ If the excluded variable X_3 happens to be correlated with one of the existing regressor (say X_2) then ε is no longer independent of X_2
- ⌘ Then OLS estimators are biased and inconsistent
- ⌘ Bias exists unless: (i) the true coefficient equals zero, or (ii) the included and omitted variables are uncorrelated

So, now we are going to talk about omitted variable bias through an example. So, let us go back to the microeconomic theory of demand function. So, when we estimate a demand function, microeconomic theory tells us that the prices of compliments and substitutes of the good are explanatory variables and they should be incorporated in the regression equation when you are trying to estimate a demand function for the commodity in the question.

But, I mean theory is theory, it is saying that you need to consider complimentary goods and the substitute goods. But does it provide the full list of complimentary goods and substitute goods? Answer is no. So, now which are the best complimentary goods and which are the best substitute goods and how many of them you are going to keep in your demand equation that is not explained by theory.

And also, even if you make a list of, say, 3 complimentary items and 4 substitute items for the commodity in question for which you want the demand equation to be estimated, you may not have data on these particular commodities and if you do not have data, then of course, you cannot include that variable in the regression.

So, now we consider that there is a true population regression function which is perfect and here I am showing you the expression for that. Of course, it is a simple version of the reality. Suppose I have only 3 explanatory variables and somehow for some reason I have omitted X3 variable, there could be reasons like I may be ignorant about the effect of that X3 variable on Y or even if I know that there is an impact of X3 on Y but I may not have data on it. So, whatever be the reason, from the regression equation that I am finally estimating from raw data I do not have X3 in the regression equation.

Now, let us see the consequences. So, now you actually end up estimating a regression equation involving two explanatory variables X1 and X2. Now, note that if the X3 variable which you have omitted if it is truly a relevant variable which actually has an effect on Y, then the portion of Y's variability that it can predict, now that component will be now embedded in the random error that you have defined for your model.

So, now note that as b_3 is not equal to 0, then b_3 times X3 is also not going to be 0. So, there is some non-zero component in the regression, but there is no variable which is taking care of that. So, it will go in the error term. So, it will be part of the error term and you know the epsilon error term will be actually β_3 times X3 plus U where U is basically new random error term which is obeying the standard assumptions for error terms that we make in classical linear regression model.

So, now you see how the omitted variable bias is emerging, so you take expectation of both hands. So basically now expected value of epsilon will be sum of two components, expected value of β_3 times X3 and expected value of U. Now U, as I said that we can assume that it is following the standard assumptions for classical linear regression model, so expected value of U will be 0. So, finally you are going to end up getting expected value of β_3 times X3 and which is a non-zero number.

So, now you see the consequence is like you do not have the error distribution which has 0 conditional mean, so the mean now could be positive, it could be negative, but it is certainly not 0. So, why this expected value of epsilon not being equal to 0 is a big deal because this comes very useful, this property comes very handy when you are trying to establish the unbiasedness of the OLS estimators.

Now we are going to talk about another dimension of the problem. So, if the excluded variable X3 happens to be linked with one of the existing regressors, say X2, then epsilon is no longer independent of X2. And what does that mean? So, your another assumption of

classical linear regression model breaks down. So, remember we said one assumption of exogeneity and there I have set this statistical criterion that covariance between my explanatory variables or the vector of explanatory variables should be not related with the vector of the error terms, which is epsilon.

So, $cov(X, \epsilon)$ is not equal to 0. But if now X_3 , which you are missing, is related with X_2 , then basically higher the root of epsilon now it is going to be impact the coefficient of beta 2 as well. So, if that is the case, if there is some correlation between X_2 and X_3 , then what happens? Then X_2 which is part of your linear regression equation that you are estimating will be correlated with the epsilon term or the error term. Why? Because that epsilon term now has that impact of that third variable. X_3 is hidden there.

So, in a nutshell, we can say that OLS estimators are biased and inconsistent. Now bias exists, unless we are very lucky and when we could be extremely lucky so that although we have forgotten to incorporate one relevant variable, but our regression results are not plagued by bias. For that we have to see two good things happening to us. So, the first good thing is that the true coefficient should be equal to 0 and the included and the omitted variables are uncorrelated.

