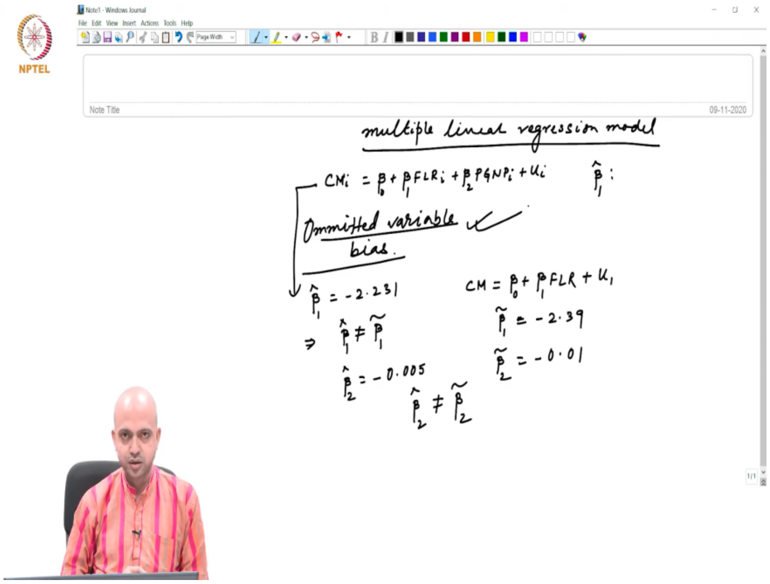
Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology Madras
Lecture 22
Multiple linear regression model and application of F Statistics Part - 1

(Refer Slide Time: 00:14)



