

**Introduction to Econometrics**  
**Professor Sabuj Kumar Mandal**  
**Department of Humanities and Social Sciences**  
**Indian Institute of Technology, Madras**

**Lecture 51**

**Relaxing the assumptions of CLRM-Autocorrelation and Heteroscedasticity Part - 2**  
 (Refer Slide Time: 0:16)

**Breusch and Godfrey Test:**

$$y_t = \alpha + \beta x_t + u_t \quad \text{--- (1)}$$

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \dots + \rho_p u_{t-p} + \epsilon_t$$

$$H_0: \rho_1 = \rho_2 = \rho_3 = \dots = \rho_p = 0$$

**step 1:** Regress eqn (1) and get  $\hat{u}_t$

**step 2:** Regress  $\hat{u}_t$  on  $\hat{u}_{t-1}, \hat{u}_{t-2}, \dots, \hat{u}_{t-p}$  &  $x_t$  ✓

**step 3:**  $(n-p) * R^2 \sim \chi^2_{df=p}$  why  $x_t$  is included in step 2?

$$= (40-6) * 0.8920$$

$$= 34 * 0.8920 = 30.238 (\text{cal } \chi^2) > 12.59$$

⇒ Reject  $H_0$   
 ⇒ Presence of auto correlation

So, this is the procedure for the Breusch and Godfrey test, you have to estimate the model, you have to get the predicted value of the error term, then you have to also create the predicted value of the previous period lag error terms and then you need to regress the  $u_t$  on its previous values and also on the explanatory variable, the explanatory variable. And then if you do so this is the result. This is the result.

(Refer Slide Time: 1:00)

Stata 16.0

```

    . reg wage productivity

    Source      SS          df           MS       Number of obs   =    40
              6274.75662      1      6274.75662      7(1, 38)       =    876.55
    Residual   272.823933      38      7.1584714      Prob > F         =    0.0000
    Total      6546.77054      39      167.866316     Adj R-squared    =    0.9574
                                         Root MSE        =    2.6755

    wage      Coef.   Std. Err.   t      P>|t|   [95% Conf. Interval]
    -----+-----
    productivity  -73.95994   -0.241048   -29.41   0.000   -66.68619   -81.24569
    _cons        29.51926   1.942346    15.20   0.000   25.58718   33.45133

    . estat bgodfrey

    Breusch-Godfrey Ljung-Box test for autocorrelation

    lags(p)      chi2      df      Prob > chi2
    -----+-----
    1             32.205      1      0.0000

    NB: no serial correlation
  
```

Now, in stata, you do not need to implement the test always manually. there is a specific command if you estimate the model reg wage and productivity then immediately after estimating the model what you need to do, you need to specifically mention a command for this Breusch and Godfrey test which is estat bgodfrey, so this is the command. I will first write the command here so that you should not forget.

(Refer Slide Time: 1:49)

Breusch and Godfrey Test:

$$y_t = \alpha + \beta x_t + u_t \quad \text{--- (1) \quad statstc \quad estat bgodfrey}$$

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \dots + \rho_p u_{t-p} + \epsilon_t$$

$$H_0: \rho_1 = \rho_2 = \rho_3 = \dots = \rho_p = 0$$

step 1: Regress eqn (1) and get  $\hat{u}_t$

step 2: Regress  $\hat{u}_t$  on  $\hat{u}_{t-1}, \hat{u}_{t-2}, \dots, \hat{u}_{t-p}$  &  $x_t$

step 3:  $(n-p) * R^2 \sim \chi^2(p)$  why  $x_t$  is included in step 2?

$$= (40-6) * 0.8920$$

$$= 34 * 0.8920 = 30.238 \text{ (cal)} > 12.59$$

$\Rightarrow$  Reject  $H_0$

$\Rightarrow$  Presence of auto correlation

So, this is the stata command estat bgodfrey.

(Refer Slide Time: 2:17)

Stata 16.0

```

    . reg wage productivity
    . estat bgodfrey
  
```

Source	SS	df	MS	Number of obs	F	Prob > F
Model	6274.75662	1	6274.75662	71	18	0.0000
Residual	272.82393	38	7.158714			0.9584
Total	6546.77854	39	167.866316			0.9574

lag(p)	chi2	df	Prob > chi2
1	32.205	1	0.0000

HB: no serial correlation

Stata 16.0

```

    . reg u1 u2 u3 u4 u5 u6 productivity
    . estat bgodfrey
  
```

Source	SS	df	MS	Number of obs	F	Prob > F
Model	171.172596	7	24.4532866	71	18	0.0000
Residual	20.722595	26	.79702234			0.8930
Total	191.895191	33	5.81503549			0.8629

lag(p)	chi2	df	Prob > chi2
1	32.205	1	0.0000

Now, see the chi square value is 32.20 and again highly significant even. But there is a difference between the test what we have conducted and slight difference what stata is reporting. See in our manual result what we did we have specified 6 lags but stata is reporting p equals to only 1, what is the number of lag? Only 1.

Now, your question is you are interested in testing higher order lags that is why we are going for Breusch and Godfrey test but stata is reporting only result with 1 period lag it is as good as the

Durbin Watson test statistic, why this is so? Now, the reason is even though Breusch and Godfrey can detect higher order autocorrelation if you look at your result.

