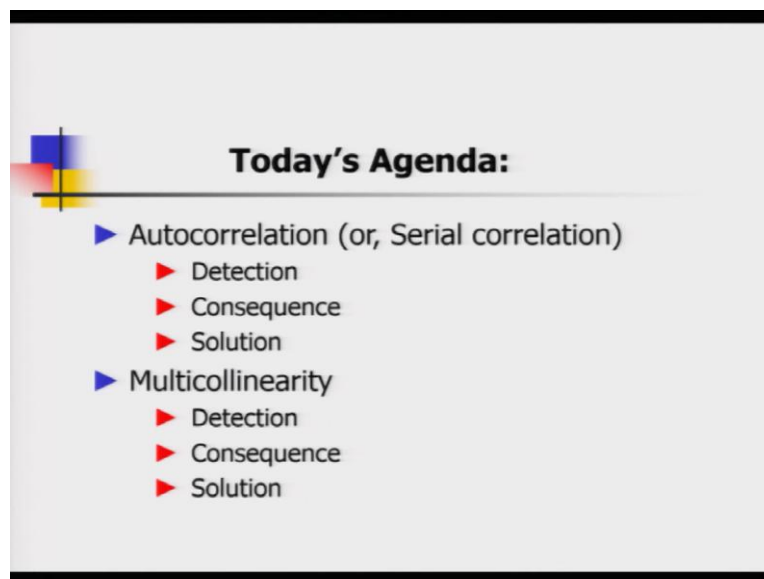


Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur
Lecture No. 31

Violations of CLRM Assumptions (Autocorrelation and Multicollinearity)

Hello friends, welcome back to the lecture series on Applied Statistics and Econometrics. So, today we are going to discuss the last topic in misspecification error and we are also going to talk about another problem sometimes which dilutes the results of linear regression analysis. So, before we start let us have a look at today's agenda items.

(Refer Slide Time: 0:39)



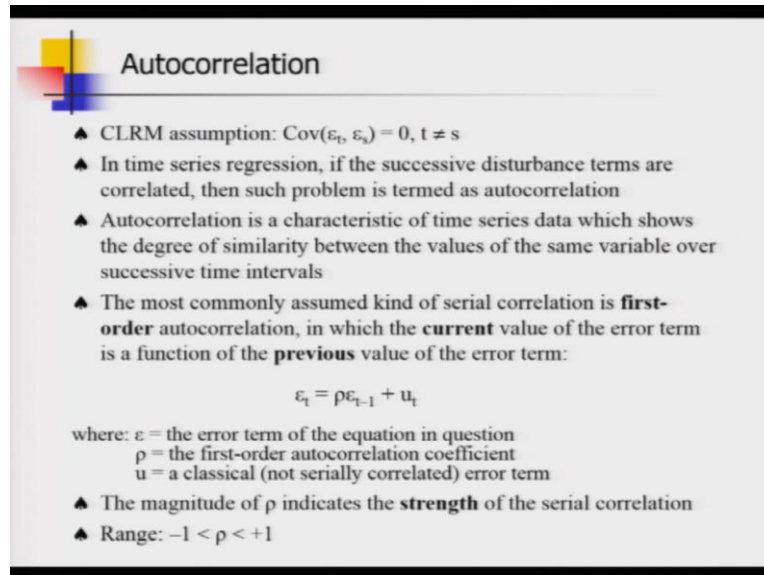
So, today we are going to talk about one kind of model misspecification error which is known as autocorrelation or serial correlation. So, we are going to talk about the detection of the problem, the nature and the consequences of the problem and we are also going to hint about one solution.

Then we are going to talk about a problem called Multicollinearity which cannot be called a model misspecification error as such but definitely when you are running regression, this kind of problem may arise. So, I decided to warn you about it by educating it to you and we will have a very brief discussion on detection and the consequences of multicollinearity problem. Finally, we will also talk about the solution of the problem.

So, if you remember we have assumed some assumptions for the classical linear regression model and one of the assumption says that there should not be any serial correlation between the error term. So, that means that if there is a data point say number fifth and if there is

another data point say number sixth or fourth or forty second, there should not be any correlation between them or in statistical terms the covariance between these two error terms should be equal to 0. So, that is basically the assumption of no serial correlation. And when we have time series data, then this problem is called autocorrelation.

(Refer Slide Time: 2:19)



Autocorrelation

- ▲ CLRM assumption: $\text{Cov}(\epsilon_t, \epsilon_s) = 0, t \neq s$
- ▲ In time series regression, if the successive disturbance terms are correlated, then such problem is termed as autocorrelation
- ▲ Autocorrelation is a characteristic of time series data which shows the degree of similarity between the values of the same variable over successive time intervals
- ▲ The most commonly assumed kind of serial correlation is **first-order** autocorrelation, in which the **current** value of the error term is a function of the **previous** value of the error term:

$$\epsilon_t = \rho\epsilon_{t-1} + u_t$$

where: ϵ = the error term of the equation in question
 ρ = the first-order autocorrelation coefficient
 u = a classical (not serially correlated) error term

- ▲ The magnitude of ρ indicates the **strength** of the serial correlation
- ▲ Range: $-1 < \rho < +1$

So, now we are going to look at this autocorrelation problem from the time series data's perspective. So, let us start with that old classical linear regression model which says that covariance between epsilon t and epsilon s is equal to 0 where t is not equal to s. So, of course, t and s are two different time points on the time scale and in time series what happens this successive disturbance terms are correlated with each other. So, in time series regression if we observe that the successive disturbance terms are correlated with each other, then this problem is called autocorrelation.

And if I want to talk little bit more about autocorrelation through a formal definition, then we can say that autocorrelation is a feature of time series data that shows the degree of similarity between the values of the same variable over successive time intervals. So, in this lecture we are going to talk about the most common form of autocorrelation which is called the first order autocorrelation so that means that epsilon t is correlated with epsilon t plus 1 and epsilon t minus 1. So, the lag between the time periods is considered to be 1.