multiple linear regression model

$CM_i = \beta_0 + \beta_1 FLR_i + \beta_2 PGNP_i + u_i$ $\quad \hat{\beta}_1 :$

Ommitted variable bias. ✓

$CM = \beta_0 + \beta_1 FLR + u_1$

$\hat{\beta}_1 = -2.231$

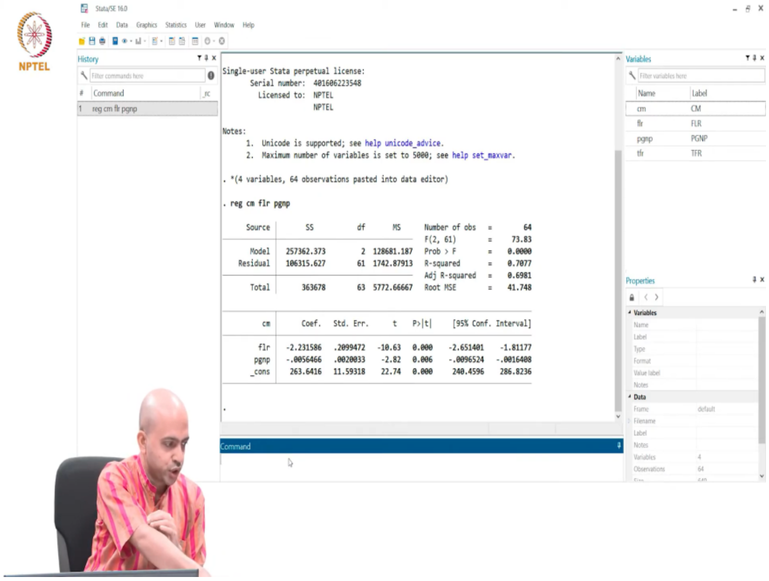
$\tilde{\beta}_1 = -2.39$

$\Rightarrow \hat{\beta}_1 \neq \tilde{\beta}_1$

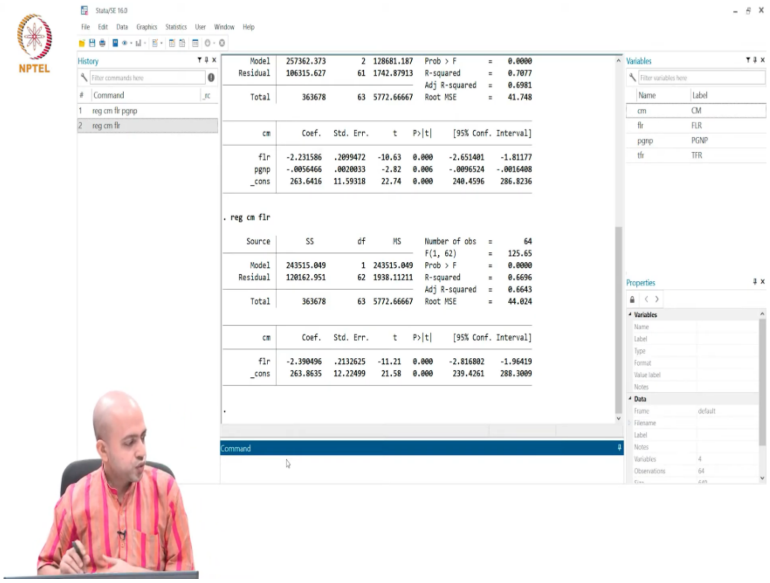
$\tilde{\beta}_2 = -0.01$

$\hat{\beta}_2 = -0.005$

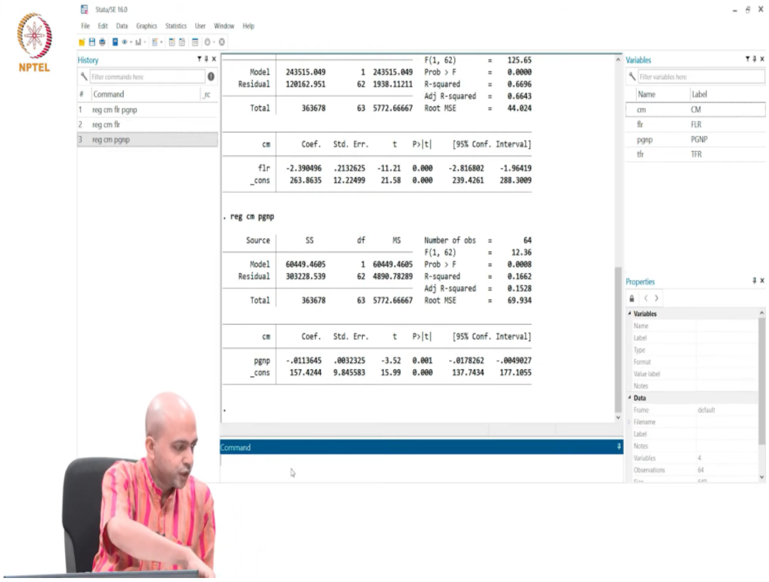
$\hat{\beta}_2 \neq \tilde{\beta}_2$



```
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 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
------------------------------------------------------------------------------

.
```



```
       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

------------------------------------------------------------------------------
          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
------------------------------------------------------------------------------

.
```



```
-------------+----------------------------------   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]
-------------+----------------------------------------------------------------
        pgnp |  -.0113645   .0032325    -3.52   0.001    -.0178262   -.0049027
       _cons |   157.4244   9.845583    15.99   0.000     137.7434    177.1055
------------------------------------------------------------------------------

.
```

Welcome once again. So we were discussing about multiple linear regression model in our previous class and the example what we were talking about is child mortality rate (that was our dependent variable) equals to beta 0 plus beta 1, female literacy rate, FLR and then beta 2 per capita GNP plus ui. And this is a state level data.

So, in our last class we learnt how to estimate this multiple linear regression model and how to interpret this beta 1 and beta 2. Once again I will just repeat. Once you estimate this beta 1 hat, then the interpretation of beta 1 hat is basically, for unit change in FLR, child mortality rate changes by beta 1 hat amount, keeping the impact of PGNP constant.

And we have also learnt how to keep the impact of PGNP constant in different steps and then we saw that actually when we keep the impact of PGNP constant following several steps and when we estimate the model at one go including all the variables, the magnitude of the beta hat turns out to be the same. That is how we justified the multiple linear regression model interpretation.