See apart from  $u_1$  that means apart from the first order lag  $u_2, u_3, u_4, u_5, u_6$  all these coefficients are actually insignificant, that is the reason stata is reporting the Breusch and Godfrey test with only 1 period lag. Had there been higher order lag, stata would have considered those number of lags. Let us say in the data you have lag up to 3 periods then stata would have reported the result with lag equals to 3.

But here, only  $u_1$  is significant that means  $u_{t-1}$  is significant,  $u_{t-2}, u_{t-3}$  all are insignificant. Since all our insignificant stata has not reported the result with higher period lags, stata has reported only this, this should be your conclusion.

So now, what we have learned? We have learned Durbin Watson test, we have also learned the Breusch and Godfrey test to test for higher order autocorrelation. Now, the next question that comes to our mind, that once you detect autocorrelation, what is the solution?

(Refer Slide Time: 5:19)

Solutions for autocorrelation

1. First check whether autocorrelation is pure or due model misspecification

$$y_t = \alpha + \beta x_t + u_t$$

↓  
prod

$$y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t$$

So, solutions for autocorrelation. Now, the first solution is what you need to check is whether autocorrelation is pure or due to model misspecification. So, that means when we were running this model  $y_t$  equals to  $\alpha + \beta x_t + u_t$  where  $y_t$  is wage and  $x_t$  is productivity, it may

so happen that in the true model there is non-linearity between wage and productivity and that is actually leading to this autocorrelation problem.

So, that means we have to check whether  $y_t$  equals to  $\alpha + \beta_1 x_1 t + \beta_2 x_2 t + u_t$ , sorry this is not  $2t$  I will say this is  $\beta_2 x_1 t^2$ , so that means I am including productivity as well as square in the model to check whether there is any model misspecification. if the autocorrelation is due to model misspecification, then once you correct that by a square term your autocorrelation problem will get automatically resolved, you do not need to do anything else. That is why econometrician first say that before applying any other medicine you just see whether autocorrelation is pure or due to model misspecification. So, how will you do that? You need to generate the productivity square.

(Refer Slide Time: 8:15)

The screenshot shows the Stata command window with the following commands and results:

```

. estat bgodfrey
Breusch-Godfrey LM test for autocorrelation
-----+-----
lags(l)      ch12      df      Prob > ch12
-----+-----
1            32.285      1      0.0000

          NB: no serial correlation

. gen x=productivity^2
. reg wage productivity x
Source      SS          df           MS       Number of obs = 40
-----+-----+-----+-----+-----
Model    6532.18781      2    3256.09391      F(2, 37) = 3482.88
Residual 34.5097227      37    .93404397      Prob > F = 0.0000
Total   6546.77954      39    167.866116      R-squared = 0.9947
                                           Adjusted R-squared = 0.9944
                                           Root MSE = .96889

wage      Coef.   Std. Err.      t    P>|t|   [95% Conf. Interval]
-----+-----+-----+-----+-----
productivity  1.94883   .0779943    24.99   0.000    1.790799   2.106861
x          -.0079546  .0004608   -15.34   0.000   -.0093311  -.0065781
_cons     -18.21817  2.91459    -5.49   0.000   -22.20473  -14.2316
  
```

So, gen, let us say I am giving productivity a different name which is let us say x and how I am defining x? This is nothing but productivity, productivity square, productivity square and then automatically my x variable is included. So now, what you need to do reg wage then productivity and you also include your x.

Now, interestingly you see the productivity is positive, coefficient of productivity is positive but productivity square is actually negative and significant that means as productivity increases wage first increases but after some point of time it is coming down, it is not increasing that much, so

there might be some reason behind this that means some kind of non-linearity might exist between wage and productivity. you might explore the reason I am not talking into that part here. So, what I will do now after this model once you estimate this modified model let us see what is the value of the Durbin Watson test statistic.

(Refer Slide Time: 9:50)

The screenshot shows the Stata command window with the following commands and results:

```

1 test time
2 reg wage productivity
3 dwstat
4 reg wage productivity
5 predict uresidual
6 gen u1-u5
7 gen u1-u1
8 gen u1-u2
9 gen u1-u3
10 gen u1-u4
11 gen u1-u5
12 reg u1-u2 u3 u4 u5 u1 productivity
13 reg wage productivity
14 estat lmtest
15 gen v=productivity^2
16 reg wage productivity v
17 dwstat
  
```

The results window displays the following statistics:

lags(p)	chi2	df	Prob > chi2
1	32.285	1	0.0000

LM: no serial correlation

Source	SS	df	MS	Number of obs =	40
Model	652.18781	2	326.09391	F(2, 37)	= 3482.88
Residual	34.5087227	37	.93484397	Prob > F	= 0.0000
Total	686.69654	39	17.6076036	R-squared	= 0.9947
				Adj R-squared	= 0.9944
				Root MSE	= .96689

	wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
productivity	1.94883	.8770613	24.39	0.000	1.798799	2.100561
v	-.0079366	.0004968	-15.34	0.000	-.0089231	-.00693
_cons	-18.21817	2.95459	-5.49	0.000	-22.20873	-18.2316

estat lmtest

Durbin-watson d-statistic( 3, 40) = 1.029979

So, dwstat. Now see, the productivity, after improving productivity square the Durbin Watson test statistic has improved a lot from 0.12292 it is now 1.02. So it has improved but still it is lower than the Durbin Watson lower value. What was the lower value we were discussing?

