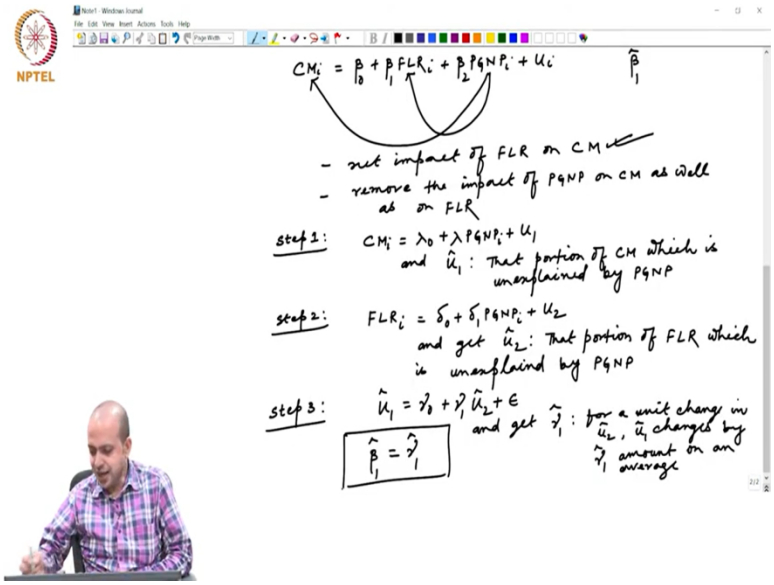
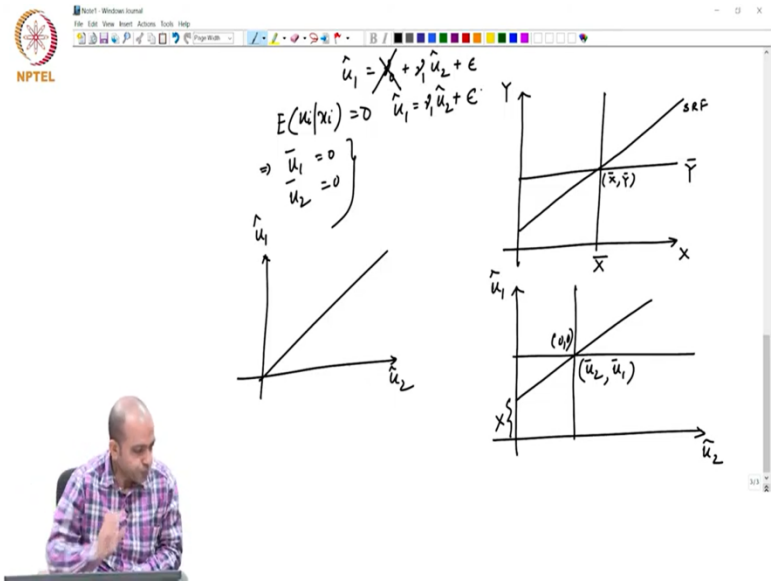
Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology Madras
Application of STATA for hypothesis testing and introduction to multiple linear regression model Part - 4

(Refer Slide Time: 00:14)



So, I will make you understand what is the mistake I have committed in step 3.

(Refer Slide Time: 00:18)



In step 3, basically what we did, u1 hat equals to gamma 0 plus gamma1 u2 hat plus epsilon. This is the equation we have specified. Now, if you recall, one of the properties of our sample regression function is that it must pass through the origin.

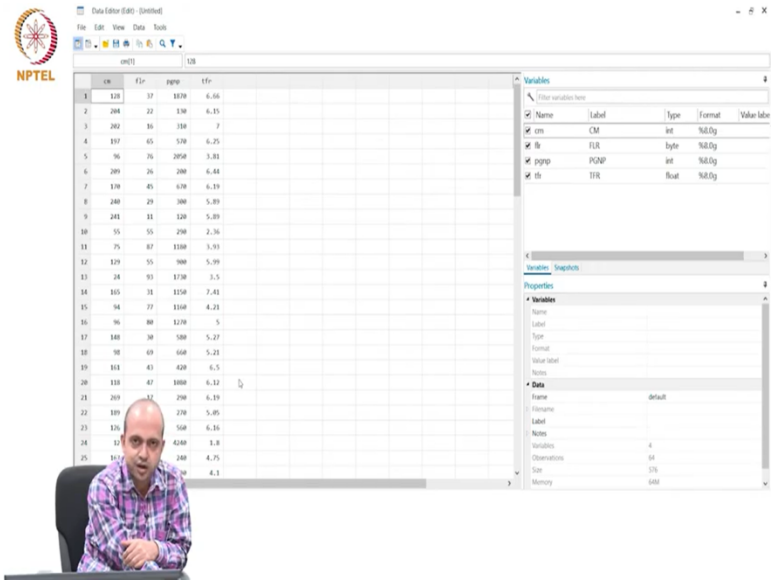
That means, this is basically x bar and y bar. This is the property that sample regression function- must pass through the sample mean or average. That means here when I am running a regression of y on x, it should pass from x bar and y bar. That is one of the important property that we have discussed. This is the SRF. Now, in this regression, what we are doing? If you replace x by u2 hat and your dependent variable here is u1 hat, then that should also pass through the origin.