So, today what we will do? We have also learnt previously the justification of FLR from the point of view of omitted variable bias. So, that means even if our objective is to estimate the impact of FLR only on child mortality rate, do I need to include PGNP also in the model? And we said that yes, even if our objective is to estimate the impact of FLR on child mortality rate, we should also include PGNP in the model so as to avoid omitted variable bias.

Now, with the same example and same data set, we will just see the severity of omitted variable bias in an empirical set up. So, let us see the omitted variable bias. So, we will take the same example. So, first what we will do? We will run the complete model which is given by CM and then FLR and PGNP. This is our complete model. And what is the model of FLR on PGNP? The coefficient of FLR is minus 2.231.

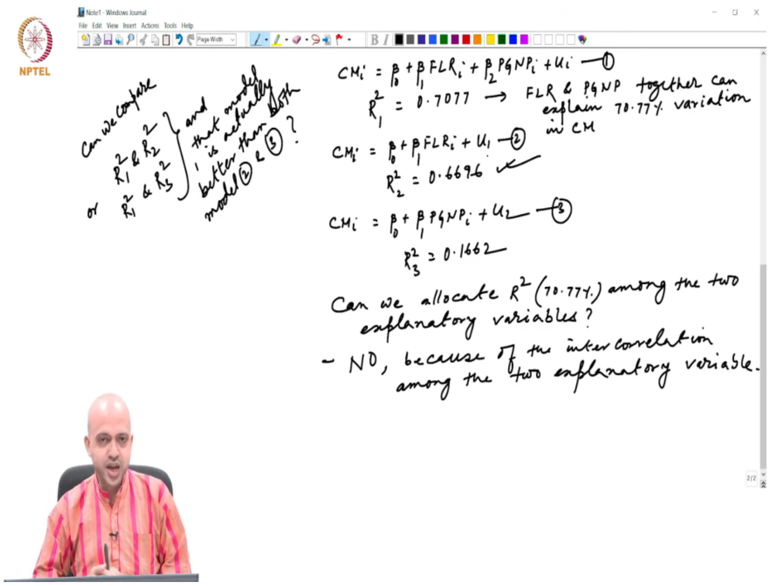
So, that means from the complete model, beta 1 hat equals to minus 2.231. Now, the next model, what we are going to estimate is CM equals to beta0 plus beta1 FLR plus let us say u1. And from this model, the estimated coefficient we will say beta1 tilde, as we have mentioned theoretically also, we see what is the value of beta1 tilde and then we will examine whether beta1 tilde is actually equals to beta1 hat or not. If they are equal, then we will say that there is no bias involved, beta 1 tilde now we are going to estimate.

So, we will include only CM (child mortality rate) on FLR. So, that means we are not including PGNP in the model. Now, we will see what is the coefficient? Coefficient is minus 2.39. So, we can easily understand from this that beta1 hat, that means this implies beta1 hat is actually not equals to beta1 tilde. It is 2.231, this is minus 2.39, that is increase in the value of the beta1 tilde. So, that means there is a bias when we include, when we do not include the model, other relevant variable in the model.

Similarly, what we can do? We will look at the estimate of PGNP also from the complete model, what is the beta2 hat? So, beta2 hat from the complete model, beta2 hat equals to, that means from PGNP is minus 0.005. And then, we will include, we will try to estimate beta2 tilde also from the incomplete model and that shows, so we will include PGNP as explanatory variable and what is the value? Minus 0.01. So, that means we can say that beta2 hat is not equal to beta2 tilde, rather than beta2 tilde becomes almost double.

So, with this example, we can understand the severity of this omitted variable bias when we do not include an important variable, explanatory variable in the model. So, that means that gives you the proper justification of multiple linear regression model. Even if our objective is to estimate the impact of a particular variable in the model, we must include other relevant or important variables as control in the model. Otherwise our model will suffer from omitted variable bias and this example clearly demonstrate the severity of the omitted variable bias.