(Refer Slide Time: 10:31)

The slide contains the following handwritten notes:

- $Code = 0.1229$
- $d_L = 1.39112$
- $du = 1.60$
- $df < d_L$
- $\rightarrow$  there is  $\rightarrow$  no autocorrelation
- $d_L \approx 2(1 - \hat{\rho})$
- $= 0.1229 < d_L$
- $d_L = ?$
- $du = ?$
- $R = \text{total no. parameters in the model}$
- $\alpha : \text{sig}(1\%, 5\%)$
- Wage =  $f(\text{productivity})$ ?
- productivity =  $f(\text{Wage})$ ?
- Factor market: Product exhaustion theorem
- $W = VMP_L = P \times MP_L$
- $\ln w_{it} = \alpha + \beta \ln \text{productivity}_{it} + u_{it}$
- DW test detects only  $(AR)$

If you remember it is 1.39 and how much value you are getting here?

(Refer Slide Time: 10:40)

The software interface displays the following statistical results:

lags(p)	chi2	df	Prob > chi2
1	32.205	1	0.0000

Model Summary:

Source	SS	df	MS	Number of obs = 40
Total	652.18781	2	326.09391	F(2, 37) = 3482.88
Residual	34.5087227	37	.93484397	Prob > F = 0.0000
				R-squared = 0.9947
				Adj R-squared = 0.9944
				Root MSE = .96689

Regression Coefficients:

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
productivity	1.94883	.0779963	24.99	0.000	1.798799 2.100861
_cons	-.0079266	.0004968	-15.34	0.000	-.0089231 -.00693

Durbin-Watson d-statistic( 3, 40) = 1.029979

It is 1.02, so it is still lower than the lower limit of Durbin Watson test so that means the autocorrelation is not due to model misspecification, it is actually a pure autocorrelation meaning the other reasons that inertia and other things that means the time stage macro-economic variables they exhibit business cycle, so on and so forth and that is why this autocorrelation is actually existing in the data. So, that means by modifying the model we cannot solve

autocorrelation problem, we need to think about other method. And now we will discuss about other method of solving autocorrelation problem.

(Refer Slide Time: 11:46)

**Solutions for autocorrelation**

1. First check whether autocorrelation is pure or due model misspecification  

$$y_t = \alpha + \beta x_t + u_t$$

$\downarrow$   
 OLS                  Prod  
 $y_t = \alpha + \beta_1 x_t + \beta_2 x_t^2 + u_t$
2.  $\rho$  differencing is applicable when  $\rho$  is known  

$$y_t = \alpha + \beta x_t + u_t \quad \text{--- (1)}$$

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

$$y_{t-1} = \alpha + \beta x_{t-1} + u_{t-1} \quad \text{--- (2)}$$

$$\text{--- (1)} - \rho \text{--- (2)} \Rightarrow (y_t - \rho y_{t-1}) = (1 - \rho)\alpha + \rho(\beta x_{t-1}) + (u_t - \rho u_{t-1}) \quad \text{--- (3)}$$

Apply OLS in (3).  
if  $\rho = 1$

So, the second method is called rho differencing, where rho is the autocorrelation coefficient. So, your model was  $y_t$  equals to  $\alpha$  plus  $\beta x_t$  plus  $u_t$  and  $u_t$  follows  $u_t$  equals to  $\rho u_{t-1}$  plus  $\epsilon_t$  that is your data generating process.

So, you need to first take the difference of this, so that means  $y_t$  minus 1 should be equal to  $\alpha$  plus  $\beta x_t$  minus 1 plus  $u_t$  minus 1. Then what you need to do,  $y_t$  minus  $\rho y_{t-1}$  equals to, so this is  $\alpha$  minus  $\rho \alpha$ ,  $\alpha$  minus  $\rho \alpha$ , so I am multiplying let us say this is equation 1, this is equation 2 and I am multiplying equation 2 by  $\rho$  and then I am taking 1 minus 2 into  $\rho$  that implies  $y_t$  minus  $\rho y_{t-1}$ , this would become  $\alpha$  minus  $\rho \alpha$  so that means if you take  $1 - \rho$  into  $\alpha$ ,  $\alpha$  plus  $\beta$  into  $x_t$  minus  $\rho$   $x_{t-1}$  plus  $u_t$  minus  $\rho u_{t-1}$ . Let us say this is 3.

And then apply OLS in 3 that is the procedure. So, that means what I am saying while OLS was not applicable in equation 1, how come OLS is still applicable in this modified equation after rho differencing? Because if you look at the data generating process for the  $u_t$  this says that  $u_t$  minus  $\rho u_{t-1}$  is actually  $\epsilon_t$  and  $\epsilon_t$  follows all the assumption of classical linear regression model.



That is the reason this error term is actually the classical error term, it does not show any autocorrelation problem. So, this is the rho differencing method so that means you need to transform your dependent as well as independent variable. How?

