



**Econometric Modelling**  
**Professor Sujata Kar**  
**Department of Management Studies**  
**Indian Institute of Technology, Roorkee**  
**Lecture 14**  
**Problems of Multicollinearity**

Hello and welcome back to the course on Econometric Modelling. So in Module 11 -13, I had discussed Multiple Regression Analysis or introduced the Multiple Regression Analysis from different dimensions including its measurement, properties, assumptions and then certain other things in like inferences and adjusted r square.

(Refer Slide Time: 00:54)

<b>Part 1: Introduction to Econometrics</b> Module 1: An Overview Module 2: Formulation of Econometric Modelling Module 3 & 4: Review of Basic Concepts Module 5 : Types of Data	<b>Part 5: Univariate Time Series Modeling</b> Module 25, 26, 27: Problem of Serial Correlation Module 28: AR, MA & ARMA Processes Module 29: Modelling Seasonal Variations
<b>Part 2: Overview of Classical Linear Regression Model</b> Module 6 & 7: Simple Regression Module 8: Assumption of Classical Linear Regression Module 9: Properties of OLS Estimators Module 10: Hypothesis Testing	<b>Part 6: Models with Binary Dependent and Independent Variables</b> Module 30 & 31: Spline Function & Categorical Variables Module 32 & 33: Probit, Logit and Multinomial Logit Models
<b>Part 3: Multiple Regression Analysis &amp; Diagnostic Tests</b> Module 11, 12 & 13: Multiple Regression <b>Module 14: Problems of Multicollinearity</b> Module 15 & 16: Omitted Variables & Parameter Stability Module 17 & 18: Problem of Heteroscedasticity	<b>Part 7: Multivariate Models</b> Module 33 & 34: Simultaneous Equations System Module 35 & 36: Introduction to VARs
<b>Part 4: Statistical Inference</b> Module 19: t-test Module 20 & 21: Wald test Module 22 & 23: F-test Module 24: Chow test	<b>Part 8: Modelling Long Run Relationships</b> Module 37, 38 & 39: Stationarity & Unit Root Testing Module 40: Basics of Cointegration

Now I will be talking about problems of Multicollinearity. This is actually not a part of the classical linear regression model or not a part of the assumptions under classical linear regression model but it is a part of the Gauss Markov Theorem. If you remember Gauss Markov theorem assumes that there should not be any collinearity among the independent variables, specifically perfect collinearity among the independent variables.

(Refer Slide Time: 01:28)


### Multicollinearity

• Model:  $y = X\beta + u$

•  $X$  is a  $n \times k$  matrix. If the rank of  $X < k$ , then uniqueness of the estimated parameters breaks down. If rank of  $X < k$ , then we can get  $Xa = 0$  where  $a$  is any constant. We have found  $\hat{\beta}$  such that  $(y - \hat{y})$  is minimized where  $X\hat{\beta} = \hat{y}$ . But since  $Xa = 0$ , we can also write  $X(\hat{\beta} + a) = \hat{y}$ . Therefore, to get unique results we need to have rank of  $X = k$ .

Perfect Multicollinearity

- Columns of  $X$  are linearly dependent such that  $\exists \theta_{k \times 1}$  for which  $X\theta = 0$ .
- The problem is OLS breaks down. We don't get  $\hat{\beta}$ . Because if OLS gives  $\tilde{\beta}$ , then  $X\tilde{\beta} = \hat{y}$ . But  $X(\tilde{\beta} + \theta) = \hat{y}$  also holds. So, we don't get unique

 NPTEL ONLINE CERTIFICATION COURSE 4

Now this issue is actually addressed here i.e. the problem of multi-collinearity. So, we again begin with our population model which is  $y$  equals  $x$  beta plus  $u$ . Now  $x$  is  $n$  by  $k$  matrix, the rank of  $x$  if less than  $k$  then the uniqueness of the estimated parameters breaks down. Now we are specifically are talking about the problem of perfect multi collinearity, what happens?

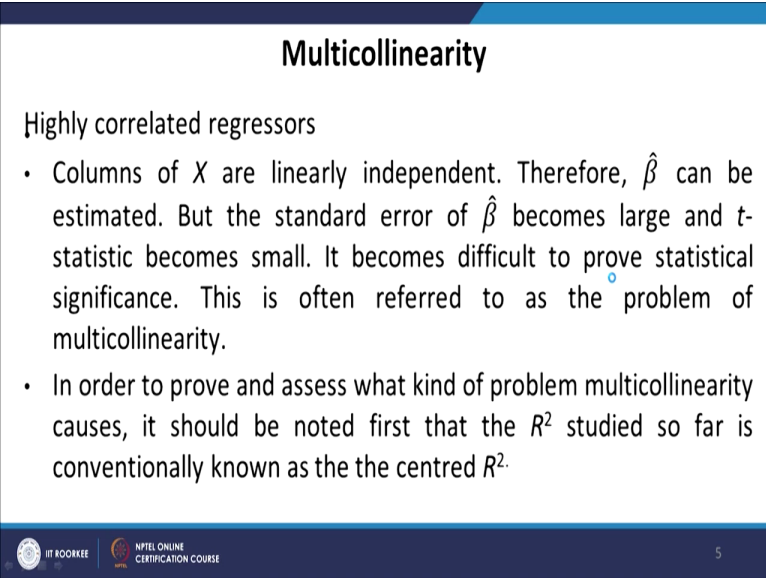
If rank of  $x$  is less than  $k$  then we can get  $x$  multiplied by  $a$  equals to  $0$  where  $a$  is any constant. So we have found  $\hat{\beta}$  such that  $y$  minus  $\hat{y}$  is minimized where  $X\hat{\beta}$  equals to  $\hat{y}$ . So this is the standard thing that we have minimized the residual sum of square in order to get a  $\hat{\beta}$  and from there we have  $X\hat{\beta}$  equals to  $\hat{y}$ . But since  $x a$  is equals to  $0$  we can also write  $X\hat{\beta} + a$  equals to  $\hat{y}$ .

