Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

Module - 7
Lecture - 56
Multicollinearity (Contd.)

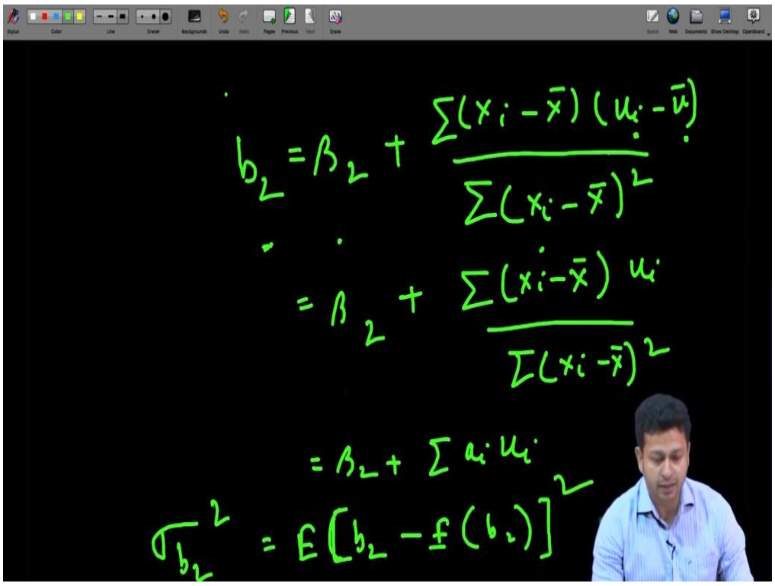
Hello and welcome back to the lecture on Applied Econometrics. We have been talking about the problem of multicollinearity. Now, we explained the concept of multicollinearity. We have also given you a proof why perfect collinearity is a problem. Now, we will try to see why exactly if I have multicollinearity; I understand perfect collinearity is a problem; but if I have like imperfect collinearity, it may be high multicollinearity, low multicollinearity; why is that going to be a problem?

If suppose, if my X variables are related to some extent, how is that going to create a problem in my regression equation? And that is what we are going to see. The major problem and that you need to remember is that, in case of multicollinearity, if the X variables are related, what happens is that, the standard error term corresponding to each of these beta coefficients that we are estimating is going to be high.

I repeat; the major problem that occurs due to the multicollinearity is that the standard error term associated with each of the beta coefficients in the regression equation are going to be high, because, if there is a high multicollinearity problem. So, depending on the extent of multicollinearity, we will have different levels of standard error. And a high multicollinearity would lead to a high standard error.

And we will see what happens if I have high standard error. So, let me first explain the standard error for the different coefficients in a regression equation. And we will use the result that we previously derived when we derived the relationship between b 2 and beta 2. So, the sample regression coefficient and the population regression coefficient. So, I will not actually derive it.

(Refer Slide Time: 02:06)



$$b_2 = \beta_2 + \frac{\sum (x_i - \bar{x})(u_i - \bar{u})}{\sum (x_i - \bar{x})^2}$$

