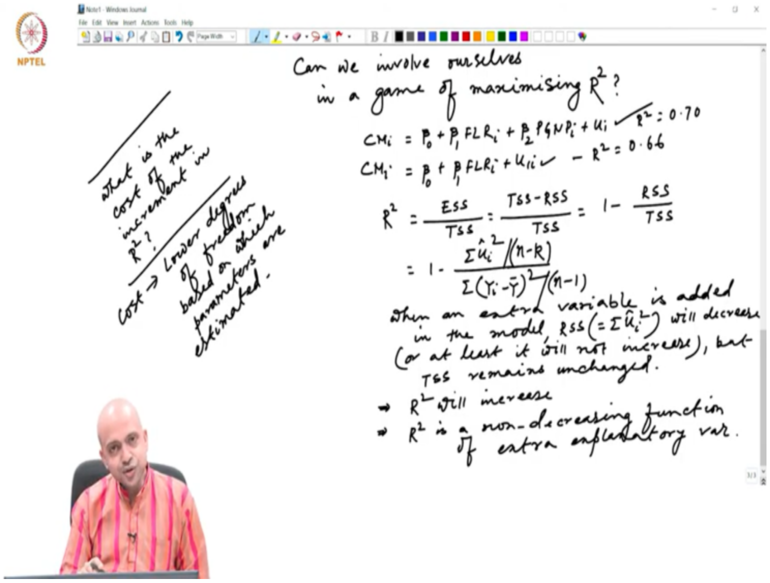
Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology Madras
Lecture 23
Multiple linear regression model and application of F Statistics Part - 2

(Refer Slide Time: 00:14)



Can we involve ourselves in a game of maximising R^2?

$CM_i = \beta_0 + \beta_1 FLR_i + \beta_2 PGNP_i + u_i$ ✓ $R^2 = 0.70$

$CM_i = \beta_0 + \beta_1 FLR_i + u_{1i}$ ✓ $- R^2 = 0.66$

$$R^2 = \frac{ESS}{TSS} = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$= 1 - \frac{\Sigma \hat{u}_i^2 / (n-k)}{\Sigma (Y_i - \bar{Y})^2 / (n-1)}$$

When an extra variable is added in the model, $RSS (= \Sigma \hat{u}_i^2)$ will decrease (or at least it will not increase), but TSS remains unchanged.

$\Rightarrow$ R^2 will increase

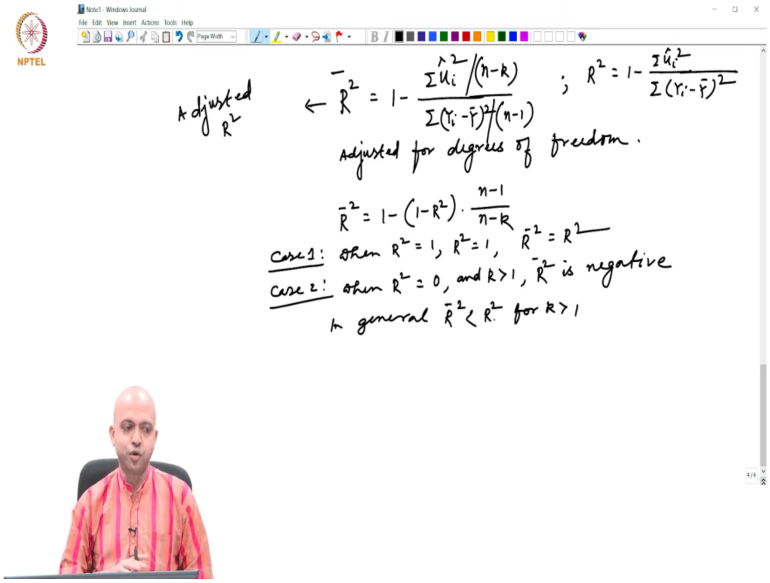
$\Rightarrow$ R^2 is a non-decreasing function of extra explanatory var.

What is the cost of the increment in R^2?

Cost $\rightarrow$ Lower degrees of freedom based on which parameters are estimated.

And if you adjust your R square in this way.