(Refer Slide Time: 08:41)



$CM_i = \beta_0 + \beta_1 FLR_i + \beta_2 PGNP_i + u_i$ — ①

$R_1^2 = 0.7077$ → FLR & PGNP together can explain 70.77% variation in CM

$CM_i = \beta_0 + \beta_1 FLR_i + u_1$ — ②

$R_2^2 = 0.6696$ ✓

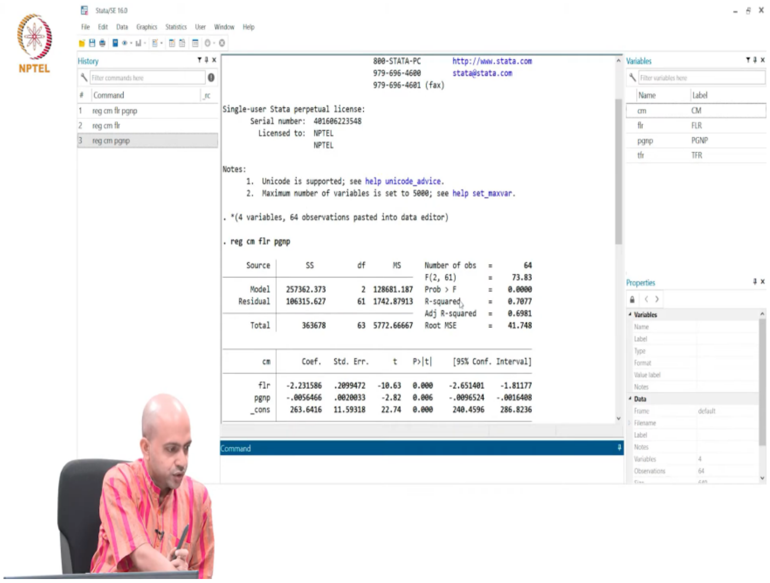
$CM_i = \beta_0 + \beta_1 PGNP_i + u_2$ — ③

$R_3^2 = 0.1662$

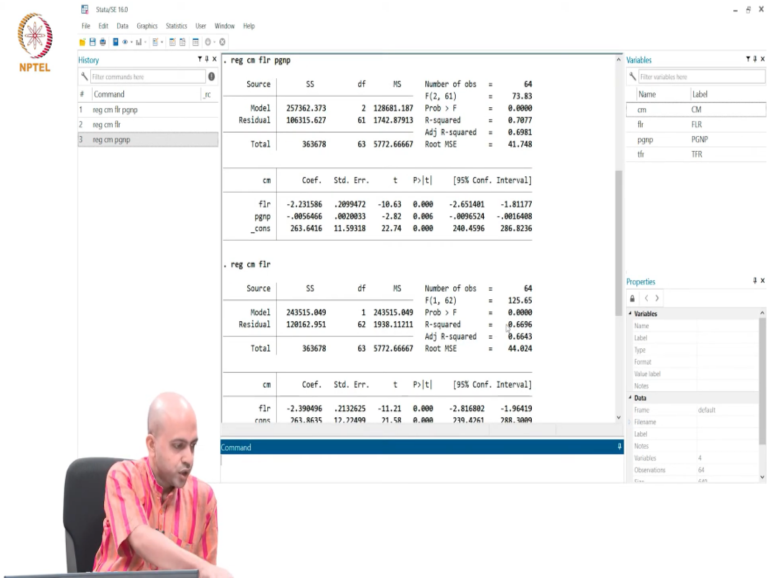
Can we allocate R^2 (70.77%) among the two explanatory variables?

- NO, because of the inter correlation among the two explanatory variable.