So, that means this should be u2 bar and u1 bar. So, that should also pass through the sample average. Now, my question is what is the value of u2 bar and u1 bar? If you recall, one of our assumptions what we mentioned is that expectation of ui given xi is equal to 0. If that is the case, following this assumption we can say that u1 bar equals to 0 and u2 bar is also equal to 0. And that means, this is the co-ordinate called 0, 0. So, that means this intercept should not be there in this equation.

So, that means when you are running an equation of y2, u2 hat on u1 hat, that regression must pass through the origin. This point is 0, 0. So, that means this intercept should not be there. If the intercept is there, that means that cannot be the co-ordinate of 0,0. So, that means in this equation, when I am specifying this term should not be 0, should not be there. So, our equation would be u1 hat equals to gamma1 u2 hat plus epsilon. No intercept should be there and that is coming from the property- one of the important property of the sample regression function- that sample regression function must pass through the origin.

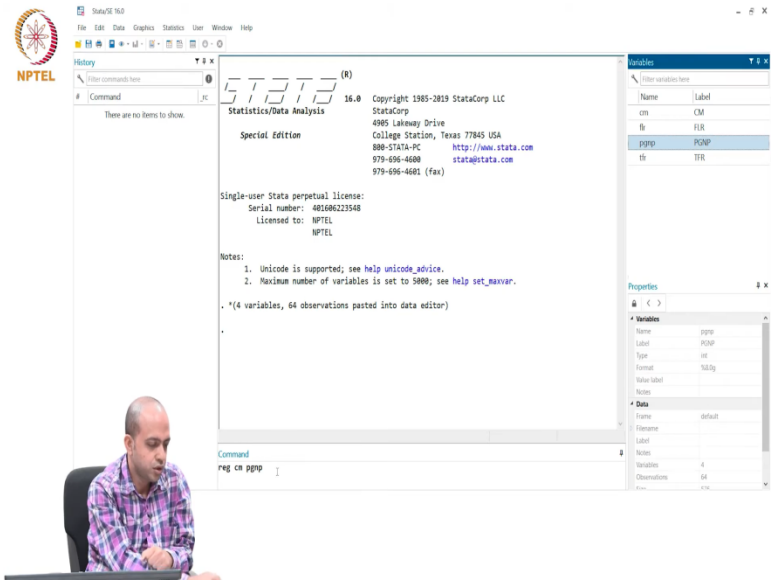
And since then u1 and u2, they both have 0 mean, the sample average of u1 and u2 is 0, so the line must pass through the origin, there should not be any intercept. So, this is basically the theoretical portion. That means this is the theory of how to keep the impact of other factor constant. Now, these steps, what we have learnt now, we will try to perform using the data set.

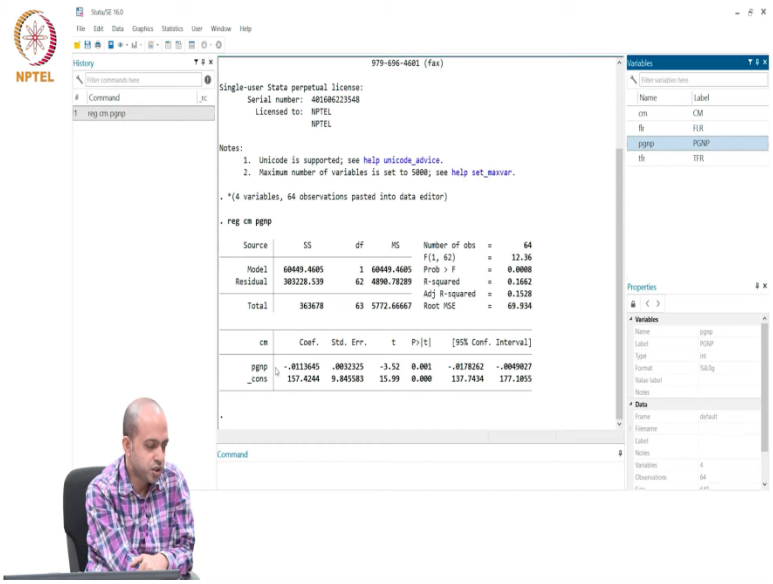
(Refer Slide Time: 04:39)



So, this is the data what we are getting. Look at this. So, first I will show you the data. This is the data, this is child mortality rate, this is female literacy rate, this is per capita GNP and this is total fertility rate, some other variable and we have state level data. So, what we will do now? We will try to follow that step.

(Refer Slide Time: 05:10)



