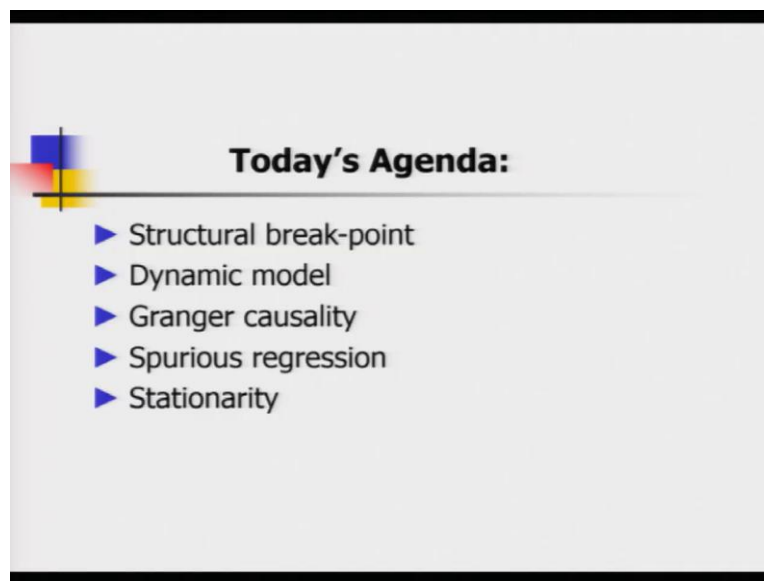


**Applied Statistics and Econometrics**  
**Professor Deep Mukherjee**  
**Department of Economic Sciences**  
**Indian Institute of Technology, Kanpur**  
**Lecture No. 32**  
**Time Series Regression with Stationary Data**

Hello friends, welcome back to the lecture series on Applied Statistics and Econometrics. So, today we are going to start our discussion on a new topic and we are going to devote to two consecutive lectures to modern times series econometrics. But before we study this modern time series econometrics let us have a look at today's agenda item.

(Refer Slide Time: 0:38)



So, first I will start with an important aspect of time series regression which is called structural brake point analysis and I am going to show you however good old friend f test can be useful to detect structural breaks. And next we are going to talk about dynamic model and we know by dynamic model I mean that we are dealing with dynamism in time series data regression analysis.

And the third we are going to talk about a very important concept in modern time series econometrics as proposed by Nobel Laureate Clive Granger that is known as Granger causality. And next we are going to discuss what is spurious regression and finally we are going to end todays discussion with introducing the notion of Stationarity.

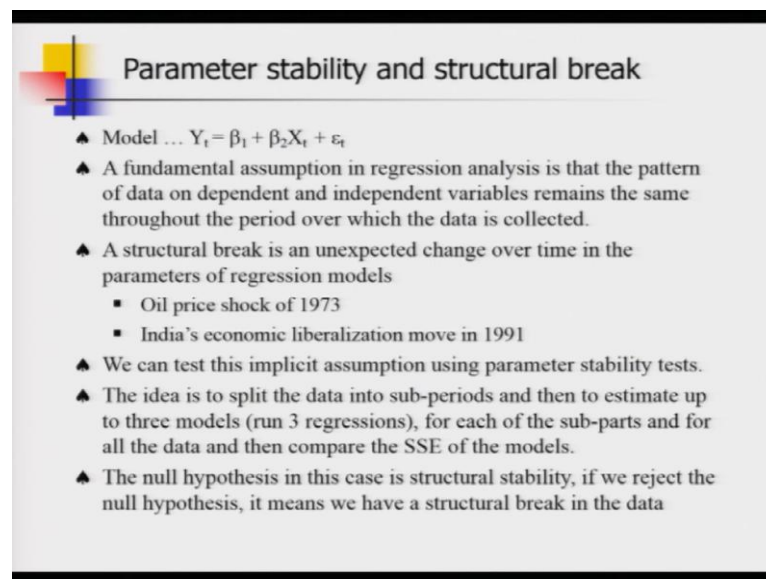
So, when we conduct simple linear regression analysis with time series data or why time series, it can be even cross section data. We always make an assumption that the parameters;

the intercept and the slope coefficient or the parameters are going to remain the same for the entire range of Y and X.

So, for the sample has it come from cross section or has it come from time series it dose not matter when you are fitting regression line using some data points you are always assuming that throughout the range of the data in the sample the coefficient values are going to be the same.

But there could be an alternative argument that if you have time series data set which is scanning over a long time, in the middle there could be some changes which could have changed the trend, the overall long run trend or pattern in the data and it may not be then rational to assume that the slope and the intercept coefficients are going to remain the same throughout the entire study period. So, basically we are talking here about there is some point where the values of the slope and intercept coefficients or the parameters they are going to change. So, this is called as structural break.

(Refer Slide Time: 3:08)



**Parameter stability and structural break**

- ▲ Model ...  $Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$
- ▲ A fundamental assumption in regression analysis is that the pattern of data on dependent and independent variables remains the same throughout the period over which the data is collected.
- ▲ A structural break is an unexpected change over time in the parameters of regression models
  - Oil price shock of 1973
  - India's economic liberalization move in 1991
- ▲ We can test this implicit assumption using parameter stability tests.
- ▲ The idea is to split the data into sub-periods and then to estimate up to three models (run 3 regressions), for each of the sub-parts and for all the data and then compare the SSE of the models.
- ▲ The null hypothesis in this case is structural stability, if we reject the null hypothesis, it means we have a structural break in the data

So, a structural break can be formally defined as an unexpected change over time in the parameters of the regression models. Now here I am going to show you two instances where this kind of huge macroeconomic shocks or changes can lead to structural break in the economy.

