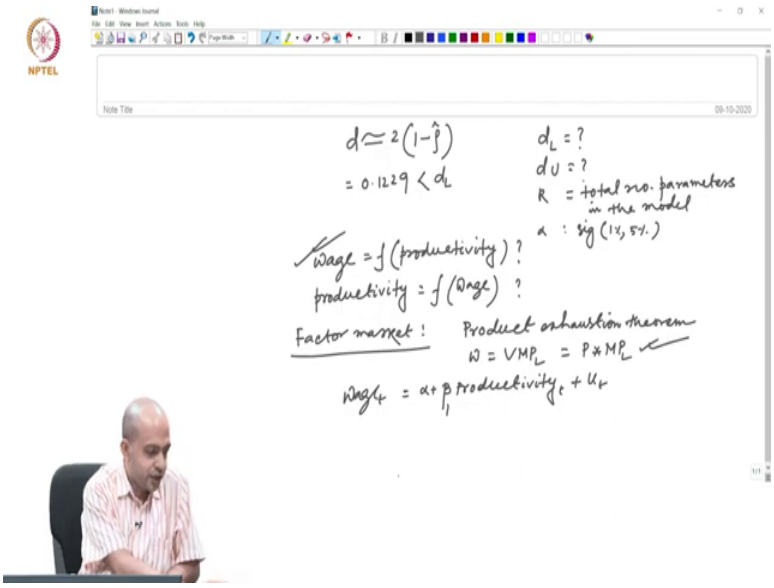
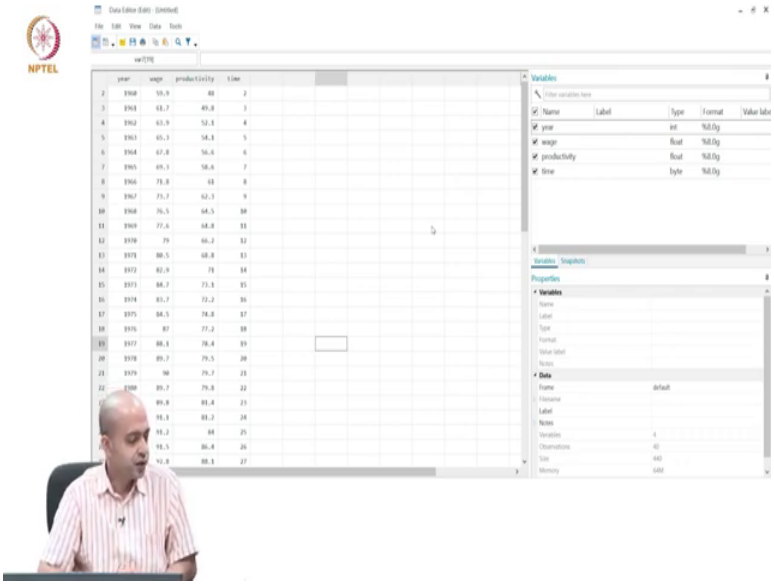
Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras
Lecture 50
Relaxing the assumptions of CLRM-Autocorrelation and Heteroscedasticity Part - 1

(Refer Slide Time: 0:17)





So, welcome once again to our discussion of autocorrelation. Yesterday we were discussing about autocorrelation problem. So, basically autocorrelation, what does it mean? It means that error terms of the two time periods ut or ut minus 1, or ut plus 1 they get correlated. And lastly,

we are talking about detection of autocorrelation by using Durbin and Watson test statistic and how we defined the Durbin and Watson test statistic, basically that was d equals to 2 into 1 minus rho hat, where rho is the first order autocorrelation coefficient as we defined earlier.

So, that means depending on the value of rho hat autocorrelation between the 2 successive periods error term will get a value of d and as we said then for a given level of significance from the Durbin Watson table what you need to find out what is dL and what is dU corresponding to k equals to total number of parameters, total number of parameters in the model, parameters in the model and a specific level of significance 1 percent, 5 percent, or 10 percent whatever.

So, let us say this calculated d equals to something around 0.1229 and as we said that when the d value, the calculated d Durbin Watson test statistic value is within the neighborhood of 2 then there is no autocorrelation, there is no autocorrelation. But as d approaches towards 0 then the evidence of autocorrelation gets more and more, we get more evidence.

So, once you get this d value what you have to check? You have to check what is the dL and what is the dU. So, if this calculated value is less than dL lower limit of Durbin Watson that means you have to reject your null. And what is my null? There is no autocorrelation. So, when no autocorrelation gets rejected then you have to say that yes your data is suffering from autocorrelation problem that is the Durbin Watson test statistic we discussed.

Now, today what we will do, we will take one dataset and then we will estimate the model and we will also see how to get the Durbin Watson test statistic value and then we will compare that test statistic value with the tabulated value. So, we will be using one data set. This is the data set, see this is the data set this is a time series data on wage and productivity starting from 1959 to 1998, you have 40 years data, 40 years data on the wage and productivity at the economy level and this data is also taken from the US.

So, what we will do, we will estimate the model using the two variables wage and productivity. Now, when there are two variables wage and productivity, firstly what we need to think, which is your dependent variable and which is the independent variable. So, we have two variables wage, should wage be a function of productivity is a function of wage that you first need to decide.