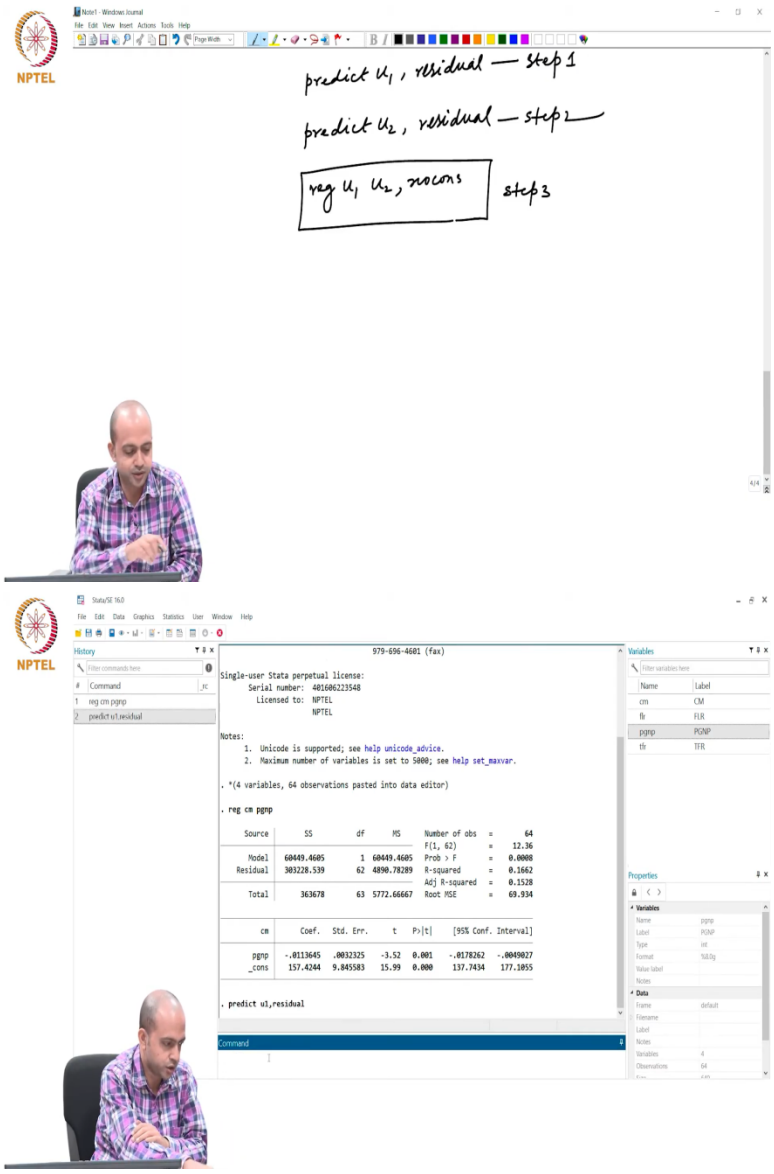


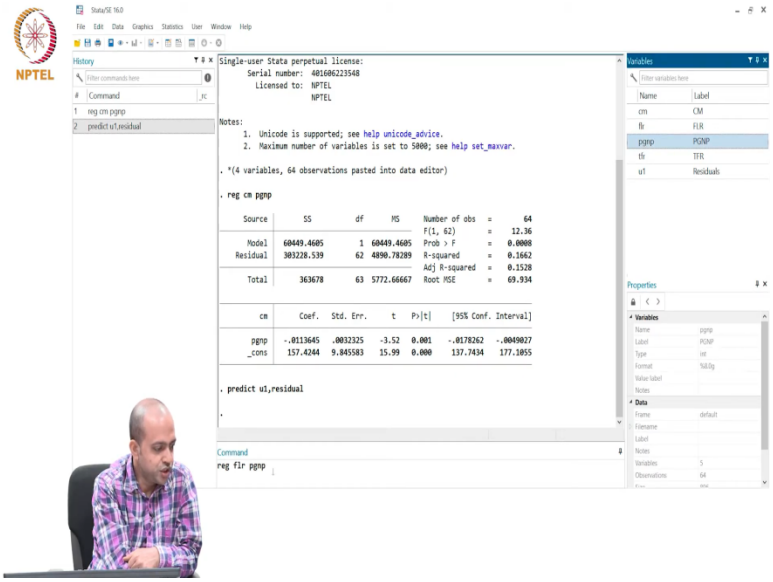
So, we have child mortality rate, FLR and PGNP and we want to get the net impact of FLR on child mortality rate. So, step 1, what we need to do? We need to run a regression where the dependent variable is CM and independent variable would be PGNP because we need to remove the impact of PGNP from both CM and FLR. Since we are interested to get net impact of FLR, we are eliminating the impact of PGNP. So, first step, we are removing the impact of PGNP of CM. How to do that?

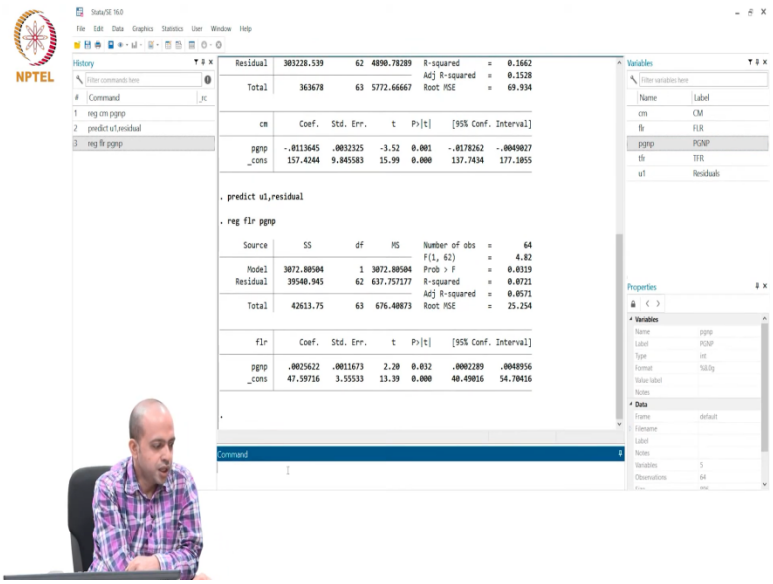
We will run a regression, **reg CM PGNP**. This is how we will run the regression. You can easily interpret this coefficient also. See there is a negative relationship between per capita GNP and CM. As income of a particular state increases, child mortality decreases by 0.01 unit on an average, which is expected also. As income increases, people become more aware, people take food, good rest and they give proper care in terms of medicine, so on and so forth. And as a result of which, child mortality rate goes down.