(Refer Slide Time: 00:21)



Adjusted R^2 $\leftarrow$ $\bar{R}^2 = 1 - \dfrac{\Sigma \hat{u}_i^2 / (n-k)}{\Sigma (Y_i - \bar{Y})^2 / (n-1)}$; $R^2 = 1 - \dfrac{\Sigma \hat{u}_i^2}{\Sigma (Y_i - \bar{Y})^2}$

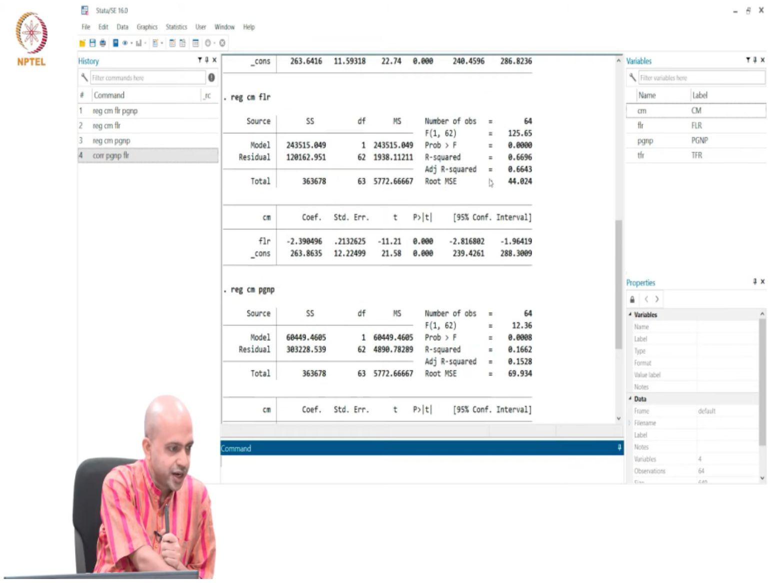
Adjusted for degrees of freedom.

$$\bar{R}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-k}$$

Case 1: When $R^2 = 1$, $R^2 = 1$, $\bar{R}^2 = R^2$

Case 2: When $R^2 = 0$, and $k > 1$, $\bar{R}^2$ is negative

In general $\bar{R}^2 < R^2$ for $k > 1$



```
       _cons    263.6416   11.59318    22.74   0.000     240.4596    286.8236

. reg cm flr

      Source |       SS           df       MS      Number of obs   =        64
-------------+----------------------------------   F(1, 62)        =    125.65
       Model |  243515.049         1  243515.049   Prob > F        =    0.0000
    Residual |  120162.951        62  1938.11211   R-squared       =    0.6696
-------------+----------------------------------   Adj R-squared   =    0.6643
       Total |      363678        63  5772.66667   Root MSE        =    44.024

------------------------------------------------------------------------------
          cm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         flr |  -2.390496   .2132625   -11.21   0.000    -2.816802    -1.96419
       _cons |   263.8635   12.22499    21.58   0.000     239.4261    288.3009
------------------------------------------------------------------------------

. 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]
-------------+----------------------------------------------------------------
```

That means when your R square is actually adjusted by, sorry, so this is 1 minus summation ui hat square divided by summation yi minus y bar whole square. Then if you adjust R square by corresponding degrees of freedom, then we will get R bar square or adjusted R square. This is called adjusted R square.

So, when you say adjusted R square, this is adjusted for degrees of freedom. So, adjusted R square, adjusted for degrees of freedom. Now, we can compare R bar square or adjusted R square of two models, two different models involving different number of explanatory variables, absolutely no problem here. Once you adjust with degrees of freedom, R bar square or adjusted R square of the two models are easily comparable.

So, while R square is not comparable across two different models, R bar square is. That is one thing we have to keep in mind. And what is the relationship between R bar square and R square? See, R bar square equals to 1 minus R square into n minus 1 by n minus k. So, from here, because R square equals to 1 minus sum of ui hat square divided by summation yi minus y bar whole square.

So, if you put this into here, then you will get a relationship like this. R bar square equals to 1 minus R square. So, 1 minus R square and this would become n minus 1 by n minus k. So, from this relationship, we can understand two different cases, let us say case 1 when R square equals to 1, then what will happen? This would become 0 and R bar square is also 1. That means R bar square equals to R square.