You first need to decide this because as we said that econometrics per se will not tell you which is dependent and which is independent variable and that is why we said if you recall that econometrics is basically the art and science. So, from the existing theory or from your own common sense you need to identify what should be the dependent and what should be the independent variable. So, we have to take help from some existing theories, that will guide us identifying dependent and independent variables.

Now, if you recall, one of the theory is that we learned from our micro economics particularly the chapter on factor market. In factor market there is one theorem which says that in a perfectly competitive market when the laborers are paid based on their value of marginal product then the value of the total product gets exhausted which is called the product exhaustion theorem.

So, that means laborer are paid W equals to V into MPL, VMPL value of marginal product. So, this is I will write VMPL value of marginal product of labor which is nothing but equals to P into MPL. So, if you multiply the marginal productivity of labor with the price level you will get the value of marginal product and that should be at the equilibrium should be equals to the wage rate, that is the equilibrium condition.

And if you do so, if the laborers are paid according to their value of marginal product then the total value of your output will get exhausted that is called product exhaustion theorem. So, this theory give some kind of guidance that actually wage is a function of productivity.

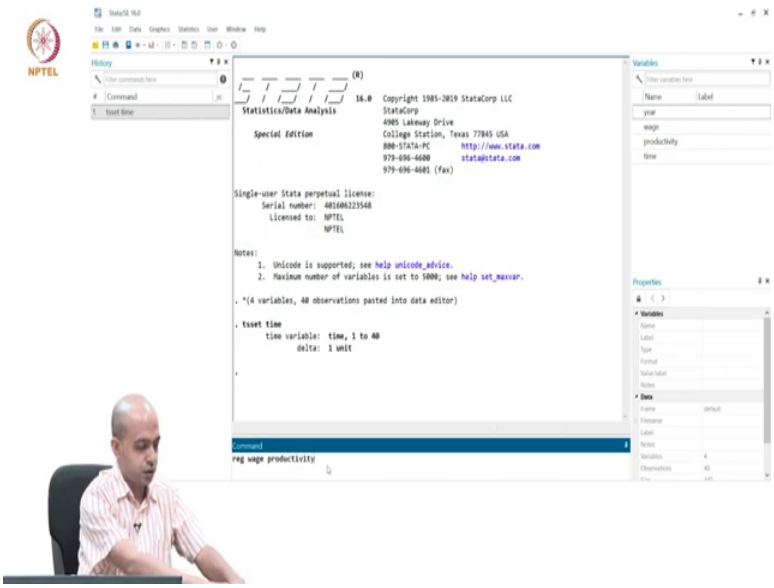
But there are counter arguments also, some economies they believe that even productivity is also depends on wage because wage gives you some kind of motivation to work more, to produce more. If the laborers are not paid enough then obviously there is a high chance that their productivity will go down.

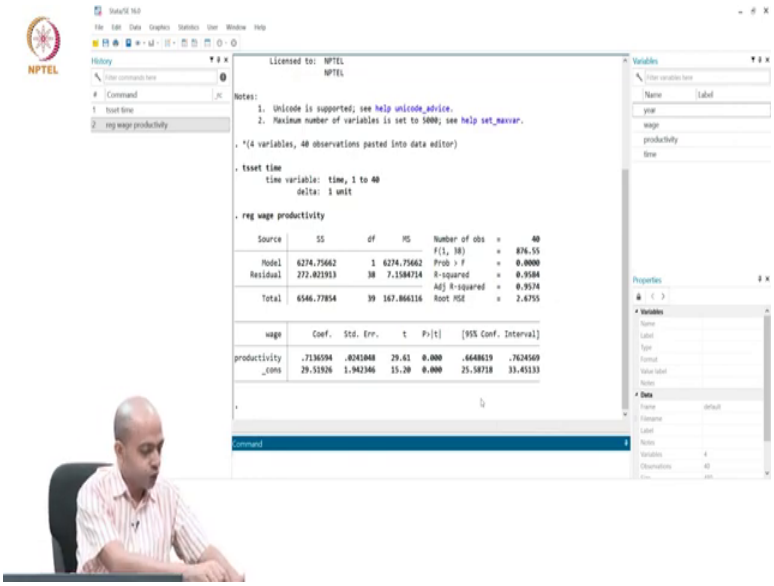
So, that is why more the wage would be the motivation to work, more would be the output per unit of labor and as a result of which your productivity may also increase. So, this type of relationship is also possible. So, that means between wage and productivity actually there is a two-way relationship but for the time being we will ignore the two-way relationship between wage and productivity.