If your dependent variable is y then the modified dependent variable would be  $y_t$  minus rho  $y_{t-1}$  and your independent variable would be transformed as  $x_t$  minus rho  $x_{t-1}$ . But the point here is this rho differencing method is applicable only when a rho is known to you but if you do not know the rho what will happen? If you know then you put the specific value of rho, if you do not know then what we need to do, we need to assume a specific value for rho. Now, if we assume, rho equals to 1 then what will happen?

(Refer Slide Time: 16:50)

The slide content includes the NPTEL logo and the following mathematical derivations on a whiteboard:

$$(y_t - \rho y_{t-1}) = (1-\rho)\alpha + \beta(x_t - \rho x_{t-1}) + (u_t - \rho u_{t-1})$$

If  $\rho = 1$

$$(y_t - y_{t-1}) = \beta(x_t - x_{t-1}) + (u_t - u_{t-1}) \rightarrow \text{First differencing}$$

$d < R^2 \Rightarrow \text{Apply first differencing}$

So, that means your rho transforming equation was  $y_t$  minus rho  $y_{t-1}$  equals to 1 minus rho into alpha plus beta  $x_t$  minus rho  $x_{t-1}$  plus  $u_t$  minus rho  $u_{t-1}$ . So, if you put rho equals to 1 that means if we assume there is perfect first order, perfect positive autocorrelation then this equation will become  $y_t$  minus  $y_{t-1}$  equals to beta  $x_t$  minus  $x_{t-1}$  plus  $u_t$  minus  $u_{t-1}$ , just by putting rho equals to 1 and this equation is called first differencing.

So, first differencing is a solution to solve autocorrelation problem when rho equals to 1. So, generally econometrician say that when d is less than R square apply first differencing. Now,

what we need to do, we need to transform the variable as  $y_t$  minus  $y_{t-1}$  and  $x_t$  as  $x_t$  minus  $x_{t-1}$ .

(Refer Slide Time: 19:19)

The screenshot shows the Stata 16.0 interface. The command window on the left contains the following commands:

```

1 test time
2 reg wage productivity
3 dwtat
4 reg dwtat dprod
5 predict resid
6 gen u1=ta
7 gen u2=ta2
8 gen u3=ta2
9 gen u4=ta2
10 gen u5=ta2
11 gen u6=ta2
12 reg u1 u2 u3 u4 u5 u6 productivity
13 reg wage productivity
14 estat lpsolby
15 gen x=productivity^2
16 reg wage productivity x
17 dwtat
18 gen dwtat=d.wage
19 gen dprod=d.productivity

```

The main window displays the results of the regression of wage on productivity:

Source	SS	df	MS	Number of obs	=	40
Model	6512.18781	2	3256.09391	F(2, 37)	=	3882.88
Residual	34.5907227	37	.934884397	Prob > F	=	0.0000
Total	6546.77854	39	167.86518	R-squared	=	0.9947
				Adj R-squared	=	0.9944
				Root MSE	=	.96689

The coefficient table is as follows:

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
productivity	1.94883	.0779943	24.99	0.000	1.798799 2.104861
x	-.0079166	.0004968	-15.94	0.000	-.0089231 -.00691
_cons	-14.21817	2.95459	-5.49	0.000	-22.20473 -10.2316

The Durbin-Watson d-statistic is 1.829979. The command window also shows the following commands:

```

. gen dwtat=d.wage
(1 missing value generated)
. gen dprod=d.productivity
(1 missing value generated)

```

So, what we will do, we will transform the variable, so we will say that  $d$  wage, this is  $d$  wage I am writing equals to  $d$  dot wage. So, what is the command I am using? Earlier I was using  $l$ ,  $l$  for lag and  $d$  for difference. So, the moment I say  $d$  wage,  $d$  wage is basically a name and  $d$  dot wage means wage  $t$  minus 1 that means basically  $y_t$  minus  $y_{t-1}$ . I have created the first difference. Similarly,  $d$  prod equals to  $d$  dot productivity. So, I have created, I have created the two, first difference of wage as well as productivity. And now if you regress  $d$  wage on  $d$  productivity then we will see.

(Refer Slide Time: 20:55)

The slide shows a hand-drawn equation in a white box on a black background:

$$(y_t - \rho y_{t-1}) = (1-\rho)\alpha + \beta(x_t - \rho x_{t-1}) + (u_t - \rho u_{t-1})$$

Below this, it says: "If  $\rho = 1$ "

$$(y_t - y_{t-1}) = \beta(x_t - x_{t-1}) + (u_t - u_{t-1}) \rightarrow \text{First differencing}$$

Below that, it says: " $d < R \Rightarrow$  Apply first differencing"

In the bottom left corner, there is a small video inset of a man with a shaved head, wearing a striped shirt, sitting at a desk and speaking.

But before we do that you, we have to again look back to our model see in the first difference equation there is no constant term. The constant term gets cancelled when you put rho equals to 1, so in any first difference equation you have to keep in mind whenever you run first difference equation you should not include constant term in the model.