$$= \beta_2 + \frac{\sum (x_i - \bar{x}) u_i}{\sum (x_i - \bar{x})^2}$$

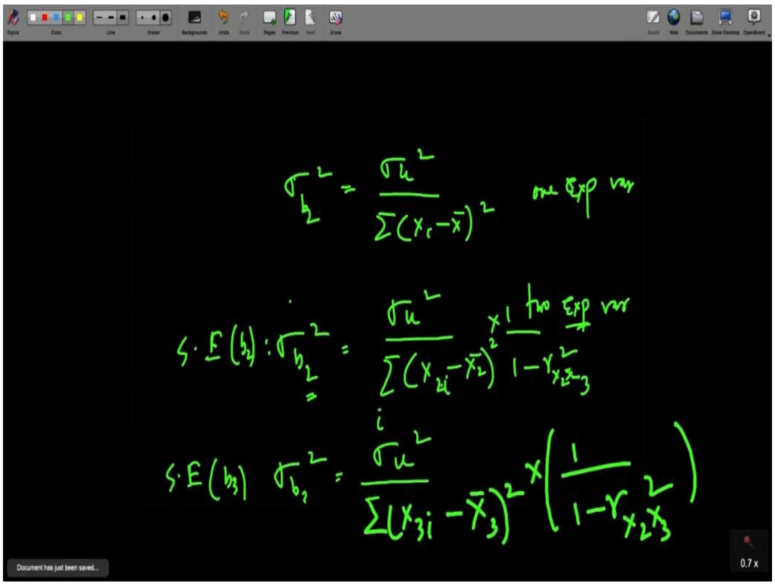
$$= \beta_2 + \sum a_i u_i$$

$$\sigma_{b_2}^2 = E[b_2 - E(b_2)]^2$$

So, you just look at your previous notes. So, where we derived it as b 2 is equal to beta 2 plus X i minus X bar into u i minus u bar by summation of X i minus X bar whole square. So, we said that the value of b 2 is actually a sum total of a sort of non-stochastic component and a stochastic component, because of the involvement of the error term here. And that is what actually gives a distribution for the b 2 that we determine.

Now, if I further simplify; so, I am not going to show you the full proof. You can again find this proof in Christopher Dougherty's textbook, chapter 3. It is going to be X i minus X bar into u i by summation X i minus X bar whole square. Now, if I actually use this relationship, I can actually write, let us say, simplify it as beta 2 plus summation of a i u i; this one. And now, if I use this value of b 2 here, and if I actually try to get a sigma square value for b 2, what I will do is, I will actually do something like this: expectation of b 2 minus of E of b 2 and a whole square of that. So, let us say I use a third bracket here. So, this is going to be the value of my sigma square.

(Refer Slide Time: 04:32)



$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \bar{X})^2} \quad \text{one Exp var}$$

$$S.E(b_2) : \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum_i (X_{2i} - \bar{X}_2)^2} \times \frac{1}{1 - r_{X_2X_3}^2} \quad \text{two Exp var}$$

$$S.E(b_3) \; \sigma_{b_3}^2 = \frac{\sigma_u^2}{\sum (X_{3i} - \bar{X}_3)^2} \times \left(\frac{1}{1 - r_{X_2X_3}^2} \right)$$

And if I simplify it, what I will get is that; I am not going to show the steps again. So, what I will get is sigma beta 2 square is equal to, it is going to be sigma u square by X i minus X bar whole square. And if I have a term; so, this is where I have like simple OLS, I have only 1 variable X; but if suppose I have 2 variables, X 2 and X 3; let us say I have only 1 explanatory variable; but if I have say 2 explanatory variables, so, what will happen is that, because there is a correlation between these 2 explanatory variable, so, the equation is going to be, in that case is, sigma u square by summation of X.

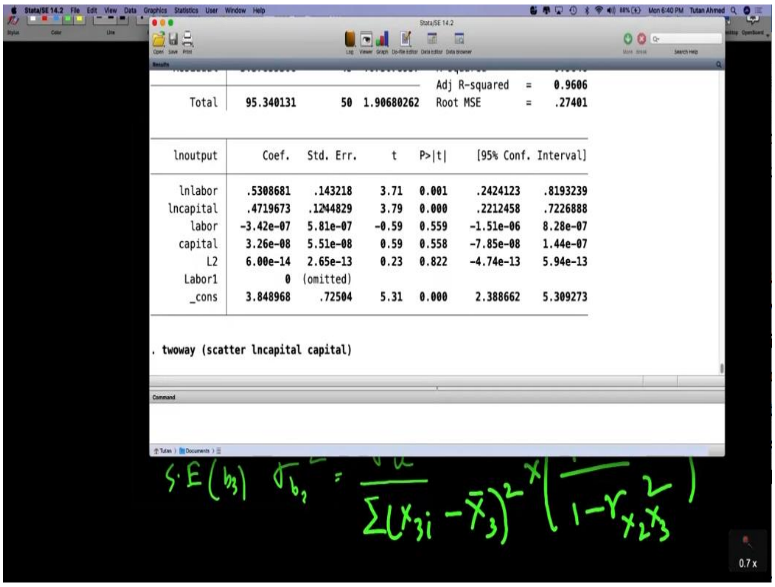
So, here I have, my only variable is; let us say I am trying to get the coefficient for X 2 variable. So, it is going to be X 2 minus X 2 bar all over i; let us say I put an i here also, over all i; square into 1 by 1 minus r X 2 X 3 square; so, where r X 2 X 3 is nothing but the correlation coefficient between these 2 variables, X 2 and X 3. Similarly, if I have to write, say the sigma square value for beta 3, so, standard error.