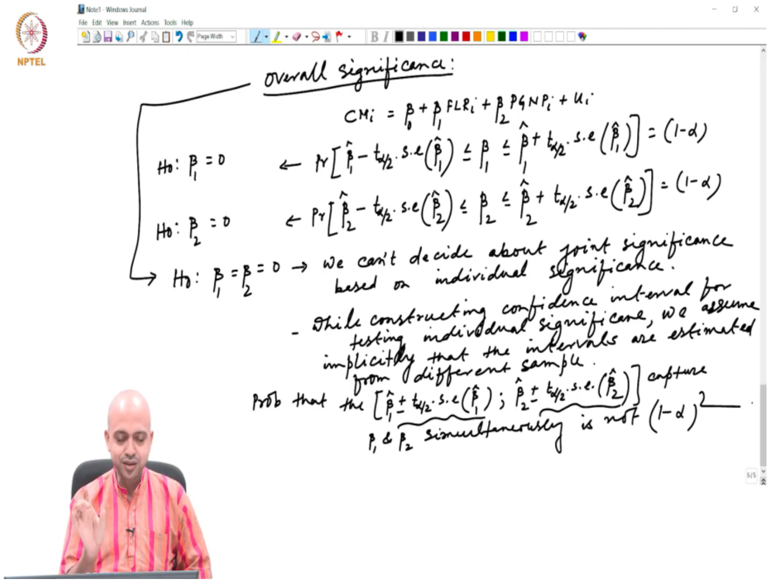
Case 2, when R square equals to 0, then what will happen? And k is also greater than 1, we can say that R bar square is actually negative. R bar square may become negative also. That means while R square lies always between 0 and 1, R bar square may become negative when R square equals to 0. Of course, that is a very, very rare possibility that in a model R square equals to 0. But theoretically, this is also possible. In general, we can say, in general R bar square is actually lower than R square for k greater than 1.

This is the relationship. That means, as additional variable is included in the model, R bar square increases less than R square and that is what we actually want. We need to give a penalty, we need to give a check to balance the R square out of this these lower degrees of freedom. So, this is the relationship between R bar square and R square. Now, the question is, among these two goodness of fit measure, which one is best?

Econometrician, they say that nothing is called better than the other one, rather, we need to report both these R squares- R square as well as adjusted R square when we compute the model. For example, when you estimate the model, if you look at, then your model is actually giving R square as well as adjusted R square.

And you see adjusted R square is always lower than the R square. From here also, you see R square equals to this and adjusted R square is actually 1.1528. So, adjusted R square is lower than this one. So, this is how we can actually get the adjusted R square measure by adjusting its degrees of freedom if at all we want to compare the explanatory power of two different models.

(Refer Slide Time: 06:34)



overall significance:

$CM_i = \beta_0 + \beta_1 FLR_i + \beta_2 PGNP_i + u_i$

$H_0: \beta_1 = 0 \leftarrow Pr[\hat{\beta}_1 - t_{\alpha/2} \cdot s.e(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2} \cdot s.e(\hat{\beta}_1)] = (1-\alpha)$

$H_0: \beta_2 = 0 \leftarrow Pr[\hat{\beta}_2 - t_{\alpha/2} \cdot s.e(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{\alpha/2} \cdot s.e(\hat{\beta}_2)] = (1-\alpha)$

$H_0: \beta_1 = \beta_2 = 0 \rightarrow$ we can't decide about joint significance based on individual significance.

- While constructing confidence interval for testing individual significance, we assume implicitly that the intervals are estimated from different sample.

Prob that the $[\underbrace{\hat{\beta}_1 \pm t_{\alpha/2} \cdot s.e(\hat{\beta}_1) ; \hat{\beta}_2 \pm t_{\alpha/2} \cdot s.e.(\hat{\beta}_2)}]$ capture β_1 & β_2 simultaneously is not $(1-\alpha)^2$.

Then, one more thing that we must learn in multiple linear regression model, that is called overall significance. So, we are running this model, child mortality rate equals to beta0 plus beta1 FLR plus beta2 PGNP, this our model. And how do we compare, how do we check the significance of this beta1 hat?

The significance of beta1 hat, we say that following the confidence interval approach, we say that we will construct an interval and we will say that probability beta1 hat minus t alpha by 2 into standard error of beta1 hat less than or equals to beta 1 hat plus t alpha by 2 into standard error of beta 1 hat, equals to 1 minus alpha.

So, we will construct the interval and we will see whether the interval captures the hypothesized value of the true population parameter. That is how we check the individual significance of the variable. Similarly, for beta 2, we say that beta 2 hat minus t alpha by 2 into standard error of beta 2 hat less than or equal to beta 2, less than equals to beta 2 hat plus t alpha by 2 into standard error of beta 2 hat is also equals to 1 minus alpha.