So, as I already explained in sentences that same thing is represented in bullet number 4, so here I am restating the same thing. So, by first order autocorrelation we mean that the current value of the error term is a function of the previous value of the error term. So, we can write

this particular equation that you are seeing here. So, ϵ_t can be represented as a linear equation as ρ times its old value ϵ_{t-1} plus these U_t .

So, here let me define what are the ϵ of course, we all know so that is the error term of the equation that comes from the original classical linear regression model and then ρ is the first order autocorrelation coefficient. Why it is called autocorrelation coefficient? So, it is a special kind of correlation coefficient which is computed on the same variable, for that we observe a different values from different time periods. So, that is why it is called autocorrelation.

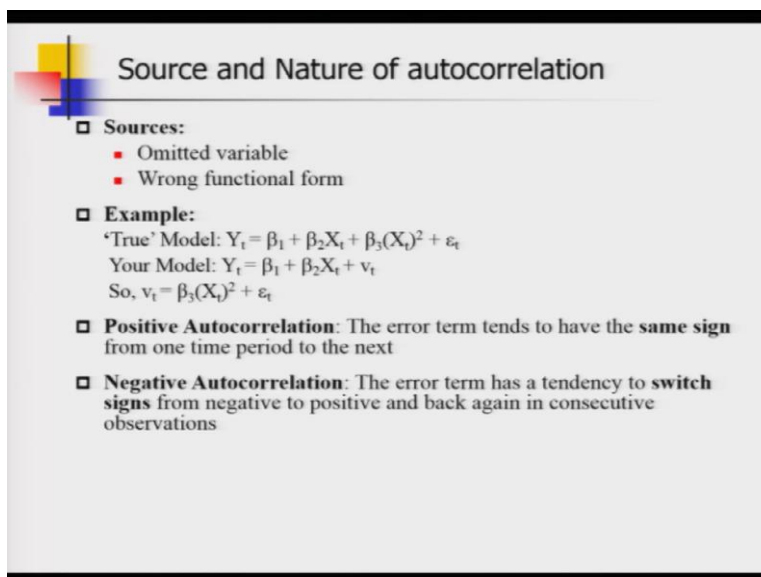
So, if you remember when we had the concept of correlation introduced in the course, we talked about correlation between two distinct variables X and Y. But here we are talking about one single variable which is ϵ here, this stochastic disturbance term and we are interested in the correlation between the past value the present value of the same variable ϵ , that is why it is called autocorrelation.

Now U is basically now another error term and that is a classical error term. So, that means that U is following or obeying all the classical linear regression model assumptions that we make on the disturbance term. So, U is not serially correlated or autocorrelated.

Now this ρ parameter in the equation measures the strength of the correlation between the past value of ϵ and the present value of the ϵ . And as it is a correlation coefficient its value lies between minus 1 and plus 1. So, needless to say that if you figure out ρ is equal to 0, then basically the linear dependence between the past and the present value of same variable ϵ breaks down and you say that there is no correlation between the past and present value of the stochastic disturbance term.

But if you see that ρ is a very high positive number, it is almost close to 1, then you say that there is a very strong dependence between the past value of the error term and the present value of the error term. Seldom we get value of ρ being negative, but theoretically this is possible.

(Refer Slide Time: 6:38)



Source and Nature of autocorrelation

- **Sources:**
 - Omitted variable
 - Wrong functional form
- **Example:**

'True' Model: $Y_t = \beta_1 + \beta_2 X_t + \beta_3 (X_t)^2 + \varepsilon_t$
Your Model: $Y_t = \beta_1 + \beta_2 X_t + v_t$
So, $v_t = \beta_3 (X_t)^2 + \varepsilon_t$
- **Positive Autocorrelation:** The error term tends to have the **same sign** from one time period to the next
- **Negative Autocorrelation:** The error term has a tendency to **switch signs** from negative to positive and back again in consecutive observations

Now, in this slide we are going to talk about the source and nature of autocorrelation. So, what could be the sources of autocorrelation? There could be 2 types of sources, one is omitted variable and the other one is wrong functional form specification. These are all model misspecification issues that we have dealt with before and you are running a time series regression.

So, if you are now missing out on some statistically relevant variable which impacts Y, then that component will be part of the error term and as this particular time series variable is evolving with time, it has a trend in it so it will be there in the epsilon, the random error component of your regression model.

So, itself it will make the that random, your component varying with time and there will be strong correlation between the past value and the present value of the random error term. And this could be the case for wrong functional form specification as well. Suppose there is a variable for which you see there is a quadratic trend after you have plotted the data with respect to time, but suppose you have not plotted the data, you started your regression by fitting a straight line, so you are basically fitting a linear trend, whereas the correct trend is a quadratic one.

So, as you are not fitting a proper polynomial to take care of the non-linear trend in the visible in the data, then that non-linear component of the trend equation will be now part of the error term epsilon. So, again you see that different values of epsilon, its past and present values of they will be correlated with each other because of this non-linear trend component

that is part of the epsilon random error now because you have wrongly specified your functional form.

Now, let us talk about this problem of autocorrelation from a simple example. Suppose we have this true model where the true relationship between Y and X is a quadratic one so you have X_t and X_t^2 both in the regression equation. But suppose you have forgotten to include this X_t^2 variable, so you are estimating a linear version of the model or restricted model which is just linear equation. So, now this term β_3 times X_t^2 will be part of the error of your estimated model.

So, here the new model that you are estimating is having this stochastic disturbance term v_t . I just want to differentiate between these two random errors because they are not the same and so expression for v_t is given also. So, you see that it is composed of the original random error coming from the true model which is ϵ_t and the non-linear part of the actual relationship between X and Y which is β_3 times X_t^2 .