(Refer Slide Time: 21:15)

The screenshot shows the Stata software interface. The main window displays the results of a regression analysis. The command window shows the following commands:

```

1. gen dlnprod=lnprod
2. reg lnprod dlnprod
3. estat ic
4. reg lnprod dlnprod
5. predict uresidual
6. gen u1=u1
7. gen u2=u2
8. gen u3=u3
9. gen u4=u4
10. gen u5=u5
11. gen u6=u6
12. reg u1 u2 u3 u4 u5 u6 lnprod
13. reg lnprod dlnprod
14. estat ic
15. gen x1=productivity*2
16. reg lnprod x1
17. estat ic
18. gen dlnprod=dlnprod
19. gen dlnprod=dlnprod
20. reg dlnprod dlnprod
21. reg dlnprod

```

The main window shows the following results:

Durbin-Watson d-statistic( 3, 40) = 1.029979

Source SS df MS Number of obs = 39

Model	18.2855082	1	18.2855082	F(1, 37)	=	21.88
Residual	33.3324534	37	.90087712	R-squared	=	0.3629
Total	52.3189536	38	1.37681457	Adj R-squared	=	0.3457
				Root MSE	=	.94915

reg dlnprod


	Source	SS	df	MS	Number of obs	=	39
dlnprod	18.2855082	1	18.2855082	F(1, 37)	=	21.88	
_cons	33.3324534	37	.90087712	R-squared	=	0.3629	
Total	52.3189536	38	1.37681457	Adj R-squared	=	0.3457	
				Root MSE	=	.94915	

reg dlnprod

	coef.	Std. Err.	t	P> t	[95% Conf. Interval]
dlnprod	.4794336	.1473985	4.39	0.000	.3795682 .5793071
_cons	.8929887	.2842564	0.32	0.748	-.4838194 .6680568

Command: reg dlnprod dlnprod

In the bottom left corner, there is a small video inset of the same man from the previous slide, speaking.



StatView HD

File Edit Data Graphics Statistics User Window Help

History

1. list time  
2. reg wage productivity  
3. dstat  
4. reg wage productivity  
5. predict cnsdual  
6. gen u1-u5  
7. gen u2-u5  
8. gen u3-u5  
9. gen u4-u5  
10. gen u5-u5  
11. gen u5-u5  
12. reg u1 u2 u3 u4 u5 productivity  
13. reg wage productivity  
14. estat lqplotly  
15. gen x productivity^2  
16. reg wage productivity x  
17. dstat  
18. gen d wage d wage  
19. gen dprod d productivity  
20. **reg d wage d productivity** \*\*\*  
21. reg dprod d productivity  
22. **estat dprod, nocns**

Source	SS	df	MS	Number of obs =	39
Model	18.995982	1	18.995982	F(1, 37)	= 21.08
Residual	33.3324534	37	.90087712	Prob > F	= 0.0000
Total	52.3284356	38	1.37881457	Adj R-squared	= 0.3457
				Root MSE	= .94915

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dprod	-.6794136	-.1479985	4.59	0.000	-.7795682	-.5792671
_cons	.8920987	.2842564	0.32	0.748	-.4839394	.6689168

**. reg d wage dprod, nocns**

Source	SS	df	MS	Number of obs =	39
Model	74.5729524	1	74.5729524	F(1, 38)	= 84.77
Residual	33.4270231	38	.879658094	Prob > F	= 0.0000
Total	107.999976	39	2.76923014	Adj R-squared	= 0.6823
				Root MSE	= .9379

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dprod	-.7199553	-.0781937	9.21	0.000	-.8616604	-.5782582

Command

Variables

Name Label

year

wage

productivity

time

u1 Residual

u2

u3

u4

u5

Properties

Variables

Name Label

Type

Format

Value Label

Display

Data

Frame default

Element



Label

Series

Variables 14

Observations 42

View

StatView HD

File Edit Data Graphics Statistics User Window Help

History

1. list time  
2. reg wage productivity  
3. dstat  
4. reg wage productivity  
5. predict cnsdual  
6. gen u1-u5  
7. gen u2-u5  
8. gen u3-u5  
9. gen u4-u5  
10. gen u5-u5  
11. gen u5-u5  
12. reg u1 u2 u3 u4 u5 productivity  
13. reg wage productivity  
14. estat lqplotly  
15. gen x productivity^2  
16. reg wage productivity x  
17. dstat  
18. gen d wage d wage  
19. gen dprod d productivity  
20. **reg d wage d productivity** \*\*\*  
21. reg dprod d productivity  
22. **estat dprod, nocns**

Source	SS	df	MS	Number of obs =	39
Model	18.995982	1	18.995982	F(1, 37)	= 21.08
Residual	33.3324534	37	.90087712	Prob > F	= 0.0000
Total	52.3284356	38	1.37881457	Adj R-squared	= 0.3457
				Root MSE	= .94915

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dprod	-.6794136	-.1479985	4.59	0.000	-.7795682	-.5792671
_cons	.8920987	.2842564	0.32	0.748	-.4839394	.6689168

**. reg d wage dprod, nocns**

Source	SS	df	MS	Number of obs =	39
Model	74.5729524	1	74.5729524	F(1, 38)	= 84.77
Residual	33.4270231	38	.879658094	Prob > F	= 0.0000
Total	107.999976	39	2.76923014	Adj R-squared	= 0.6823
				Root MSE	= .9379

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dprod	-.7199553	-.0781937	9.21	0.000	-.8616604	-.5782582

