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

Module - 7
Lecture - 52
Multicollinearity

Hello and welcome back to the lecture on Applied Econometrics. So, today we are going to cover an important topic called multicollinearity. Now, when we spoke about the Gauss-Markov assumptions regarding regressions, so we talked about different properties that we need to satisfy. So, multicollinearity is one among them. So, interesting thing about multicollinearity is that it is not something that we have to ensure that there is no multicollinearity, because there will always be some multicollinearity in a regression equation, but we have to ensure that the degree of multicollinearity is within some limit.

So, we will explain what is multicollinearity and why we are more bothered about the degree as such for multicollinearity. So, multicollinearity, it arises because of the fact that we have many explanatory variables which might themselves be related, and usually that is the case. So, when we talk about any dependent variable, let us say we talk about kid’s education. So, when we talk about kid’s education, there are different explanatory variables; for example, father's education, father's income, and there are so many other variables and which are related.

So, for a kid's education, it is very important that; usually we will see that the father's education and kid's education are related, because a father who is educated, he knows the value of education, and he will want his kid to also be educated. So, they are correlated; but at the same time, the father needs to have enough income to send his kid to a good school, for example.

So, all these variables are important, but at the same time, what happens is that father's education and father's income is also related, because father's income was basically a dependent variable of father's education. So, this way, all these variables are related. All these explanatory variables are related amongst themselves, and at the same time, they are all explaining the new dependent variable which is the kid’s education.

So, you take any other example and you will see the complex nature of this reality. So, everywhere, you will see that all these different explanatory variables which appears to be very, like directly related to the dependent variable; we will see that the independent variable themselves are related.

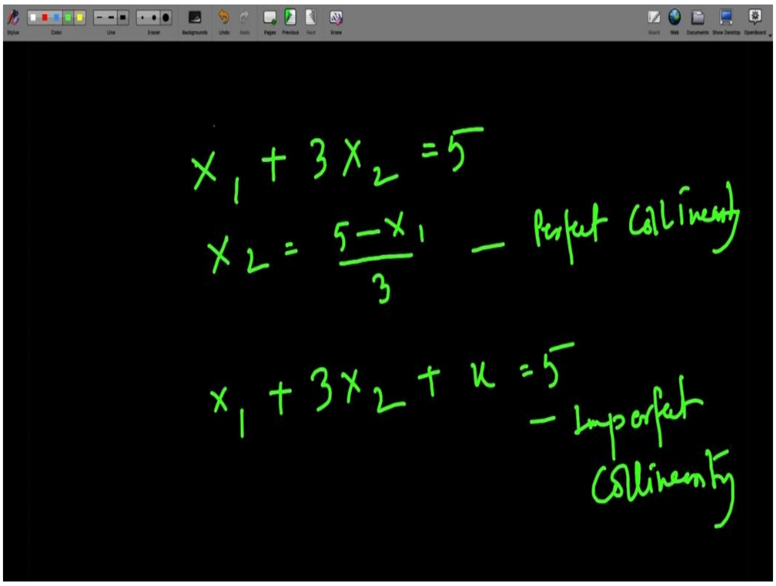
And we will do an example where we will see the number of hours worked by female in the labor force, how they are related to different variables like husband's education, husband's wage, husband's age, the female's parents' education, female's parents' income, the number of kids that female has, the number of kids who are below certain age group, number of kids above certain age group, the tax the family pays and so forth.

And again, if we look at the independent variables, we will see that the independent variable themselves are very highly related. For example, husband's education and husband's income, or husband's income and husband's wage, father's education and mother's education. And so, all the variables that we will see as the X variable, we will see that they are quite clearly, they are also related.

So, that is why, because the reality is some complex, we will see all the explanatory variables, they are actually themselves are related. So, multicollinearity is essentially a measurement of how much this sort of relationships are there among the independent variables, among the X variables. So, that is what we see in multicollinearity. And if we have multicollinearity, we never say that we are violating Gauss-Markov assumption.

We have a problem when we have perfect collinearity, or we have a problem when we have very high multicollinearity. And we are going to see, what do we mean by perfect collinearity? What do we mean by imperfect collinearity? What do we mean by high or low multicollinearity? So, let us give a couple of examples. So, first let me explain what do I mean by a perfect and imperfect collinearity. So, when I am dealing with perfect collinearity, so, perfect collinearity will normally come from equation like, let us say X 1 + 3 X 2 = 5.