(Refer Slide Time: 40:50)

Omitted Variable Bias

- ⌘ The direction of the bias depends on the estimators as well as the covariance between the regressors and the omitted variables
- ⌘ A positive covariance of the omitted variable with both a X and the Y will lead the OLS estimate of the included regressor's coefficient to be greater than the true value of that coefficient
- ⌘ Suppose we have two explanatory variables

	Included and Omitted variables are positively correlated	Included and Omitted variables are negatively correlated
Omitted variable is positively correlated with Dep. Var. Y	Positive Bias	Negative bias
Omitted variable is negatively correlated with Dep. Var. Y	Negative bias	Positive Bias

- ⌘ The actual magnitude of bias is very hard to detect
- ⌘ There exists no statistical test that detects omitted variable biases

Now in this slide we are going to talk about omitted variable bias in a bit more detail. So, the direction of the bias depends on the estimators as well as the covariance between the regressors and the omitted variables because I told you in the last slide only that suppose we are ignoring and omitting X_3 and X_2 and X_3 are somehow correlated and then now your epsilon and X_2 they are going to be correlated via this route of X_3 .

So, you see that the covariance between regressors those which are staying in the final regression equation and those which are omitted variables which are now part of your error, the covariance between these two are going to play a major role. So, a positive covariance of the omitted variable with both an explanatory variables X and the dependent variable Y will lead the OLS estimate of the included regressor's coefficient to be greater than the true value of the coefficient. So, basically there is an upward bias in this context.

Now, in this slide I am going to show you simple table which is going to talk about the nature of the bias. But in addition, I should also tell you that it is very difficult to measure the extent of the bias. So, let us have this simple table where we are talking about only two explanatory variables, so here I am showing you different possible cases. So, let us look at different cells where you see this nature of the bias reported in colour red.

So, let us talk about the cell 1 which is basically the northeast one. So, here we see two conditions are given, so the first is saying that included and the omitted variables are positively correlated. And not only that, we also have to satisfy one more condition and that is given in the first row, so that is saying that omitted variable is positively correlated with dependent variable Y . So, only in that case you will get positive bias.

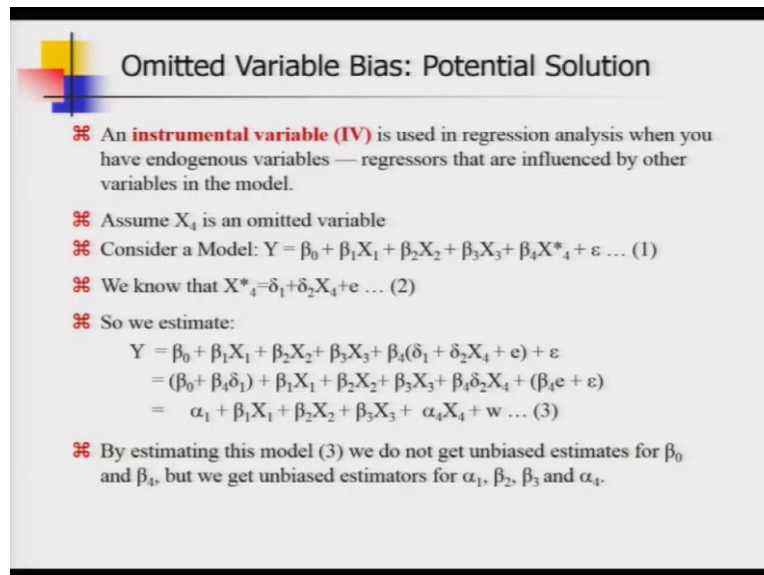
Let us now you know go in to look at one case of negative bias and then I will stop because rest you can follow. So, now let us concentrate on the second column which is saying that included and omitted variables are negatively correlated. And now let us assume that my omitted variables is positively correlated with depended variables Y which is given in first row.

So, again, in that case you will see that bias will come but the direction will be different and this time we are going to get negative bias. Similarity, you can see different combinations are going to lead to different types of bias. Now the bad part of omitted variable base is that although we know there could be a problem but you there is no hypothesis testing tool to help us so that we can detect that in our regression model whether we have a problem of omitted variable bias or not.

So, what to do? Well, here you have to use your grey scale. So, to do lots of thinking, you have to consult the previous literature, you have to figure out what Wiseman have talked about in your particular regression problem that we are dealing with, whether someone has suspected a presence of omitted variable bias which is very relevant for the regression problem that we currently handling.

So, from the literature you have to guess and also you have to use your intuition to figure out whether there is omitted variable bias or not in the regression equation that you are specifying. But suppose your intuition says that, well there could be one variable which I am omitting or may be the literature is suggesting you that the model specification that you currently have maybe suffering from omitted variable bias, then how do you solve for it? So, in the last slide of today's lecture we are doing to discuss about it.