These are basically, these are nothing but the standard error for beta 2. So, this is the standard of a beta 2, and it has this formula. And similarly, I can write standard error for beta 3, which is the coefficient for X 3. And that is going to be beta 3 square. It is going to be, so, variance of the error term and summation of X 3i. Now, variable I am concerned is X 3 square into; the other term remains same; 1 by 1 minus r X 2 X 3 square.

So, essentially, you see that we have; basically, what you are doing? You are multiplying a term in case of multiple explanatory variable. So, here I have 2 explanatory variables. So, what I am doing is that, I am multiplying the term 1 by 1 minus r X 2 X 3 square. And now, let us try to understand. So, this is basically the form of standard error. And again, if you want to see the proof, I would recommend you consult the Christopher Dougherty's textbook, chapter 3.

Now, let us try to understand what is happening here. Suppose the correlation between X 2 and X 3 is very high; so, R square value is very high. So, if that happens, what will happen? So, if the R square value is very high, so, 1 minus R square value is going to be very low. And if 1 minus R square value is going to be very low, what will happen is the standard error of beta sigma, beta 2 or beta 3, whichever you take, that value is going to be very high. And if the standard error of beta is going to be very high; so, I normally; if you remember a regression, how a regression looks like; so, you have;

(Refer Slide Time: 08:21)

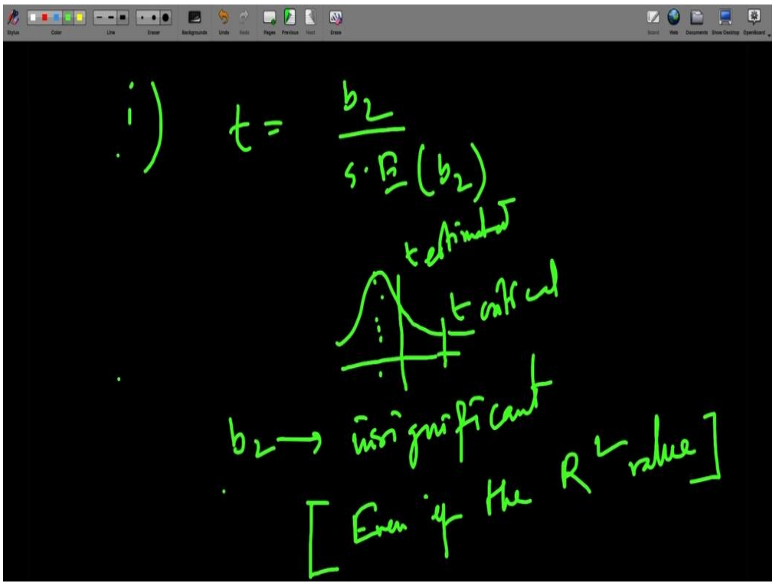


```
                                                   Adj R-squared   =    0.9606
       Total    95.340131        50  1.90680262   Root MSE        =    .27401

    lnoutput       Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

     lnlabor    .5308681    .143218     3.71   0.001     .2424123    .8193239
   lncapital    .4719673   .1244829     3.79   0.000     .2212458    .7226888
       labor   -3.42e-07   5.81e-07    -0.59   0.559    -1.51e-06    8.28e-07
     capital    3.26e-08   5.51e-08     0.59   0.558    -7.85e-08    1.44e-07
          L2    6.00e-14   2.65e-13     0.23   0.822    -4.74e-13    5.94e-13
      Labor1           0  (omitted)
       _cons    3.848968     .72504     5.31   0.000     2.388662    5.309273

. twoway (scatter lncapital capital)
```

Stata/SE 14.2
Results
Command
$S.E(b_3)\ \sigma_{b_2} = \frac{}{\sum(X_{3i}-\bar{X}_3)^2} \times \left(\frac{}{1-r_{x_2x_3}^2}\right)$
0.7 x

I am taking a previous regression that we have run. So, this is a coefficient, this is a standard error and this is your t value. So, you get your; so, let me actually write down the implication.