Therefore to get unique results we need to have the rank of  $X$ ,  $X$  equals to  $k$  otherwise this problem of uniqueness will break down. So columns of  $X$  are linearly independent if the columns of  $x$  are linearly dependent such that there exists some  $\theta$  of dimension  $k$  by  $1$  for which  $X\theta$  equals to  $0$  then the problem of OLS breaks down we do not get  $\hat{\beta}$  because if OLS gives us  $\tilde{\beta}$  then  $X\tilde{\beta}$  equals to  $\hat{y}$ .

But  $X\tilde{\beta} + \theta$  is also equals to  $\hat{y}$  and that is why we do not get unique estimates, right? So this is the problem of perfect multicollinearity. Most of the statistical packages actually

do not even provide results if there is perfect multicollinearity. So this is a problem which is actually taken care of by statistical packages which are used to estimate multiple regression models.

(Refer Slide Time: 03:40)



**Multicollinearity**

Highly correlated regressors

- Columns of  $X$  are linearly independent. Therefore,  $\hat{\beta}$  can be estimated. But the standard error of  $\hat{\beta}$  becomes large and  $t$ -statistic becomes small. It becomes difficult to prove statistical significance. This is often referred to as the problem of multicollinearity.
- In order to prove and assess what kind of problem multicollinearity causes, it should be noted first that the  $R^2$  studied so far is conventionally known as the the centred  $R^2$ .

BIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 5

But the problem is actually of concern when we have simple multicollinearity so not perfect multicollinearity but multicollinearity of some level which actually implies highly correlated regressors. Columns of  $X$  are linearly independent so not a problem therefore beta hat can be estimated but the standard error of beta hat becomes large and  $t$  statistic becomes small. It becomes difficult to prove statistical significance this is often referred to as the problem of multicollinearity.

In order to prove and assess what kind of problem multicollinearity causes, it should be noted for us that the  $r$  square studied so far is conventionally known as the centered  $r$  square. So now we are going to introduce something which is called uncentered  $r$  square and explain that what kind of problem we might associate or might have or might associate with the problem of multicollinearity.

(Refer Slide Time: 04:44)

### Multicollinearity

- The centred  $R^2$  is written as  $R^2 = \frac{(\hat{y} - \bar{y})'(\hat{y} - \bar{y})}{(y - \bar{y})'(y - \bar{y})}$  where  $(\hat{y} - \bar{y})'(\hat{y} - \bar{y})$  is ESS and  $(y - \bar{y})'(y - \bar{y})$  is TSS.
- The uncentred  $R^2$  is calculated as  $R_{uc}^2 = \frac{\hat{y}'\hat{y}}{y'y}$  where  $y'y = \hat{y}'\hat{y} + \hat{u}'\hat{u}$ .
- Now, let us split  $X$  into  $X_1$  and  $X_2$  such that  $x_1: n \times 1, X_2: n \times (k-1)$ .
- We need to show that if  $x_1$  is highly correlated with  $X_2$ ; i.e.  $x_1$  almost lies in the column space of  $X_2$ , then SE of  $\hat{\beta}_1$  will be very large.
- To show that first construct  $M_2 = I - X_2(X_2'X_2)^{-1}X_2'$ .

$M = I - X(X'X)^{-1}X'$   
 $\frac{\beta - \hat{\beta}}{SE(\hat{\beta})}$

So we define the centered r square as below:

$$R^2 = \frac{(\hat{y} - \bar{y})'(\hat{y} - \bar{y})}{(y - \bar{y})'(y - \bar{y})}$$

This is centered around the mean value of the series. So  $\hat{y}$  minus  $\bar{y}$  prime multiplied by  $\hat{y}$  minus  $\bar{y}$  this is my explained sum of square and what it is explaining? It is explaining the deviation of the estimated  $\hat{y}$  from the mean of the series  $\bar{y}$ , so it is centered around the mean values and that is why it is called centered r square.

Similarly in the denominator i have  $y$  minus  $\bar{y}$  prime multiplied by  $y$  minus  $\bar{y}$ , so the original values are also centered around their mean values. This is our total sum of squares. Now uncentered r square is calculated as

$$R_{uc}^2 = \frac{\hat{y}'\hat{y}}{y'y}$$

You can see that it is not centered around any value and that is why it is called uncentered r square.

Now let us split the dependent variable or the matrix of dependent variable  $x$  into two components; one is one single variable  $X_1$  and the rest are included in  $X_2$ . So  $X_1$  is a  $n$  by  $1$  vector there is only one variable of course corresponding to that variable I have  $n$  observations and  $X_2$  contains  $n$  by  $k$  minus  $1$  variable provided that I have total  $k$  variables including the constant term. So  $X_2$  is of dimension  $n$  by  $k$  minus  $1$ .

We need to show that if  $X_1$  is highly correlated with  $X_2$  that is  $X_1$  almost lies in the column space of  $X_2$ . Then standard error of  $\beta_1$  hat will be very large. If  $X_1$  that is this variable is highly correlated with all the other variables or some of the other variables then the standard error of  $\beta_1$  hat will be very large and what will be the problem? If you remember while discussing  $t$  inferences in the context of simple regression analysis or even in the context of multiple regression analysis what do we have?

In the numerator, we have  $\beta_1$  hat minus the null hypothesis value of the population parameter  $\beta_1$  divided by the standard error of  $\beta_1$  hat. So if the standard error of  $\beta_1$  hat becomes large it goes up. If the standard error  $\beta_1$  hat becomes very large then the entire thing becomes very small and then it is actually difficult to reject the null hypothesis because it will fall in one of the or most of in the left side rejection region.