(Refer Slide Time: 45:16)



Omitted Variable Bias: Potential Solution

- ⌘ An **instrumental variable (IV)** is used in regression analysis when you have endogenous variables — regressors that are influenced by other variables in the model.
- ⌘ Assume X_4 is an omitted variable
- ⌘ Consider a Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \dots (1)$
- ⌘ We know that $X_4 = \delta_1 + \delta_2 X_4 + e \dots (2)$
- ⌘ So we estimate:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 (\delta_1 + \delta_2 X_4 + e) + \varepsilon$$

$$= (\beta_0 + \beta_4 \delta_1) + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 \delta_2 X_4 + (\beta_4 e + \varepsilon)$$

$$= \alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \alpha_4 X_4 + w \dots (3)$$
- ⌘ By estimating this model (3) we do not get unbiased estimates for β_0 and β_4 , but we get unbiased estimators for $\alpha_1, \beta_2, \beta_3$ and α_4 .

So, here as a potential solution, econometricians have come up with a novel idea called instrumental variable. And an instrumental variable is used in regression when you have endogenous variables. Now what do we mean by endogenous variables? Remember when we made the assumptions, we said all our X variables are given so they are exogenous in nature. So, basically a change in Y will not have any impact on Xs. And if that breaks down, we can say that our explanatory variables are endogenous in nature. So, their values are actually determined within the model, not outside the model.

So, the endogenous variables are basically the regressors which are influenced by our variables in the model. So, as you see the values of some of the explanatory variables may be determined within the model by other variables. Now, suppose we have one regression model, we are specifying a 3 explanatory variable linear regression model, but there could be literature which is suggesting that well there could be another fourth variable which may have played a role in explaining the variation of Y and it could have determined values of Y but you are not incorporating it, let that variable be denoted by X4.

So, now X_4 is an omitted variable, we are not taking care of that variable in the linear regression model to start with. Now you consider an alternative model where you have the first 3 explanatory variables but now you have one variable, X^* , for which you do not have any observations. So, this X^* you can say that this is basically the omitted variable, it is omitted that is why I am putting an asterisk on the top of X so that I can differentiate it from the other explanatory variables that actually are being used to get the coefficient estimates from raw data. So, X^* is basically the hidden impact of that omitted variable X_4 .

Now we are going to make use of one trick, suppose we do not have data on X^* , even if we know that we are missing out on that variable but we just cannot get data on that particular variable, so what do you do? So, you have to figure out some additional variable which is pretty much linked with the variable that you are missing. So, basically you have to get a proxy variable for X^* .

And now let us assume that there is some variable X_4 which is basically the proxy variable for X^* . So, basically X_4 you can say it is an instrumental variable which is used as an instrument for the true omitted variable X^* . And as they are doing the same job by explaining the variation of Y , they are denoted by the same symbol X_4 . So now you see we have two equations 1 and 2, now we are going to write the final model equation that we are going to estimate.

So, we write the mother regression model and now we are going to make use of this equation 2, so we are going to replace X^* from 1 by the expression in equation 2. So, if we do so, then you finally get an equation that is equation 3. Now note that we land up getting a completely new intercept term. So, here if you look at the second line for the expression for Y , the new intercept term has become $\beta_0 + \beta_4 \delta_1$. And I am introducing a new symbol to represent that in simple manner and that is α_1 .

And now also note that the coefficient for the instrumental variable or the proxy variable X_4 is now product of two different coefficients, the coefficient of original omitted variable, that is β_4 , and it is multiplied δ_2 which is establishing the linkage between the true omitted variable X^* and the proxy variable or instrumental variable X_4 . So, that product is also being denoted by a new symbol α_4 .

So, by estimating these model equation number 3, I mean, we do not get the unbiased estimates for β_0 and β_4 , but we get unbiased estimators for α_1 , β_2 , β_3 and α_4 . Why are we so sure about that? Because now you see a new error term has

been introduced in this model and that is w here if you can look at the equation number 3 again and note what is w ? W is β_4 times e plus ϵ .

Now note that if you take expectation of Y , then β_4 times e , that is going to be 0 because from equation 2 we can see that the e random error term is going to obey all the classical linear regression model properties. And ϵ is anyway obeying the classical linear regression properties. So, if you take expected value of W it is going to be 0.

So, we end our discussion on instrumental variables here. Next time we are going to come back with very nice case study on instrumental variables. Thank you, see you then, bye.