So, if you are talking about that oil price shock of 1973 that actually shook the world and not only it shook the US economy, but in many other economies where hard hit by this oil price

shock of 1973. So, one can expect that there could be some change in the parameter values or regression coefficients before and after 1973.

Now, from India we can talk about one such instance in 1991 the England government rolled out economic liberalization and that was a huge economic policy change and it was pre-meditated not like the oil price shock which shook the entire world. This economic liberalization of India changed Indian macroeconomic conditions only. So, one off course can assume that before and after 1991 the coefficient of time series regression models may have changed. So, this two are very good potential cases of structural brake points.

So, we can test this implicit assumption that parameter values are remaining stable over the entire range of data and for that we can make use of f test. So, let me talk about the philosophy beforehand and then we will go to the steps of this f test So, here is the philosophy.

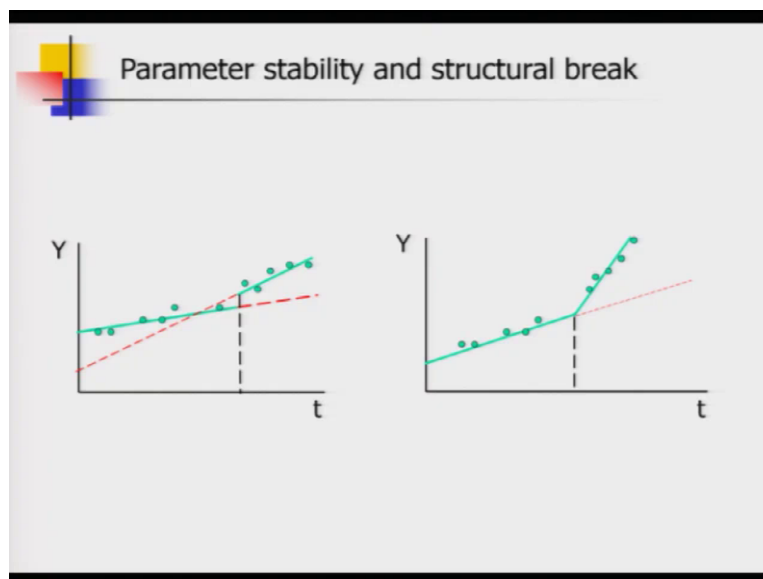
So, you see a dataset and then you that they are could be a structural break point in the middle of the data, it must not be in the exact middle of the data somewhere in the data and you suspect that there could be heterogeneity in slope and the intrinsic parameters around there particular time point.

So, basically you are talking about running to different regressions So, you call that the period which is after the structural break point all the potential, the structural break point you call that period 2 and the time period which is before that structural break point you call that pre break point time period or period 1. So, you have to run two separate regressions and of course, you are going to gate two different sets of intercept and slope coefficient values.

Now, you are statistically test whether these set of coefficients that you obtain from your period 1 or pre break point regression is statistically different from the coefficients said that you have obtain from period 2 or post break point period regression. So, in total, to conduct a test, we are going to run 3 different regression models, one is for the entire dataset and one is for the pre break point dataset and one for the post break point dataset and for each of the sub parts and for the combined data, we are going to compare the sum of square errors of the regression models.

And what could be the null hypothesis? The null hypothesis in this case is structural stability that is the statusco and if we reject the null hypothesis, then that means that we have a structural break in the time series data.

(Refer Slide Time: 7:07)



Now here, in this slide I am going to show you some graphs which will make this parameter stability and structural break much clear in your mind. So, here I am showing you two different cases, let us focus on the first diagram which is at the left hand side and here you see I am measuring time along the horizontal axis and the value of the economic variable along the vertical axis and these green circles, these are the data points for different time periods and you see that actually one can suspect that there is some break point in the time series at some time period or some time point such that there could be two regressions lines passing through this scatter plot.

And if why fit two regressions lines, the sum of square errors actually will be low. So, in other words the fit will improve. So, here you see that the green straight lines are actually talking about the fitted straight lines which are basically fitted based on period 1 and period 2 data points only and of course, they could be extended towards both side and that extensions are marked with red coloured dashed lines.

Now, we move to the second diagram. The second diagram also tells us the same story but with some part difference because if you compare the first diagram with the second one, the only difference is that here the way these two straight lines for period 1 and period 2 are model such that there is a continuity here.

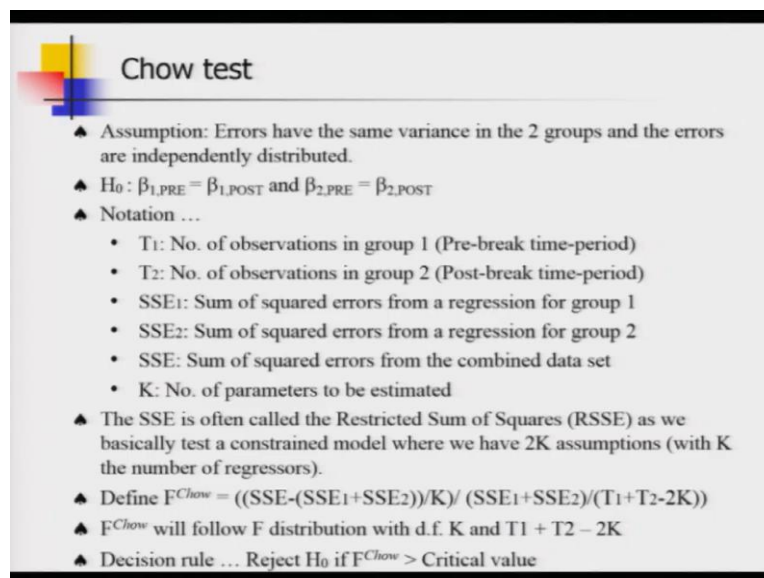
So, there is a kink point and after that the slope changes, but there is a continuity. So, if someone is interested to find out the value of the dependent variable Y at the exact break point, then some value could be generated from the spline regression. This is called spline regression.