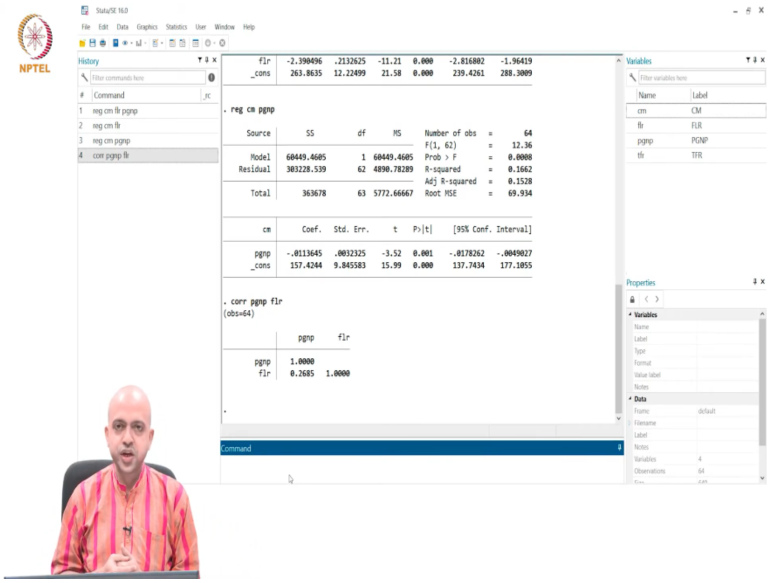
Can we compare R_1^2 & R_2^2 or R_1^2 & R_3^2 } and that model is actually better than both model ② & ③ ?



```
                     800-STATA-PC        http://www.stata.com
                     979-696-4600        stata@stata.com
                     979-696-4601 (fax)

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 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 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

------------------------------------------------------------------------------
          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
```



```
1  reg cm flr pgnp
2  reg cm flr
3  reg cm pgnp
4  corr pgnp flr
```

```
         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]
-------------+----------------------------------------------------------------
        pgnp |  -.0113645   .0032325    -3.52   0.001    -.0178262   -.0049027
       _cons |   157.4244   9.845583    15.99   0.000     137.7434    177.1055
------------------------------------------------------------------------------

. corr pgnp flr
(obs=64)

             |     pgnp      flr
-------------+------------------
        pgnp |   1.0000
         flr |   0.2685   1.0000


.
```

Now, one more thing that also happens when we do not include, what happens to the goodness of fit? So, from the complete model, beta0 plus beta1 FLR plus beta2 PGNP plus ui. So, from the complete model we will try to get the R square and we will call it as R square 1. So, what is the R square value from the complete model, the goodness of fit? If we look at the complete model, then we see the R square value is 0.7077.

And then, when we run the other 2 models, CM i equals to let us say beta0 plus beta1 FLR plus u1, what is the R square value? Let us say this is called R square2, and then we will run this model also, CM equals to beta0 plus beta1 PGNP plus u2 and we will call this R square as let us say R square3. So, we will now report the R square2 and R square3 and we will see.

So, when we include only FLR in the model and then we see the R square value is 0.6696, this is 0.6696. And then when we include only PGNP in the model, then we get the R square value is 0.1662. Now, in a complete model, R square value is 0.7077, that means we can say that FLR and PGNP, these two variables together, they can explain 70.77 percent variation in child mortality rate. So, from this, what we can say that, FLR and PGNP together can explain 70.77 percent variation in child mortality rate.

Now, a natural question comes to our mind; out of these 70.77 percent explanation, how much of this is explained by FLR and how much of this is explained by PGNP? So, one obviously would tend to say that out of this 70.77, probably 66.96 percent is explained by FLR and remaining would be PGNP. But if you look at, 70.77 minus 66.69 is not actually 16.62 percent.

So, that means we cannot actually allocate the R square among these two explanatory variables. What I say? I will just repeat here. PGNP and FLR together can explain 70.77 percent variation in child mortality rate because that is the R square from the complete model. Now, a natural question comes to our mind. Out of this 70.77 percent, how much percentage of this R square is due to FLR and how much of this R square is due to PGNP? Can we allocate this 70.77 percent among these two variables?

So, if we would like to allocate these 70.77 among these two variables, so when we include FLR in the model, then we see that 60.96 percent variation is explained by FLR itself. So, that means 70.77 percent minus this percentage should be explained by PGNP, but actually if we look at, PGNP is explaining much more than that. So, that means we cannot say, we cannot actually, we cannot, can we allocate R square which is 70.77 percent among the two explanatory variables? The answer is actually no, we cannot.