(Refer Slide Time: 04:32)



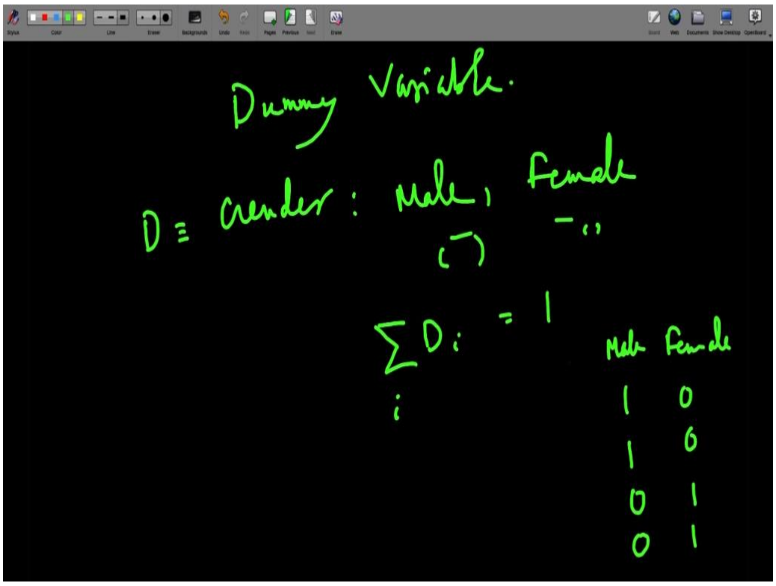
So, they are like, it is a deterministic relationship. There is no stochastic component here. So, every time you basically get, if you want to express X 2 in terms of X 1, it is going to be 5 minus X 1 by 3. So, all the time, the relationship is going to be like that. So, you do not have any component of randomness. So, you always get for a certain value of X 1, you always get a certain value of X 2. So, this is a case of perfect collinearity.

Whereas, in case of imperfect collinearity, how we write it as, X 1 + 3 X 2 plus some stochastic component, let us say some error term u is equal to 5. So, in this case, you always have certain error term that will stop the equation from being a deterministic equation, and that is why you can run a regression equation. And when you have this sort of an equation, you can actually get a regression equation; we can also see the case of; this is a case of imperfect collinearity.

And only in case of imperfect collinearity, we have this problem of multicollinearity that we see; imperfect collinearity. So, we will do some examples to see what is perfect collinearity, what is imperfect collinearity, where multicollinearity may arise; and some of the cues that we get at this stage, we will use when we actually elaborate different examples of multicollinearity. So, let me give you an example.

(Refer Slide Time: 06:27)

1.



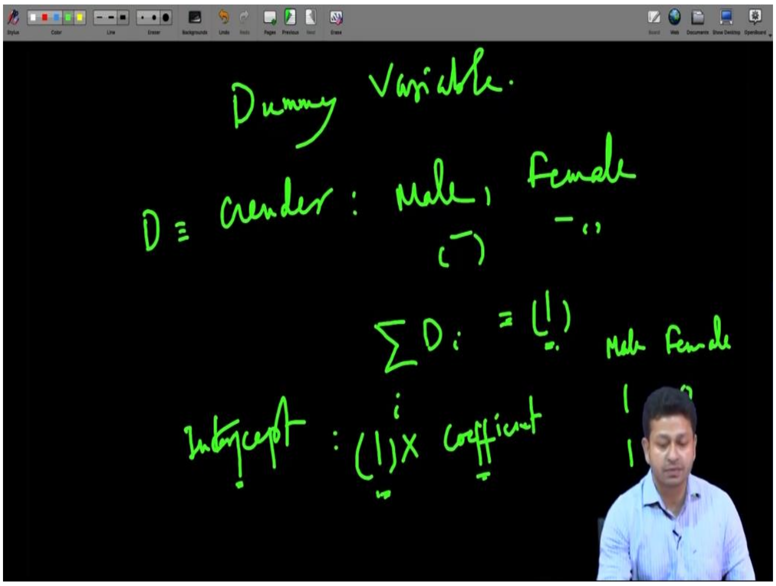
So, when we dealt with the concept of dummy variable, you remember; let us say, if there is a dummy variable, let us say gender, and you have 2 categories, male and female. Now, we always used one particular category as the reference or base category or base dummy. For example, if I use female as a reference or base category, so, we used the male in the regression equation, and we saw how male is changing with respect to the female.

So, that is how we explain a dummy variable. Now, if I include both male and female in the regression equation, so, what we will see is that, your regression equation will actually stop because there is a problem of collinearity; it will give you an error message that there is a problem of collinearity, perfect collinearity. Now, perfect collinearity here in the case of dummy variable, comes from the fact that; so, like, if I define this dummy as D, the moment you add all the different values of dummies; so, let us say D i over all i; so, it will always give me a value of 1, because if someone is male, then he is not female, or if someone is female, she is not male.