We are not going to cover spline regression. In this course, but I am just coining the term, I am introducing you to this new term so that if you are interested you can study it later. But in the first diagram you see there is a gap between these two green straight lines, so for a particular time period  $t$  naught where you are suspecting structural break point, for that actually the value of  $Y$  is not defined or you can report two different values, one coming from the regression line for period 1 and one from the period 2.

Now, note that how do you know that there is a structural break point? So, many times your investigation for existence of structural break point is guided by the economic information. So, if you know that some major thing has happened, then you can say that when that particular event happened there could be some structural break around that time point.

But this is quite arbitrary of course, so ideally speaking you should test for several break points around that particular data when you know that something major event has happened which could have changed many many economic functions.

(Refer Slide Time: 10:41)



**Chow test**

- ▲ Assumption: Errors have the same variance in the 2 groups and the errors are independently distributed.
- ▲  $H_0: \beta_{1,PRE} = \beta_{1,POST}$  and  $\beta_{2,PRE} = \beta_{2,POST}$
- ▲ Notation ...
  - $T_1$ : No. of observations in group 1 (Pre-break time-period)
  - $T_2$ : No. of observations in group 2 (Post-break time-period)
  - $SSE_1$ : Sum of squared errors from a regression for group 1
  - $SSE_2$ : Sum of squared errors from a regression for group 2
  - $SSE$ : Sum of squared errors from the combined data set
  - $K$ : No. of parameters to be estimated
- ▲ The SSE is often called the Restricted Sum of Squares (RSSE) as we basically test a constrained model where we have  $2K$  assumptions (with  $K$  the number of regressors).
- ▲ Define  $F^{Chow} = ((SSE - (SSE_1 + SSE_2)) / K) / ((SSE_1 + SSE_2) / (T_1 + T_2 - 2K))$
- ▲  $F^{Chow}$  will follow F distribution with d.f.  $K$  and  $T_1 + T_2 - 2K$
- ▲ Decision rule ... Reject  $H_0$  if  $F^{Chow} > \text{Critical value}$

Now, we are going to talk about the statistical testing to determine whether there is a break point or not at some supposed point. So, what you have to do before you conduct Chow test, you have to first identify a potential or a set of potential break point time periods and then you have to conduct the Chow break point test. So, if you have one time period in mind, then you have to run Chow test only once but if you have multiple break points in mind, then you have to run multiple Chow break point times, one for each potential break point.

But note that Chow test also comes with some assumption and I think this is quite rigid and strong assumption, but nevertheless Chow test is very popular in applied econometric world, so we are covering this Chow test with this association clearly stated at the very beginning. So, Chow assumes that errors have the same variance in both groups  $p$  and post break point and the errors are independently distribution.

So, when we say that errors have same variance in two time periods  $p$  and post break point, then actually we are referring to the assumption of homoskedasticity. And when we are saying that the errors are independent of each other, then we are referring to the autocorrelation assumption. So, here Chow assumes that there is homoskedasticity among the errors and he is also assuming that there is 0 serial or autocorrelation.

Now it is time to set the null hypothesis to move on. So, our null hypothesis says that  $\beta_1$  pre is equal to  $\beta_1$  post. So,  $p$  and post are talking about two time periods before and after the break point respectively and similarly we can also set a null hypothesis for  $\beta_2$  pre equals to  $\beta_2$  post. Same interpretation is valid as we interpreted the case of  $\beta_1$  and note that these are not two different hypothesis, this is basically a joint hypothesis because if there is no structural change, then actually both will occur simultaneously.

So, what do I mean? I say that then  $\beta_1$  will be the same in pre and post break point periods and  $\beta_2$  will also be same in the same and post break point periods. If at least one of them differs from one period to the other, then we can say there could be some structural change. So, we are actually conducting a joint hypothesis testing here. And of course, to solve a joint hypothesis testing, we are going to make use of  $f$  test.

But before we write down our  $f$  test statistic, let us introduce some notation and of course, later on you will see that the test statistic is somewhat different from what we have seen earlier. So, for that I am introducing new notations. So,  $T_1$  is the number of observations in group 1 which is the pre break time period.  $T_2$  is the number of observations or data point in group 2 which is the post break time period.  $SSE_1$  is the sum of squared errors from regression from group 1.

$SSE_2$  is sum of squares errors from a regression for group 2 and then there is a combined  $SSE$  which gives me the sum of squared errors from the combined dataset and  $K$  is the number of parameters to be estimated. So, if there are small  $k$  number of explanatory variables and one intercept, then this capital  $K$  will be equal to small  $k$  plus 1. This  $SSE$

which is estimated from the combined data is often called a restricted sum of squares. As we basically test a constrained model where we have  $2K$  assumptions imposed and this is because we have  $K$  number of regressors.