Now to show that how these things happen that y presence of high correlation between  $X_1$  and the other variables contained in  $X_2$  causes the problem of multicollinearity, we will first consider or reconsider the projection matrices that I had discussed in the previous to the previous module. So there I defined one orthogonal projection matrix  $M_1$  which was  $i$  minus  $x_1$  into  $x_1$  bar  $x_1$  inverse  $x_1$  prime.

Similarly, I have the orthogonal projection matrix  $M_2$  here defined in terms of  $x_2$ , so  $i$  minus  $x_2$  into  $x_2$  prime  $x_2$  inverse  $x_2$  prime.

(Refer Slide Time: 08:28)

### Multicollinearity

Following the F-W-L theorem  $M_2 = I - X_2(X_2'X_2)^{-1}X_2'$

1. we run a regression of  $x_1$  on  $X_2$  and collect the residuals  $M_2 x_1$ .
2. Run a regression of  $y$  on  $X_2$  and collect the residuals  $M_2 y$ .
3. Regress  $M_2 y$  on  $M_2 x_1$  to obtain  $\hat{\beta}_1$ .



Therefore,  $\hat{\beta}_1 = (x_1' M_2 x_1)^{-1} (x_1' M_2 y)$   $\hat{\beta} = (X'X)^{-1} X'Y$

$$\text{Var}(\hat{\beta}_1) = \sigma_u^2 (x_1' M_2 x_1)^{-1}$$

$$= \sigma_u^2 [x_1' x_1 - x_1' X_2 (X_2' X_2)^{-1} X_2' x_1]^{-1}$$

$$= \sigma_u^2 \left[ x_1' x_1 \left( 1 - \frac{x_1' X_2 (X_2' X_2)^{-1} X_2' x_1}{x_1' x_1} \right) \right]^{-1}$$

$\hat{\beta} = \sigma^2 (X'X)^{-1}$



7

Following the Frisch-Waugh Lovell theorem we run a regression of  $x_1$  on  $x_2$  and collect the residuals  $M_2 X_1$ , so we are trying to find out how much of  $X_1$  is explained by  $X_2$  that is all the other independent variables. In order to do that first running a regression of  $X_1$  on  $X_2$  and collecting the residuals then we run a regression of  $y$  on  $X_2$  and collect the residuals  $M_2 y$ .

Now if we regress  $M_2 y$  on  $M_2 X_1$  then we will be obtaining the estimated value of  $\beta_1$  hat. Now why we are taking this indirect route following Frisch-Waugh Lovell theorem? Because we have a purpose to prove that this correlation between  $X_1$  and  $X_2$  can be troublesome for testing the significance or while testing the significance of  $\beta_1$  hat. So we write  $\beta_1$  hat as see you remember  $\beta$  hat is  $x' x^{-1} x' y$  where  $x$  is the independent variable,  $y$  is the dependent variable. (Refer to slide time 08:28)

So here in this context, my independent variable is  $M_2 x_1$  and my dependent variable is  $M_2 y$  so in a similar fashion when I estimate  $\beta_1$  hat  $x' x^{-1}$  will be replaced with  $(x_1)' M_2 X_1$  prime which is  $X_1$  prime  $M_2$ ,  $M_2 X_1$ . So again  $M_2 x_1$  whole inverse multiplied by  $x_1$  prime  $M_2 y$  so this actually is  $x_1$  prime  $M_2$  prime  $M_2 x_1$  inverse  $x_1$  prime  $M_2$  prime  $M_2 y$  right, if you remember  $M_2$  is a symmetric matrix so  $M_2$  prime  $M_2$  is actually  $M_2$ ,  $M_2$  and they are also idempotent matrices. So  $M_2$ ,  $M_2$  is actually equal to  $M_2$  and that is why we have this expression and this expression right.



Now if I consider the variance of beta 1 hat then again by drawing analogy if you remember the variance of the beta hat is sigma square x prime x inverse right. So similarly, this is my independent variable so it would be  $M_2 \times 1$  prime  $M_2 \times 1$  whole inverse and  $M_2 \times 1$  prime  $M_2 \times 1$  whole inverses this expression which is equivalent to this expression so that is how we have sigma square u x 1 prime  $M_2 \times 1$  inverse.

After that I replace the value of  $M_2$ , I write the expression of  $M_2$  here,  $M_2$  is  $i$  minus  $x_2$ ,  $x_2$  prime  $x_2$  inverse  $x_2$  prime. So when I replace the value of  $M_2$  here I have  $x_1$  prime multiplied by  $i$  multiplied by  $x_1$ . So  $x_1$  prime  $x_1$  and then this expression is coming in between  $x_1$  prime and  $x_1$  whole inverse taking out  $x_1$  prime  $x_1$  common out I have  $1$  minus the whole expression here divided by  $x_1$  prime,  $x_1$ .

(Refer Slide Time: 12:02)

### Multicollinearity



- Now, the uncentred  $R^2$  from a regression of  $x_1$  on  $X_2$  will be
- $R_{uc}^2 = \frac{\hat{x}'_1 \hat{x}_1}{x'_1 x_1}$   $R_{uc}^2 = \frac{\hat{y}y}{yy}$   $y = x\beta + u$
- Similar to  $\hat{y} = X\hat{\beta} = P_X y$  where  $P_X = X(X'X)^{-1}X'$ , here we can have  $P_2 = X_2(X_2'X_2)^{-1}X_2'$   $M_2 = I - P_2$
- Now  $\hat{x}'_1 \hat{x}_1 = (P_2 x_1)'(P_2 x_1) = x'_1 P_2' P_2 x_1 = x'_1 P_2 x_1$   
 $= x'_1 X_2 (X_2' X_2)^{-1} X_2' x_1$


IIT ROORKEE

NTEL ONLINE  
CERTIFICATION COURSE
8

### Multicollinearity

- $$\text{Hence, } \text{Var}(\hat{\beta}_1) = \sigma_u^2 \left[ x_1' x_1 \left( 1 - \frac{x_1' x_2 (x_2' x_2)^{-1} x_2' x_1}{x_1' x_1} \right) \right]^{-1} \quad \left( 1 - \frac{\hat{x}_1' x_1}{x_1' x_1} \right)$$

$$= \sigma_u^2 [x_1' x_1 (1 - R_{uc}^2)]^{-1} \quad (1) \quad R_{uc}^2$$
- The above expression shows that the  $\text{Var}(\hat{\beta}_1)$  of depends on three factors, the error variance,  $\sigma_u^2$ ,  $x_1' x_1$  and  $R_{uc}^2$ .
- Note that  $R_{uc}^2$  can be also denoted by  $R_j^2$  where it refers to the  $R^2$  obtained by regressing any independent variable  $j$  on the rest of the regressors.