And if I represent this dummy with respect, using this values 0 1; let us say female is the base dummy, so I put 0. So, if someone is female, if female is 0, then the person has to be male; so, it is 1. And if a person is male, then that person has to be female; if not male, then the person has to be female. And if the person is not male, that person has to be female, and so forth. So, if I add these values, it will always give me a value of 1, because 0 plus 1 is always 1.

Now, the problem that we have in case of dummy variable we have seen is that, this summation of dummy which gives me a coefficient of 1 is actually perfectly linearly related with the coefficient of the intercept terms. So, intercept term always gives you a certain value, but that intercept term could also be like, represented as 1 into certain, the coefficients.

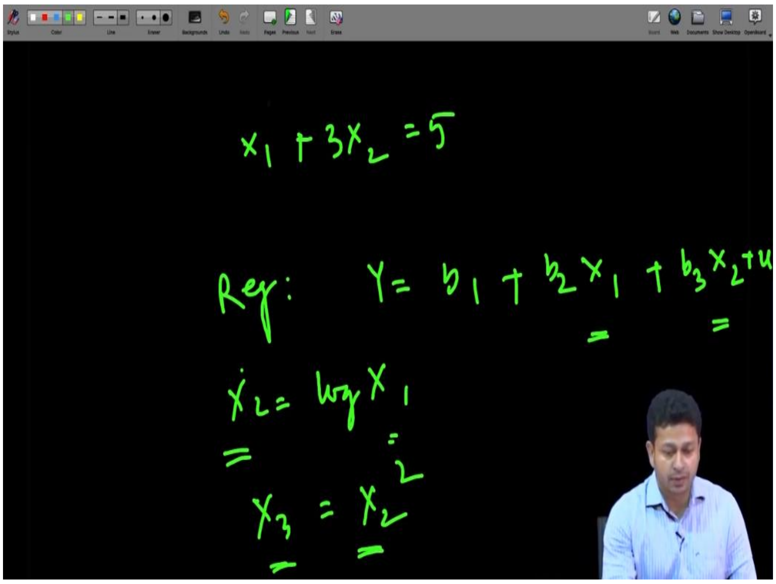
(Refer Slide Time: 08:38)



So, the value of the variable is 1 and it has certain coefficient value. So, now, because of that, so, in case of dummy, it has 1 as a variable value, the X value, and it has a coefficient value. Now, because both of these variables; so, dummy variable and intercept term, both of these has this 1 as the value of the variable; so, it will always give rise to a problem of a perfect collinearity.

So, because 2 variables, they are having the same value; so, they are varying in the same manner; so, that is why they have a problem of perfect collinearity. So, similarly, if I have; the example we have written;

(Refer Slide Time: 09:31)



So, if I use X 1 + 3X 2 = 5. And if I have in my regression, let us say regression equation where I write Y is equal to beta 1 plus beta 2 X 1 plus beta 3 X 2; and I can add a u, error term. If I run this regression equation, so, what will happen is that, because X 1 and X 2, if they are perfectly linearly related, so, my regression equation actually dropped one of these variables.

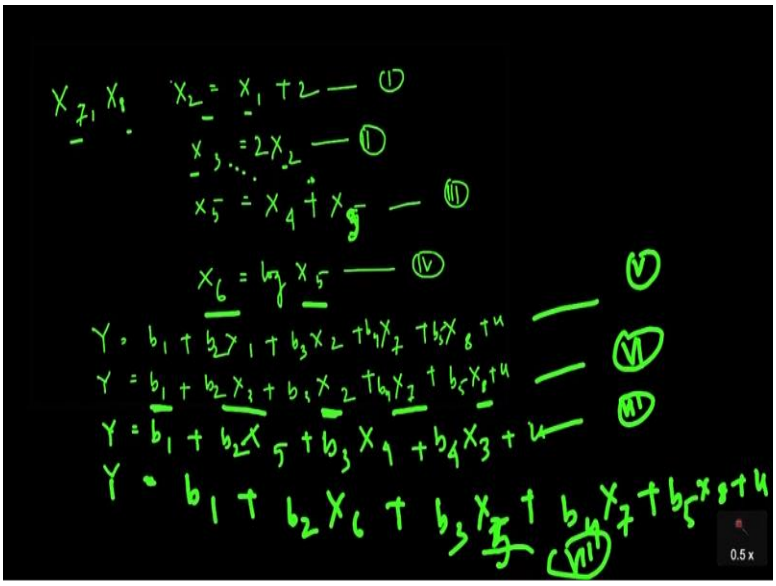
So, either it will say error message that it cannot run the regression, because X 1 and X 2 are perfectly linearly related or it will drop one of these variables. If you are using a software like Stata, so, it will drop one of these variables; or for that matter, R and other softwares also, they are smart enough to do that. So, it will drop one of these variables and it will actually run the regression for you; but there will be a drop of 1 variables, and that is because of the problem of perfect collinearity; but as such, there is no problem.