(Refer Slide Time: 08:34)



i) $t = \frac{b_2}{s.E(b_2)}$

$b_2 \rightarrow$ insignificant

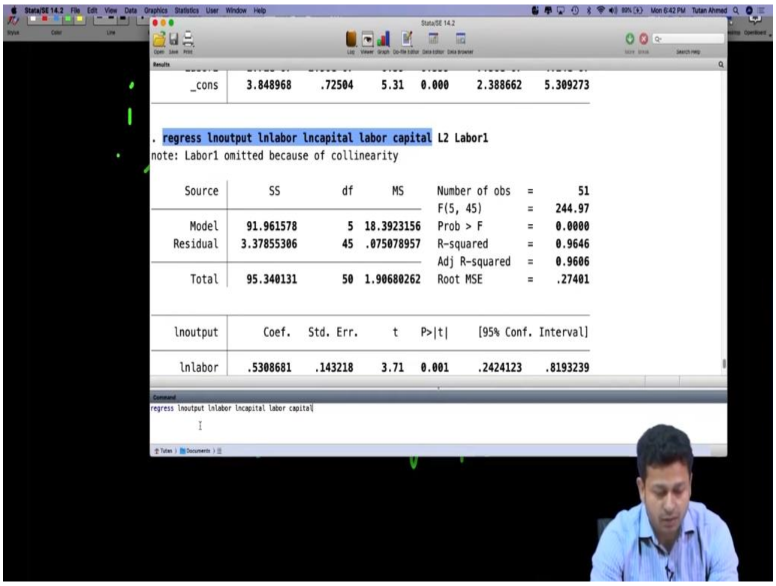
[Even if the R^2 value]

So, basically, my t distribution, t value is nothing but my beta 2 by my standard error of beta 2. So, that is how we define it. Now, if my standard of beta 2 is very high, what we are going to have is very low t value. And if I have very low t value, what will happen is that, this t value that we estimate is going to be somewhere near here; t value, let us say t estimated. And it is going to be, if I have like a higher standard error, it is going to be somewhere near to this mean.

And that will mean that my t critical, it is going to be left with a t critical, which would mean that my t critical, so, basically, my coefficient is going to be; if that is the case, my beta coefficient, the corresponding b 2 is going to be insignificant. So, the problem here is that, if I have high standard error, so, even if my B 2 is actually important, it is actually explaining the regression equation.

Just because of the fact that I have a high standard error, I will actually see a very low b 2 value; it is going to be insignificant, it is going to give me an insignificant beta; not low value, but low t value. And because of the low t value, I am going to get an insignificant beta 2 or b 2. So, low t value means high P value. So, we will see an insignificant value for the particular beta coefficients.

(Refer Slide Time: 10:13)

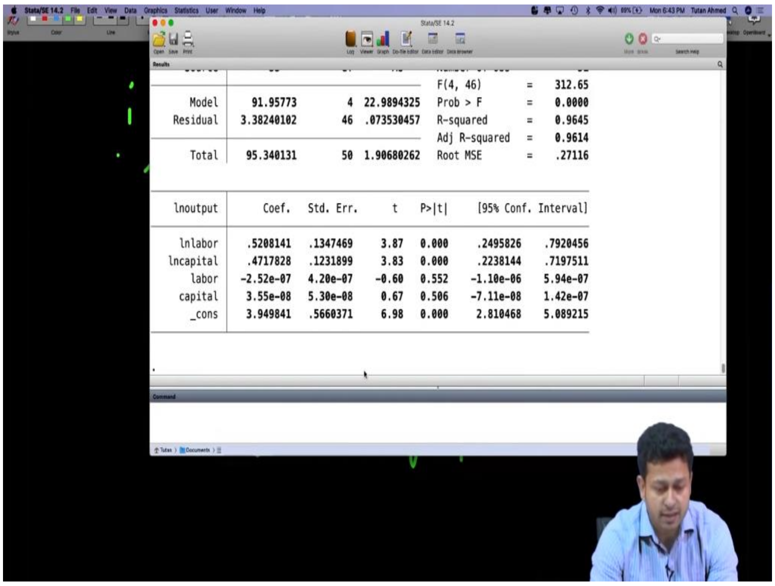


```
       _cons     3.848968     .72504     5.31   0.000     2.388662    5.309273

. regress lnoutput lnlabor lncapital labor capital L2 Labor1
note: Labor1 omitted because of collinearity

      Source         SS           df       MS      Number of obs   =        51
                                                   F(5, 45)        =    244.97
       Model     91.961578         5  18.3923156   Prob > F        =    0.0000
    Residual    3.37855306        45  .075078957   R-squared       =    0.9646
                                                   Adj R-squared   =    0.9606
       Total     95.340131        50  1.90680262   Root MSE        =    .27401

    lnoutput        Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

     lnlabor     .5308681    .143218     3.71   0.001     .2424123    .8193239
```

```
regress lnoutput lnlabor lncapital labor capital
```