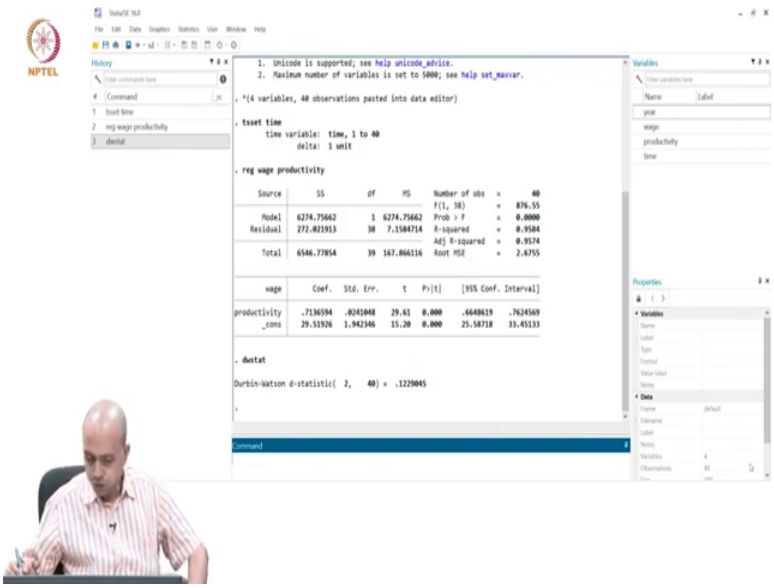
That means the simultaneity between wage and productivity though it is a possibility for the timing we are just ignoring that and we are considering that wage actually depends on productivity and then we are writing this type of model, wage of the t^{th} period equals to alpha plus beta1 productivity plus ut, this is the model we are going to estimate.

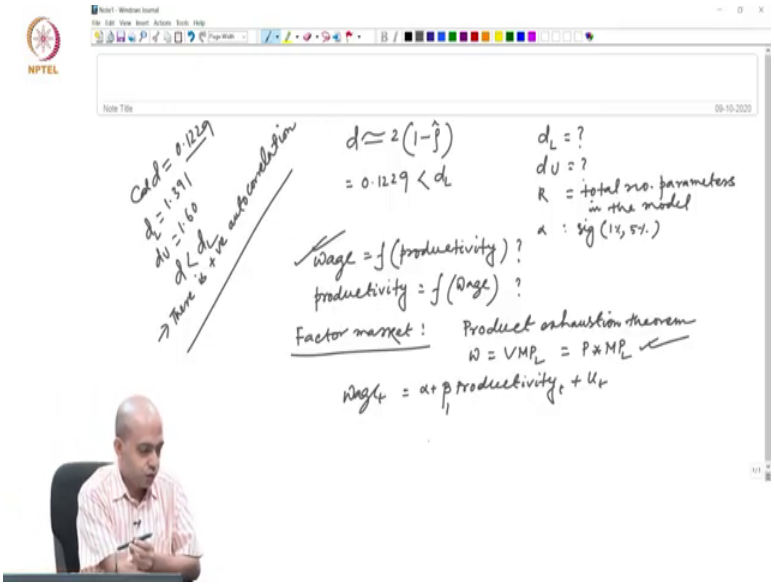
And after estimation we will see whether there is autocorrelation problem in this data set that means whether this ut is actually correlated with ut minus 1 or not. So, first we will estimate and then we will discuss about the Durbin Watson test statistic. So, let us look at the data set, so this is the data, we have 1960 to 1998, so we have some 40 years data and we will now estimate the model.

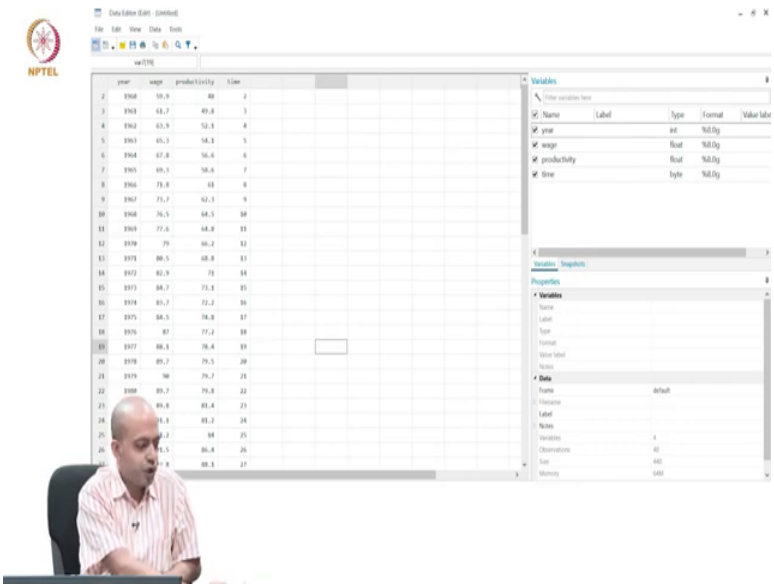
(Refer Slide Time: 10:09)




NPTEL
Stata/SE 16.0
File Edit Data Graphics Statistics User Window Help
History
Command
1 tsset time
2 reg wage productivity
Licensed to: NPTEL
NPTEL
Notes:
1. Unicode is supported; see help unicode_advice.
2. Maximum number of variables is set to 5000; see help set_maxvar.
. *(4 variables, 40 observations pasted into data editor)
. tsset time
time variable: time, 1 to 40
delta: 1 unit
. reg wage productivity
Source | SS df MS Number of obs = 40
F(1, 38) = 876.55
Model | 6274.75662 1 6274.75662 Prob > F = 0.0000
Residual | 272.021913 38 7.1584714 R-squared = 0.9584
Adj R-squared = 0.9574
Total | 6546.77854 39 167.866116 Root MSE = 2.6755
wage | Coef. Std. Err. t P>|t| [95% Conf. Interval]
productivity | .7136594 .0241048 29.61 0.000 .6648619 .7624569
_cons | 29.51926 1.942346 15.20 0.000 25.58718 33.45133
Command
Variables
Name Label
year
wage
productivity
time
Properties