If you have some amount of stochastic component involved, then there is no perfect collinearity and you will have a good regression equation. So, then the question of degree of multicollinearity will come if you have the stochastic regression equation. Here we have to kind of see that there are different type; we will do couple of examples. And one thing we have to remember that, when we do not have; I mean, there could be relationship, basically a very defined relationship, but the relationship is not linear.

So, I am talking about collinearity. So, if the relationship is not linear, then you do not have; you still have the problem multicollinearity, but you do not have the problem of perfect collinearity. So, let us say, if I represent, my X 2 is equal to log of X 1. So, here the relationship is pretty defined, but they are not linear. So, the moment it is not linear, you do not have perfect collinearity problem, you still have the collinearity problem, the multicollinearity problem, but it is not the perfect collinearity problem.

Or you can write maybe X 3 is equal to say X 2 square or so forth. So, then again, it is not a linear relationship; so, here, again, you will not have perfect collinearity problem, but you will have multicollinearity; you will have imperfect collinearity. So, that is what we are going to see in next few examples. So, let us do some examples. So, let me actually create a small quiz for you, and we will try to answer the quiz.

(Refer Slide Time: 12:18)



We will do a hands on, and then we will get back to this and we will try to answer the quiz. So, my first question is that, suppose I have my X 2, let us say; X 2 = X 1 + 2; my equation number 1. Then I have let us say my X 3 = 2 X 2; equation number 2. Then I introduce some random variable X 4. And let us say my X 5 = X 4 + X 5. Then I have my X 6. My X 6 is, let us say I will define X 6 as log of X 5. So, these are like my 4 equations.

Now, my question to you is, if I run a regression, and the regression equation actually takes care of the collinearity, and it actually drops variable if there is a collinearity problem, so, how many variables will be dropped in each of this equation if my regression equation is of this form? Let us say I have 2 other variables, X 7 and X 8. And for each of this equation, my regression equation is like that: Y is equal to beta 1 plus beta 2 X 1 plus beta 3 X 2 plus X 7 plus X 8.

So, this is 1 equation, let us say. Then you have, you run another regression equation, Y is equal to beta 1 plus beta 2 X 3, and then beta 3 X 2; I am just taking the variable from all these above equations; and then I add X 7, X 8. Let us say I have to introduce some other coefficient; so, beta 4 beta 5. And we can also add some error term here; so, here, again, I will add beta 4 beta 5 X 8 plus some error term. I minimise it.

Let me write couple of more equations. So, Y is equal to, in this case is going to be beta 1 plus beta 2 X 5, beta 3 X 4, and beta 4 X 5 X 3. I use the X 3 from here to here; so, X 3. So, that is going to be beta 4 X 3. The equations might look a little messy, but it is actually very simple thing what I am doing. What I am just trying to do is, I am basically introducing; in any regression equation, I am having these 2 terms, and I am incorporating all the variables that I am writing in the different equations.

So, for every different regression equation, I am using different variables from the different equations. In this regression equation, I also have to add; actually, I have not added X 7 and X 4, but it does not matter really; let us say we omit X 7 and X 8. And then, the last equation, let us say I write Y is equal to; have to minimise the size further; I hope it is still visible; Y is equal to beta 1 plus beta 2 X 6 plus beta 3 X 7 or beta 4 X 7 beta 5 X 8 plus some error term.

Now, I have all these equations. I have written very long equations. Now, let us say I also name them; this is 5, this is 6, this is 7 and this is 8. Now, I will let you guess. If I run regression for all these different equations, so, how many variables will be dropped? So, in the first equation; I will actually try to brainstorm here, and then we will do a hands on to show you what exactly is happening. So, let us look at equation 5.

And in equation 5, I have X 2 and X 1 are there; they have a perfect linear relationship; so, they are perfectly collinear. And because of that, definitely 1 variable has to be dropped. So, either X 1 or X 2 will be dropped, and the regression will run. So, there will be 1 variable drop in this case. Similarly, for equation 2, it is again a perfect linear relationship between X 3 and X 2. So, whenever you run a regression, 1 variable will get dropped.