So, what restriction actually we are imposing? When we are imposing this restriction on the entire dataset, we actually mean to say that the parameters are remaining stable in this two data time period. So, that is the restriction that we are going to impose when we are combining the entire dataset and running one regression.

Now, I am going to define the  $F$  statistic which is somewhat different from the previous  $f$  statistics. It is called  $F$  Chow and the formula is given here and you know how to compute  $SSE$ , so I think it will not be very difficult for you to compute this value of the  $F$  Chow test statistic.

Now,  $F$  Chow test statistic will follow an  $F$  distribution with two degrees of freedom capital  $K$  and  $T_1$  plus  $T_2$  minus  $2K$  and you know the decision rules, so you reject null hypothesis if the calculated value of  $F$  which is  $F$  Chow is greater than the critical value found from the  $F$  table.

Now, we are going to talk about dynamic econometric models. So far we have dealt with static model, so it does not matter whether we are dealing with cross section data or time series data. We always have a model where the past values of any variable is not affecting the current dependent variable which we are trying to explain.

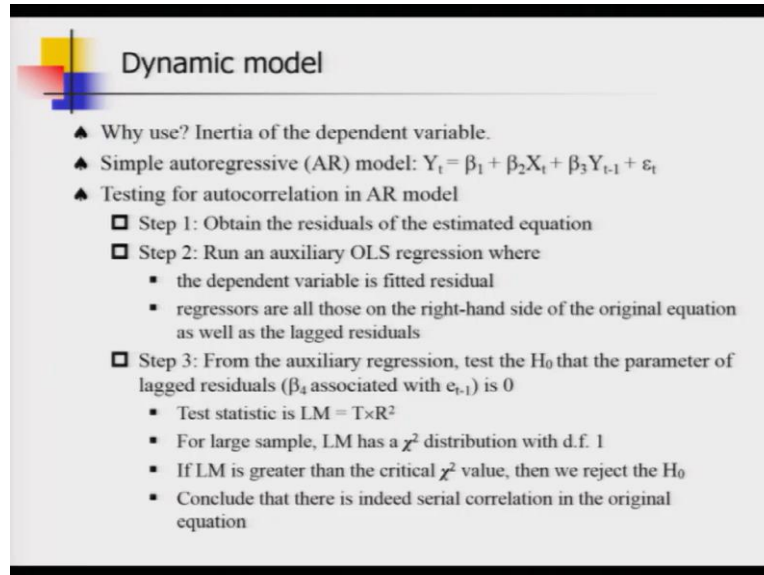
But if you make use of the past values of the dependent variable to explain the variation in the dependent variable or you want to predict the future value of the dependent variable  $Y$ , then actually you have a dynamic econometric models. So, you are making use of the lagged dependent variables.

And this modelling is also very popular in the fields of macroeconomics, international trade, development, growth, so I am not going to talk about a lot about this models because of course, in 50 minutes time, I cannot do justice to all sort of models if I introduce them one by one.

But I just want you to expose to new areas which you can venture yourself if you are truly interested in this particular idea and if you find useful for your research projects or your business projects, so it is not a bad idea to have a overall view of the major time series

models, so with that objective I am going to briefly talk about dynamic econometric model as well.

(Refer Slide Time: 18:03)



**Dynamic model**

- ▲ Why use? Inertia of the dependent variable.
- ▲ Simple autoregressive (AR) model:  $Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + \varepsilon_t$
- ▲ Testing for autocorrelation in AR model
  - Step 1: Obtain the residuals of the estimated equation
  - Step 2: Run an auxiliary OLS regression where
    - the dependent variable is fitted residual
    - regressors are all those on the right-hand side of the original equation as well as the lagged residuals
  - Step 3: From the auxiliary regression, test the  $H_0$  that the parameter of lagged residuals ( $\beta_4$  associated with  $e_{t-1}$ ) is 0
    - Test statistic is  $LM = T \times R^2$
    - For large sample, LM has a  $\chi^2$  distribution with d.f. 1
    - If LM is greater than the critical  $\chi^2$  value, then we reject the  $H_0$
    - Conclude that there is indeed serial correlation in the original equation

Now, why do we use dynamic econometric model? So, every economic variable has some inertia and if we stuck at the starting model, then that inertia is not going to be modelled. So, if you want to model that inertia, then actually you have to make use of the dynamic econometric models. Dynamic econometric models could be of 2 actually 3 major types and although I am going to get into detail, but it is not a bad idea to know what are the main types of dynamic econometric model there in the literature.

So, one is called autoregressive model, it is abbreviated as AR and what is it I am going to show you in the next slide and then there is distributed lag model, the abbreviation is DL and then finally there is a model which is combination of the first two so autoregressive distributed lag ARDL model. So, in this particular lecture we are going to have a glimpse of these types of dynamic models. So, let us first have a look at the autoregressive model.

So, now we are going to deal with a very simple, I think the simplest possible autoregressive model which says that the current value of the dependent variable Y depends on the current value of the explanatory variable X and the last periods value of the dependent variable Y. So, you can write a linear regression equation as  $Y_t$  equals to  $\beta_1$  plus  $\beta_2$  times  $X_t$  plus  $\beta_3$  times  $Y_{t-1}$  plus epsilon.