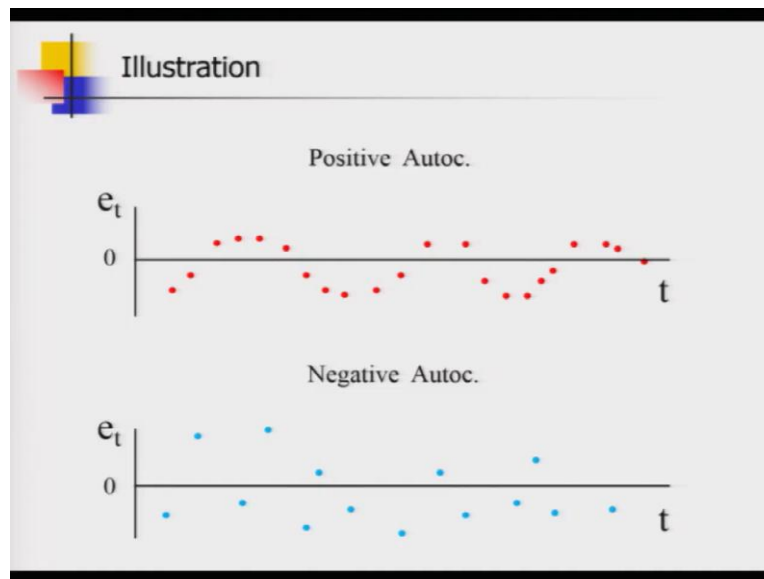
So, now we are going to talk about the nature of autocorrelation problem. So, I have already told you that the measure of autocorrelation which is the rho parameter, it can take value between minus 1 and plus 1. So, if rho takes value between 0 and plus 1, then we can say that there is positive autocorrelation and if rho takes value between 0 and minus 1, then we can call that rho or the autocorrelation is of negative type.

So, what is positive and what is negative, it is going to be clear from the formal definition and I am also going to show you two diagrams which will be the illustration of these two types of autocorrelation.

So, now let us have a look at formal definition for autocorrelation. We have already spoken about positive and negative autocorrelation by looking at the value of the rho parameter and here in this slide I am showing you the formal definition for both case.

So, in the case of positive autocorrelation, the error term tends to have the same sign from one time period to the next and in the negative autocorrelation case, the error term has a tendency to do switch signs from negative to positive and back again in consecutive observations. So, these two definitions will be clear from the pictures that I am going to show here in the next slide.

(Refer Slide Time: 11:34)



So, the first picture in the slide is going to talk about the positive autocorrelation. So, here along the horizontal axis I am measuring different time points and along the vertical axis I am measuring the fitted residuals which is denoted by e_t . So, what is the fitted residual? So, if you have that simple linear regression model based on time series data, say y_t equals to β_1 plus β_2 times X_t plus ϵ_t , you can adopt the OLS technique and then you will get the measures or proxies for the slope and intercept coefficients.

So, after you get those fitted values of \hat{y}_t , then you can detect that fitted value from the original observations y_t and that will give you a measure or proxy or fitted value for ϵ_t these are the realised values of the random variable ϵ_t and they can be denoted by this symbols e_t . So, now here you look at the slide, the first diagram here, there is a sine cosine type wave that we see in the scatter plot. So, different points that you see here in this diagram the red bullets they are corresponding to different time periods.

And as you see that if suppose I start with a negative ϵ_t next ϵ_t is also negative one, then it crosses that 0 line and then it becomes positive, but you see then next 3 e_t s are also positive, then it again goes down below 0 and then there are consecutive 5 time periods for which again I observe $\hat{\epsilon}_t$ or e_t to be negative.

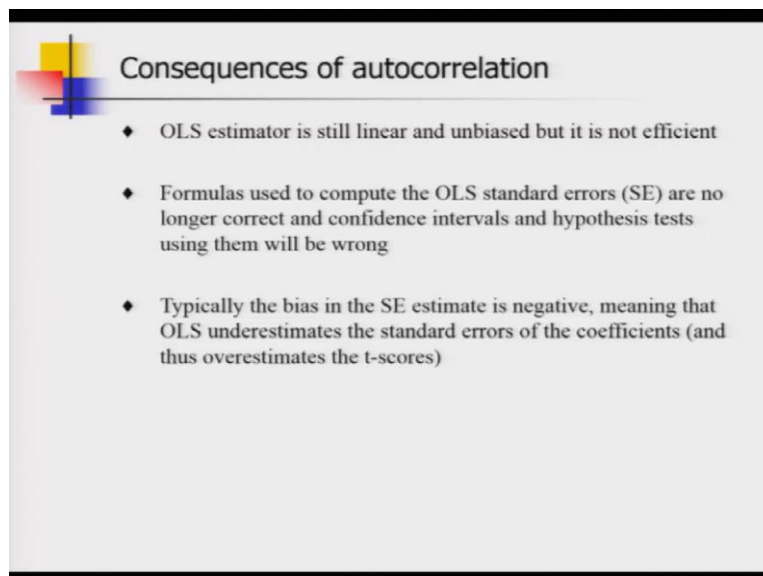
So, actually if you start with a positive number, then you will see a series of positive numbers, then it will fall down close to 0, then eventually it will become negative, then you will see a series of negative numbers, then it will cross the 0 line and then it will be again a series of positive numbers, so there will be a sine cosine type wave of the fitted residual

values that is observed over a time period. In that case you can say that there is positive autocorrelation in the variable.

Now, let us look about the negative autocorrelation case. Here, you see the dots are sky blue coloured bullets and these bullet points are all corresponding to different time periods. Now you see here the pattern is very different from the previous diagram. So, suppose I start with negative value of e_t for time period 1, the next time period 2 I see the next value of e_t is positive and then if I now move on to time period 3, then I see the e_t value if again gone down below 0 and it is a negative one.

But in the fourth time period, it has again risen and now it is a positive number. In the fifth period it is again negative. So, you see it is constantly changing the signs. So, if in one period it is positive, the very next period it is negative, then it again becomes positive and this way if you see the they are constantly switching their signs, then you are convinced that in your data you have a case of negative autocorrelation. So, we have spoken in detail about the detection of autocorrelation from graph.