So, this is by confidence interval approach and we can also check the significance by t statistic. You look at the calculated t statistic of FLR and PGNP and compare the tabulated value and we decide whether the variable is significant or not. what was our null hypothesis? Null hypothesis was, beta1 equals to 0 and in this case, our null hypothesis is beta2 equals to 0.

But now, when we talk about overall significance of the model, let us say that our null hypothesis is beta1 equals to beta2, this becomes 0 or not, this is our null hypothesis. For testing overall significance of the model, this is our null hypothesis. That means, what we are saying, combinedly, FLR and PGNP do not have any significant impact on child mortality rate.

So, obviously we can understand one thing from this model. If we have already tested the individual significance of the variable following the t statistic or this interval approach, what is the necessity of putting this null hypothesis once again? Cannot we say conclude anything about this on the basis of the individual significance of the variable? That is the question that we are asking here.

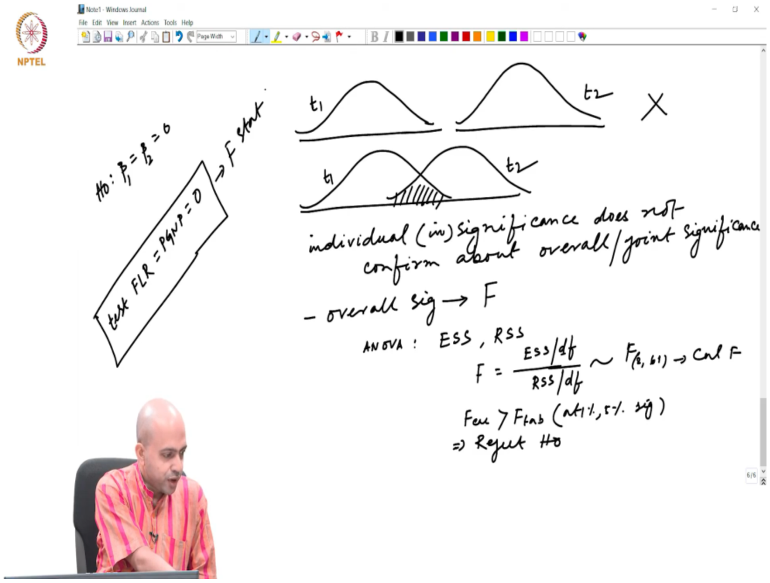
What is the question we are saying? That individually, we have tested the significance of beta 1 and beta 2, then what is the need of testing the combined significance of beta 1 and beta 2? Cannot we say, that since these variables are individually significant or insignificant, they are jointly also significant or insignificant? The answer is no.

We cannot decide about joint significance or overall significance based on individual significance. What is the reason? The reason we can understand clearly from the confidence interval approach. While constructing confidence interval for testing individual significance, we assume implicitly that these intervals are estimated based on a different sample, that the intervals are estimated from different samples.

However, they are not. When we test the joint significance, that means based on these probability statement, we cannot say that beta1 hat plus or minus t alpha by 2 into standard error of beta1 hat, this interval and this interval beta 2 hat, t alpha by 2 into standard error of beta 2 hat. Probability that the intervals, this and this capture beta 1 and beta 2 simultaneously is not 1 minus alpha square, that we cannot say. Please try to understand what I am saying.

So, if we try to infer something about overall significance based on the individual significance, that means indirectly, what we are saying? That probability this interval will capture the hypothesized value of the true population parameter is 1 minus alpha. But that does not mean that these two intervals jointly capture the true value of beta1 and beta2 is 1 minus alpha square because these two intervals are actually not independent. They are constructed based on the same sample and that is why we cannot infer anything about the joint significance based on the individual significance.

(Refer Slide Time: 16:21)