So, why this is called autoregressive? Because you see Y is regressed on its own previous or past time period's values. So, that is why it is called autoregressive and why it order 1?



Because you are just making use of the previous period. So, if I annual data and if you go back 2 time periods and if you have a both  $Y_t$  minus 1 and  $Y_t$  minus 2 as explanatory variables or regressors in the regression equation, then actually you will have an AR(2) model.

Now, of course, in the last lecture only we have dealt with the case of autocorrelation. So, when you are running a time series regression, autocorrelation could be big big issue. So, in this dynamic econometric model also we should check for autocorrelation. It does not matter, do not think that as we are adding the past values of the dependent variable  $Y_t$  minus 1 or  $Y_t$  minus 2, we can get rid of the autocorrelation problem. So, autocorrelation must be checked to see whether you got a sound regression result or not.

So, how do we test for autocorrelation? Now, note that in the last lecture only we have learnt about a test and that is the Durbin-Watson test. Can I make use of that test here? The answer is no. Why? Because if you remember from last lecture I have told that Durbin-Watson makes a couple of assumptions, and one of the assumptions says that you cannot make use of the lagged dependent variable as a regressor in the regression equation for which you are testing autocorrelation.

But here, this is precisely what we are doing. So, there must be some solution to this problem, if we want to test for autocorrelation in this dynamic econometric models. And fortunately there is one solution, so we are going to study a particular use of a test and this test we have already seen in the course and let me see whether you can recollect when I am going through the test procedures.

So, in step one we have to obtain the residuals of the estimated equation so the first task is of course, you have to get the coefficient estimates for  $\beta_1$ ,  $\beta_2$  and  $\beta_3$  via OLS method and then of course you can generate an estimate for  $\epsilon_t$  and you can call it  $\epsilon_t$  or  $\hat{\epsilon}_t$  whichever way you like and then in step 2 you have to now run an auxiliary OLS regression where the dependent variable is the fitted residual so  $\epsilon_t$  or  $\hat{\epsilon}_t$  will become the dependent variable and then you have to add the regressors and these regressors are all those on the right hand side of TV original equations.

So, I mean  $X$  and  $Y_t$  minus 1 in this case and the lagged residual. So, basically you have to also incorporate  $\epsilon_{t-1}$  as an explanatory variable and that is why it is called auxiliary regression. In step 3 now you have to make use of the findings from the auxiliary regression

in step 2. So, from the auxiliary regression, note down the R square value of that auxiliary regression and we are going to make use of that goodness of fit statistic here in the step 3.

In step 3 we will now test a null hypothesis that the parameter of the lagged residuals variable, so that is  $\rho$  so the parameter associated is  $\beta_4$  you can say that is equal to 0. So, we set a null hypothesis that  $\beta_4$  is equal to 0. Now you have to define the test statistic and the test statistic is LM.

So, hopefully you remember we have actually studied this LM test statistic in the context of heteroskedasticity and this formula is not new to you, if you remember the discussion. So, that is capital T times R squared that you obtained from the auxiliary regression and capital T of course, is the total data points or the number of time periods in your dataset.

So, for a large sample, LM has a Chi square distribution with degrees of freedom 1. Why we are setting a degrees of freedom 1 here? Because note, what parameter restriction we are imposing when we are conducting this test. So, we are setting a 1 restriction which is on  $\beta_4$  that is associated with the fitted value, lagged fitted value explanatory variable, so degrees of freedom will be only 1.

Now you can to set a decision rule and if LM is greater than the critical Chi square value found from the statistical table, then you reject the null hypothesis and if you reject the null hypothesis, then you can conclude that there is indeed serial correlation or autocorrelation in the original equation.

Now, we are going to move on to a new topic or area in time series econometrics and this is very important because this area was introduced by one person who own Nobel Prize for this tremendous contribution in time series econometrics. Yes, I am talking about Clive Granger and he actually introduced this concept called Granger causality in time series econometrics and that is what we are going to study next.

I am not going to talk an Granger causality with a whole lot of details. I just want to share the basic philosophy with which Granger stated discussion about this and he proposed a test and all. And this is very appealing. Why? Because so far in the course we have run many many regressions, we have talked about the regression, summary statistics and all, but we have not talked about causality. And in fact, if you remember, I have warned you that even if you have a very good fit from your regression, you cannot say that your Xs are causing Y.

So, Granger, for the first time in time series econometric said that something could be done to find out or establish causality between two variables which are evolving over time and as this was the first time talked about in the time series econometrics, you can very well guess the importance of this concept. So, let us talk about Granger causality.

(Refer Slide Time: 27:01)

**Granger causality**

- ▲ Clive Granger argued that causality in time series econometrics can be tested.
- ▲ Granger causality is a phenomenon in which one time series variable consistently and predictably changes before another variable.
- ▲ A variable X Granger-causes another variable Y if predictions of the value of Y based on its own past values and on the past values of X are better than predictions of Y based only on Y's own past values.
- ▲ First estimate ARDL model
 
$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \dots + \beta_p Y_{t-p} + \alpha_1 X_{t-1} + \dots + \alpha_p X_{t-p} + \varepsilon_t$$
- ▲ Then test the  $H_0$  that the coefficients of the lagged Xs (the  $\alpha$ s) jointly equal zero  $\rightarrow$  F test
- ▲ If we reject this  $H_0$ , then X Granger-causes Y
- ▲ Then estimate the model for checking bi-directional causality
 
$$X_t = \beta_0 + \beta_1 X_{t-1} + \dots + \beta_p X_{t-p} + \alpha_1 Y_{t-1} + \dots + \alpha_p Y_{t-p} + \varepsilon_t$$

So, formally speaking Granger causality is a phenomenon in which one time series variable consistently and predictably changes before another variable. So, there is another term that some variable Granger-causes another variable, so let us have a formal definition for that. So, a variable X Granger-causes another variable Y if predictions of the value of Y based on its own past values and on the past values of X are better than the predictions of Y based on only on Ys own past values.