If I actually run this regression equation; so, let me actually go back and run this regression equation. Here I am just going to include this ln output, ln labor, ln capital, labor and capital. So, essentially, labor, capital, and the logarithmic 2 variables for log capital, log labor; these are my explanatory variables.

(Refer Slide Time: 10:38)



```
                                                   F(4, 46)        =    312.65
       Model      91.95773         4  22.9894325   Prob > F        =    0.0000
    Residual    3.38240102        46  .073530457   R-squared       =    0.9645
                                                   Adj R-squared   =    0.9614
       Total     95.340131        50  1.90680262   Root MSE        =    .27116

    lnoutput        Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

     lnlabor     .5208141   .1347469     3.87   0.000     .2495826    .7920456
   lncapital     .4717828   .1231899     3.83   0.000     .2238144    .7197511
       labor    -2.52e-07   4.20e-07    -0.60   0.552    -1.10e-06    5.94e-07
     capital     3.55e-08   5.30e-08     0.67   0.506    -7.11e-08    1.42e-07
       _cons     3.949841   .5660371     6.98   0.000     2.810468    5.089215
```

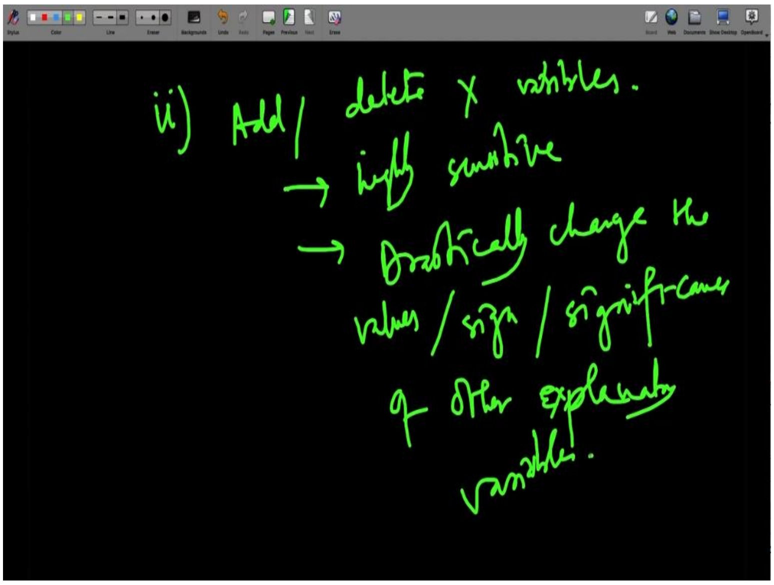
If I run this, I will see that for labor and capital, the t values are pretty low, like 0.6, 0.67, and the P values are quite high. We know that for low t value, we will have a corresponding high P value. Now, so, these 2 variables are going to be insignificant. And the reason here is that, there is a high multicollinearity problem. So, because I have labor, capital and at the same time, I have log of labor and log of capital.

So, this is something that is going to give me an insignificant value for beta. What happens is that, because my model is well fitted, so, because the variables are actually explaining the error, so, my explained sum of error is actually going to be pretty high. And if that is the case, then it is actually explaining most of the variations in the total sum of square. So, it is going to give me a very high R square.