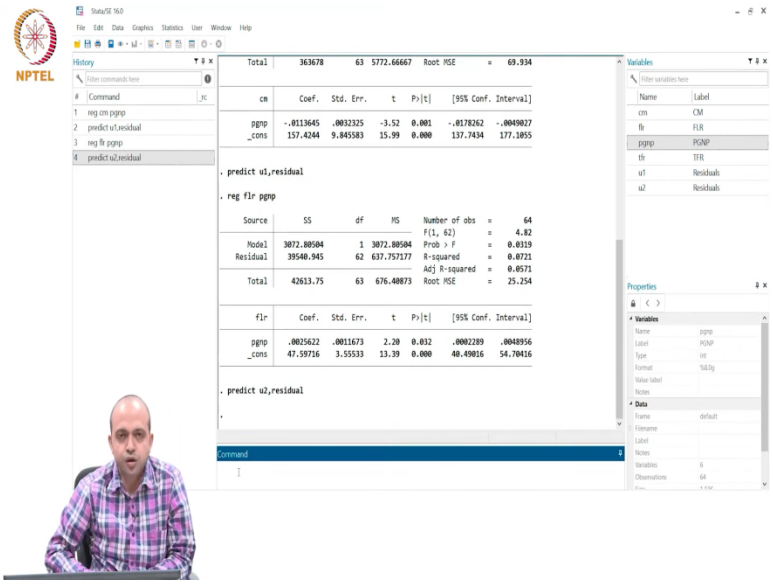
So, once we estimate this model, then what we need to do? From this regression, we need to collect the predicted value of the error term. Now, to collect the predicted value of the error term u1 hat from this regression, what we need to do? We need to put a specific command. We need to give a specific command and the command is very simple. I am writing the command also.

(Refer Slide Time: 07:13)




NPTEL
Stata/SE 16.0
File Edit Data Graphics Statistics User Window Help
History
Filter commands here
Command
1 reg cm pgnp
2 predict u1,residual
Single-user Stata perpetual license:
Serial number: 401606223548
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, 64 observations pasted into data editor)
. reg cm pgnp
Source SS df MS Number of obs = 64
F(1, 62) = 12.36
Model 60449.4605 1 60449.4605 Prob > F = 0.0008
Residual 303228.539 62 4890.78289 R-squared = 0.1662
Adj R-squared = 0.1528
Total 363678 63 5772.66667 Root MSE = 69.934
cm Coef. Std. Err. t P>|t| [95% Conf. Interval]
pgnp -.0113645 .0032325 -3.52 0.001 -.0178262 -.0049027
_cons 157.4244 9.845583 15.99 0.000 137.7434 177.1055
. predict u1,residual
.
Command
reg flr pgnp
Variables
Filter variables here
Name Label
cm CM
flr FLR
pgnp PGNP
tfr TFR
u1 Residuals
Properties
Variables
Name pgnp
Label PGNP
Type int
Format %8.0g
Value label
Notes
Data
Frame default
Filename
Label
Notes
Variables 5
Observations 64





The command is **predict**, and then you have to give some name. Predict command is always used for predicting something. That means when I am writing u1 hat, u2 hat, y hat, the hat means the predicted value, that is why predict command. What you are predicting? Some name you are giving. Predict u1. What is u1 basically? Residual. This is the command. Similarly, for u2, **predict u2, residual**. This is the command. So, now we will put this command here. We will say, this is predict u1, residual.

Now, the moment I put the command, Stata will immediately predict the u1 hat and look at here, u1 is already included as a variable now. Whether it is predicted or not, how will you check? You

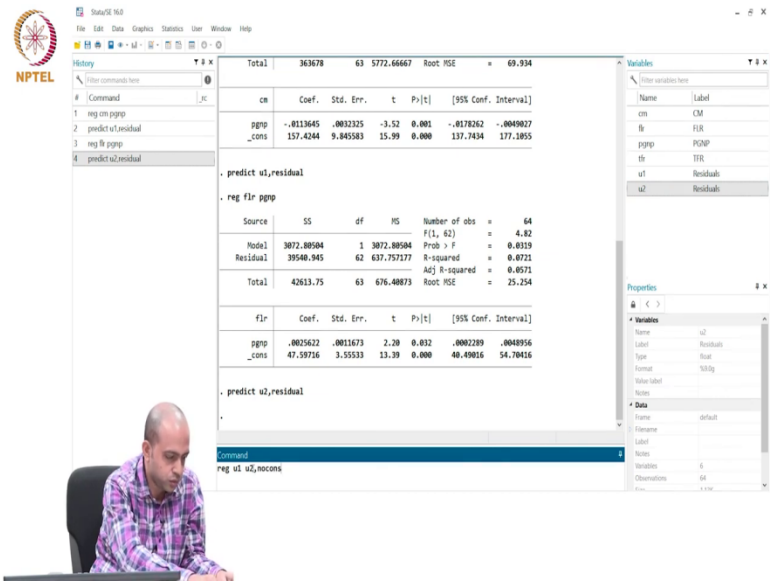
just look at in the variable list, if the variable is appearing, that means Stata has already predicted the variable and stored it also.