9

Now the uncentered r square from a regression of  $x_1$  on  $x_2$  will be  $\hat{x}_1' \hat{x}_1$  divided by  $x_1' x_1$ . This is simply because the uncentered r square from a regression of  $y$  on  $x$  where the population model is  $y = x\beta$ , we had written that the uncentred r square is  $\hat{y}' \hat{y}$  divided by  $y' y$ . So in a similar fashion when we are regressing  $x_1$  on  $x_2$  then our estimated  $x_1$  will be denoted by  $\hat{x}_1$ .

So we have  $\hat{x}_1' \hat{x}_1$  divided by  $x_1' x_1$ , so similar to  $\hat{y}' \hat{y} = x\beta$  equals to  $\hat{y}' \hat{y}$  equals to  $p' x y$  where  $p'$  is where we can have another projection matrix and that is  $p_2$ , so  $p_2' M_2$  is actually  $i$  minus  $p_2$  the way  $m$  used to be equals to  $i$  minus  $p$ . So this is how we define  $p_2$  another projection matrix. And if you remember that  $p' x y$  equals to  $\hat{x}' \hat{y}$  equals to  $\hat{y}' \hat{y}$  equals to  $y' p x y$  equals to  $\hat{y}' \hat{y}$ .

So now this is something which we are going to utilize next, So  $\hat{x}_1' \hat{x}_1$  will be basically  $p_2' x_1$  prime  $p_2' x_1$  since  $\hat{y}' \hat{y}$  equals to  $p' x y$   $\hat{x}_1$  will be equals to  $p_2' x_2$  or rather  $p_2' x_1$ . So  $p_2' x_1$  prime multiplied by  $p_2' x_1$ . Now expanding I have  $x_1' p_2$  prime  $p_2' x_1$  which is  $x_1' p_2' x_1$  and now we replace the value of  $p_2$  so this is the value of  $p_2$  or expression for  $p_2$  (refer slide time 12:02).

So  $x_1' p_2$ ,  $x_2' p_2$  inverse  $x_2' p_2$ . This is what I have got for  $\hat{x}_1' \hat{x}_1$  hat right now. This was my variance expression in the previous to previous slide I have just



rewritten it here and then you can see that this expression is exactly the expression here which is  $\hat{x}_1' \hat{x}_1$ . So I can always write it here, this is in the bracket i will be having  $1 - \hat{x}_1' \hat{x}_1$  divided by  $\hat{x}_1' \hat{x}_1$ .


Now remember this was actually the uncentered R square when we are regressing  $x_1$  on  $x_2$  so this expression is replaced with uncentered R square,  $R^2_{uc}$  and the rest of the things remain the same, we call it equation 1 (refer slide time 12:02). So the above expression shows that the variance of  $\hat{\beta}_1$  depends on 3 factors; first of all the error variance, then the second thing is  $\hat{x}_1' \hat{x}_1$  which is the total sum of square when we are regressing  $x_1$  on the rest of the independent variables and the uncentered r square.


So these are the 3 things on which the variance of  $\hat{\beta}_1$  depends. Note that  $R^2_{uc}$  can also be denoted by  $R_j^2$  where it refers to the R square obtained by regressing any independent variable  $j$  on the rest of the regressor, we can also call it auxiliary regression of  $x_1$  on the rest of the independent variables.

(Refer Slide Time: 15:46)

### Multicollinearity

- In equation (1), a larger  $\sigma_u^2$  means larger variances for the OLS estimators.
- Because more “noise” in the equation makes it more difficult to estimate the partial effect of any of the independent variables on  $y$ , this is reflected in higher variances for the OLS slope estimators.
- Second, the larger the total variation in  $x_1$  is, the smaller is  $\text{Var}(\hat{\beta}_1)$ .
- Thus, everything else being equal, for estimating  $\beta_1$ , we prefer to have as much sample variation in  $x_1$  as possible.

 IIT ROORKEE

 NPTEL ONLINE  
CERTIFICATION COURSE

10

## Multicollinearity

- $$\text{Hence, } \text{Var}(\hat{\beta}_1) = \sigma_u^2 \left[ x_1' x_1 \left( 1 - \frac{x_1' x_2 (x_2' x_2)^{-1} x_2' x_1}{x_1' x_1} \right) \right]^{-1} \quad \left( 1 - \frac{\hat{\lambda}' \lambda}{x_1' x_1} \right)$$

$$= \sigma_u^2 [x_1' x_1 (1 - R_{uc}^2)]^{-1} = \frac{\sigma_u^2}{x_1' x_1 (1 - R_{uc}^2)} \quad (1)$$
- The above expression shows that the  $\text{Var}(\hat{\beta}_1)$  depends on three factors, the error variance,  $\sigma_u^2$ ,  $x_1' x_1$  and  $R_{uc}^2$ .
- Note that  $R_{uc}^2$  can be also denoted by  $R_j^2$  where it refers to the  $R^2$  obtained by regressing any independent variable  $j$  on the rest of the regressors.



