

Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 23
Understanding Multicollinearity

Welcome to the lecture 7 part A. In this lecture, we are going to talk about Multicollinearity.

(Refer Slide Time: 00:28)

What is multicollinearity?

- ▶ Consider the standard linear model
$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$
where $\epsilon \sim N(0, \sigma^2 I_n)$ and $n > p$
- ▶ This implies $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 I_n)$
- ▶ The least square estimator of β is $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$
- ▶ The sampling distribution of $\hat{\beta}$ is
$$\hat{\beta} \sim N_p(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$$

NPTEL logo in the top right corner. A video inset in the bottom right corner shows Prof. Sourish Das speaking into a microphone, with the 'cmj' logo below it.

We consider the linear model y equal to X beta plus epsilon, where epsilon follow normal distribution 0 sigma square I_n and n is greater than p . So, that means, sample size is greater than number of features that we have in our data set ok. Now, if this is the case, then we can show that y follow normal X beta sigma square I_n , ok. It is a N variate normal. The least square estimator of beta will be beta hat which is X transpose X inverse X transpose y .

(Refer Slide Time: 01:20)

▶ This implies $y \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$

▶ The least square estimator of β is $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

▶ The sampling distribution of $\hat{\beta}$ is

$\hat{\beta} \sim N_p(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$

$E(\hat{\beta}) = \beta$ $V(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$


NPTEL

cmj

What is multicollinearity?

Now, we in the previous lecture, we also discussed that beta hat follow p variate normal with mean as expectation of beta hat as mean as beta and variance of beta hat is sigma square X transpose X inverse.


(Refer Slide Time: 01:47)



What is multicollinearity?

$$X = \begin{pmatrix} X_1 & \dots & X_p \\ | & & | \end{pmatrix}$$
$$\text{Cor}(X_i, X_j) = 1$$


- ▶ If correlation between two predictors of X is 1, that means one column is exactly dependent on other, that will result $\det(X^T X) = 0$
- ▶ Hence $X^T X$ will not be invertible, (because $(X^T X)^{-1} = \frac{\text{Adj}(X^T X)}{\det(X^T X)}$)
- ▶ In such case unique solution does not exist.



Now, the first question we will ask ourselves is what is multicollinearity? If correlation between two predictors, there are p many predictors so, X is typically X_1 to X_p ok. If any of the two predictors, any of the two predictors have correlation exactly 1. So, correlation between say i -th predictor and the j -th predictor is exactly 1. That means one column is exactly dependent on other and that will result determinant of $X^T X$ to be 0.

So, it immediately means that $X^T X$ is not invertible because what is $X^T X$? $X^T X$ is and $X^T X$ inverse is adjugate of $X^T X$ divided by determinant of $X^T X$ and if it is 0 so that means, $X^T X$ is not invertible. In such case, unique solution does not exist right.


(Refer Slide Time: 03:05)



Why multicollinearity is a problem? $(X^T X)^{-1} = \frac{\text{Adj}(X^T X)}{|X^T X| = \delta}$


- ▶ If correlation between two predictors of X is nearly 1 or -1, **but not exactly 1.**
- ▶ For example $\text{cor}(X_i, X_j) = 0.99$ - what happens then?
- ▶ $\det(X^T X) = \delta > 0$, where δ is a very small value.
- ▶ $X^T X$ is invertible - but every element of $(X^T X)^{-1}$ will be **very large.**
- ▶ Unique solution $\hat{\beta}$ exists but $\text{Cov}(\hat{\beta}) = \sigma^2(X^T X)^{-1}$ will be extremely large - so standard error will be very large.

Hence valid statistical inference cannot be implemented



Now, in such situation in now, this is what I am giving you a extreme situation ok. Now, most of the time, you will never probably give put one if one column is exactly linearly dependent on other.

(Refer Slide Time: 03:31)



▶ Hence $X^T X$ will not be invertible, (because

$$(X^T X)^{-1} = \frac{\text{Adj}(X^T X)}{\det(X^T X)}$$


▶ In such case unique solution does not exist.

$X_i = c X_j$

cmj

Why multicollinearity is a problem?

▶ If correlation between two predictors of X is nearly 1 or -1, **but not exactly 1**.



So, that means, effectively you can write X_i equal to some constant of X_j right so, and if we know that there are some columns which is exactly linearly dependent on another column, then we will not put that column in our data set at all. So, this is like extreme example.