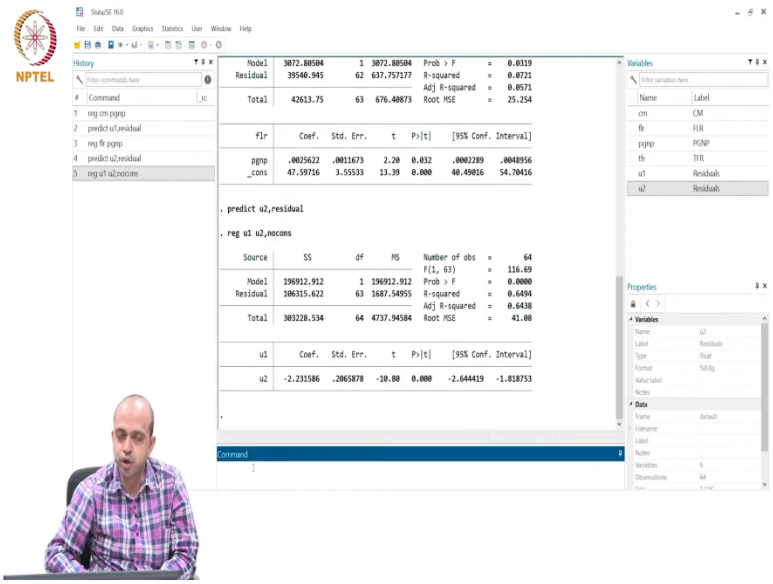
Now, what is the next step? In the next step, we need to remove the impact of PGNP on FLR. So, again regress FLR on PGNP, this is how. And the impact is, look at here, positive that means as per capita income GNP increases, FLR also increases. And the variable is significant at what percentage level can you think of from the p value? P value is 0.032.

So, if you multiply this value with 100, what will come? 3.2 which is greater than 1 but less than 5. That means this PGNP is significant at 5 percent level. This is how you have to interpret this coefficient. So, again from this, we need to predict and same command, predict u2 and then you put residual. So, once you put this command, then Stata will immediately predict the residual also. And how will you check whether it is predicted or not? Look at here, u2 is also included.

Now, what we have to do? We have to run a regression that means in step 3, u1 hat is regressed on u2 hat with no constant term. And what would be the command for that in step 3? Reg u1 on u2 and you need to specify Stata that I do not want any constant term. And what is the command for that? We need to specify **'no cons'**, this is the command in step 3. This is step 2, this is step 3, step 1. In step 1, we will regress and then we will collect this.

(Refer Slide Time: 11:47)