**. estat dstat**

Durbin-Watson d-statistic( 1, 39) = 1.589653

Command

Variables

Name Label

year

wage

productivity

time

u1 Residual

u2

u3

u4

u5

Properties

Variables

Name Label

Type

Format

Value Label

Display

Data

Frame default

Element


Label

Series

Variables 14

Observations 42

View



The slide features a hand-drawn diagram on a whiteboard background. At the top left is the NPTEL logo. The diagram shows the following equations and notes:

$$(y_t - \beta y_{t-1}) = (1-\beta)\alpha + \beta(x_t - \beta x_{t-1}) + (u_t - \beta u_{t-1})$$

If  $\beta = 1$

$$(y_t - y_{t-1}) = \alpha(x_t - x_{t-1}) + (u_t - u_{t-1}) \rightarrow \text{First differencing}$$

$d < R \Rightarrow$  Apply first differencing in original equation

Below the whiteboard, there is a video inset showing a man in a striped shirt speaking into a microphone.

So, now you reg d wage d prod then no cons. this is the command. Now, here again after running this first difference equation what is the Durbin Watson value?

Durbin Watson value has improved a lot 1.5096 in the first difference, so the Durbin Watson value is 1.50 which is actually now if you go back which is if you go back then you will see that your value was Durbin Watson lower value was something around 1.34 or something, so that means it has now improved a lot.

(Refer Slide Time: 24:18)

The screenshot shows the Stata command window with the following commands and output:

```
1. test time
2. reg wage productivity
3. dstat
4. reg wage productivity
5. predict considual
6. gen u1=ua
7. gen u2=ut
8. gen u3=ua2
9. gen u4=ua3
10. gen u5=ua4
11. gen u6=ua5
12. reg u1 u2 u3 u4 u5 u6 productivity
13. reg wage productivity
14. estat lqplot
15. gen x=productivity^2
16. reg wage productivity x
17. dstat
18. gen d wage-d wage
19. gen dprod-d productivity
20. reg dprod-d productivity
21. lstat
22. lsdresid
23. lsdresid
24. lsdresid
25. lsdresid
26. lsdresid
27. lsdresid
28. lsdresid
29. lsdresid
30. lsdresid
31. lsdresid
32. lsdresid
33. lsdresid
34. lsdresid
35. lsdresid
36. lsdresid
37. lsdresid
38. lsdresid
39. lsdresid
40. lsdresid
41. lsdresid
42. lsdresid
43. lsdresid
44. lsdresid
45. lsdresid
46. lsdresid
47. lsdresid
48. lsdresid
49. lsdresid
50. lsdresid
51. lsdresid
52. lsdresid
53. lsdresid
54. lsdresid
55. lsdresid
56. lsdresid
57. lsdresid
58. lsdresid
59. lsdresid
60. lsdresid
61. lsdresid
62. lsdresid
63. lsdresid
64. lsdresid
65. lsdresid
66. lsdresid
67. lsdresid
68. lsdresid
69. lsdresid
70. lsdresid
71. lsdresid
72. lsdresid
73. lsdresid
74. lsdresid
75. lsdresid
76. lsdresid
77. lsdresid
78. lsdresid
79. lsdresid
80. lsdresid
81. lsdresid
82. lsdresid
83. lsdresid
84. lsdresid
85. lsdresid
86. lsdresid
87. lsdresid
88. lsdresid
89. lsdresid
90. lsdresid
91. lsdresid
92. lsdresid
93. lsdresid
94. lsdresid
95. lsdresid
96. lsdresid
97. lsdresid
98. lsdresid
99. lsdresid
100. lsdresid
```

	change	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
dprod		.7199553	.0781937	9.21	0.000	.5616404 .8782502

Output for `reg wage productivity`:

```
. reg wage productivity
Source          SS           df           MS           Number of obs   =    40
                ===============+-----+-----+-----+-----+-----+-----
                Total         6546.77954    39          167.866116     F(1, 38)         =   876.55
                Model         6274.75662     1          6274.75662     Prob > F          =  0.0000
                Residual      272.022913     38          7.1584714      R-squared         =  0.9584
                Total         6546.77954    39          167.866116     Adj R-squared     =  0.9524
                Model         6274.75662     1          6274.75662     Root MSE         =  2.6755
```

	wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
productivity		.7138594	.0241848	29.41	0.000	.6648619 .7624569
_cons		29.51926	1.942346	15.20	0.000	25.58718 33.45133

Output for `reg dprod-d productivity`:

```
. reg dprod-d productivity
Source          SS           df           MS           Number of obs   =    40
                ===============+-----+-----+-----+-----+-----+-----
                Total         6546.77954    39          167.866116     F(1, 38)         =   876.55
                Model         6274.75662     1          6274.75662     Prob > F          =  0.0000
                Residual      272.022913     38          7.1584714      R-squared         =  0.9584
                Total         6546.77954    39          167.866116     Adj R-squared     =  0.9524
                Model         6274.75662     1          6274.75662     Root MSE         =  2.6755
```

	dprod-d	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
dprod-d		.7138594	.0241848	29.41	0.000	.6648619 .7624569
_cons		29.51926	1.942346	15.20	0.000	25.58718 33.45133

Durbin-Watson d-statistic( 1, 39) = 1.509653

Durbin-Watson d-statistic( 2, 40) = .1229845

If you look at the original equation that means this equation you have to see reg wage and productivity where we have run that equation, if you look at the reg wage productivity, reg wage and productivity and then dw stat yeah, so this is lower than the R square and first difference is applicable.