So, now let me talk about this bookish definition in plain simple language. When you are trying to predict future values of one dependent variable Y, then what are the explanatory variables that you can think about? One is of course, if you want to run a static model, then you can get some values of some potential explanatory variables. Now if you want to make your model dynamic, then there could be 3 different alternatives and what are they?

So, as I told you that there are 3 types of dynamic econometric model, so if you want to predict future values of Y, you and if you want to make your model an autoregressive type, then you can have the p number of past time periods value on the dependent variable itself. So, basically you are going to make use of  $Y_{t-1}$   $Y_{t-2}$  dot dot dot  $Y_{t-p}$  as explanatory variables, if you do not have any other information on other explanatory variables.

But suppose you are lucky, you have some explanatory variables and for them also you can have a time series data, now you have this luxury to make use of distributed lagged model. What is distributed lagged model? So, you take a 1 or 2 Xs and then not only you are making use of the current period value of that explanatory variable and link that with the current period value with the dependent variable, but you are also making use of the past values of the explanatory variables to see whether they can explain the variability or they can determine the value of the dependent variable in this particular period.

So, in symbolic language basically I am adding some new additional regressors in my regression equation. So, if I have say for the simplicity if I have only 1 explanatory variable, so then apart from  $X_t$  that will automatically be counted in regression equation. I will be now adding  $X_{t-1}$ ,  $X_{t-2}$  and dot dot dot  $X_{t-p}$  if I am going back  $p$  time periods. So, that is basically the distributed lagged model. And not only that you can have a combination of both.

So, you can go back to the in the past up to  $p$  time periods and you can have  $Y_{t-1}$  up to  $Y_{t-p}$  and  $X_{t-1}$  up to  $X_{t-p}$ . So, this is a general ARDL model or you can call it autoregressive distributed lagged model that is the full form. And now with this general model, let us see what Granger is saying. So, Granger is saying that you have to first estimate an ARDL model and the lag selection actually is in your hand. So, what will be the value of  $p$  you decide.

So, if you want to have the simplest possible ARDL model, then you will have only  $\beta_0 + \beta_1 Y_{t-1} + \alpha_1 X_{t-1}$  and then of course, the random error term  $\epsilon_t$ . So, this is the simplest possible ARDL model, this is called ARDL of 1 1 model.

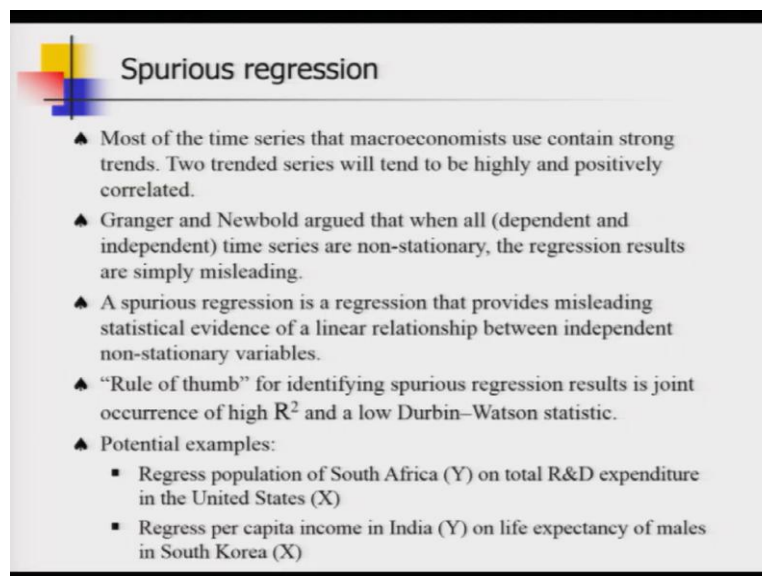
Now, you test the null hypothesis that the coefficients of the lagged Xs jointly equal to 0 or not. So, basically if the null hypothesis is true, then that means that if you are throwing  $p$  past values of the explanatory variable  $X$ , then jointly they are unable to determine the value of  $Y$  or they fail to explain the variability in the  $Y$  variable. So, that is what this null hypothesis says.

So, basically we can say that if we are able to reject this null hypothesis by conducting an  $F$  test, then we can say that actually the variable  $X$  Granger-causes  $Y$ . Now the very interesting point is that of course,  $F$  test you can conduct, it is very easy to conduct, but Granger has some more twist in the story and he told that actually the causality could be bidirectional.

So, we have here shown how X can impact Y, but there could be the other way round that Y can also X. So, basically you have to run another regression now from the different perspective and then check whether that is actually indeed the case or not.

So, here in the slide the last equation I am showing that you need to make use of to check bidirectional causality and then you need to set the similar kind of null hypothesis and conduct f test and then again here if you reject the null hypothesis in the second regression, second ARDL model, then you can say that Y Granger-causes X and in that case you can say that this causality is bidirectional. So, X Granger-causes Y and Y Granger-causes X.