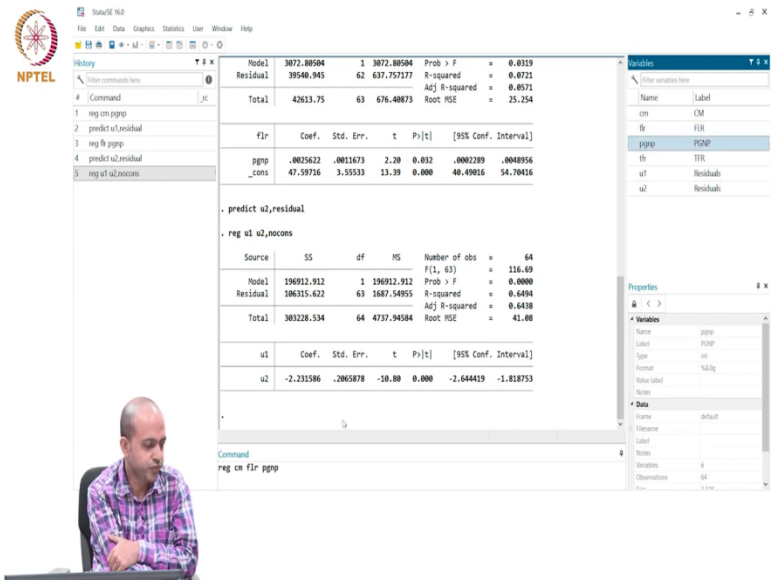


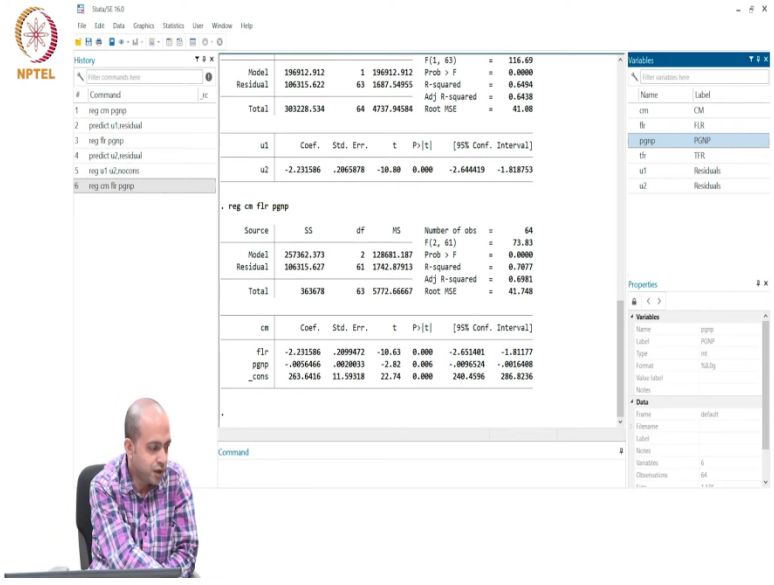
In step 3, what we need to do, we will regress u1 on u2 and then we will put no cons. So, what is the coefficient here? Coefficient of u2 is minus 2.231. So, that means for a unit change in u2, u1 changes by minus 2.231 unit on an average. But then, what is its relation with the original equation? Since u2 hat is basically the purified value of FLR and u1 is the purified value of child mortality rate, that means, the interpretation of this would be when FLR increases by 1 unit, then child mortality rate decreases by 2.231 unit.

So, u2 this 2.231 is basically the net impact of FLR on child mortality rate. Since this is negative, we can say that female literacy rate and child mortality rate, they are negatively related and which is very true also. As we said in the beginning that as female literacy rate increases, then mothers become more aware of what type of proper care they need to take during pregnancy and after delivery, child mortality rate will decrease. That is why the negative sign.

So, this is the step by step procedure what we should follow to keep the impact of other factor constant. But that does not mean that whenever we want to get the net impact of a particular variable in a multiple regression model, we will follow all these steps. Rather, at a single step itself, we can get the net impact. That is the beauty of this statistical software. How to do that? What was our original model? Our original model was CM equals to beta0 plus beta1 FLR plus beta2 PGNP.

(Refer Slide Time: 14:36)





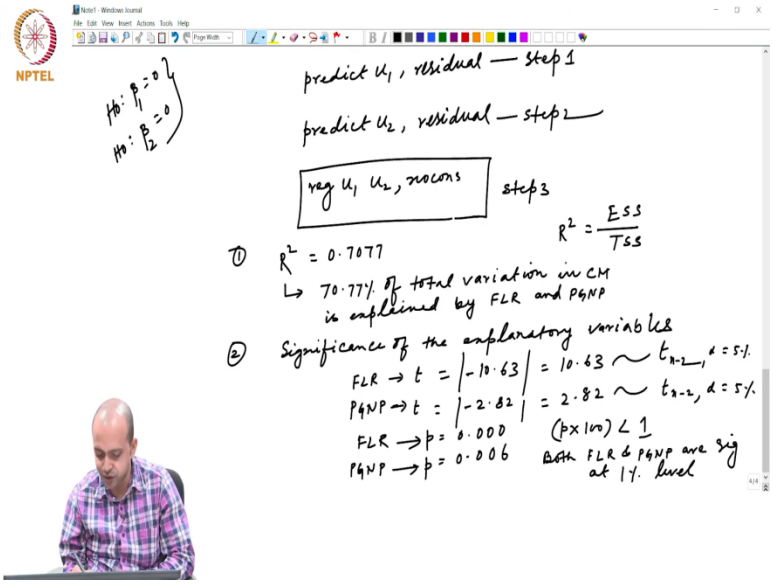
So, if you specify that regression, how will you specify? Our dependent variable was CM, then independent variables were FLR and PGNP. This is our complete model that we wanted to run and if we run this, now look at the beauty. What is the coefficient of FLR? minus 2.231. What was the coefficient what we got following the steps?

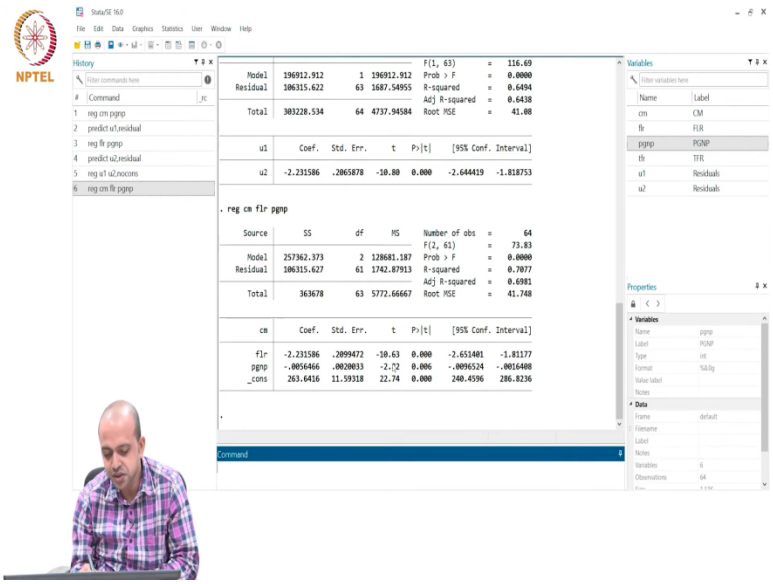
If you follow, that is also minus 2.231586. So, this coefficient is exactly following with the coefficient what we got earlier. So, that means what we can say? That the theoretically what we have derived and empirically what we have estimated, both are matching. So, in multiple linear regression model, the coefficient that we get is basically the net impact.