NPTEL
Stata/SE 16.0
File Edit Data Graphics Statistics User Window Help
History
Command
1 tsset time
2 reg wage productivity
3 dwstat
1. Unicode is supported; see help unicode_advice.
2. Maximum number of variables is set to 5000; see help set_maxvar.
. *(4 variables, 40 observations pasted into data editor)
. tsset time
time variable: time, 1 to 40
delta: 1 unit
. reg wage productivity
Source | SS df MS Number of obs = 40
F(1, 38) = 876.55
Model | 6274.75662 1 6274.75662 Prob > F = 0.0000
Residual | 272.021913 38 7.1584714 R-squared = 0.9584
Adj R-squared = 0.9574
Total | 6546.77854 39 167.866116 Root MSE = 2.6755
wage | Coef. Std. Err. t P>|t| [95% Conf. Interval]
productivity | .7136594 .0241048 29.61 0.000 .6648619 .7624569
_cons | 29.51926 1.942346 15.20 0.000 25.58718 33.45133
. dwstat
Durbin-Watson d-statistic(2, 40) = .1229045
Command
Variables
Name Label
year
wage
productivity
time
Properties





Now, since this is the time series data and you are trying to estimate the model and after estimation you also want to get the Durbin Watson test statistic then what you need to do first, you have to explain that this data is a time series data and you have to put a specific command for that, unless you use that specific command stata will not be able to recognize this data set as a time series data.

And there is a small command. What is that command? The command is ts set means time series set, so you have to specify what is your time variable here, if you look at I have created the time

variable as time only and time starts from 1, 2, 3, like that otherwise you can put year also, year is your time variable, whatever, either ts set year or t set time, both will work, both will work.

So, ts set time that is the first thing you need to do, you need to specify the data as a time series data by the specific command called ts set time. ts set means time series set and time is basically the time variable. Now, we will estimate the model reg wage and then productivity.

And see your productivity is highly significant because the P value is 0.000 so that means it is almost, it is significant at 1 percent level highly significant. And r square is 0.9584. high r square but still this result might be of less use unless we check for autocorrelation problem.

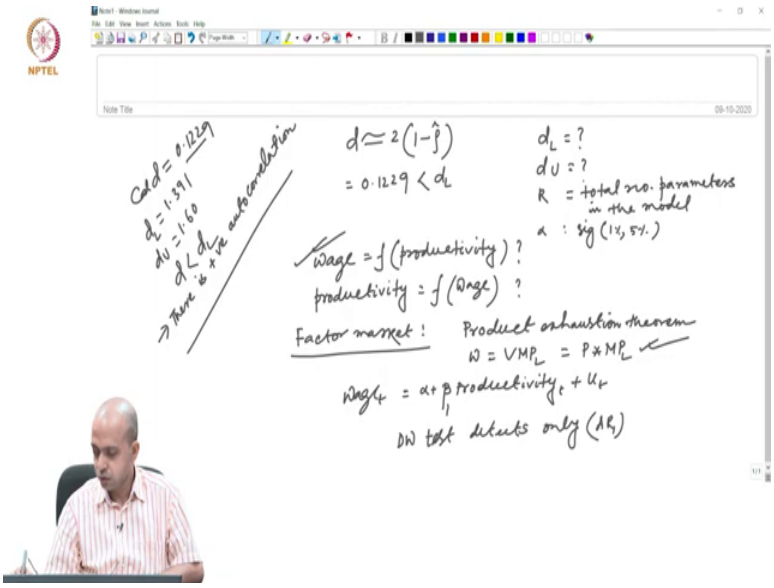
How will you do that? To detect autocorrelation we need to get the Durbin Watson test statistic which says dwstat, this is the command, this is the command, Durbin Watson test statistic is 1229. So, as I said the Durbin Watson test statistic value depends on what is the total number of observation and what is the total number of parameters to be estimated from the model.

There are two parameters, productivity and constant term. So, that means stata is showing the Durbin Watson test statistic value which is 2 and 40.. Now, if you go back to the Durbin Watson table, and if you specify n equals to 40 and k equals to 2 then you will see that your dL Durbin Watson value is almost 10 point something, Durbin Watson value is around 10.

And from there immediately you have to go to the table, unless you go to the table you will not be able to understand. So, when the number of observation is 40 and then you have calculated value is 0.12 so this is this is your calculated value, this is a calculated value calculated value of d is 0.12 and then what is your dL at 5 percent level of significance, if you look at dL actually equals to dL equals to you see it is 1.391. And what is dU? dU is 1.60, so that means this d is actually lower than dL, from the value itself you can understand that it is tending towards 0, it is not tending towards 2. Since calculated d is lower than the Durbin Watson lower limit the decision is that there is autocorrelation, there is positive autocorrelation, that is how you have to take your decision.So, this is how you have to get it from the Durbin Watson test statistic.