(Refer Slide Time: 15:30)



The slide is titled "Consequences of autocorrelation" and features a decorative graphic of overlapping colored squares (yellow, red, blue) on the left side. It contains three bullet points:

- ◆ OLS estimator is still linear and unbiased but it is not efficient
- ◆ Formulas used to compute the OLS standard errors (SE) are no longer correct and confidence intervals and hypothesis tests using them will be wrong
- ◆ Typically the bias in the SE estimate is negative, meaning that OLS underestimates the standard errors of the coefficients (and thus overestimates the t-scores)

Now, let us look at the consequences of autocorrelation. So, why are we so much bothered about autocorrelation? If we ignore autocorrelation, then what will go wrong? So, that I am going to list here in terms of 3 bullet points in this slide. So, you see the OLS estimators that you get that you run OLS and you get this coefficient estimates, they are still going to be linear and unbiased, there is no problem with that, but there will be a problem with the variance part and that is why the OLS estimators are not going to be of minimum variance nature.

So, the minimum variance property is breaking down and that is why the OLS estimators is no longer the most efficient estimator. Now, why we are saying that it is not efficient because formulas used to compute the OLS standard errors which are abbreviated as SE are no longer the correct one and the confidence intervals and all the hypothesis tests that use them will give you wrong inference. So, the typically the bias in the standard error estimate is negative, so that means that the OLS underestimates the standard errors of the coefficients and of course, the consequence is that it will overestimate the t scores.

So, basically you will draw wrong conclusion by looking at the t value and the associated probability values. When you are judging whether a particular coefficient associated with a variable is statistically significant or not, you generally tend to get higher t values. So, then there is high chance that you will declare a particular coefficient significant when it is actually not.

(Refer Slide Time: 17:29)

First order autocorrelation

- ▲ Regression Equation: $Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$
- ▲ AR(1) model: $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$ where $-1 < \rho < 1$
- ▲ Assumptions on u_t
 - Mean: $E(u_t) = 0$
 - Variance: $E(u_t^2) = \sigma_u^2$
 - Covariance: $\text{COV}(u_t, u_s) = 0, t \neq s$
- ▲ These assumptions about u_t imply the following about ε_t :
 - Mean: $E(\varepsilon_t) = 0$
 - Variance: $E(\varepsilon_t^2) = \sigma_\varepsilon^2 / (1 - \rho^2)$
 - Covariance: $\text{COV}(\varepsilon_t, \varepsilon_{t-k}) = \rho^k$

Now, in this slide I am going to talk about the most popular and common case of autocorrelation and that is the first order positive autocorrelation. But the theoretical framework that I am going to show you it is general enough so that it can also take care of the negative autocorrelation if you come across one. So, let us start with that same old regression based on time series data that says that Y_t equals to beta 1 plus beta 2 times X_t plus epsilon t.

Now, you propose an autoregressive of order 1 model that is abbreviated as AR(1) model and that model says that you can have another model for your stochastic disturbance term epsilon t that is there in the original regression equation and now you write that this epsilon t is linearly dependent on its past value of the last year. So, the dependence is reflected in the

equation that I have written here. So, here you see that there is no intercept term, but there is this ρ which is the autocorrelation coefficient between the past value and the present value of the variable ϵ_t and this ρ is multiplied with the past value of the stochastic disturbance term which is ϵ_{t-1} .

And of course, there will be some stochastic disturbance term U_t which is following or obeying the classical linear regression model assumptions. So, here you note that I am still not imposing any sign restriction on ρ , so ρ can take any value between minus 1 and plus 1.

Now, as I said that this new random error variable U_t obeys the classical linear regression model assumptions. Let us have a look at those assumptions explicitly because we are going to make use of these assumptions to find out the properties of ϵ_t which is serially correlated.

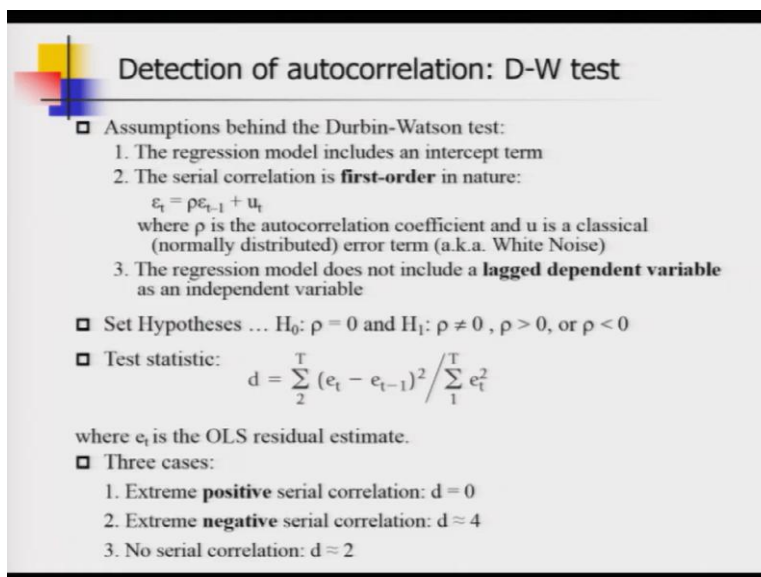
Now, let us first look at the mean, so if the classical linear regression models on the distribution of error terms holding, then we can write expected value for U_t is equal to 0, the variance will be σ^2_U which is constant, it is not varying with respect to time and then we have covariance term $\text{COV}(U_t, U_s)$ where $t \neq s$ equal to 0. So, you see for U_t there is no autocorrelation.

But now let us see how these assumptions on U are reflected in the properties of ϵ . We can show by calculations, I am not showing you the calculations because this is not a theory course and I have not shown proofs for many things. I have limited time so I am not going to derive this thing step by step, I am only going to report the final results.

So, the mean expected value of ϵ_t will be equal to 0, the variance expected value of ϵ_t^2 will be $\sigma^2_\epsilon / (1 - \rho^2)$ and finally the covariance term this $\text{COV}(\epsilon_t, \epsilon_{t-k})$ shall be equal to ρ^k to the power k .