Why this is no? Because you see, if you allocate 60.96 percent on FLR, then the remaining 70.77 minus this much should have been explained by PGNP, but PGNP is explaining much more than that. So, that means it is actually not possible to allocate this R square. Why? Because of the intercorrelation among the two explanatory variable, if you calculate this intercorrelation, among these two you can see what is the intercorrelation between PGNP and FLR, you can easily calculate that.

What is the command for that? corr PGNP and FLR if you give, then you see that 0.2685 is their intercorrelation. Because of this intercorrelation, we cannot say how much percentage of this R square is due to FLR and how much percentage of this R square is due to PGNP. That is one thing we have to keep in mind.

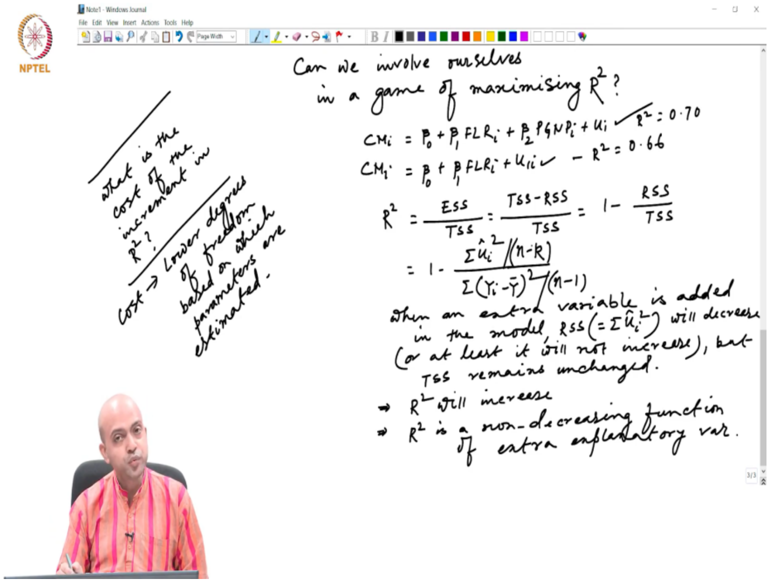
But one thing what we observe that compared to the partial models, when we include one more variable in the model, R square is actually increasing. So, whenever, let us say that this is our initial model CM is the dependent variable and FLR is the independent one. And then R square is 0.6696. If we include PGNP also in the model, then the complete R square, complete model gives R square equals to 70.77.

So, adding one extra variable is giving additional R square in the model. Similarly, when we have only PGNP in the model, R square is 0.1662, however when we add FLR in the model, that gives R square equals to 0.7077. Now, the question here that comes to our mind; can we compare these two R square, that means can we compare R1 square and R2 square or R1 square and R3 square and say that R1 square is better, let us say this is our model 1, this is model 2, this is model 3.

So, can we compare R1 square and R2 square or R1 square with R3 square and say that model 1 is actually greater than both model 2 and 3? This is a question; this is an interesting question. So, compared to these 2 partial models, our complete model gives better R square. So, since complete model gives better R square, when we compare 2 and 1, we see that there is some change in the R square value, that means explanatory power of the model increases when we add PGNP also in the model, because R square increases from 0.6696 to 0.7077.

Similarly, compared to 3 and 1, when we add FLR also in the model, R square increases from 0.1662 to 0.7077. Because of this comparison, can we say that model 1 is actually better than model 2 and 3? What we are posing is can we compare R1 square with R2 square and R3 square? And say and decide which model is better than the other one? The answer is again no. We cannot decide the suitability of a particular model based on only R square. What is the reason? Let us try to explain that.

(Refer Slide Time: 20:37)



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

cost → Lower degrees of freedom based on which parameters are estimated.

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.

⇒ R^2 will increase

⇒ R^2 is a non-decreasing function of extra explanatory var.