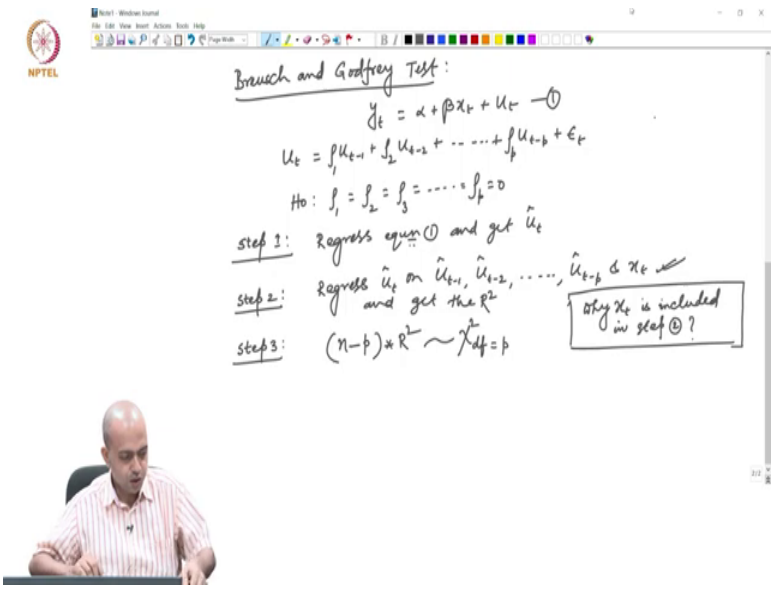
So, that means even though your result shows around 95 percent of your total variation in wage you can explain by productivity that is of less use, this particular coefficient 0.7136 this of less use because the data is suffering from autocorrelation problem, Durbin Watson test statistic.

(Refer Slide Time: 17:19)



Now, once you estimate Durbin Watson test statistic and think about the autocorrelation, the next thing what we need to think is, see Durbin Watson test statistic, Durbin Watson test statistic can detect dw test, can detects if you remember detects only AR1, only AR1 that means it can detect only first order autocorrelation. We do not know whether there is higher order autocorrelation also in the equation. So, to check for the higher order autocorrelation what we need to do is we we need to go for another improved test of autocorrelation which was suggested by Breusch and Godfrey.

(Refer Slide Time: 18:19)



So, now we will talk about Breusch and Godfrey test. So, let us once again write our original model as y equals to alpha plus beta xt plus ut and the data generating process for the ut is now ut equals to rho 1 ut minus 1 plus rho 2 ut minus 2 plus rho p ut minus p plus epsilon t and epsilon t follows all the assumption of classical regression model, so that means this epsilon t is actually the classical error term, classical error term.

So, that means now we assume that the data generating process is such that the error term is actually correlated with all its lags values and how many lags are there? p lags, p number of lags. So, what is our hypothesis here? Our null hypothesis is that rho 1 equals to rho 2 equals to rho 3 equals to rho p equals to 0. Since, our claim is the presence of autocorrelation the null is that there is no autocorrelation that means all these autocorrelation coefficients are actually 0.

Let us say this is our equation 1. Now, what are the steps for this Breusch and Godfrey test, so there are certain steps of this tests, let us say that step 1, so in step 1 what you need to do, you just regress equation 1 and get ut hat that is step 1. Then in step 2 regress ut hat on ut minus 1 hat, ut minus 2 hat and ut minus p hat and xt, that is step 2. So, that means I am regressing ut on all its previous values as well as the explanatory variable xt.

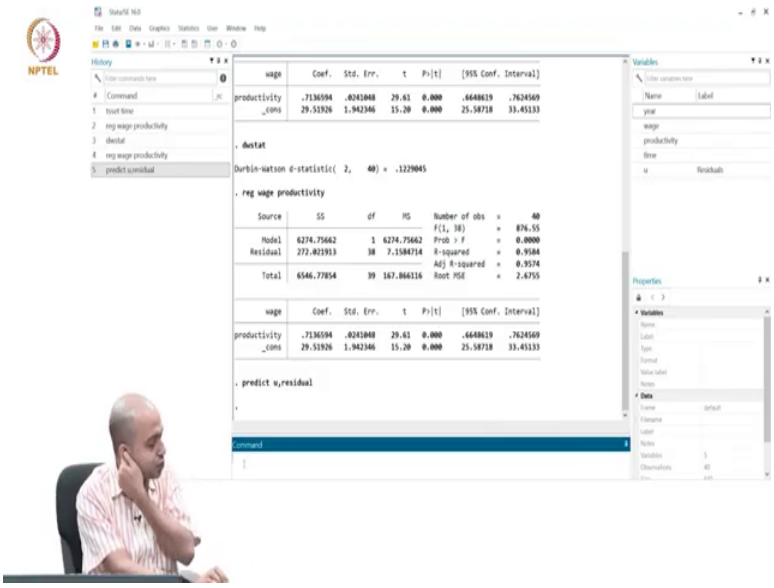
And then step 3, regress on this and get the R square. So, you regress ut on all its previous values as well as xt and get the R square. Then in step 3 what you need to do, you need to multiply this R square with n minus p, where p is the number of lag, n is the total number of observation and this actually follows a chi square distribution with degrees of freedom equals to p.

If your calculated chi square again is greater than the tabulated chi square at 1 percent and 5 percent level of significance, then you have to reject your null and what is your null, null is this. If the calculated chi square is less than the tabulated one then you do not reject the null, so this is the procedure.