Now remember this covariance thing which is ρ^k , I am going to come back to this thing later when we will be discussing the case of time series data analysis. The modern version of that.

(Refer Slide Time: 21:15)



Detection of autocorrelation: D-W test

- Assumptions behind the Durbin-Watson test:
 1. The regression model includes an intercept term
 2. The serial correlation is **first-order** in nature:
$$e_t = \rho e_{t-1} + u_t$$
where ρ is the autocorrelation coefficient and u is a classical (normally distributed) error term (a.k.a. White Noise)
 3. The regression model does not include a **lagged dependent variable** as an independent variable
- Set Hypotheses ... $H_0: \rho = 0$ and $H_1: \rho \neq 0, \rho > 0, \text{ or } \rho < 0$
- Test statistic:
$$d = \frac{\sum_2^T (e_t - e_{t-1})^2}{\sum_1^T e_t^2}$$
where e_t is the OLS residual estimate.
- Three cases:
 1. Extreme **positive** serial correlation: $d = 0$
 2. Extreme **negative** serial correlation: $d \approx 4$
 3. No serial correlation: $d \approx 2$

So, in this slide again we are going to discuss the detection of autocorrelation via hypothesis testing. So, although we have detected the nature of autocorrelation from graphs, but that is a summary measure that is a graphical summary of the residuals. You have to be statistically sure than indeed there is positive or negative autocorrelation in the residuals.

So, for that you have to make use of a test called or known as Darwin-Watson test. It is a slightly complicated test compared to our old friends like t and f but needless to say whenever you are running time series regression in any statistical package, routinely they report Darwin-Watson's statistic value and the corresponding p value.

So, it is important to understand what the software output actually says, so that is why I am introducing this Darwin-Watson test although it is not in the family of t f chi square tests that we have discussed in the course. First of all note that there will be assumptions to be made if you want to conduct Darwin-Watson test. First assumptions says that the regression model should have an intercept term and I am talking about the original regression model of course.

Then the second assumptions says that the serial correlation is of first order in nature and that is also we have discussed and the third assumptions says that the regression does not include a lagged dependent variable as an independent variable. So, this is a new thing that we are introducing here so let us have a discussion on that. What do we mean by lagged dependent variable? Many times we assume that when we are modelling an economic variable, it is partially explained by its past values.

And sometimes when you are running a regression to model Y_t , you can add Y_t minus 1, Y_t minus 2 as explanatory variables because you believe that as these are the values coming from the past although they are on the same variable, but as they are coming from the past periods, they are truly exogenous, so you can throw them in the regression model.

So, that if you involve this kind of variable in the model which are basically the past values of the variable that you are explaining that is your dependent variable that is called the lagged dependent variable and the values of the lagged dependent variables are the previous data points of the same dependent variable. Now you do not have such scenario if you want to conduct a Durbin-Watson test.

Now, it is time to say hypotheses. So, we will start with the null hypothesis, we want to test null hypothesis that ρ equals to 0 for obvious reason we want to be sure whether there is autocorrelation or not. Autocorrelation if it is there, then it is going to be either negative or positive but it will certainly not be 0. So, null will be set at ρ equals to 0. And the alternative as you say I am showing anything can happen, it could be positive, it could be negative so that I why we can write ρ not equal to 0. So, actually we are proposing a two sided test.

Now, what is the test statistic? So, test statistic is denoted by Durbin-Watson d , a small d . In some textbooks capital D is also used. So, this is the formula, so you see you have to run that original regression and that basically regressing Y_t on X_t and you obtain the parameter and slope coefficient estimates.

Then you obtain the fitted values of Y and then you take the difference Y_t and \hat{Y}_t and that is basically the proxy measure or estimate of that random error ϵ_t you call it e_t that we have discussed also but I am just reminding you again. And this e_t we are going to make use when we are defining our Durbin-Watson d test statistic.

So, here in the formula, you see we are going to make use of that e_t that I just explained how to construct and then there could be 3 possible cases that you can encounter when you are looking at the values of potential values of this d statistic. So, if there is extreme positive serial correlation, then you can get d equal to 0, if you find extreme negative serial correlation, then you can get d almost equal to 4 and if there is no serial or autocorrelation, then you will get a value close to 2. And these are all results coming from econometric analysis.

I am not going to explain why for d equals to 2, you will see no serial correlation and all. I am not going to show you any proof, but later in one of the slides I am going to talk about the relationship between the Durbin-Watson test statistic d and the proxy measure of this autocorrelation coefficient and from there you will get some idea why this d values are taking such values and we are saying that by looking at the value of d we can actually infer whether there is positive or negative or 0 autocorrelation.

(Refer Slide Time: 27:04)

Detection of autocorrelation: D-W test	
☞ Case 1: Set up the test hypotheses and decision rule:	
$H_0: \rho \leq 0$	(no positive serial correlation)
$H_A: \rho > 0$	(positive serial correlation)
if $d < d_L$	Reject H_0
if $d > d_U$	Do not reject H_0
if $d_L \leq d \leq d_U$	Inconclusive
☞ Case 2: Set up the test hypotheses and decision rule:	
$H_0: \rho = 0$	(no serial correlation)
$H_A: \rho \neq 0$	(serial correlation)
if $d < d_L$	Reject H_0
if $d > 4 - d_L$	Reject H_0
if $4 - d_U > d > d_U$	Do Not Reject H_0
Otherwise	Inconclusive

Now, in this slide we are going to continue with that Durbin-Watson test. So, we are going to talk about two different cases here and case 1 says that we are going to test for positive serial autocorrelation and then you set up your null hypothesis saying that well, ρ could be negative or 0 and alternative H_A you can say that ρ is positive and then you translate this hypothesis in terms of the decision rules so that you can decide whether to go with the null hypothesis or the alternative hypothesis.