More reliable, more reasonable example is that correlation between two predictors nearly 1 or minus 1, but not exactly 1. It will not be exactly 1. So, correlation between say X_i and X_j is 0.99. What happens then? So, that case what will happen is determinant of $X^T X$ is some small delta value where delta is positive. So, determinant of $X^T X$ may be very small value, delta is a very small value ok. It is a very small value.

So, $X^T X$ will be invertible there will be no problem, $X^T X$ will be invertible, but every element of $X^T X$ will be very large. Every element because what

is $X^T X$ inverse, $X^T X$ inverse is adjugate of $X^T X$ and divided by determinant of $X^T X$.

Now, if $X^T X$ determinant of $X^T X$ is very small value, then basically you are dividing every element of adjugate of $X^T X$ by a very small value. Effectively that will make the $X^T X$ inverse very large.

(Refer Slide Time: 05:33)

▶ $\det(\mathbf{X}^T \mathbf{X}) = \delta > 0$, where δ is a very small value.

▶ $\mathbf{X}^T \mathbf{X}$ is invertible - but every element of $(\mathbf{X}^T \mathbf{X})^{-1}$ will be very large.

▶ Unique solution $\hat{\beta}$ exists but $\text{Cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$ will be extremely large - so standard error will be very large.
Hence valid statistical inference cannot be implemented.

NPTEL

cmj

Correlated Predictors

So, what will happen in this case is basically unique solution of beta hat exist, but the covariance of beta hat will be extremely large. So, every element of covariance of X beta hat will be extremely large. So, the standard error so, naturally the standard error of beta hat will be very large. Hence, valid statistical inference cannot be implemented.

(Refer Slide Time: 06:00)

▶ We consider simple no-intercept model: *mtcars*

$$\text{mpg} = \beta_1 \text{wt} + \beta_2 \text{drat} + \epsilon$$

▶ $\rho(\text{wt}, \text{drat}) = -0.71$

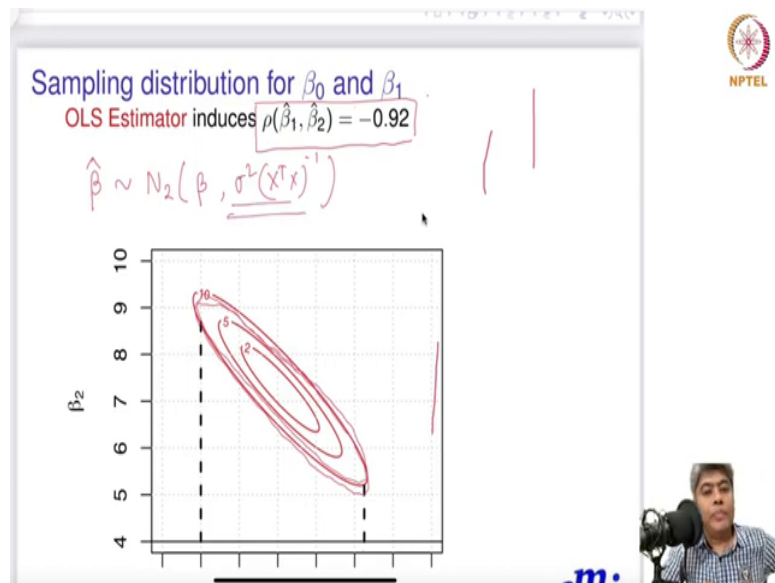
Sampling distribution for β_0 and β_1

OLS Estimator

What does it mean actually? Let us try to understand with a simple day, simple toy example. So, I am considering a no intercept model of the I am considering the empty curves dataset. It is a simple small toy dataset, but very good dataset to understand the multicollinearity. I am using I am considering a no intercept model ok, just to understand the concept. I mean probably I will not use in real when we will I will do a proper analysis, but just to understand the geometry of the concept.