$H_0 : \beta_1 = \beta_2 = 0$ → F stat

$\boxed{\text{test FLR} = \text{PGNP} = 0}$

individual (in) significance does not confirm about overall/joint significance

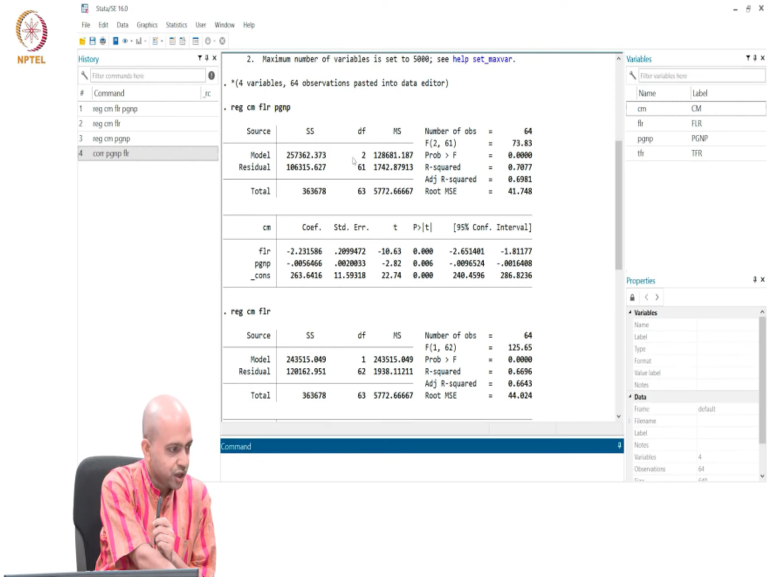
- overall sig → F

ANOVA : ESS , RSS

$F = \dfrac{ESS/df}{RSS/df} \sim F_{(2,61)} \to \text{Cal } F$

$F_{cal} > F_{tab}$ (at 1%, 5% sig)

⇒ Reject H_0



#	Command	_rc
1	reg cm flr pgnp	
2	reg cm flr	
3	reg cm pgnp	
4	corr pgnp flr	

Name	Label
cm	CM
flr	FLR
pgnp	PGNP
tfr	TFR

```
    2.  Maximum number of variables is set to 5000; see help set_maxvar.

. *(4 variables, 64 observations pasted into data editor)

. reg cm flr pgnp

      Source |       SS           df       MS      Number of obs   =        64
-------------+----------------------------------   F(2, 61)        =     73.83
       Model |  257362.373         2  128681.187   Prob > F        =    0.0000
    Residual |  106315.627        61  1742.87913   R-squared       =    0.7077
-------------+----------------------------------   Adj R-squared   =    0.6981
       Total |      363678        63  5772.66667   Root MSE        =    41.748

------------------------------------------------------------------------------
          cm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         flr |  -2.231586   .2099472   -10.63   0.000    -2.651401    -1.81177
        pgnp |  -.0056466   .0020033    -2.82   0.006    -.0096524   -.0016408
       _cons |   263.6416   11.59318    22.74   0.000     240.4596    286.8236
------------------------------------------------------------------------------

. reg cm flr

      Source |       SS           df       MS      Number of obs   =        64
-------------+----------------------------------   F(1, 62)        =    125.65
       Model |  243515.049         1  243515.049   Prob > F        =    0.0000
    Residual |  120162.951        62  1938.11211   R-squared       =    0.6696
-------------+----------------------------------   Adj R-squared   =    0.6643
       Total |      363678        63  5772.66667   Root MSE        =    44.024
```

So, in terms of this diagram, let us say that this is our t distribution - this is distribution for t1 and then, distribution for t2 is actually, if we say, this is not the case. If these two, if this is the case, then we could have said that just because of individually they are significant, jointly also variables would be insignificant.

Because the probability statement that the intervals capture the true population parameter 1 minus alpha square is valid, if and only if these two distributions are independent. Rather, in reality, what happens? If this is your t1, then t2 also like this. This is t1 and this is t2 and this is the overlapping area. Because these intervals are constructed based on the same sample, that is why individual significance or insignificance, I will write, does not confirm about overall significance, overall or joint significance.

So that means for overall significance, we need to have a several different statistic. Overall significance requires a different statistic and that is F. If you recall initially we said that when you estimate the model, the Stata will supply two different types of statistic, one is t and another one is F. So, when you have only one explanatory variable in the model, t and F- we cannot distinguish their role, but when you have multiple variable in your model, that means in the context of multiple linear regression model, t and F have different roles. While t, they indicate individual significance, F in that context will give you the overall significance of the model.

And how does this, how is this F statistic defined? If you estimate the model from our ANOVA table, what we get? We get ESS, we get RSS. Now, if we take a ratio of this ESS and RSS adjusted by their degrees of freedom, then this F is equals to actually ESS by its degrees of freedom and then RSS by its degrees of freedom. And that will follow a F distribution with degrees of freedom for the numerator and denominator, whatever is specified.