So, you reject null hypothesis if you observe that your calculated value of d is less than the tabulated value d_L so there is a Durbin-Watson test statistic table and from there you can read the critical values. So, if the d value is less than d_L , then you reject your null hypothesis and you do not reject null hypothesis if you observe your d is greater than d_U . But if the calculated value lies between d_L and d_U , then the test result is inconclusive. So, basically what happens, Durbin-Watson test statistic table is fairly complicated one.

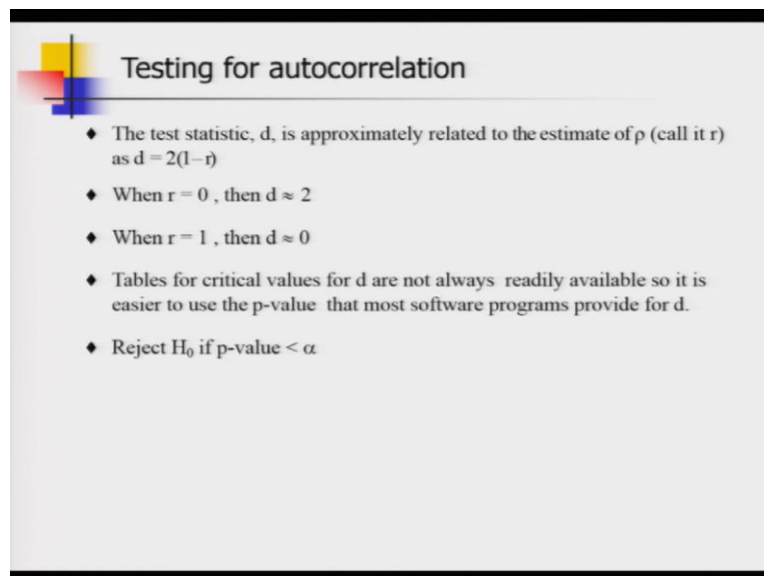
So, for each case they are reporting two numbers d_L and d_U and upper limit and one lower limit. So, if you are conclusively in one of the regions where the calculated value is either greater than the d_U value or lower than the d_L value, then you are safe. So, then you can say

that I can make a decision, but if you fall in the middle zone, so which is the inconclusive zone, so, in that case you cannot make a decision. So, this is very different from what we have encounter so far in the course because always we were able to make conclusive decisions by conducting t test, f test or chi square test, but that is not the case with Darwin-Watson.

So, I do not want to spend more time on Darwin-Watson because this is a very common test in econometrics but it is not very straight forward to understand so I skip by showing this slide once more to you and talking about the second case, but I am also going to tell you that you do not worry about this critical values and so much, there is an easier way out to read the software output and take a call whether you will go with null hypothesis or you will not go with the null hypothesis.

So, now let us look at case 2 which says that I am going to go for a two tail test. So, although I am saying that my null hypothesis is of no autocorrelation or no serial correlation, but my alternative hypothesis says that there could be either positive or negative serial correlation. And again I am showing you the value ranges for the regression rule and let us now move on to the second slide where I am going to talk about a simpler way to handle the Darwin-Watson test because this is very difficult to remember.

(Refer Slide Time: 30:38)



Testing for autocorrelation

- ◆ The test statistic, d , is approximately related to the estimate of ρ (call it r) as $d = 2(1-r)$
- ◆ When $r = 0$, then $d \approx 2$
- ◆ When $r = 1$, then $d \approx 0$
- ◆ Tables for critical values for d are not always readily available so it is easier to use the p-value that most software programs provide for d .
- ◆ Reject H_0 if $p\text{-value} < \alpha$

So, in this slide I am going to continue the discussion on the Darwin-Watson test. So, here I am now going to show you relationship between the Darwin-Watson test statistic and that proxy for autocorrelation coefficient. So, if you can get an estimate for the unknown parameter value rho, suppose we call it small r , then it can be written that there is a linear

relationship between d and r and an equation is shown here. So, small d equals to 2 times 1 minus r .

Now, you see from this particular relationship between Darwin-Watson d and proxy measure for autocorrelation coefficient we can actually figure out different values of d statistic for different cases of autocorrelation. So, let us now first look at the case of 0 autocorrelation. So, if there is 0 autocorrelation we expect the r will take on the value 0 so if you plug 0 in that equation, you are going to get 2. And if you plug the value of r equals to 1 that means that there is perfect positive autocorrelation, then you are going to get a d value of 0.

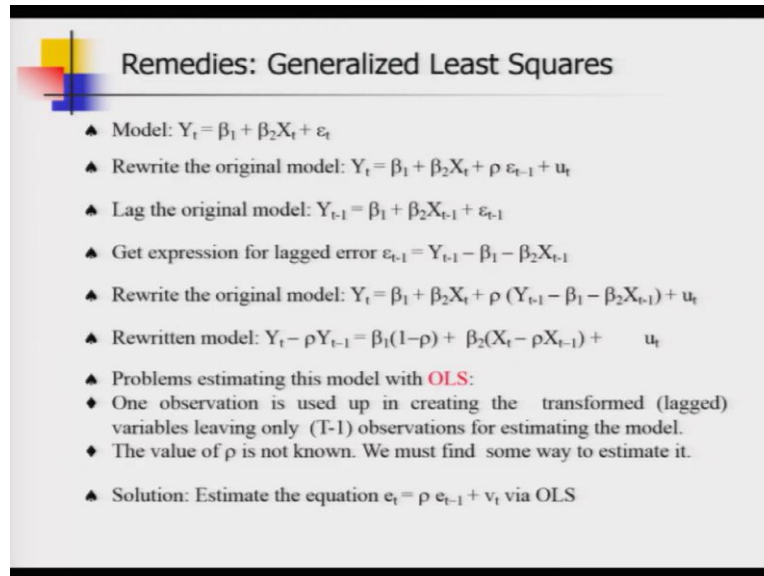
So, by looking at the calculated d statistic you can actually infer what is there in the data. So, if you get a value of d equal to 2, then you know or if you get a value very close to 2, then you know that there is no autocorrelation that exists in the residuals. So, your time series regression is autocorrelation free. And if you find d very close to 0, a very small number, then you can be sure of the fact that there is very strong positive autocorrelation and in fact, it could lead to a value of autocorrelation coefficient equals to 1.