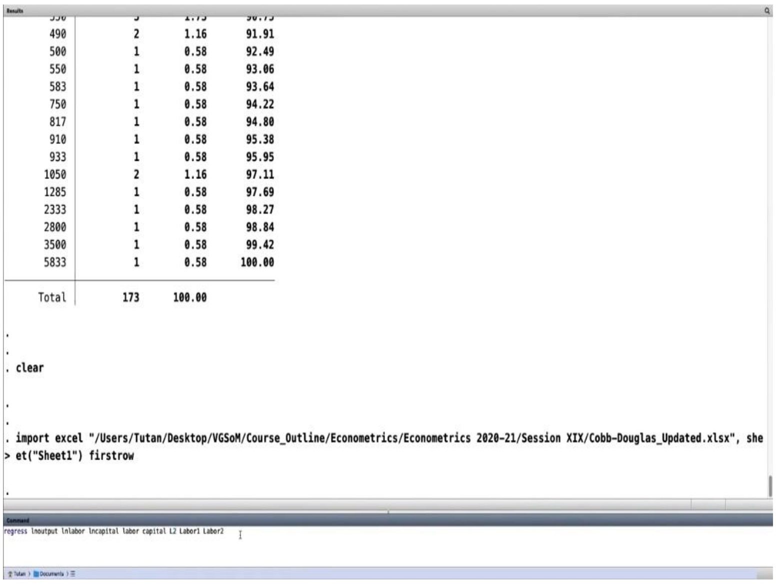
So, it depends; Stata or other softwares can decide which variable it will drop. So, either X 2 or X 3 will get dropped, and 1 variable will be dropped. So, either it is going to be X 3 or it is going to be X 2 here, any of these. So, essentially, this regression equation will show 1 coefficient for the intercept term, and 1 from either X 2 or X 3, and then 2 for X 7 and X 8. So, there is no problem with X 7, X 8, because they are not linearly related with any of this.

Now, let us see what happens for equation 7. Equation 7, you have X 5 and X 4 and X 3. So, they are all linearly related, whereas X 5 is a sum of X 4 and X 3. Now, here you have like X 3 and X 4, they are independent; they are not related; or they are not perfectly linearly related; but X 5 is. So, that is why, because of this dependence, it will drop one of these variables, either X 5 or it will drop X 3 or X 4.

So, it will again drop 1 variable in equation 7 here. And finally, in equation 8, the last equation, here again I have my X 6 and X 5. There is a direct relationship, but they are not linear, so, there is a logarithmic relationship. So, here, there will be no drop of any of these explanatory variables. All the explanatory variables will be there because of the fact that in this relationship, in equation 4, X 6 and X 5, they are not linearly related.

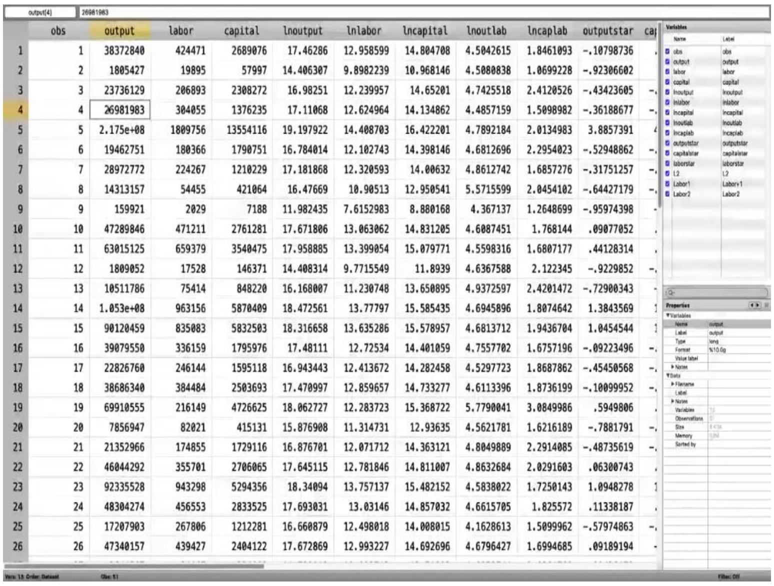
So, that is basically the idea of a perfectly linear relationship and imperfect relationship basically. So, in case of perfect collinearity, there is a problem. And in case of imperfect collinearity, there is the problem of multicollinearity, and we have to see to what extent we have the multicollinearity problem. Now, let us do one hands on example, to get an idea how it is happening.

(Refer Slide Time: 20:05)



So, let me actually open a data set; and I am using the software Stata, Stata 14, and I am using a very common sort of problem or common sort of equation that we have in macroeconomics that is Cobb-Douglas production function.