But the only problem with the first difference is that we are specifying a specific value for rho hat which is 1 and also the moment you take first difference, the interpretation of the var, of the coefficient we were interested to see the impact of  $x_t$  on  $y_t$  but ultimately what we are getting the impact of  $\Delta y_t$  on  $\Delta x_t$  that is the problem. But anyway, the severity of autocorrelation is largely reduced.

(Refer Slide Time: 25:26)

NPTEL

$$(y_t - \rho y_{t-1}) = (1-\rho)\alpha + \rho(x_t - \rho x_{t-1}) + (u_t - \rho u_{t-1})$$

If  $\rho = 1$

$$(y_t - y_{t-1}) = \rho(x_t - x_{t-1}) + (u_t - u_{t-1}) \rightarrow \text{First differencing}$$

$d < 2 \Rightarrow$  Apply first differencing in original equation

When  $\rho$  is unknown

A.  $d \approx 2(1-\hat{\rho})$

B.  $\hat{u}_t = \hat{\rho} \hat{u}_{t-1} + \epsilon_t \Rightarrow \hat{\rho}$

There is no guarantee that this  $\hat{\rho}$  is the best

The next question what we need to do is if we do not know the value of when this rho is unknown what actually you can do instead of putting rho equals to 1 there are two ways by which you can get rho. See if you recall we have a relationship like this is 2 into 1 minus rho hat, so the moment you know Durbin Watson test statistic using this relationship you can get rho hat, so you estimate the model, get your Durbin Watson test statistic and then put the Durbin Watson value here and get your rho hat, that is method number 1.

Method number 2 is you estimate the model, get your  $u_t$  hat and then run this regression reg this equals to rho  $u_t$  minus 1 plus epsilon t, so that means if you regress  $u_t$  hat on  $u_t$  minus 1 hat from there also you will get rho hat that is also possible, so when rho is unknown these are the two ways by which you can get rho hat but please keep in mind the relationship  $d$  is almost equals to 2 into 1 minus rho hat is applicable when you have sufficiently large number of time periods.

Here the number of time periods is 40 which is large but if you have 10, 15 years data then this is not actually applicable this method you have to use the method number B this one, you can simply regress  $u_t$  on  $u_t$  minus 1 and get your rho hat. Now, econometrician say sometimes that in this method how do you know that the rho hat is the best method? Because you estimate the model regress  $u_t$  hat on  $u_t$  minus 1 hat, collect that rho hat and using that rho hat to transform the variable, so there is no guarantee that this rho hat is the best.

(Refer Slide Time: 28:29)

**Cochrane-Orcutt procedure:**

st 1.  $y_t = \alpha + \beta x_t + u_t$

st 2.  $u_t = \rho u_{t-1} + \epsilon_t \Rightarrow \hat{\rho}_1$

st 3.  $(y_t - \hat{\rho}_1 y_{t-1})$  &  $(x_t - \hat{\rho}_1 x_{t-1})$   
 $= y_t^*$  &  $x_t^*$

st 4. Reg  $y_t^*$  on  $x_t^*$  and get  $u_t^*$   
 $u_t^* = \rho_2 u_{t-1}^* + \epsilon_t \Rightarrow \hat{\rho}_2$

$\hat{\rho}_1 \dots \hat{\rho}_2 \dots \hat{\rho}_3$  will continue until  
 $|\hat{\rho}_2 - \hat{\rho}_3| < 0.001$

That is why they sometimes suggest to iterate this procedure for one or two couple of more rounds so and that procedure is called Cochrane Orcutt procedure. What they say? You run this regression  $y_t$  equals to alpha plus beta  $x_t$  plus  $u_t$  and then you regress  $u_t$  on  $\rho u_{t-1}$  plus epsilon  $t$  and then from there you collect your  $\rho_1$   $u_{t-1}$   $\rho_1$  hat.

Then what you do this is step 1, this is step 2, then in step 3 what you do you transform your dependent and independent variable with this  $\rho_1$  hat  $x_t$  minus  $\rho_1$  hat  $x_{t-1}$  so I have changed my this and let us say this is called  $y_t^*$  and let us say this is called  $x_t^*$ . Then what I need to do? I need to regress this  $y_t^*$  in step 4, reg  $y_t^*$  on  $x_t^*$  and get  $u_t^*$ .

Then what you need to do? Again, run  $u_t^*$  hat equals to  $\rho_2 u_{t-1}^*$  plus epsilon  $t$  and get  $\rho_2$  hat. Again, you take this  $\rho_2$  hat to transform your dependent and independent variable. So, that means you will get  $\rho_1$  hat,  $\rho_2$  hat,  $\rho_3$  hat in every successive iteration and I will stop when the two successive periods  $\rho$  hat is actually equal and this will continue until  $\rho_2$  hat minus  $\rho_3$  hat equals to  $\rho_3$  hat or  $\rho_2$  hat and the  $\rho_3$  hat this difference is less than 0.001. And I will take that  $\rho$  hat to transform the variable, I will take that  $\rho$  hat to transform the variable instead of using this  $\rho_1$  hat which is coming in the first stage. Is this clear?