Now, we are going to talk about an easier way to handle Darwin-Watson. Now Darwin-Watson is a very important test and it is routinely done by all statistical software when you are running a time series regression. So, those softwares are reporting the p value for the test statistic and you forget about those critical values like d_L , d_U because it is difficult to remember and not only it is difficult to remember, the Darwin-Watson table is also very limited, so if you consult an econometrics textbook, you are not going to get all possible cases of d_L d_U values.

So, if you get a case which is relevant for your statistical investigation, you may not get a particular combination of d_L and d_U and then take a decision. So, the best way to get around this d_L d_U issue is by looking at the p value and then we all know how you are going to make a decision.

So, that we have learnt in the case of hypothesis testing and we have applied the same rule for t test, chi square test, f test everything that we have done, so it is not new. So, we are going to reject the null hypothesis if you observe a p value which is less than the chosen level of significance α .

(Refer Slide Time: 34:22)



Remedies: Generalized Least Squares

- ▲ Model: $Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$
- ▲ Rewrite the original model: $Y_t = \beta_1 + \beta_2 X_t + \rho \varepsilon_{t-1} + u_t$
- ▲ Lag the original model: $Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + \varepsilon_{t-1}$
- ▲ Get expression for lagged error $\varepsilon_{t-1} = Y_{t-1} - \beta_1 - \beta_2 X_{t-1}$
- ▲ Rewrite the original model: $Y_t = \beta_1 + \beta_2 X_t + \rho (Y_{t-1} - \beta_1 - \beta_2 X_{t-1}) + u_t$
- ▲ Rewritten model: $Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + u_t$
- ▲ Problems estimating this model with **OLS**:
 - ◆ One observation is used up in creating the transformed (lagged) variables leaving only $(T-1)$ observations for estimating the model.
 - ◆ The value of ρ is not known. We must find some way to estimate it.
- ▲ Solution: Estimate the equation $e_t = \rho e_{t-1} + v_t$ via OLS

So, suppose your Darwin-Watson and the graph tells you that there is autocorrelation what will you do? So, will you not run any regression involving time series data? Absolutely, not. There are some solutions of you come across this autocorrelation problem and we are going to talk about one type of solution because we do not have time. I have only one lecture devoted to this particular problem, so I will not cover the other solution that one can use. But here the solution I am going to show you is a very popular one and it is called generalized least squares method.

So, you start with the original regression model and that is Y_t equals to β_1 plus β_2 times X_t plus ε_t . Now, you rewrite the original model by using that assumption that we made that AR(1) assumptions. So, that says that the random error term, ε_t that is there in the regression model that follows an AR(1) process. So, we plug that expression for or in place of ε_t and we get this rewritten original model.

Now you lag the original model by one period. So, by lagging means that we have to go to the past. So, you see the t subscript is now changed to t minus 1 as we are going 1 period back in the past. So, similarly, we have to make that adjustment for the variables, but not for the coefficients, so you see X_t becomes X_{t-1} and ε_t will become ε_{t-1} . Now you get an expression for the lagged error term ε_{t-1} and that is simple you have to just switch expression from one side to the other.

Now you plug these lagged error ε_{t-1} term that you obtained in fourth bullet point, you take that expression and place that in the expression or equation that you see in the second bullet point. So, you see after ρ there is ε_{t-1} and ε_{t-1} is not

known to us so we are finding a proxy for ϵ_{t-1} and we are replacing ϵ_{t-1} by that proxy expression. So, now you rewrite the original model and it becomes somewhat complicated looking. So, now you see Y_t equals to $\beta_1 + \beta_2 X_t + \rho Y_{t-1} - \beta_1 - \beta_2 X_{t-1}$.

And then finally you have this new random error which is U_t which obeys the classical linear regression model assumptions. Now, you rewrite it further, what you do, you take the y values in one side. So, now you have $Y_t - \rho Y_{t-1}$ equals to $\beta_1 + \beta_2 X_t - \rho X_{t-1} + \epsilon_t$. Of course plus ϵ_t . Now you see what it is happening, the intercept has changed to a new thing, the dependent variable y is in some kind of a difference form and the explanatory variable X is also in some kind of a difference form.

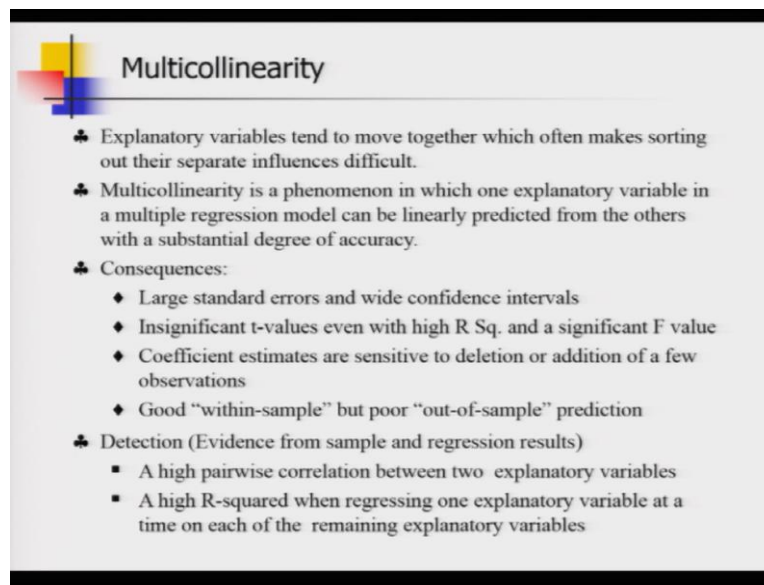
So, now if we want to run OLS with these data, can we do it? Although we have done all these algebra, but still we cannot apply OLS on this particular equation. There are two different reasons why we cannot do OLS here. So, one is the observation is used up the first one is a pity problem so you lose out on one observation that is eaten up in creating the transformed or lagged variables leaving only $T - 1$ observations to run your OLS model, but it is a pity problem, if you have say 40 years of data or 40 months of data, then you have a large enough sample because the sample size is greater than 30 and you can sacrifice one observation and then conduct OLS that is not a big problem.