So, now if we go back to our example. So, when we take this model, this is our complete model and we look at the ANOVA table. What is model sum of square? The model sum of square value is this, 257362.373. And what is the degrees of freedom? It is 2. ESS is defined as TSS minus RSS; TSS is 63 because total number of observations is 64 and RSS is 61. So, 63 minus 61 is actually 2.

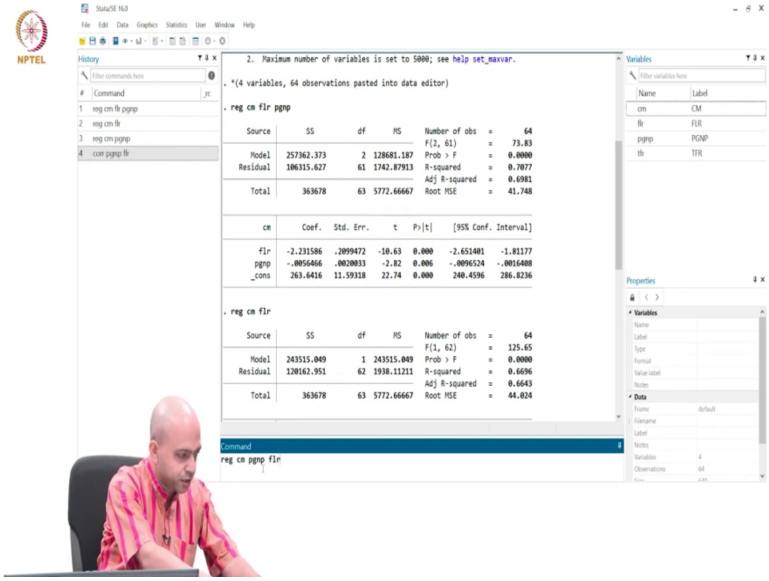
So, how is this F is defined then? These models if you take, a ratio of this MS, so how this is MS? MS is basically SS by its degrees of freedom. So, we have two MS and if we take the ratio of these two, so that means MS for the first and MS for the second, then we will arrive at the F value and what would be the value? This value would be 73.63. That is nothing but this 128681 divided by 1742. If we take ratio of this, that will give you our F statistic.

What would be the degrees of freedom? Degrees of freedom would be 2 and 61. So, 2 and 61 in this particular case. So, that means we need to first define the degrees of freedom for the ESS, degrees of freedom from the RSS and then what we need to do? We need to compare, this is called calculated F and then we need to compare this calculated F with the tabulated value. And if the calculated F is greater than tabulated F at let us say 1 percent or 5 percent significance then we have to reject H naught.

And what is H naught? H naught was beta1 equals to beta2 equals to 0. Once this is rejected, we will establish the alternative, that means yes the variable they have their joint significance also. But we need not compare the F value with the tabulated one, as we said earlier, if we go back with the p value. So, the moment we estimate the model, your Stata is also reporting F value with corresponding p value.

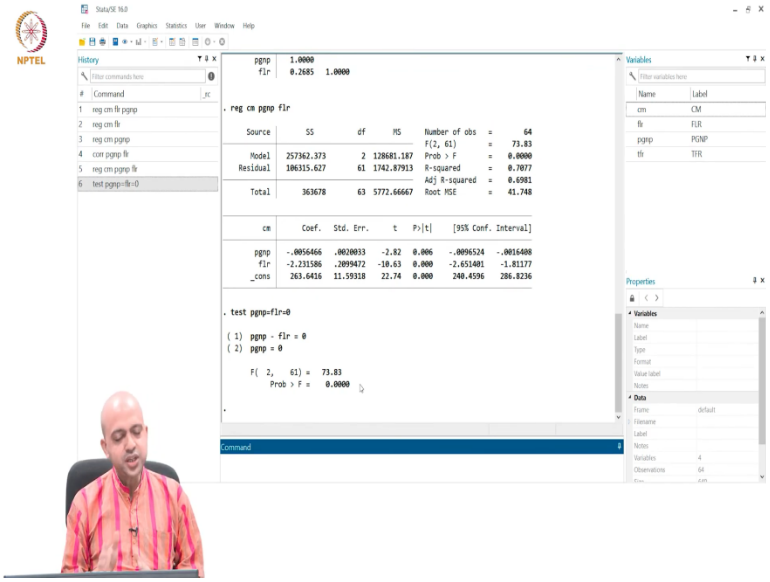
If you look at the p value, p value is 0.0000. If you multiply that by 100, that will still be less than 1, that means we will say that our model is significant at 1 percent level. In Stata, we have a specific command for implementing F statistic and the command is the test command; test, you need to put FLR and then you need to put PGNP and then you can put 0. This is a test command for F statistic. Test command is a very, very powerful command, we can test the individual significance of the variable from this.