So in equation 1; a larger sigma u square means larger variances for the OLS estimators. I show you the expression once again. This has an inverse right which means this actually can be written as sigma square u divided by x 1 prime x 1 into 1 minus R u c square right, so sigma square u is in the numerator x1 prime x1 multiplied by 1 minus R u c square is actually in the denominator. So larger sigma square u means larger variances for the OLS estimators because more noise in the equation makes it more difficult to estimate the partial effect of any of the independent variables on y, this is reflected in higher variances for the OLS slope estimators?

Second, the larger the total variation in x1 is the smaller is the variance of beta 1 hat. So if you remember x1 prime x1 measuring the total variation in x1 is in the denominator. So larger the value of the total variation in x1 smaller will be the variance of beta 1. Thus everything else being equal for estimating beta 1. We prefer to have as much sample variation in x1 as possible.

(Refer Slide Time: 17:13)

## Multicollinearity

- Remember that  $y'y$  was termed as TSS. Drawing analogy,  $x_1'x_1$  can be called  $TSS_1$  and generalizing it to  $j$ th independent variable, we may call it  $TSS_j$ .
- Although it is rarely possible for us to choose the sample values of the independent variables, there is a way to increase the sample variation in each of the independent variables: increase the sample size. In fact, when sampling randomly from a population,  $TSS_j$  increases without bound as the sample size gets larger and larger. This is the component of the variance that systematically depends on the sample size.



## Multicollinearity

- Because  $R^2$  measures goodness-of-fit, a value of  $R^2_j$  close to one indicates that most variations in  $x_j$  is explained by the rest of the independent variables (say  $X_2$ ). This means that  $x_j$  and the rest of the variables (or some of them) are highly correlated.
- If  $R^2_j$  increases to one, then  $x_j$  would mostly lie in the column space of  $X_2$  and  $\text{Var}(\hat{\beta}_j)$  will be very large.
- Thus, a high degree of linear relationship between  $x_j$  and  $X_2$  can lead to large variances for the OLS slope estimators.
- Alternatively, for a given  $\sigma_u^2$  and  $TSS_j$ , the smallest  $\text{Var}(\hat{\beta}_j)$  is obtained when  $R^2_j = 0$  which happens if, and only if,  $x_j$  has zero sample correlation with every other independent variable.



Now remember that  $y'$  was termed as the total sum of square, drawing analogy  $x_1'$  can be called total sum of square 1. That is a total sum of square when we are regressing the first variable or variable  $x_1$  on the rest of the independent variables.

And generalizing it to the  $j$ th independent variable we may call it  $TSS_j$ . Although it is rarely possible for us to choose the sample values of the independent variables, there is a way to increase the sample variation in each of the independent variables, increase the sample size. So in fact when sampling randomly from a population  $TSS_j$  increases without bound, the sample

size gets larger and larger. This is the component of the variance that systematically depends on the sample size.

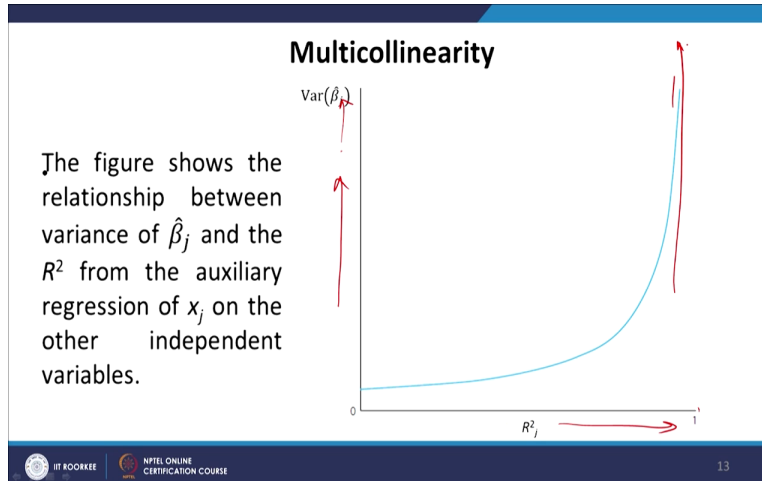
And that is why this problem can always be handled or managed by increasing the sample size. So this is a systematic component that can be actually handled by increasing the sample size because  $R^2$  measures goodness of fit a value of  $R^2$  close to 1 indicates that most variations in  $x_j$  is explained by the rest of the independent variables say  $x_2$ . This means that  $x_j$  and the rest of the variables or some of them are highly correlated.

So higher the value of  $R^2$ , the closer it is to 1 which implies that  $x_1$  is actually explained by the rest of the independent variables. This implies that these independent variables are highly correlated. If  $R^2$  increases to 1 then  $x_j$  would mostly lie in this column in the column space of  $x_2$  and the variance of  $\hat{\beta}_j$  will be very large. Thus a high degree of linear relationship between  $x_j$  and  $x_2$  can lead to large variances for the OLS slope estimators.

Alternatively, for a given  $\sigma_u^2$  and  $tss_j$ , the smallest variance of  $\hat{\beta}_j$  is obtained when  $R^2$  equals 0. See a small variance of  $\hat{\beta}_j$  is a desirable property even though multicollinearity actually does not violate the assumption of efficiency but because of multicollinearity variance of  $\hat{\beta}_j$  tends to increase so this is not a desirable property.

Alternatively when  $R^2$  is actually close to 0, then this implies that very little of  $x_1$  is explained by the other independent variables,  $x_j$  has zero sample correlation with other independent variables and as a result of which we may think that there is little multicollinearity among the independent variables or at least between  $x_1$  and rest of the independent variables.

(Refer Slide Time: 20:19)



So this figure actually shows the relationship between variance of beta j hat and the R square from the auxiliary regression of x j on the other independent variables. As R j squared approaches 1 variance of beta j hat actually keeps on increasing right, we have not mentioned any specific value because it can be anything but then this shows that R j square approaching 1 increases the variance of beta j hat.