(Refer Slide Time: 20:30)



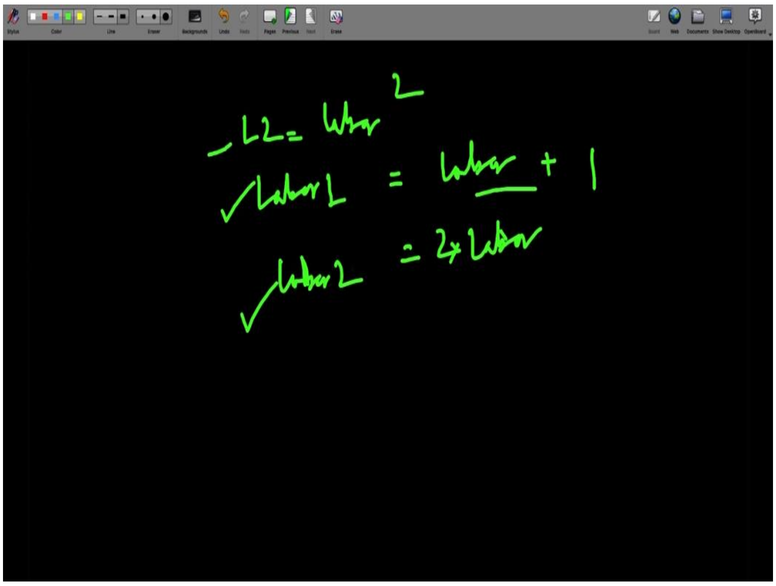
	obs	output	labor	capital	lnoutput	lnlabor	lncapital	lnoutlab	lncaplab	outputstar
1	1	38372840	424471	2689076	17.46286	12.958599	14.804708	4.5042615	1.8461093	-.10798736
2	2	1805427	19895	57997	14.406307	9.8982239	10.968146	4.5080838	1.0699228	-.92306602
3	3	23736129	206893	2308272	16.98251	12.239957	14.65201	4.7425518	2.4120526	-.43423605
4	4	26981983	304055	1376235	17.11068	12.624964	14.134862	4.4857159	1.5098982	-.36188677
5	5	2.175e+08	1809756	13554116	19.197922	14.408703	16.422201	4.7892184	2.0134983	3.8857391
6	6	19462751	180366	1790751	16.784014	12.102743	14.398146	4.6812696	2.2954023	-.52948862
7	7	28972772	224267	1210229	17.181868	12.320593	14.00632	4.8612742	1.6857276	-.31751257
8	8	14313157	54455	421064	16.47669	10.90513	12.950541	5.5715599	2.0454102	-.64427179
9	9	159921	2029	7188	11.982435	7.6152983	8.880168	4.367137	1.2648699	-.95974398
10	10	47289846	471211	2761281	17.671806	13.063062	14.831205	4.6087451	1.768144	.09077052
11	11	63015125	659379	3540475	17.958885	13.399054	15.079771	4.5598316	1.6807177	.44128314
12	12	1809052	17528	146371	14.408314	9.7715549	11.8939	4.6367588	2.122345	-.9229852
13	13	10511786	75414	848220	16.168007	11.230748	13.650895	4.9372597	2.4201472	-.72900343
14	14	1.053e+08	963156	5870409	18.472561	13.77797	15.585435	4.6945896	1.8074642	1.3843569
15	15	90120459	835083	5832503	18.316658	13.635286	15.578957	4.6813712	1.9436704	1.0454544
16	16	39079550	336159	1795976	17.48111	12.72534	14.401059	4.7557702	1.6757196	-.09223496
17	17	22826760	246144	1595118	16.943443	12.413672	14.282458	4.5297723	1.8687862	-.45450568
18	18	38686340	384484	2503693	17.470997	12.859657	14.733277	4.6113396	1.8736199	-.10099952
19	19	69910555	216149	4726625	18.062727	12.283723	15.368722	5.7790041	3.0849986	.5949806
20	20	7856947	82021	415131	15.876908	11.314731	12.93635	4.5621781	1.6216189	-.7881791
21	21	21352966	174855	1729116	16.876701	12.071712	14.363121	4.8049889	2.2914085	-.48735619
22	22	46044292	355701	2706065	17.645115	12.781846	14.811007	4.8632684	2.0291603	.06300743
23	23	92335528	943298	5294356	18.34094	13.757137	15.482152	4.5838022	1.7250143	1.0948278
24	24	48304274	456553	2833525	17.693031	13.03146	14.857032	4.6615705	1.825572	.11338187
25	25	17207903	267806	1212281	16.660879	12.498018	14.008015	4.1628613	1.5099962	-.57974863
26	26	47340157	439427	2404122	17.672869	12.993227	14.692696	4.6796427	1.6994685	.09189194

Now, we know that in Cobb-Douglas production function, output and labor and capital, they are related. And I have like this linear values for output, labor and capital. And at the same time, I have logarithmic value ln output, ln labor and ln capital. So, for this purpose, we just need to remember these variables. We do not need other variables now. And here, what I have done is, I have included other variables.