Now, one thing you need to keep in mind I am interested in autocorrelation that means whether ut depends on all its previous values or not but here in step 2 why I have included this xt also? Can you think of? Why I am writing the question for your thinking why xt is included in step 2? This is your question for thinking. Why xt is included in step 2 when my objective is to check only autocorrelation?

So, this is the Breusch and Godfrey test, this is called, this is a more generalized test. Why this is so? Because I have assumed a more generalized data generating process for the ut that is why ut actually depends on all its previous values and the lag is up to p. Now, for this data what I was talking about, what will do, we will now see whether after specifying sufficient large number of lag also there is autocorrelation or not. So, we will implement the Breusch and Godfrey test using the same data set.

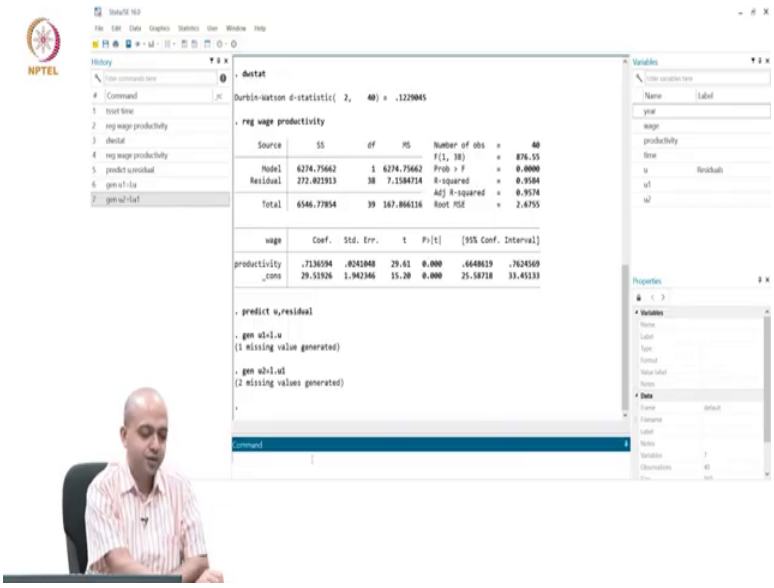
(Refer Slide Time: 25:10)



So, once again I will write the main regression reg wage productivity and then first step is as I said you have to get the residual term and what is the command for this? If you remember this is called predict, predict command, predict u and then residual. So, this is the command. So, that means you have now predicted u. If you see here the moment you successfully predict the variable would be added here.

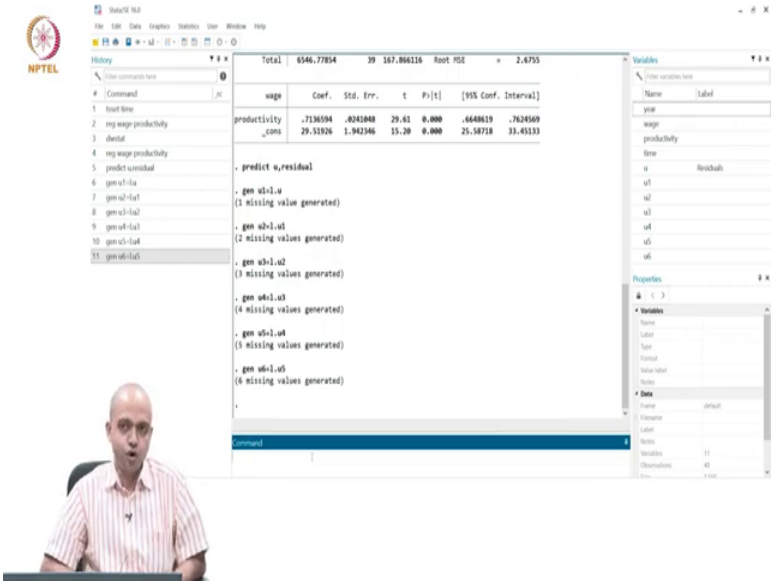
And then what you need to do is you need to get the all the previous values of ut that means you have to get ut minus 1, ut minus 2, ut minus 3, dot, dot, dot, ut minus p hat. Let us assume that we are interested in p equals to 6 that means we will see up to 6th order lag whether the correlation coefficient is significant or not. So, lag we are assuming p equals to 6. So, here is a specific command again I will be using for that to get the lag value and what is the lag value?

(Refer Slide Time: 26:41)