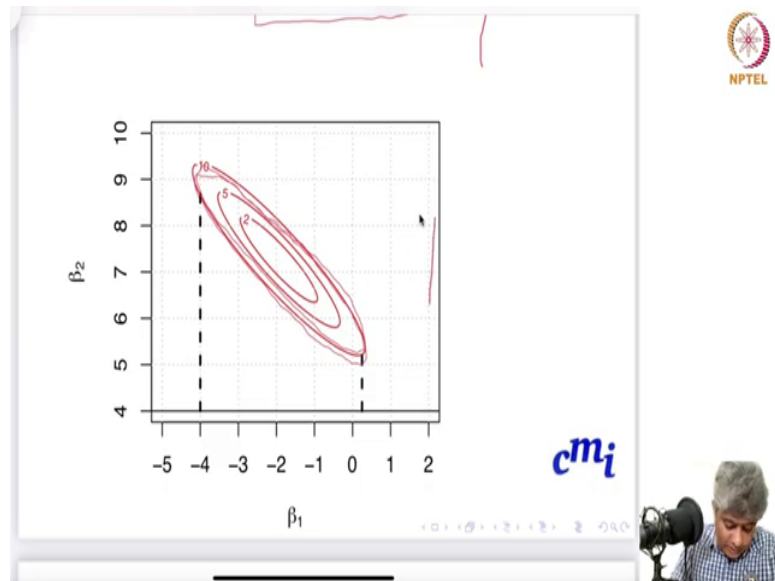
I am using a simple no intercept model where miles per gallon is a function of weight and their axial ratio. So, mpg equal to beta 1 weight plus beta 2 weight plus epsilon, I know correlation between weight and rear axial ratio is negative 0.71 ok not very bad like 0.99 or something, but on a higher side.

(Refer Slide Time: 07:02)



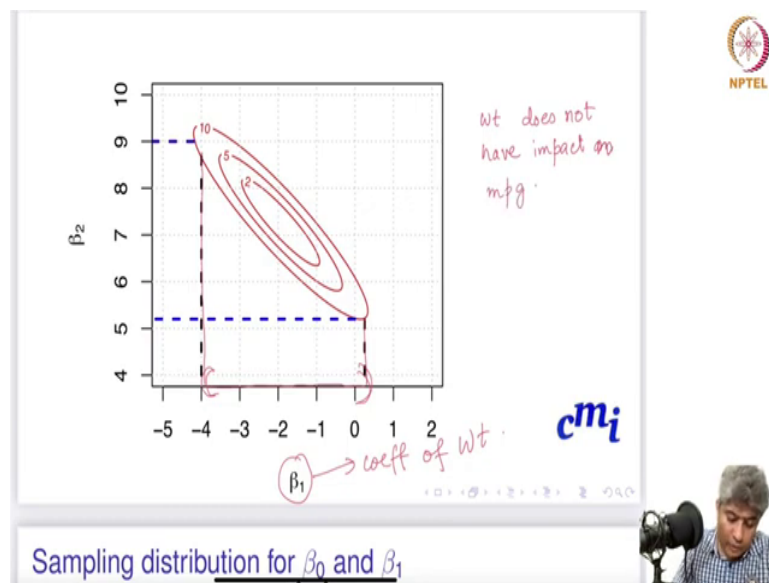
Now, what is happening if you estimate the beta 1 hat and beta 2 hat OLS estimator correlation will be 0.92? Ok. So, what is happening here?

(Refer Slide Time: 07:18)



So, we can see this is very correlation is very high.

(Refer Slide Time: 07:36)

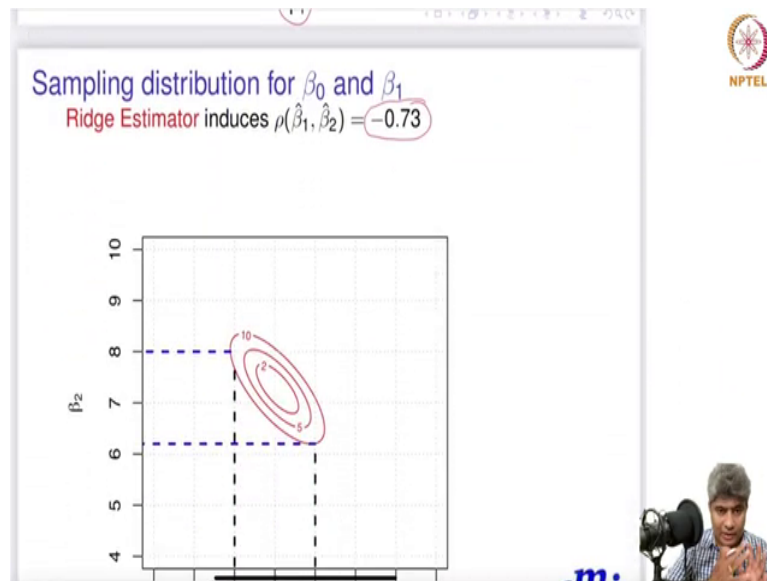


Now, the what is happening is basically the correlation is whatever the correlation that you are seeing between X_1 and X_2 that is getting induced in that is getting induced in the sampling distribution of $\hat{\beta}$, because what is the sampling distribution of $\hat{\beta}$? You see $\hat{\beta}$ is following p variate normal in this case it will be p variate 2 variate normal with $\beta \sigma^2 X^T X^{-1}$.