(Refer Slide Time: 33:11)



**Spurious regression**

- ▲ Most of the time series that macroeconomists use contain strong trends. Two trended series will tend to be highly and positively correlated.
- ▲ Granger and Newbold argued that when all (dependent and independent) time series are non-stationary, the regression results are simply misleading.
- ▲ A spurious regression is a regression that provides misleading statistical evidence of a linear relationship between independent non-stationary variables.
- ▲ “Rule of thumb” for identifying spurious regression results is joint occurrence of high  $R^2$  and a low Durbin–Watson statistic.
- ▲ Potential examples:
  - Regress population of South Africa (Y) on total R&D expenditure in the United States (X)
  - Regress per capita income in India (Y) on life expectancy of males in South Korea (X)

Now, we are going to talk about something very interesting and that is known as spurious regression. So, what is spurious regression? So, if I want to tell you something in Lemman’s language which you will not forget, so I can say that spurious regression is something which does not make any sense.

So, of course, we can now get into the theoretical details and all, but it is better that we keep the discussion simple here and let us talk about some example first before we actually go to the theoretical aspect of spurious regression, why we get it and what are the consequences, etcetera.

So, although the bullet point for examples is placed at the bottom of the slide, but let us start from that bullet point which is placed at the bottom of the slide. So, here I am talking about some potential examples. Suppose you got some time series data and you are regressing population of South Africa on total research and development expenditure made in the United

States. Now, you think about another regression that you can run, you get time series data and you regression per capita income in India on life expectancy of males in South Korea.

Now, you just think about this funny situation. Do you assume that there is any relation between this Y and X that I am talking about in this two context? Absolutely not, but very surprisingly if you collect time series data and if you make a time series plot, you will see that with time both X and Ys are moving in the same direction and now if you want to make a scatter plot of these two variables Y and X, then you may be utterly surprise to figure out that it is showing some scatter of data points which is indicating towards a positive correlation.

Now, of course, if you have that kind of scatter in this two dimension of Y and X, you can pass a straight line and fit a straight line, but that straight line is totally meaningless because you cannot get any head or tail out of that regression equation because there is no statistical dependence between these two variables, it is actually showing you a very false picture of the reality.

But, the question is, why then you are getting this kind of scatter plot or if you in fact, run this regression, if you are crazy and if you run this kind of regression, then you may be able to see a very high R squared value and the f value is also significant, so statistically speaking also, you say that my regression model is perfect.

But actually what is happening, there is some other variable which is controlling the movement of these two variables Y and X. So, you see I have chosen examples which actually have some embedded time trend in them. So, although, these two variables Y and X, they are not statistically related by any wildest dream of yours, but time component is common in this two variables.

So, of course, when you plot them side by side or if you want to find out some relation, degree of association between these two types of variables, then you will get some indication that there is some correlation or association and that is actually coming from the common trend. So, these are the examples of spurious regression.

Now let us talk about some theoretical aspects of spurious regression. So, now we are back to the serious business of time series econometrics. So, in the 1960s and 70s when econometrics was getting popularity, many many macroeconomists actually ran time series regression because those days I am talking about 1950s and 60s, a cross sectional data were very difficult to get by, forget about panel data and time series was the most available data for a

person who is interested in quantitative economics and data analysis and drawing inference from real life observations.

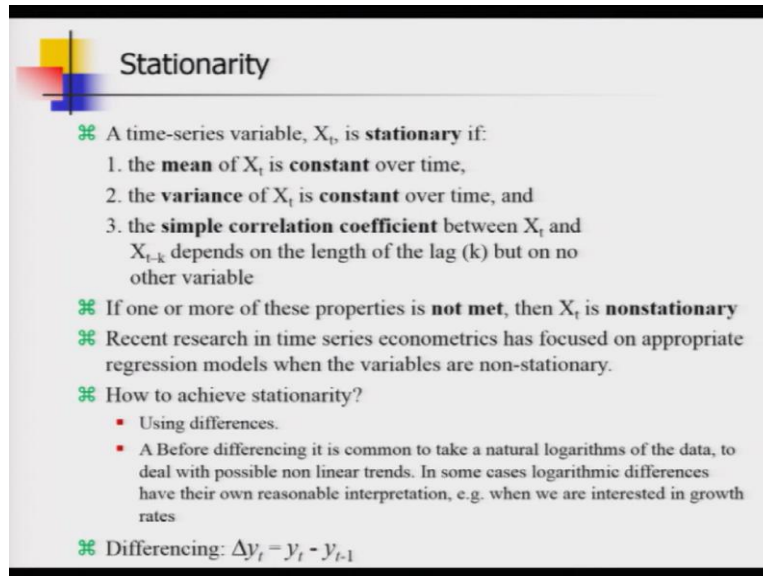
So, many macroeconomists they have used time series on different macroeconomic variables and ran simple OLS. And to their surprise they have obtained very high value of R square. So, the excellent news for them, but later on when Clive Granger came up with this idea of spurious regression with his colleague Newbold, then everybody was shocked. So, again we come back to Granger and it was him with his colleague Newbold, who actually introduced this concept of spurious regression.

So, Granger and Newbold argued that when all dependent and independent time series variables are non-stationary in nature, then regression results are simply misleading and they are spurious in nature. Now Granger and Newbold now here talking about a term which we do not know, so they are talking about non-stationarity. What is it wait for a minute we are going to discuss the definition of stationarity and non-stationary very soon.