So, the statistical software Stata is actually doing everything for us. So, that means all those steps, step 1, step 2, step 3, what we discussed earlier to keep the other factor constant, Stata is basically doing that and Stata is supplying us directly the net impact of each and every variable. That is the beauty of this multiple linear regression model. And that is the beauty of this software. But we must know all the steps, what we actually should do to understand the interpretation of multiple linear regression model in a better way.

If we do not know the steps, then we may not be able to appreciate what we are actually getting from this. This is how a multiple linear regression model is estimated and then, what we will do? As we have learnt the hypothesis testing part and goodness of fit measure previously, we will try to understand all those result from here. First of all, what is the R square here? The R square here is 0.7077. What does it mean? Can you remember? Do you recall the meaning of R square? So, R square is 0.7077.

(Refer Slide Time: 17:40)





First of all, R square is 0.7077. What is the meaning of this? The meaning of this, as R square equals to all, we defined earlier, R square is ESS by TSS. So, that means we can say, the meaning is 70.77 percent of total variation in CM is explained by FLR and PGNP. These are the 2 explanatory variables we have included in our model. So, that means we can say that out of total variation, 70.77 percent of total variation in CM is explained by the 2 explanatory variables FLR and PGNP. And that is what the goodness of fit is- 70.77 percent.

So, our model can explain 70.77 percent variation in child mortality rate. That is the interpretation of R square. This is the first thing we need. Then secondly, the significance of the explanatory variable. How will you check? So, first of all for FLR, what is the t value? Minus 10.63. So, corresponding to FLR, the t value equals to minus 10.63. So, that means what we need to do? We need to take modulus of this. It would become 10.63.

And this is called calculated t value and this we should compare with the tabulated one and what should be the degrees of freedom? Total number of observations is 64 and as we said, the t will always follow a t distribution with n-2 degrees of freedom. So that means 62 degrees of freedom and 5 percent, 1 percent and 10 percent, this we have to compare from the table.

So, alpha equals to 5 percent, we need to get the value. And since this is 10.63, even without comparing also, any value beyond 3 or 4, by rule of thumb we can say that this calculated value would be greater than the tabulated one. And since this is greater than the tabulated one, we have to reject our null hypothesis. What was our null hypothesis?

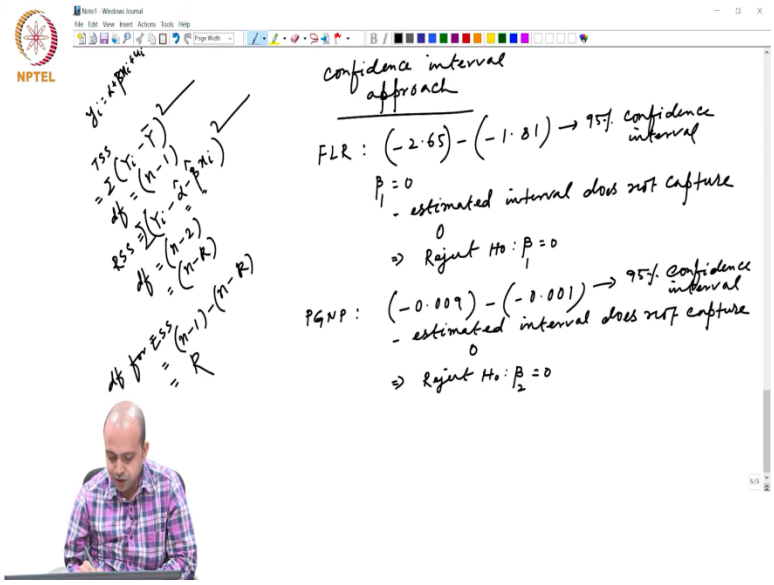
Here we can say that we have two null hypotheses; first one is beta1 equals to 0 and second null hypothesis is beta2 equals to 0. These are the two null hypotheses. So, that means neither FLR nor PGNP has any impact on child mortality rate. If this is greater than this, then we have to reject the null. We have to say that, yes FLR has significant impact on child mortality rate. This is by level of significance approach.

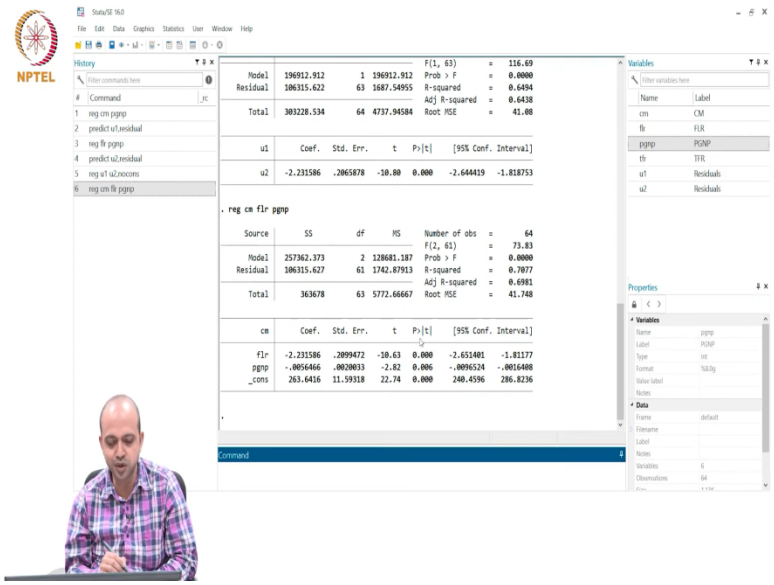
Similarly, for the PGNP, what is the t value? Minus 2.82. So, that means for PGNP, what is the value? Minus 2.82. This also, you take modulus, then it would become 2.82 and will follow then n-2 degrees of freedom, alpha equals to 5 percent, we need to compare. Since it is 2.82, we need to compare whether it is actually greater than the tabulated or not. So, that means going by this t values, we need to compare with the tabulated value.