So, essentially, what will happen is, if the model is actually, the variables are highly related, but again that they are actually explaining the Y variable, even if the R square value is high. So, they are actually going to give me a high R square value, whereas, I will have the coefficients insignificant. So, that is basically something that we need to remember in case of multicollinearity.

So, this is something, one symptom that we will see. A variable that I feel should be actually explaining the Y variable, that is actually insignificant; but at the same time, I am seeing a high R square value.

(Refer Slide Time: 12:20)



ii) Add/ delete X variables.

→ highly sensitive

→ Drastically change the values / sign / significance of other explanatory variables.

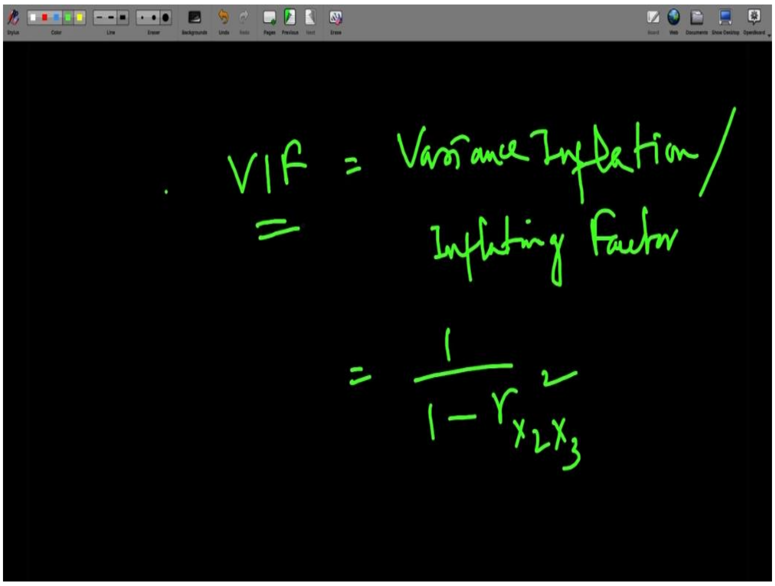
Now, let me tell you couple of more symptoms when you actually have this multicollinearity problem. So, what will happen if you add or maybe delete one variable, delete one, say X variables? What will happen is, if they are highly correlated, so, what will happen is that, this term here, this r X Y, this correlation coefficient term here, this is going to change, because they are actually dependent on which other variables are present there.

And if there is a high correlation coefficient among the variables, so, this value is going to be high. And the moment you change one variable, it is going to change. And the moment it is going to change, my standard error is going to change. And if my standard error is going to change, my t value is also going to change drastically. And it means, the moment you include a variable or exclude a X variable, it will automatically change the coefficients, the level of significance or the significance of the coefficients, the explanatory variables.

So, that is another, basically, a case where if you add or delete variable, you will see is highly sensitive, because it will drastically change the values or the sign or the significance of other explanatory variables. And it can also happen if you basically bring in more data. So, if you bring in more data or if you reduce some data, so, it will also be very sensitive to any sort of change.

So, if you see this kind of sensitivity existing in your model, you can actually presume that there might be a problem of multicollinearity. So, these are some of the symptoms for multicollinearity. And we will just conclude this lecture by saying that, usually, we have used this term, 1 by 1 minus r square X 2 X 3.

(Refer Slide Time: 14:49)



VIF = Variance Inflation/
= Inflating Factor

$$= \frac{1}{1 - r^2_{X_2X_3}}$$

So, usually, this term is also referred as VIF, variance inflation; or some books use inflating, like Gujarati use inflating, whereas Dougherty would use inflation; inflating factor. So, mathematically we write it 1 by 1 minus r X 2 X 3 square. And we will see the importance of this VIF when we talk about multicollinearity. So, with this, we end this particular lecture where we have explained the problem of high standard error, which is essentially the crucial problem when we deal with multicollinearity. And with this, we end the lecture. In the next lecture, we are going to see some other problems and how we are going to remedy the problems of multicollinearity.