So, L2 is basically a square of the value for the labor. So, basically, L2 is L square. And Labor 1, this variable I have created. I have created labor; so, I have defined the variable here; labor plus 1 is my labor 1. So, I have just added a value of 1. So, it is again a linear relationship between labor and labor 1. So, labor 1 is equal to labor plus 1. And here I have, in labor 2, what I have done is, I have basically multiplied labor with 2.

So, the value of labor, whatever I have, I have multiplied 2 into labor. And this L2 is L square, and labor 2 is 2 into labor. So, L2 is a nonlinear relationship; it is square term. And this labor 1 and labor 2, they are linear relationship. So, basically, if I write them down, so, how this will look like is this.

(Refer Slide Time: 22:04)

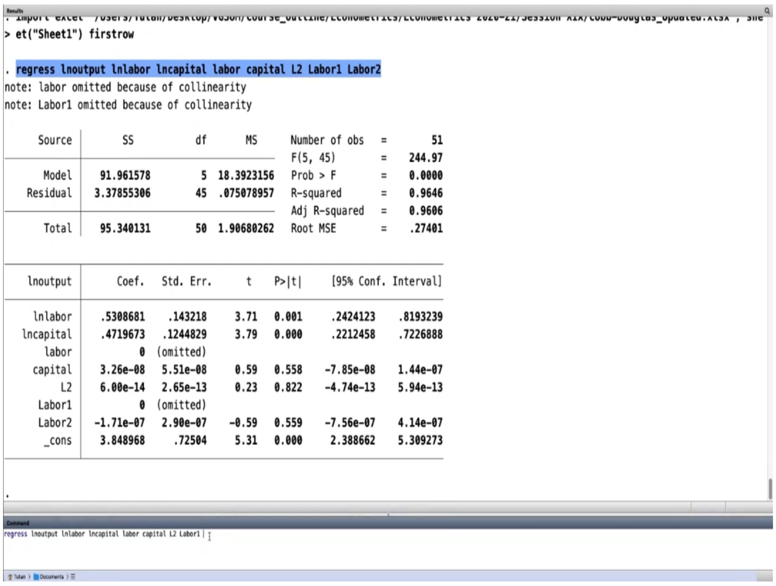


So, the variable L2 is basically labor square. Labor 1 is equal to labor plus 1; and labor 2 is equal to 2 into labor; just to avoid confusion, we have named it this way, but just to avoid confusion, I have just written it again. Now, if I run a regression with all these variables; now, we have to say using our previous concepts where, how many variables my Stata program will actually drop if I run a regression? And the command is going to be; actually, use a do file here or I can just simply write the command here.

So, if I write; so, the command is regress. And let us say my regression, I am doing ln output on output, on log of output, with ln labor, ln capital, just labor, just capital; then I am using L2 which is square of the labor value; and then labor 1 which is labor plus 1; and labor 2 which is labor into 2. Now, in this equation, before I actually run the regression, I will ask you to guess how many of the explanatory variables will get dropped.

So, based on the previous concepts, so, L2 is a square, labor square; so, it will not get dropped, it is fine. This one, there is no problem. Here it is a problem because it is a linear relationship and perfectly linear relationship. And similarly, this one also, it is a perfect linear relationship. So, basically, if I run the regression, there has to be 2 variable drops. So, 2 variables have to be dropped.

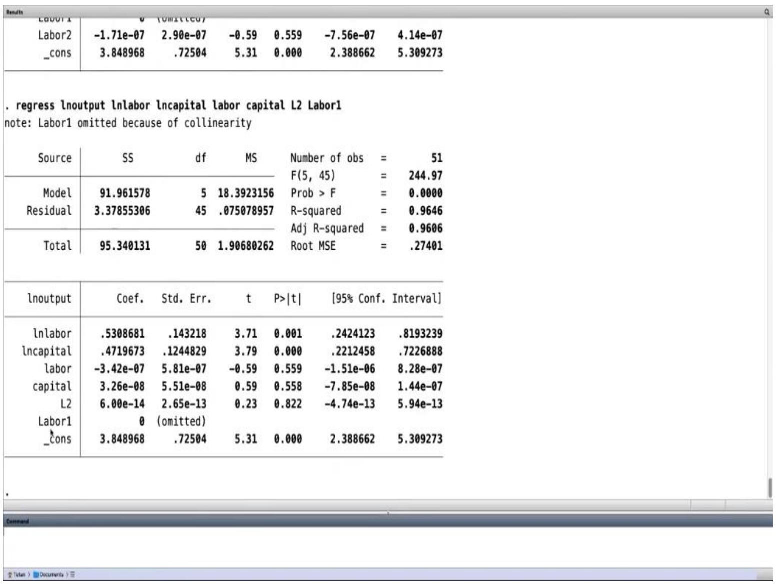
(Refer Slide Time: 24:07)


```
> et("Sheet1") firstrow

. regress lnoutput lnlabor lncapital labor capital L2 Labor1 Labor2
note: labor omitted because of collinearity
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
   lncapital |   .4719673   .1244829     3.79   0.000     .2212458    .7226888
       labor |          0  (omitted)
     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)
      Labor2 |  -1.71e-07   2.90e-07    -0.59   0.559    -7.56e-07    4.14e-07
       _cons |   3.848968     .72504     5.31   0.000     2.388662    5.309273
------------------------------------------------------------------------------
```