So, as of now let us think about a rule of thumb which will help us to detect whether actually we have run a spurious regression or not using times series data. So, the popular rule of thumb says that if you have run a regression and we are observing very high R value but very low Durbin-Watson test statistic value, then actually you have a potential case for spurious regression.

So, now we are going to talk about the definition of stationarity and this is the last slide for today's lecture, I am just going to introduce the term and in the next lecture I am going to develop on this concept. So, without wasting time let us look at definition for stationarity.

(Refer Slide Time: 40:08)



### Stationarity

- ☞ A time-series variable,  $X_t$ , is **stationary** if:
  1. the **mean** of  $X_t$  is **constant** over time,
  2. the **variance** of  $X_t$  is **constant** over time, and
  3. the **simple correlation coefficient** between  $X_t$  and  $X_{t-k}$  depends on the length of the lag ( $k$ ) but on no other variable
- ☞ If one or more of these properties is **not met**, then  $X_t$  is **nonstationary**
- ☞ Recent research in time series econometrics has focused on appropriate regression models when the variables are non-stationary.
- ☞ How to achieve stationarity?
  - Using differences.
  - A Before differencing it is common to take a natural logarithms of the data, to deal with possible non linear trends. In some cases logarithmic differences have their own reasonable interpretation, e.g. when we are interested in growth rates
- ☞ Differencing:  $\Delta y_t = y_t - y_{t-1}$

A time series variable,  $X$ , is stationary if 3 conditions are made So, first condition is saying the mean of  $X_t$  is constant over time, the second point is the variance of  $X_t$  is constant over time and the third point says the simple correlation coefficient between  $X_t$  and  $X_{t-k}$  so you have to take the lag value, you have to go back in the past  $k$  time periods to find  $X_{t-k}$  and if you now find the simple correlation coefficient between the current value and  $k$  period past value of the same variable, this this correlation coefficient depends on the length of the lag that you have chosen that is  $k$ , but on no other variable.

So, if these 3 conditions are made, then we can call a variable is stationary. So, if one of these conditions or properties is not made, then you can say that time series variable is non-stationary. Now, in recent times, the time series econometric research actually has focused on developing the regression models for non-stationary data because as I have told you that generally when you are dealing with time series data, there will be some trained component or trained element in it.

So, if you remember our lectures in the first part of the course when we talked about the classical time series analysis, so I refer to the decomposition of a time series variable in train seasonality, cyclical and random noise, these 4 components. So, trained is there, so be it a multiplicative kind of functional form or it can be additional, additive type of form, but trained is always there in a time series data when you have data variable that you are observing for a long time period, when you have too many numbers of time points on the time plane.



So, basically in that case, then Granger and Newbold are saying that mostly the macroeconomic time series are all non-stationary because you probably will find that either the mean is varying over time or you may find that the variance maybe varying over time or you can say that if you go back say  $k$  period in the past and if you compute the correlation coefficient that may not only depend on the order of lag that you have chosen here  $k$ , it may vary with something else.

So, there could be violations of these 3 conditions that I have laid out in this slide. And it is actually the reality. So, most of the macroeconomic time series variables are non-stationary. And whatever regression models that we have done in today's lecture, they are all basically assuming that you have a stationary time series data. So, if stationarity breaks down, if it not there in your data, then what to do?

So, it is a big area of research for econometricians and in last 20-30 years this area has developed tremendously. So, there are regression models which are going to handle non-stationary data and we are going to have some idea about this literature in the next lecture. As of now let us see what we can do with non-stationarity data. So, if non-stationarity is there, there are some remedies that you can quickly adopt and let us have a look at them.

So, econometricians suggest that you can take difference to avoid non-stationarity in fact, if you have non-stationary data, if you take differences, then actually there is a chance that it will become stationary. What do we mean by differencing by the way? So, for that you have to come to the very last bullet point in this slide and here I am showing you the difference operation. So, the formula for differencing exercise is shown and you see that there is this delta Greek symbol that I have introduced here. So, that is called delta or difference operator.

So, if I apply this operator in front of this time series value  $y_t$ , then actually I mean  $y_t$  minus  $y_{t-1}$ . So, basically a delta operator calculates the difference between the current value and the past value of the same variable. But note that here we are talking about one order of differencing.

In the next lecture I am going to generalize these concept  $m$  then we will talk more about it, but even before you take this differencing step, there is another simple manipulation that you can actually apply to your data and by doing that many empirical researchers have claimed that it worked.

So, there is not much theoretical justification for this technique or this trick, but it is useful. So, you can take log of the time series variable that you are interesting into study and then you can take difference, then you will get much much better result and most likely you are going to obtain stationarity.

Now, why you have to take log because logarithm is a monotonic transformation. So, the basic trend in the variable values is not going to be changed and also the natural logarithm helps you to suppress or reduce the extent of heteroskedasticity in the data. So, there are some benefits of logarithm transformation.

So, if you take logarithm transformation, another benefits could be from the interpretation side. So, if you take logarithm of some variable, then you can actually talk about growth rates and all and they are of some interest to applied econometricians especially in macro and growth theory.

So, these are the tricks that you can adopt to make a non-stationarity data a stationary one and we stop here, in the next lecture and that is going to be the last lecture on time series econometrics. We are going to develop on the topics or the concepts that we have learned here. So, come back and join me for one more time. Thank you, bye.