So, this procedure Cochrane Orcutt Procedure is nothing but repeating the procedure for couple of more rounds, more iteration and then take that particular  $\rho$  when two successive periods  $\rho$



hat actually converges to each other or the difference between the two rho hat is less than 0.001. And in the process how will you do that?

(Refer Slide Time: 33:01)

Stata 16.0

```

    . reg wage productivity
    . dstat
    Durbin-Watson d-statistic( 1, 39) = 1.509653

    . reg wage productivity
    Source      SS      df      MS      Number of obs = 40
    Model      6274.75662    1    6274.75662    Prob > F      = 0.0000
    Residual   272.82393    38    7.1584714    R-squared     = 0.9584
    Total      6546.77954    39    167.866316    Adj R-squared = 0.9574
    Root MSE   = 2.6755

    . reg wage productivity
    wage      Coef.  Std. Err.   t    P>|t|   [95% Conf. Interval]
    productivity  .7136594  .0241048   29.61  0.000   .6648619   .7624569
    _cons       29.51926  1.942366   15.20  0.000   25.58718   33.45133

    . dstat
    Durbin-Watson d-statistic( 2, 40) = .1229045
  
```

Command: prais wage productivity,corr

Stata 16.0

```

    . dstat
    Durbin-Watson d-statistic( 1, 39) = 1.509653

    . reg wage productivity
    Source      SS      df      MS      Number of obs = 40
    Model      6274.75662    1    6274.75662    Prob > F      = 0.0000
    Residual   272.82393    38    7.1584714    R-squared     = 0.9584
    Total      6546.77954    39    167.866316    Adj R-squared = 0.9574
    Root MSE   = 2.6755

    . reg wage productivity
    wage      Coef.  Std. Err.   t    P>|t|   [95% Conf. Interval]
    productivity  .7136594  .0241048   29.61  0.000   .6648619   .7624569
    _cons       29.51926  1.942366   15.20  0.000   25.58718   33.45133

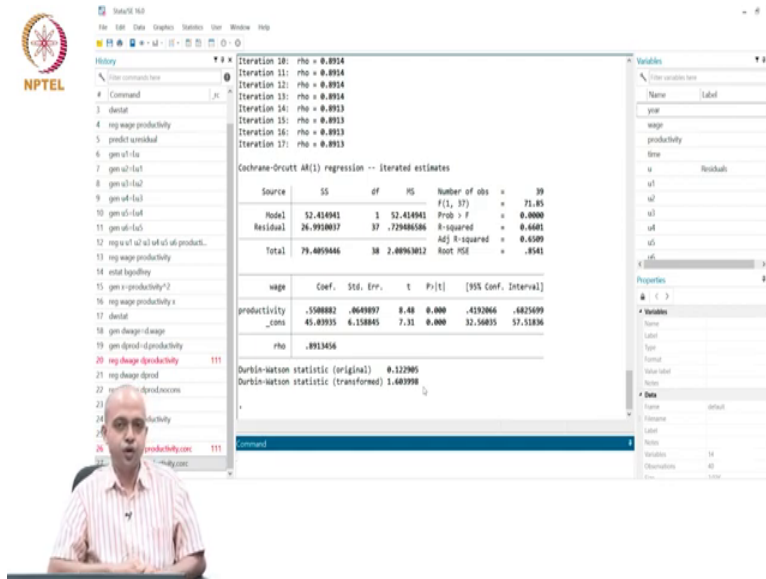
    . dstat
    Durbin-Watson d-statistic( 2, 40) = .1229045

    . prais wage productivity,corr
    variable wage not found
    r(111);

    Command: prais wage productivity,corr
  
```

In the data if you want to apply this Cochrane Orcutt Procedure to correct your yt and xt, transform your yt and xt and then run the regression, problem of autocorrelation will come down a lot, I will show you the command, this is again a specific command, prais wage productivity in corc, this is the command.

(Refer Slide Time: 34:09)



Now, look at what stata is doing? Stata is running 17 rounds of iteration and then lastly stata is taking 0.8913 as the best rho. And that rho is used to transform your yt that means yt is now yt minus rho yt minus 1 xt is now xt minus rho xt minus 1 and then the regression is run using this rho this is the regression output and you see the Durbin Watson original value which was 0.122 now it is 1.60.

So, you have solved actually the autocorrelation problem and you have got the best result, this is how you can solve your autocorrelation problem by a generalized method of rho differencing suggested by Cochrane and Orcutt Procedure. So, with this we are closing our discussion on autocorrelation, we have discussed about what is autocorrelation, what are the consequences, consequence is basically the standard error gets disturbed, efficiency property get disturbed and as a result of which you have problem in your hypothesis testing.

And then we have discussed about how to detect autocorrelation using Durbin Watson as well as Breusch and Godfrey test and lastly, we discussed about solution, rho differencing and first differencing. And then we lastly discussed about Cochrane Orcutt Procedure which is an iterative procedure to get the best rho to transform your dependent and independent variables. With this the discussion of autocorrelation is over. Thank you.