(Refer Slide Time: 20:53)

### Multicollinearity

- Since multicollinearity violates none of the CLRM assumptions, the “problem” of Multicollinearity is not really well-defined.
- It does not define how close  $R_j^2$  has to be to one in order to consider a possible Multicollinearity problem. Because even if  $R_j^2$  is close to say 0.9, whether that inflates  $\text{Var}(\hat{\beta}_j)$  or not that also depends on  $\sigma_u^2$  and  $TSS_j$ .
- Because for statistical inference what matters is how big  $\hat{\beta}_j$  is to its variance.
- Just as a large value of  $R_j^2$  can cause a large  $\text{Var}(\hat{\beta}_j)$ , so can a small value of  $TSS_j$ .

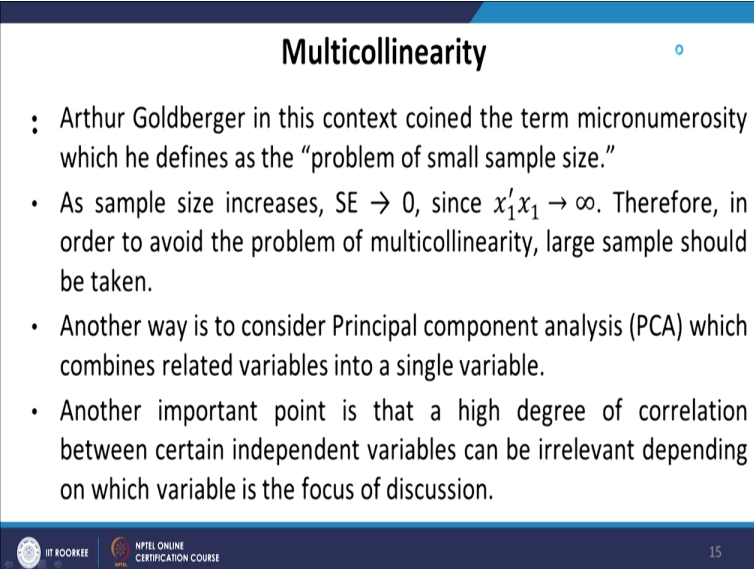
14

Since multicollinearity violates none of the classical linear regression model assumptions the problem of multicollinearity is not really well defined. It does not define how close R j squared has to be to 1 in order to consider a possible multicollinearity problem.

Because even if  $R^2$  is close to say, 0.9, so which is actually close to 1, whether that inflates variance of  $\hat{\beta}_j$  or not depends on the other two factors; that is the variance of the error term as well as the total sum of square obtained from the regression of  $x_1$  on the other independent variables. So because for statistical inference what matters is how big  $\hat{\beta}_j$  is to its variance just a large value of  $R^2$  actually cannot always you know lead to a large variance of  $\hat{\beta}_j$ .

So it is possible that even if  $R^2$  is large, the variance of  $\hat{\beta}_j$  is small because of the other two factors, the other two factors could be offsetting factors as well. So since we have three components we are determining the value of the variance of  $\hat{\beta}_j$  of course one important component is  $R^2$  but then there can be other offsetting factors as well. Since our main concern is on the value of the variance of  $\hat{\beta}_j$ , the problem of multicollinearity is actually not that well defined or well-structured.

(Refer Slide Time: 22:32)



**Multicollinearity**

- Arthur Goldberger in this context coined the term micronumerosity which he defines as the “problem of small sample size.”
- As sample size increases,  $SE \rightarrow 0$ , since  $x_1'x_1 \rightarrow \infty$ . Therefore, in order to avoid the problem of multicollinearity, large sample should be taken.
- Another way is to consider Principal component analysis (PCA) which combines related variables into a single variable.
- Another important point is that a high degree of correlation between certain independent variables can be irrelevant depending on which variable is the focus of discussion.

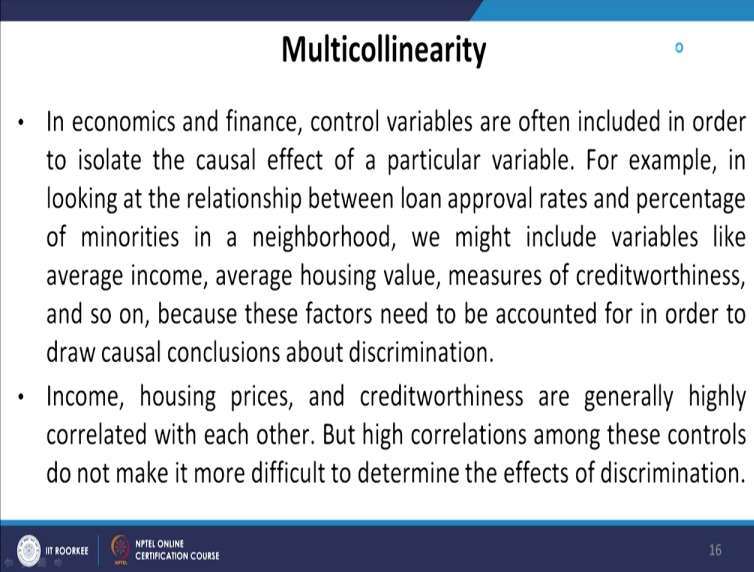
MIT ROOKEE | INTEL ONLINE CERTIFICATION COURSE | 15

Arthur Goldberger in his context coined the term micronumerosity which he defines as the problem of small sample size. As the sample size increases standard error tends to 0. Since  $x_1'x_1$  tends to infinity that is TSS becomes infinity, large or larger. Therefore in order to avoid the problem of multicollinearity large sample should be taken. So one solution is to take a

large sample where an increase in  $x_1$  prime  $x_1$  would actually offset any possibility of a high  $r_j$  square value.