Now, correlation between the weight and rear axial ratio this will get reflected this will get reflected in the correlation between the 2 of course, adjusted with for sigma squared, but this as a result because the high correlation the whole thing become very large and as a result what is happening is if you do a 95 percent confidence interval, see β_1 was the correlation β_1 was coefficient of weight ok.

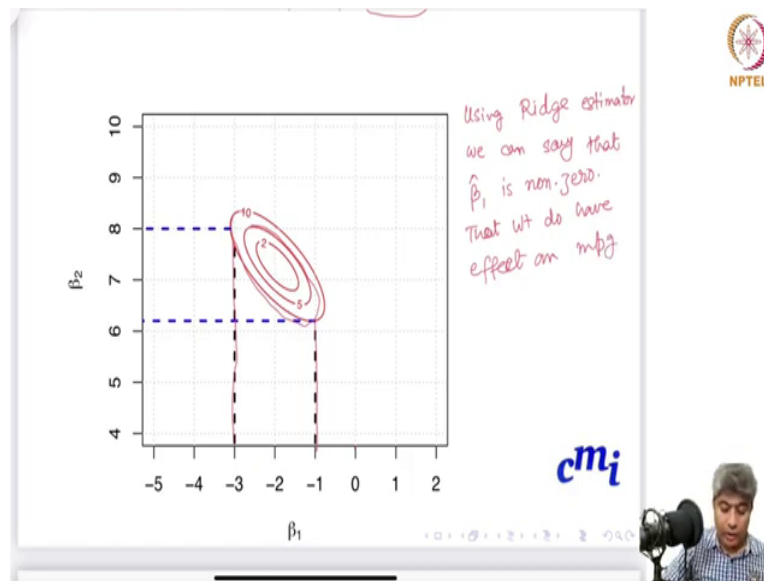
Now, it becomes so big it includes the confidence interval includes the 0. So, based on the our statistical inference will say then beta 1 does not is include 0 is possible value of beta 1. So, weight does not have effect on mpg. So, weight does not have impact on mpg ok.

(Refer Slide Time: 09:50)



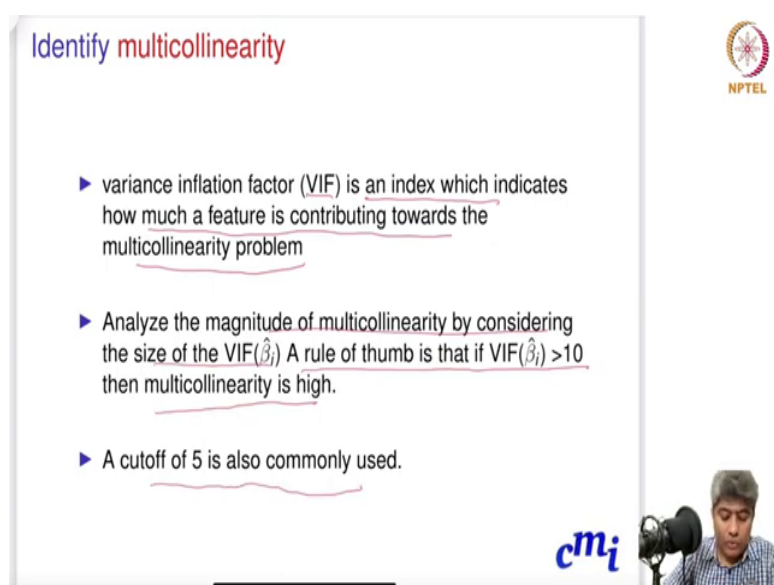
Now, instead of OLS estimator if we use the ridge estimator what happens is ridge estimator, we will talk about it how ridge estimator is being calculated, the ridge the correlation between ridge estimator is negative reduced to point negative 0.73. The correlation between the OLS estimator was very high negative 0.92 and as a result the it was getting very tight. Now, ridge estimator reduce the correlation.

(Refer Slide Time: 10:28)



So, that means, this tightness is now reduced it is now here now the confidence interval does not include 0 anymore. So, using ridge estimator we can say using ridge estimator we can say that beta 1 is non-zero that is weight do have effect on mpg ok.