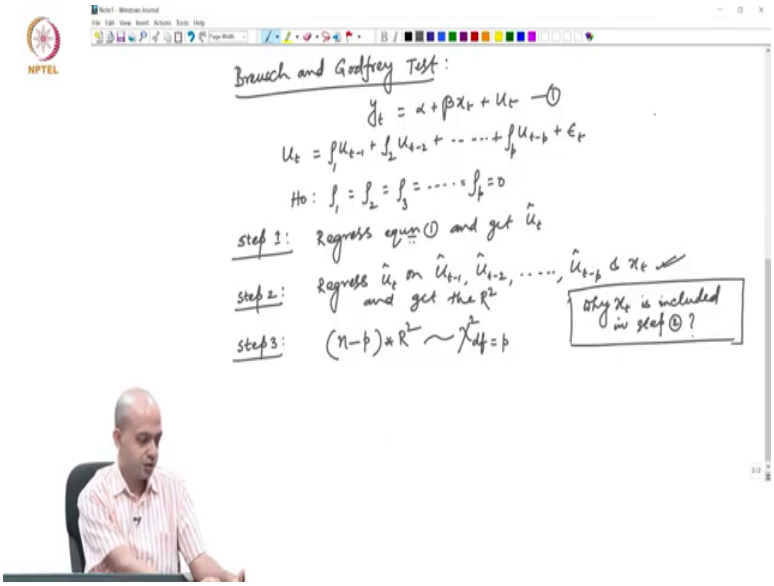
Again, we have to use the gen command, gen let us say I am giving a name called u1 for the first order lag and what is the command l dot u, this command you have to remember, l dot u is the command to get first order lag, then you put enter. So, that means u1 is actually ut minus 1 hat. Similarly, I am putting u2 for the second order lag, and what would be the command for that if you apply your common sense? You have already generated u1 that means ut minus 1 hat. Now, how will you get ut minus 2 hat? Simple put l dot u1 that is all.

(Refer Slide Time: 27:34)



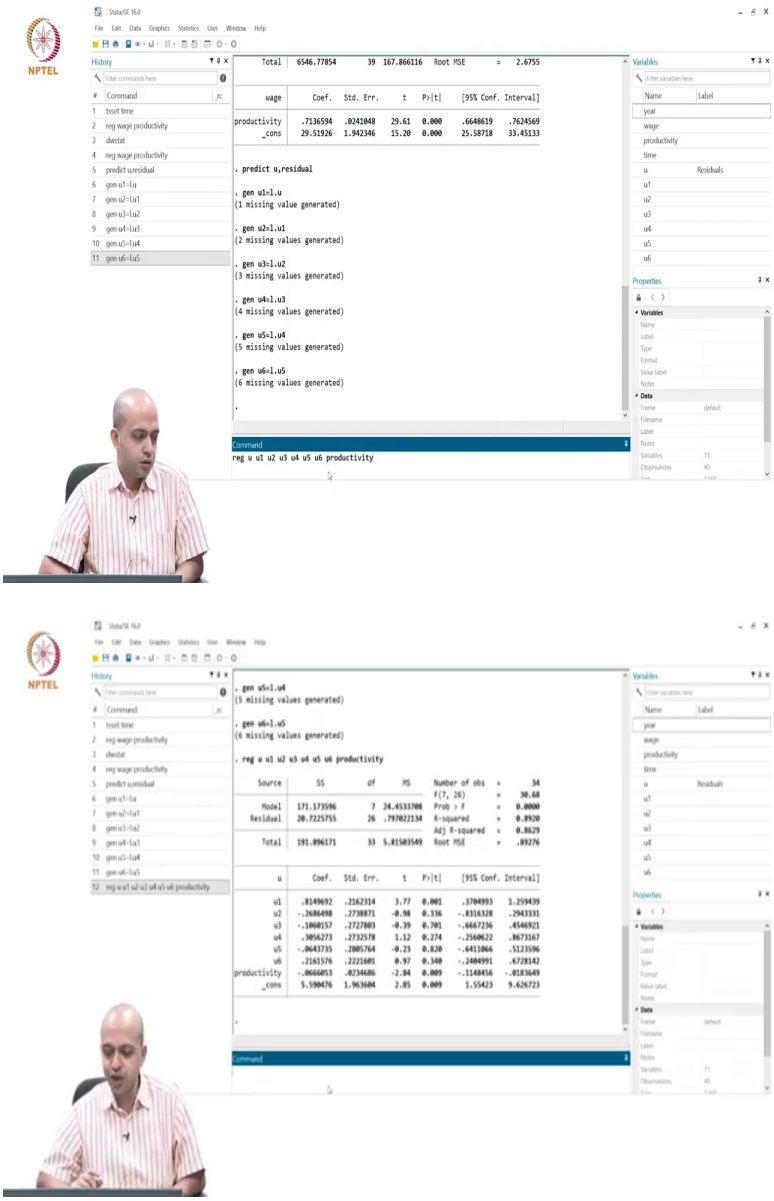
Similarly, gen u3 equals to l dot u2, similarly gen u4 equals to l dot u3, gen u5 equals to l dot u4 and then gen u6 equals to l dot u5. So, we have generated ut minus 1 hat, ut minus 2 hat, ut minus 3 hat, ut minus 4 hat, ut minus 5 hat and ut minus 6 hat, all the lagged values of the error term we have created by using l dot command, l for lag if you remember, l dot a particular variable means this data will create the lag value for that, l for lag.