Another way is to consider principal component analysis which combines related variables into a single variable. Another important point is that a high degree of correlation between certain independent variables can be irrelevant depending on which variable is the focus of our discussion okay.

(Refer Slide Time: 23:31)



**Multicollinearity**

- In economics and finance, control variables are often included in order to isolate the causal effect of a particular variable. For example, in looking at the relationship between loan approval rates and percentage of minorities in a neighborhood, we might include variables like average income, average housing value, measures of creditworthiness, and so on, because these factors need to be accounted for in order to draw causal conclusions about discrimination.
- Income, housing prices, and creditworthiness are generally highly correlated with each other. But high correlations among these controls do not make it more difficult to determine the effects of discrimination.

HT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 16

So in economics and finance control variables are often included in order to isolate the causal effect of a particular variable. For example in looking at the relationship between loan approval rates and percentage of minorities in a neighborhood we might include variables like average income, average housing value, measures of creditworthiness, and so on, because these factors need to be accounted for in order to draw a causal conclusion about discrimination. So we are actually looking into the relationship or looking at the relationship between loan approval rates and the percentage of minorities in a neighborhood.

So trying to ideally find out whether there is any discrimination in loan approval depending on whether an individual belongs to a particular minority or not. So in this context we also include certain demographic variables like income and the prices of their houses and their credit

worthiness etc. So income, housing prices, and credit worthiness all are included as control variables because they can also impact the loan approval rate, right.

Now they are actually highly correlated with each other but high correlations among these controls do not make it more difficult to determine the effects of discrimination because our focus is on something different right. We need to include these control variables because they also impact and they might have also you know contributed to the determination of in loan approval rate.

So these cannot be ignored. If these are ignored then that would lead to some other problems in regression analysis but it is quite possible that they are highly co-linear among themselves of course not perfectly collinear but highly correlated but still, they need to be included and that does not hamper our main focus because they actually control variables and our focused or variable of focus is something different.

(Refer Slide Time: 25:48)

**Multicollinearity**

Statistics designed to diagnose the severity of Multicollinearity is prone to misuse because there is no guidelines regarding which value of  $R_j^2$  and  $\text{Var}(\hat{\beta}_j)$  can be considered as indicative of Multicollinearity. Nevertheless, the most commonly used statistic is the Variance Inflation Factor (VIF) and Tolerance

- VIF for the  $j$ th coefficient is calculated as 
$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$
- VIF is precisely the term in  $\text{Var}(\hat{\beta}_j)$  that determines the correlation between  $x_j$  and the other explanatory variables.

NPTEL ONLINE CERTIFICATION COURSE 17

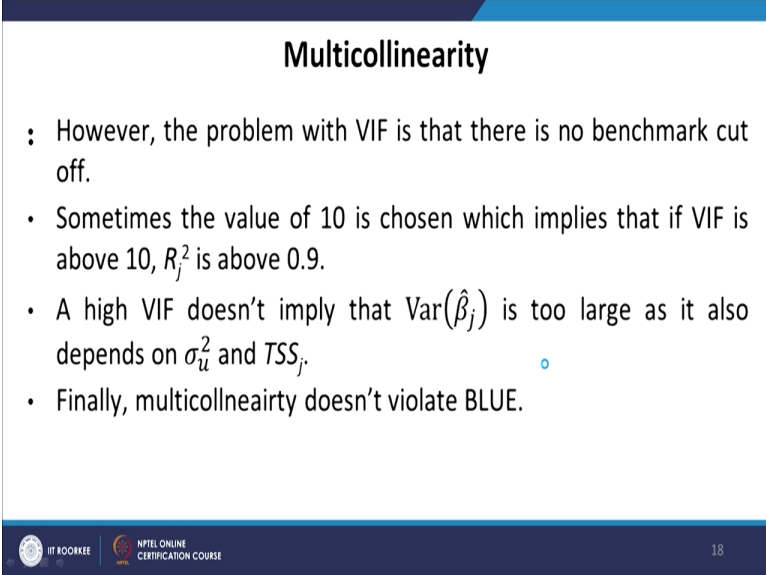
Statistics designed to diagnose the severity of multicollinearity is prone to misuse because there is no guidelines regarding which value of  $R_j^2$  or variance of  $\hat{\beta}_j$  can be considered as indicative of multicollinearity. Nevertheless, the most commonly used statistic is the variance inflation factor and tolerance. Tolerance is just the reciprocal of VIF. So VIF for the  $j$ th



coefficient is calculated as  $1 / (1 - R_j^2)$ . So this is again the regression or the  $R^2$  from the auxiliary regression of the  $j$ th independent variable on the rest of the independent variables.

VIF is precisely the term in variance of  $\hat{\beta}_j$  that determines the correlation between  $x_j$  and the other explanatory variables. So instead of looking at variance of  $\hat{\beta}_j$  we are focusing on only that component which is related to the correlation among the independent variables. So this is given by a measure called VIF.

(Refer Slide Time: 26:55)



**Multicollinearity**

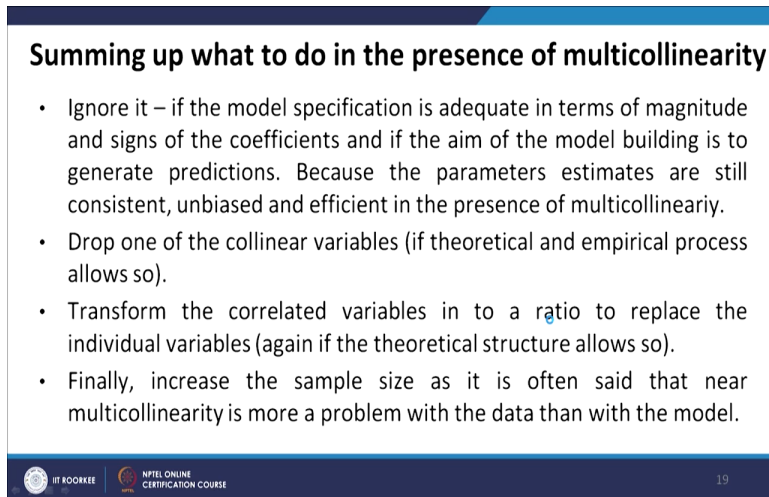
- However, the problem with VIF is that there is no benchmark cut off.
- Sometimes the value of 10 is chosen which implies that if VIF is above 10,  $R_j^2$  is above 0.9.
- A high VIF doesn't imply that  $\text{Var}(\hat{\beta}_j)$  is too large as it also depends on  $\sigma_u^2$  and  $TSS_j$ .
- Finally, multicollinearity doesn't violate BLUE.

MIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE | 18

However the problem with VIF is that there is no benchmark cut-off, sometimes the value of 10 is chosen which implies that if VIF is above 10  $R_j^2$  is above 0.9. A high VIF does not imply that the variance of  $\hat{\beta}_j$  is too large as it also depends on  $\sigma_u^2$  and  $TSS_j$ , this has already been discussed.

Finally, multicollinearity does not violate BLUE that is the Best Linear Unbiased Estimator. This is primarily because as we have mentioned in the beginning itself that it is actually not part of the Classical Linear Regression Model Assumptions. So that is how probably the problem remains less structured.

(Refer Slide Time: 27:44)



**Summing up what to do in the presence of multicollinearity**

- Ignore it – if the model specification is adequate in terms of magnitude and signs of the coefficients and if the aim of the model building is to generate predictions. Because the parameters estimates are still consistent, unbiased and efficient in the presence of multicollinearity.
- Drop one of the collinear variables (if theoretical and empirical process allows so).
- Transform the correlated variables in to a ratio to replace the individual variables (again if the theoretical structure allows so).
- Finally, increase the sample size as it is often said that near multicollinearity is more a problem with the data than with the model.

IT KOOKEE NPTEL ONLINE CERTIFICATION COURSE 19

Now summing up what to do in the presence of multicollinearity, first of all, ignore it if the model specification is adequate in terms of magnitude and signs of the coefficients. If the aim of the model building is to generate predictions because the parameter estimates are still consistent unbiased and efficient in the presence of multicollinearity.

Predictions can always be generated and predictions are always better when we have more and more independent variables regardless of whether they are highly correlated with each other or not. So depending on what is my focus or goal, I can ignore some, certain amount of presence of multicollinearity among the variables.

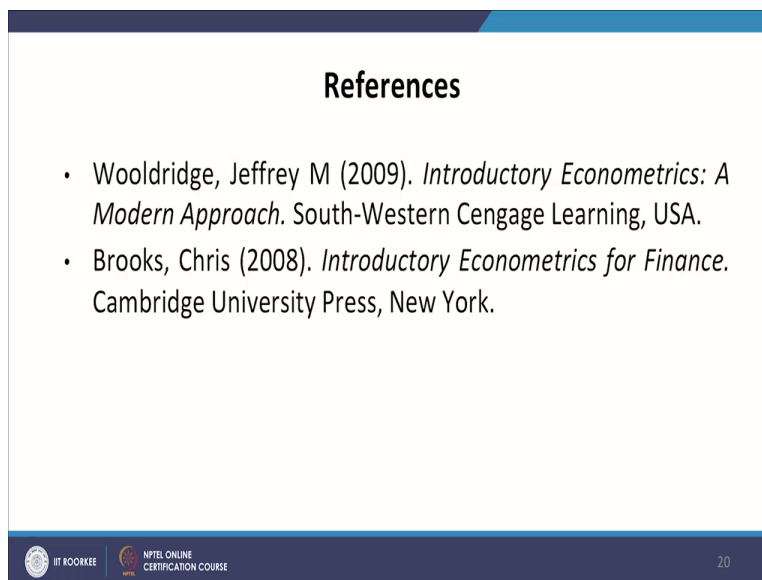
The second is, drop one of the collinear variables if theoretical empirical process allows us to do so, That is, it is not a very important variable which needs to be retained in the model specification. Third, transform the correlated variables into a ratio to replace the individual variables. Again if the theoretical structure allows us to do so. So instead of having individual variables, we can actually have a ratio of the two variables.

And finally, increase the sample size as it is often said that near multicollinearity is more a problem with the data than with the model. So model specification actually has nothing to do with multicollinearity. Multicollinearity could be present in the form of data that we collect, of course certain variable specifications may entail multicollinearity but as has already been

discussed and, also theoretically proven that by increasing the sample size we can actually partially take care of the problem of multicollinearity.

Because the presence of multicollinearity is going to inflate the variance and having a larger sample size or a large sample size we can have an opposite impact on the variance. So inflation in variance will be taken care of by an increase in the total sum of the square from the auxiliary regression of one independent variable on the rest of the independent variables. So that is all about multicollinearity.

(Refer Slide Time: 30:12)



The slide is titled "References" and contains two bullet points. The first bullet point is: "Wooldridge, Jeffrey M (2009). *Introductory Econometrics: A Modern Approach*. South-Western Cengage Learning, USA." The second bullet point is: "Brooks, Chris (2008). *Introductory Econometrics for Finance*. Cambridge University Press, New York." The slide has a blue header and footer. The footer contains the logos for BIT ROORKEE and NITEL ONLINE CERTIFICATION COURSE, along with the page number 20.

**References**

- Wooldridge, Jeffrey M (2009). *Introductory Econometrics: A Modern Approach*. South-Western Cengage Learning, USA.
- Brooks, Chris (2008). *Introductory Econometrics for Finance*. Cambridge University Press, New York.

BIT ROORKEE | NITEL ONLINE CERTIFICATION COURSE | 20

These are the references that I have followed. Thank you.