(Refer Slide Time: 11:32)



Identify multicollinearity

- ▶ variance inflation factor (VIF) is an index which indicates how much a feature is contributing towards the multicollinearity problem
- ▶ Analyze the magnitude of multicollinearity by considering the size of the $VIF(\hat{\beta}_i)$. A rule of thumb is that if $VIF(\hat{\beta}_i) > 10$ then multicollinearity is high.
- ▶ A cutoff of 5 is also commonly used.

NPTEL

cm_i

So, I hope the concept of the multicollinearity is clear now. So, because of the in summary we can say that because of the you know extreme correlation between the predictors the that correlation between the predictor get induced in the sampling distribution of the OLS estimator. As a result the sampling distribution become very tight and as a result the standard error or the margin of error for each estimator become very large.



So, the standard error or margin of error get inflated and statistical we cannot do a proper statistical inference with such kind of estimators. But we have to do; that means, some correction and these estimator is one such correction for and this problem is known as multicollinearity.

Now, for this kind of problem multicollinearity problem we have to do some kind of correction these estimator is one such estimator which actually do the correction and because

of the ridge correction the correlation between the coefficients the sampling distribution reduces and as a result the overall standard error reduces. Next question is how I identify if my data suffers from multicollinearity? There are many different ways to identify one kind of cool proof method is variance inflation factor.

Variance Inflation Factor VIF is an index which indicates how much feature is contributing towards multicollinearity problem. So, analyze the magnitude of the multicollinearity by considering the size of the variance inflation factor or rule of thumb is if the variance inflation factor is greater than 10 then definitely there is a high multicollinearity. Cutoff 5 is also commonly used.

(Refer Slide Time: 13:58)



Variance Inflation Factor


- ▶ Consider the linear regression model
$$\mathbf{y} = \mathbf{X}\beta + \epsilon.$$
- ▶ The standard error of $\hat{\beta}_j$ is
$$se(\hat{\beta}_j) = \sqrt{s^2(\mathbf{X}^T \mathbf{X})^{-1}_{jj}}$$
- ▶ It turns out that variance of $\hat{\beta}_j$ can be expressed as
$$\text{Var}(\hat{\beta}_j) = \frac{s^2}{(n-1)\text{Var}(X_j)} \frac{1}{1-R_j^2}$$

where R_j^2 is the multiple R^2 of X_j on $\{X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_p\}$, i.e.,

Now, how typically what is the motivation of variance inflation factor? Consider a linear regression model y equal to X beta plus epsilon standard error of beta j is this ok. So, you just

take the X transpose X inverse the j j-th diagonal element of the X transpose X inverse multiply with the sample variance and take the square root of that. It turns out that the variance of beta j can be also expressed as this that a square by n minus 1 variance of X j times 1 minus 1 by 1 minus R j square.

(Refer Slide Time: 14:48)



▶ The standard error of $\hat{\beta}_j$ is

$$se(\hat{\beta}_j) = \sqrt{s^2(\mathbf{X}^T \mathbf{X})_{jj}^{-1}}$$


▶ It turns out that variance of $\hat{\beta}_j$ can be expressed as

$$\text{Var}(\hat{\beta}_j) = \frac{s^2}{(n-1)\text{Var}(X_j)} \frac{1}{1-R_j^2}$$

where R_j^2 is the multiple R^2 of X_j on $\{X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_p\}$, i.e.,


$$X_j = \gamma_0 + \gamma_1 X_1 + \dots + \gamma_{j-1} X_{j-1} + \gamma_{j+1} X_{j+1} + \dots + \gamma_p X_p + \epsilon$$

Variance Inflation Factor



Now, what is R j square? R j square is multiple R square of X j on X 1 to X j minus 1 and X j plus 1 to X p. So, basically you define a you set up a simple regression model of where dependent variable X j and equal to gamma naught plus gamma 1 X 1 plus dot dot dot gamma j minus 1 X j minus 1 and gamma j plus 1 X j plus 1 to gamma p X p plus epsilon. You fit that model and get the multiple R square for that model and that models R j square is this R j square.

(Refer Slide Time: 15:30)




▶ The standard error of $\hat{\beta}_j$ is

$$se(\hat{\beta}_j) = \sqrt{s^2(\mathbf{X}^T \mathbf{X})_{jj}^{-1}}$$

▶ It turns out that variance of $\hat{\beta}_j$ can be expressed as

$$\text{Var}(\hat{\beta}_j) = \frac{s^2}{(n-1)\text{Var}(X_j)} \cdot \frac{1}{1-R_j^2}$$


▶ The term $\frac{1}{1-R_j^2}$ is known as the VIF of j^{th} predictor.