And if we run it, we will see that actually there are 2 variables Stata has decided to drop. So, out of these 3, labor, labor 1 and labor 2, it will drop 2 variables out of these 3 variables. So, Stata has chosen to drop labor and labor 1. Now, let me actually again rewrite the code. If I just change the variables a little bit, like if I omit, say labor 2. So, how many variables will be dropped here? So, now the collinearity, the problem is only between labor and labor 1.

Then L square, labor square has no problem, because it is not a linear relationship; though it is a direct relationship, but it is not a linear relationship. So, here, there will be 1 variable drop.

(Refer Slide Time: 24:57)

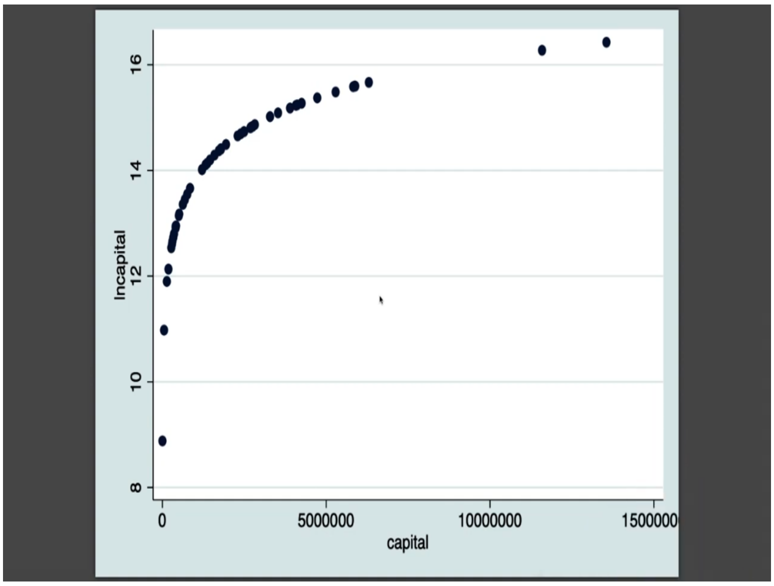

```
      Labor2 |  -1.71e-07   2.90e-07    -0.59   0.559    -7.56e-07    4.14e-07
       _cons |   3.848968     .72504     5.31   0.000     2.388662    5.309273
------------------------------------------------------------------------------

. 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
   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
------------------------------------------------------------------------------
```

So, we will see that Stata has decided to drop labor 1. So, this is how we basically try to understand if Stata is omitting variables. So, we will see that there is the problem of perfect collinearity. And of course, we have seen in the case of dummy variable; I am not repeating here; how we can actually see if you have all the dummy categories included, we will see that Stata will decide to drop one of these variable because of the perfect collinearity problem.

So, that is about the perfect collinearity. And when we talk about imperfect collinearity, so, basically, most of the cases, we will see imperfect collinearity. And again, I said like, in imperfect collinearity, there are multicollinearity problem which is matter of a degree. So, there are always some sort of imperfect collinearity and there is always some sort of multicollinearity because of that; so, we will actually be concerned if the multicollinearity is more than certain extent.

So, the last part of this; let me actually show you. When you try to understand the problem of multicollinearity, what you do is; another way to actually understand is that, you can simply plot scatter plotter; capital; and we will see.

(Refer Slide Time: 26:18)



So, it is always a good idea to actually see what sort of relationships are there among the variables. It is of course, it will be quite highly related because it is a log of capital in the y-axis. So, there will be a very definite sort of relationship. So, when you try to understand the problem of multicollinearity, what you do is, you actually plot all these different X variables and see how they are related.

And if you see some sort of good relationship, like a very discernible pattern, so, then you can claim that okay, there are some multicollinearity which might be present. So, this is another way to understand if there are multicollinearity. Now, I will end the lecture here, the introduction of multicollinearity, and we will see what are the implications of multicollinearity on the regression coefficient on the R-square term and so forth in the next few lectures. Thank you.