(Refer Slide Time: 24:40)



```
    2.  Maximum number of variables is set to 5000; see help set_maxvar.

. *(4 variables, 64 observations pasted into data editor)

. reg cm flr pgnp

      Source |       SS           df       MS      Number of obs   =        64
-------------+----------------------------------   F(2, 61)        =     73.83
       Model |  257362.373         2  128681.187   Prob > F        =    0.0000
    Residual |  106315.627        61  1742.87913   R-squared       =    0.7077
-------------+----------------------------------   Adj R-squared   =    0.6981
       Total |      363678        63  5772.66667   Root MSE        =    41.748

------------------------------------------------------------------------------
          cm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         flr |  -2.231586   .2099472   -10.63   0.000    -2.651401    -1.81177
        pgnp |  -.0056466   .0020033    -2.82   0.006    -.0096524   -.0016408
       _cons |   263.6416   11.59318    22.74   0.000     240.4596    286.8236
------------------------------------------------------------------------------

. reg cm flr

      Source |       SS           df       MS      Number of obs   =        64
-------------+----------------------------------   F(1, 62)        =    125.65
       Model |  243515.049         1  243515.049   Prob > F        =    0.0000
    Residual |  120162.951        62  1938.11211   R-squared       =    0.6696
-------------+----------------------------------   Adj R-squared   =    0.6643
       Total |      363678        63  5772.66667   Root MSE        =    44.024
```

```
reg cm pgnp flr
```



```
        pgnp |   1.0000
         flr |   0.2685   1.0000


. reg cm pgnp flr

      Source |       SS           df       MS      Number of obs   =        64
-------------+----------------------------------   F(2, 61)        =     73.83
       Model |  257362.373         2  128681.187   Prob > F        =    0.0000
    Residual |  106315.627        61  1742.87913   R-squared       =    0.7077
-------------+----------------------------------   Adj R-squared   =    0.6981
       Total |      363678        63  5772.66667   Root MSE        =    41.748

------------------------------------------------------------------------------
          cm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pgnp |  -.0056466   .0020033    -2.82   0.006    -.0096524   -.0016408
         flr |  -2.231586   .2099472   -10.63   0.000    -2.651401    -1.81177
       _cons |   263.6416   11.59318    22.74   0.000     240.4596    286.8236
------------------------------------------------------------------------------

. test pgnp=flr=0

 ( 1)  pgnp - flr = 0
 ( 2)  pgnp = 0

       F(  2,    61) =   73.83
            Prob > F =    0.0000

.
```

So, if you estimate the model once again, reg cm PGNP and then FLR, this is a complete model and now you have to put the test command. test PGNP equals to FLR equals to 0. This is the command. If you put enter, then you see the Stata is also reporting you the F statistic and the value is exactly same.

So, what we have computed manually by taking ESS by degrees of freedom, RSS by its degrees of freedom and then ratio of that, the same value is arriving if we put test command in Stata. And the value is 63.83 with the same degrees of freedom 2 and 61. And the corresponding F value is 0.0000.

So, with this, we can establish the overall significance of the model. So, by now then we have understood the importance of F statistic over the t, while t gives individual significance of the model, F gives you the overall significance of the model. And later on when we discuss about multi co-linearity, there we will see in presence of multi co-linearity, even if your variables are individually insignificant, you will see that still variables are overall significant.

That means t statistics, they become unreliable in the context of multi co-linearity and we must rely on the F statistic to check the joint significance of the model. So, when the variables are significant, there is absolutely no problem, your overall significance will also be there even though theoretically we cannot infer directly from individual to overall, theoretically that is incorrect. But you will still get overall significance of the model.

But in case of variables, most of your variables are insignificant, you will still get overall significance of the model in case of multi co-linearity. That is how we need to understand this F statistic very clearly, which is basically a ratio of the two MS defined by ESS by its degrees of freedom, then whole divided by RSS by its degrees of freedom, the overall significance of the model.

So, with this we are closing our discussion today and next day we will discuss about some other important hypothesis testing in the context of multiple linear regression model. Thank you.