So that means can we involve ourselves in a game, I will call it game, in a game of maximizing R square? Apparently, it may look like we will keep on adding variables in the model as long as our R square increases. So, we are basically trying to compare these two models, CMi equals to beta0 plus beta1 FLR plus beta2 PGNP, let us say this is ui and then, the second model CM equals to beta0 plus only beta1 FLR plus let us say u1i. These are the 2 models we are comparing.

So, if we add additional variable in the model, R square will definitely increase. Why this is so? What is the measurement of R square? R square is basically ESS by TSS and this we can again manipulate as TSS minus RSS and then divide it by TSS which is equals to 1 minus RSS by TSS. And what is RSS? RSS as we know, summation of ui hat square divided by, what is TSS? yi minus y bar whole square. This is basically our R square.

Now, when we add one extra variable in the model, what will happen to our ui hat square or RSS? The residual sum of square will definitely decrease or at least, it will not increase. So, that means when an additional, when an extra variable is added in the model, RSS which is actually equals to summation ui hat square, RSS will decrease or at least it will not increase.

And what will happen to TSS? But, TSS remains unchanged. So, if the denominator remains unchanged and the numerator decreases due to an additional variable in the model, what will happen to R square? So, R square therefore, R square will increase when additional variable in the model. So, we can say that R square is basically a non-decreasing function of any additional or extra explanatory variable.

So, this says R square is a non-decreasing function of extra explanatory variable. So, this is the thing that happens. When we add additional variable, explanatory power of the model increases, R square will increase. But, does that mean that we will keep on adding variable in the model and we will say that this model is better than this one because this gives higher R square value. In this model R square equals to 0.70 and here it is R square equals to 0.66. We cannot say that this model is better.

Why? Because this model involves, these two models, these two R square are not directly comparable. Why this is so? Because in the first model, we have two explanatory variables while in the second model, we have only one explanatory variable even though the dependent variable of both the models are same, that is child mortality rate.

So, when two models involves different number of explanatory variable, how can you say that this model is better than this? Yes, we are getting additional R square but in the process we are also losing something. In this world nothing comes free of cost. So, that means a natural question comes to our mind. This additional explanatory power comes with the cost of what? That is the question.

So, the question that we would like to answer here, the incremental value in R square, what is the cost of the increment in R square? This is an important question that I am asking. Is there any cost involved? When we add one extra variable to experience little higher level of R square? The cost is there and what is the cost? If we recall, every additional variable we include in our model, we use extra degrees of freedom.

If we recall that if you include too many variables in the model, then the degrees of freedom, based on which your beta1 hat, beta2 hat, these parameters are estimated, would be lesser and lesser. And if you have too few observations freely moving to estimate your parameters, then the reliability of the parameters comes down drastically.

So, that means the cost is in terms of, what is the cost? Cost is lower degrees of freedom based on which, parameters are estimated. This is the cost. The more variable we add in the model, lesser and lesser would be degrees of freedom in that model, based on which our parameters beta1 hat, beta2 hat and so on, they are estimated. Lower the degrees of freedom, lower would be reliability of the parameters.

So, that is one thing we keep, we have to keep in our mind. So, that means we need to give some penalty to the R square measure while adding the additional variable in the model. And what is the penalty that we give? The penalty that you have to give is the degrees of freedom. That means we have to adjust this R square, that means the numerator and denominator with their corresponding degrees of freedom. What is the degrees of freedom for this RSS as you know? This is n minus k.

What is the degrees of freedom for the denominator? This is n minus 1. That means now from this we can understand that since the numerator involves n minus k, as you include more variables in the model, k will also increase. And as a result of which this numerator will also come down, R square will also come down. That means R square will increase but not like the previous cases. Earlier, it was increasing more but now I have adjusted the R square, that means the denominator and the numerator both, by their corresponding degrees of freedom.

Since the denominator is not depending on the parameters, see here it is n minus k, so that means it remains again unchanged. But the numerator, it is the degrees of freedom n minus k, more the explanatory variable you add in the model, more would be the value of k, so that means n minus k will decrease and that means that this summation ui hat square will come down. And that gives some penalty on the addition of explanatory variable.