But well, it is a big problem when you have only 10-12 years of data because then losing out on one observation, on observation means that you are sacrificing a lot of information from the data. But the more serious problem is that the value of ρ is not known because ρ is an unknown population parameter. We do not have much idea about ρ so without knowing the value of ρ , how can I calculate these variables, these transformed variables like $Y_t - \rho Y_{t-1}$ and $X_t - \rho X_{t-1}$.

So, we have to get a proxy measure for ρ and there are ways to get proxy measure for ρ and we are going to discuss one such measure. So, we have to go back to that original model and you have to run the linear regression via OLS method, you have to generate the residuals and these residuals are e_t , so these are $Y_t - \hat{Y}_t$. And now you lag these residual by 1 period so you get e_{t-1} and then you regress these residuals e_t on its one period previous value that is e_{t-1} .

So, you will get a coefficient from that linear regression and that is going to be rho and if you estimate this regression equation via OLS method, you are going to get rho hat, a proxy measure and estimate for the unknown population parameter rho and that we have already called r and you can now make use of this coefficient estimate and you can transform the variables and then you are all set to run your OLS on the transformed model.

(Refer Slide Time: 40:54)



Multicollinearity

- ♣ Explanatory variables tend to move together which often makes sorting out their separate influences difficult.
- ♣ Multicollinearity is a phenomenon in which one explanatory variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.
- ♣ Consequences:
 - ◆ Large standard errors and wide confidence intervals
 - ◆ Insignificant t-values even with high R Sq. and a significant F value
 - ◆ Coefficient estimates are sensitive to deletion or addition of a few observations
 - ◆ Good “within-sample” but poor “out-of-sample” prediction
- ♣ Detection (Evidence from sample and regression results)
 - A high pairwise correlation between two explanatory variables
 - A high R-squared when regressing one explanatory variable at a time on each of the remaining explanatory variables

So, now in the last slide of today’s lecture I am going to focus on multicollinearity. Now, multicollinearity is not a misspecification issue as I have already told you when we began today’s lecture. It is basically a data issue. So, sometimes you may have a regression model specified where you have some explanatory variables which are related to each other.

In socioeconomic context it, many explanatory variables are somewhat related to each other, but the problem is that the mild relationship is okay but if the relationship is very strong and if you keep both, then that becomes a problem and that actually leads to the multicollinearity problem.

So, here we are going to talk about the formal definition of multicollinearity, we are going to talk about how to detect multicollinearity and the consequences of the multicollinearity finally I am going to end the discussion by pointing out some solutions. So, multicollinearity can be formally defined: it is also sometimes called collinearity in some textbooks. So, it is a phenomenon in which one explanatory variable in a multiple linear regression model can be linearly predicted from the others with a substantial degree of accuracy.

So, what does it say? So, it says that suppose you have 1 explanatory variable X_1 and there is another explanatory variable X_3 and you suspect that X_1 could be modelled through X_3 because there is strong association between these two variables. So, if you run X_1 on X_3 , so if you run OLS regression of X_1 on X_3 , then you are going to see a very significant f value, a very significant t statistic value and a very good r square value and if that happens, then you have a multicollinearity problem.

So, here I have already told you how to detect multicollinearity problem, so let us have a look at the slide and then continue the discussion. So, coming back to the detection issue, there is another way to detect multicollinearity problem and this time you do not have to run a regression, you can compute the Pearson correlation coefficients between different explanatory variables and if you notice high pair wise correlation between two explanatory variables, then you can suspect that there could be multicollinearity there in your regression model.

Now, what are the consequences of multicollinearity? Well, I am going to report some of the theoretical findings, I am not going to provide much justification or any proof, but I think you should remember this. So, here under the third bullet point, I am showing you these 4 sub bullet points which are talking about the consequences of multicollinearity if it exists at all. So, first of all you are expected to get large standard errors and wide confidence intervals. So, related consequences is that you will get insignificant t values even with high R Square values and significant f value.

And thirdly, coefficient estimates are sensitive to deletion or addition of a few observation. And last but not the least, the if multicollinearity exists, then my OLS estimators are good in prediction when you are doing that in in the sample. So, basically you are talking about the within-in sample prediction, but generally they do a very poor job when you are going for out of sample prediction.

So, now let us talk about the solution for these multicollinearity problem. now, it is not a kind of model misspecification issue that will require a serious bit of adjustment in the model and you do not have to change the model's functional form or you maximum you can do you can drop one particular collinear variable that is strongly related with another explanatory variable, but sometimes you may not have to even drop the explanatory variable because if you think that that variable is important and you should keep it inside the regression, then what you need to do?

You need to get more data because sometimes when you increase the sample size and you get more data, this association between these two potential collinear variables they become diluted. So, in that case you will not have a multicollinearity issue. So, actually what happens, you should look at the correlation matrix first before you run your regression and then you look at the values of the Pearson correlation coefficient, if the absolute value of the Pearson correlation coefficient is less than 0.775, then I think you do not have to worry about multicollinearity problem.

It becomes a serious problem when this absolute value of Pearson correlation coefficient is above say 0.75 or 0.8. So, that is at, that is what we have observed in empirical research. So, this is it for the time being. So, I end today's discussion by giving hint what we are going to discuss in the next two lectures for this week.

So, next two lectures are going to be devoted to modern time series data analysis. We are going to talk about method which is very widely known when dealing with time series data that is called Box and Jenkins method. So, in relation to that we are going to also talk about some regression models. So, see you then, thank you.