(Refer Slide Time: 28:51)



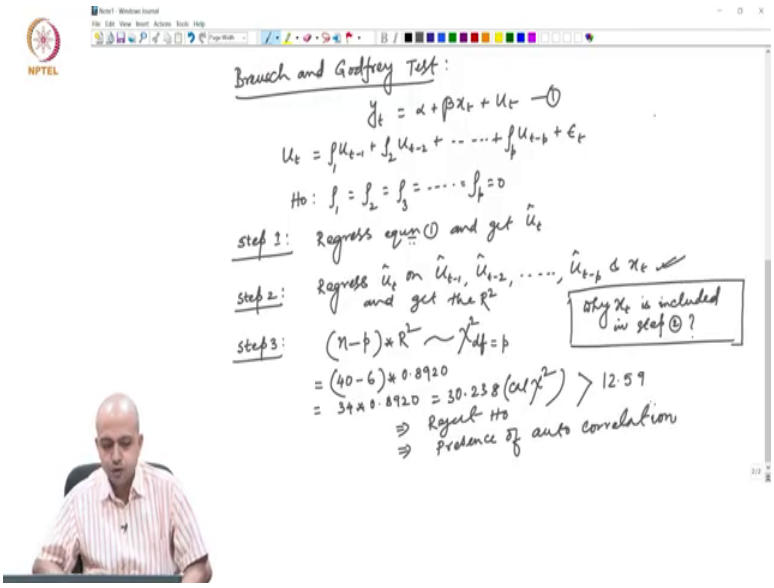
Now, what we have to do, we have to run that particular equation, just look at the equation once again, this is the regression we have to run ut should be regressed on all its previous value, so that means you are in step 2. ut is regressed on its previous values as well as xt.

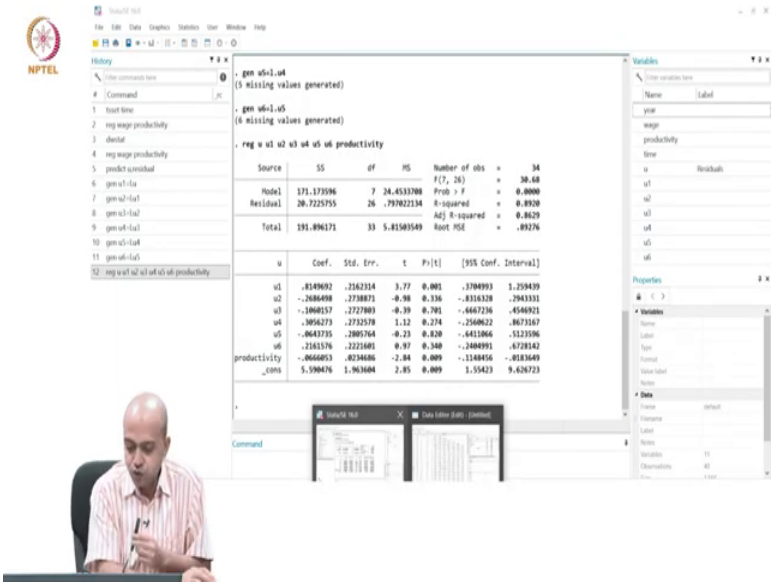
(Refer Slide Time: 29:18)

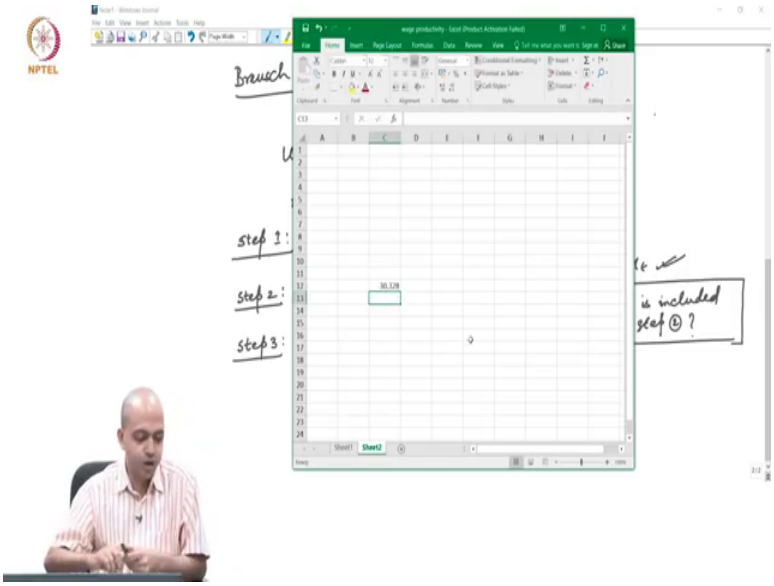


So, reg u on u1, u2, u3, u4, u5 and u6 and then you need to include productivity also. Now, if you put enter, now this is the value, this is the output and from here what is the R square? R square is 0.8920 and that should be multiplied with n minus p.

(Refer Slide Time: 30:17)







So, that means n minus p is basically n equals to 40, p equals to 6 and that should be multiplied by 0.89. What is the value? what is the value of R square? 8920. that means 34 multiplied by 0.8920 and if you multiply that how much it will come? 0.8920.

So, 0.8920 equals to 0.8920 multiplied by how much, if you multiply that by 34, so 30.238, to equals to 30.238, so this is the calculated chi square and this calculated chi square you need to now compare with the chi square tabulated value. So, from the table if you see, degrees of freedom is 6. you will see the chi square is around 12.59.

So, that means when you get chi square equals to 30 from there itself you can understand that this is a quite higher value and there is higher probability so the 5 percent level of significance the chi square is 12.9, even at 1 percent level also it is 16.81 which is greater than 12.59. That means what is your decision? This implies reject reject H naught. And what does it say? Presence of autocorrelation.