Implementation

We can calculate p different VIF's (one for each features):

▶ First we run an ordinary least square regression that has



Now, the term 1 minus R_j square is known as the variance inflation factor of the j -th predictor ok.

(Refer Slide Time: 15:42)

Implementation

We can calculate p different VIF's (one for each features):

- ▶ First we run an ordinary least square regression that has X_i as a function of all the other explanatory variables in the first equation. If $i = 1$,



$$X_1 = \gamma_0 + \gamma_2 X_2 + \dots + \gamma_p X_p + \varepsilon,$$

- ▶ The VIF of β_i would be

$$VIF(i) = \frac{1}{1 - R_i^2},$$

where R_i^2 is the coefficient of determination of the regression equation in step one, with the feature X_i as response, and all other features on the right hand side.

▶ If $VIF(i) > 10$, then multicollinearity is high. A cutoff of 5 is also used.



So, question is how we can calculate p different VIF's one for each feature? First, we run the ordinary least square regression that has X_i as a function of other explanatory variables. Say suppose if i equal to 1 we fit X_1 as a function of $\gamma_0 + \gamma_2 X_2 + \dots + \gamma_p X_p + \varepsilon$ and then you calculate the variance inflation factor β_i would be simply this guy $1 / (1 - R_i^2)$.

(Refer Slide Time: 16:26)

X_i as a function of all the other explanatory variables in the first equation. If $i = 1$,

$$X_1 = \gamma_0 + \gamma_2 X_2 + \dots + \gamma_p X_p + \varepsilon,$$



▶ The VIF of β_i would be

$$VIF(i) = \frac{1}{1 - R_i^2},$$

where R_i^2 is the coefficient of determination of the regression equation in step one, with the feature X_i as response, and all other features on the right hand side.


▶ If $VIF(i) > 10$ then multicollinearity is high. A cutoff of 5 is also commonly used.

Implementation




Once you have that if it is greater than 10 you call it there is a multicollinearity, if it is greater than 5 then that is also can be used for multicollinearity.

(Refer Slide Time: 16:39)



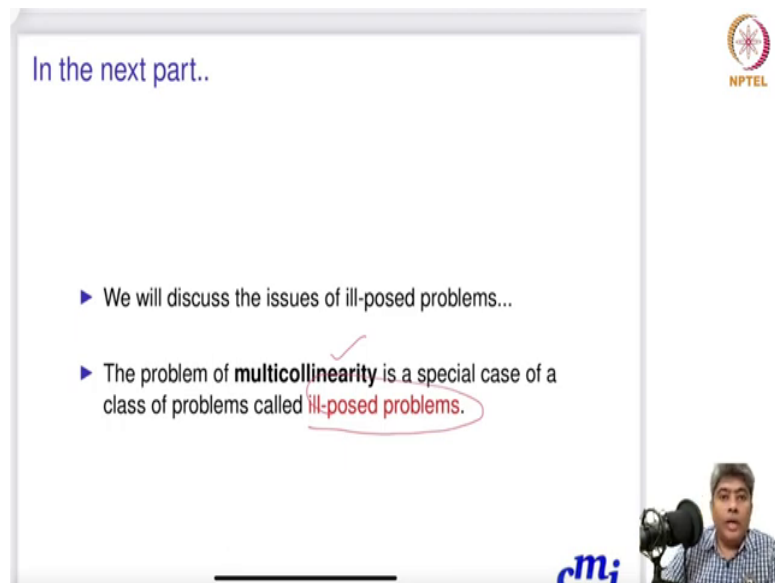
Implementation

- ▶ In R, the function `vif` in `car` package implements the variance inflation factor.
- ▶ In Python, the function `variance_inflation_factor` in `statmodels` can be used to identify the multicollinearity.



So, in R there is a function called VIF in the car package which implements the variance inflation factor. In Python, the function variance inflation factor in statmodels package can be used to identify the multicollinearity.

(Refer Slide Time: 17:07)



In the next part..

- ▶ We will discuss the issues of ill-posed problems...
- ▶ The problem of **multicollinearity** is a special case of a class of problems called **ill-posed problems**.

NPTEL

mi

Next video we will discuss issues of ill-posed problem and the problem of multicollinearity is a special case of a class of problems called ill-posed problems ok. So, we will stop here and see you in the next video where we will discuss the class of ill-posed problem.

Thank you. Bye.