But as I said earlier, Stata is also giving us the p value which is called exact level of significance. And look at the p value, p value is 0.000 and 0.006 for PGNP. So, if you multiply by 100, both the values are less than 1. What I said? So, that means p equals to first of all 0.000 for PGNP and p equals to 0.006 for, this is for PGNP and this is for FLR.

So, we have to multiply p with 100 and for both of these cases, this is less than 1. This is 0.0, this is 0.06, both are less than 1. So, that means using the p value, we can easily understand both FLR and PGNP are significant at 1 percent level. So, following the p approach, we do not need to compare this tabulated with the calculated one. So, this is called level of significance approach. Now, what we will do?

(Refer Slide Time: 24:28)





We will see the confidence interval approach. There, corresponding to FLR, what is the confidence interval that we got? The confidence interval is minus 2.65 to minus 1.81. That means we can easily understand this particular interval does not capture the hypothesized value of the population parameter. What was the hypothesized value? We have hypothesized beta1 is 0.

So, this interval does not capture 0. So, that means estimated interval does not capture 0. Then what is the implication? Reject the null hypothesis since the interval fails to capture the hypothesized value of the population parameter. So reject H_0 which is beta1 equals to 0. That means, when we reject this, we will say that beta 1 is actually significantly different from 0. So, that means FLR has significant impact on child mortality rate.

Then what is the interval for PGNP? PGNP's interval is minus 0.009 to minus 0.001. This interval also, since both are negative, it does not capture 0. So, that means estimated interval does not capture 0. So, reject H_0 which is beta2 equals to 0. And when it is rejected, we will say that even PGNP also has significant impact on child mortality rate.

But following this interval estimation approach, we can only say that whether the variable is significant at 5 percent level or not because this is 95 percent confidence interval. So, Stata will always give only 95 percent confidence interval. This is also 95 percent confidence interval. If we want to know whether it is significant at 1 percent, obviously if it is 5 percent, then automatically it would become significant at 10 percent, it is absolutely no problem as we said earlier. But whether it is significant at 1 percent level or not, to understand that, we need to construct the interval at 99 percent confidence interval.

But following the p value, we can easily say that it is significant even at 1 percent level. So, that is the advantage of using p value over t and this confidence interval. This is how we can interpret the coefficients actually. Now, what about this ANOVA table? From this ANOVA table again, as we said this particular table is called ANOVA- Analysis of Variance. So Stata is again supplying model sum of square, residual sum of square and total sum of square. That means when the total sum of square is basically this, 363678, that is called total sum of square or TSS.

TSS equals to 363678, that is summation yi minus y bar whole square. Out of this, how much our model is able to explain? 257362.373, that is called ESS or in Stata's language it is called model sum of square. And what is the remaining? Remaining portion is called residual sum of square or RSS that is, 106315.

And then, again corresponding to degrees of freedom, what would be the degrees of freedom for TSS? Total number of observation is 64 and once again I would like to revise you that TSS is basically summation yi minus y bar whole square. Since there is only one restriction imposed, this is called degrees of freedom for TSS and should be n-1. And what is the degrees of freedom for RSS?

The RSS equals to yi minus alpha hat minus beta hat xi square. Our model was yi equals to alpha plus beta xi plus ui. So, I would like to remind you, before we estimate RSS, we must estimate alpha hat and beta hat. So, we are putting 2 restrictions here in terms of alpha hat and beta hat,

that is why degrees of freedom for RSS would be n-2. Or in general, we can say that this is n-k where k is total number of parameters to be estimated from the model.

And then, you can easily say that degrees of freedom for ESS would be (n-1) – (n-k) which is equal to k. k is the degrees of freedom where k is the total number of parameters. So, in this model, we have 3 parameters to be estimated, alpha, beta and this constant term. That is why you see the degrees of freedom for model is n-k, k equals to 3, 1 2 3, that is why it is 61. (64-3)

And if you take this 63 minus 61 equals to 2. Is this clear? k equals to 3 here, because three parameters we are going to estimate; two regression coefficient and one constant term. That is why RSS degrees of freedom is n minus k. 64 minus 3 equals to 61. And then 63 minus 61 equals to 2. So, this is how we have to understand the ANOVA table which clearly shows the decomposition of total variation in our yi.

But one sum of certain other measures which are also supplied by Stata which is another F statistic, look at the F statistic value which is 73.83 and then, another R square which is called adjusted R square. These things we have not yet discussed. We will understand the importance of F statistic and adjusted R square in our next class. So, in our next class, we will discuss how to use these measures; F statistics and R square, adjusted R square which is also supplied by Stata. That way, will discuss in our next class. Thank you.