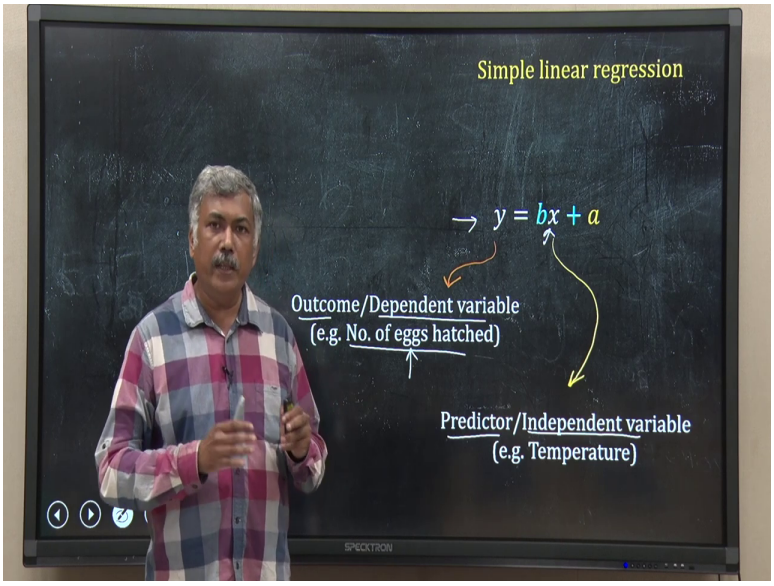
Data Analysis for Biologists
Professor Biplab Bose
Department of Bioscience & Bioengineering
Mehta Family School of Data Science & Artificial Intelligence
Indian Institute of Technology, Guwahati
Lecture 30
Multiple Linear Regression

Hello everyone, welcome back to our course, in this lecture, I will discuss Multiple Linear Regression or in short multiple regression. Earlier we have done linear regression. So, what do we do in linear regression?

(Refer Slide Time: 00:54)

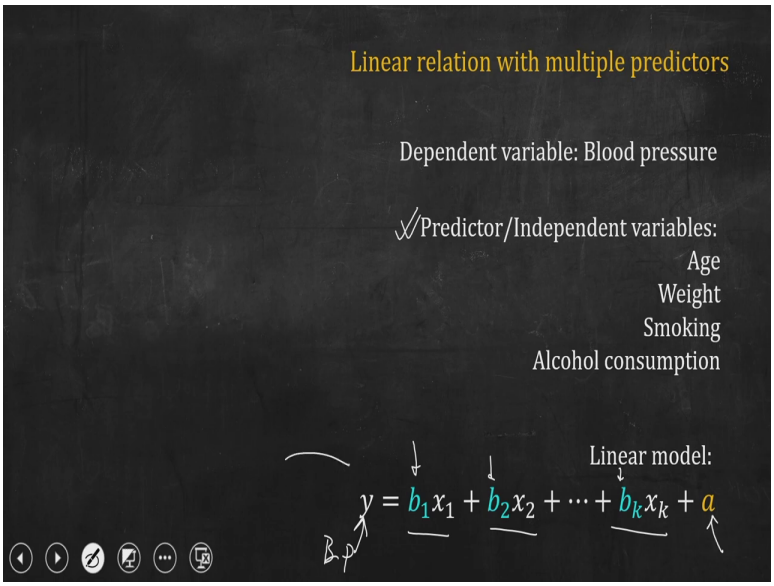


$$y = bx + a$$

So, in linear regression, we have one predictor and one dependent variable. So, and I believe that in between these two we have a linear relation. For example, we may have y equal to bx plus a, where x is the my x is the predictor or independent variable whereas, y is the outcome or the dependent variable and they have a linear relation between them.

For example, suppose you are studying you know hatching of eggs of a particular bird and you may know or you may apprehend that temperature is a factor. So, temperature is a predictor or independent variable whereas, the number of eggs hatched is a outcome or dependent variable and you perform you generate the data then you perform linear regression. Now, we have learned in last three lectures.

(Refer Slide Time: 01:59)



$$y = b_1x_1 + b_2x_2 + \cdots + b_kx_k + a$$

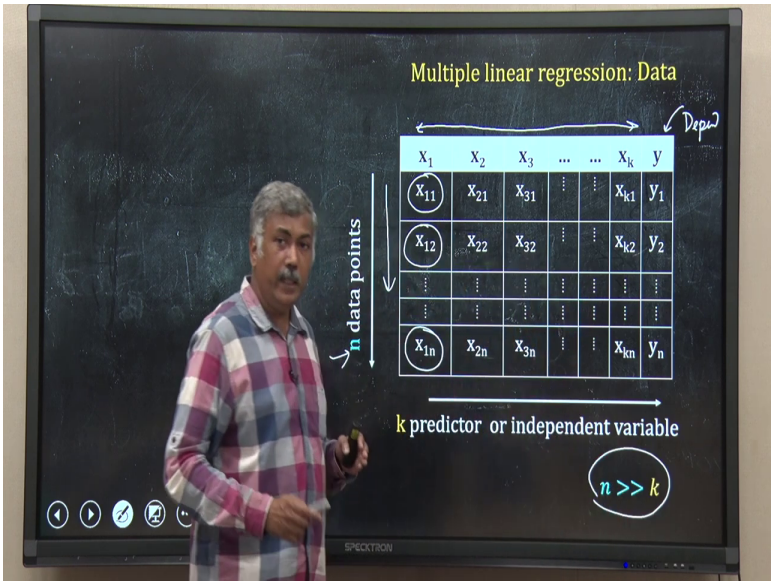
Now, in some cases you will realize that the dependent variable does not depend upon only one predictor or one independent variable. It depends upon multiple independent variables. For example, suppose you are doing some clinical studies and you may be interested in how the blood pressure of those volunteers are getting affected by some other clinical parameter and in that case, the blood pressure is your dependent variable.

Whereas, you may have multiple independent variables or multiple predictor some of them I have listed here for example, you can have age is a predictor, weight of the person may also affect the blood pressure, smoking, how many cigarettes that person smokes, and alcohol consumption can also be another independent variable or predictor. Now, you may believe that these all these four predictor has some sort of linear relation with blood pressure, the dependent variable, then what is our model? Yes, we can create a linear model in this case also. So, in this case y that is the blood pressure.

So, y is equal to b1x1 b2x2 class bkxk this x1, x2 up to xk are my predictors and b1, b2 bk, all these are coefficient or parameters of my model, and at the end, I have inter set a, which is also a coefficient or parameter. So, this is my linear model y equal to b1x1 plus b2x2 plus up to bkxk plus the intercept a. Now, you want to create this model out of a data. So, that is where multiple

linear regression or in short multiple regression comes into work. So, for this type of work, what will be the data format?

(Refer Slide Time: 03:50)



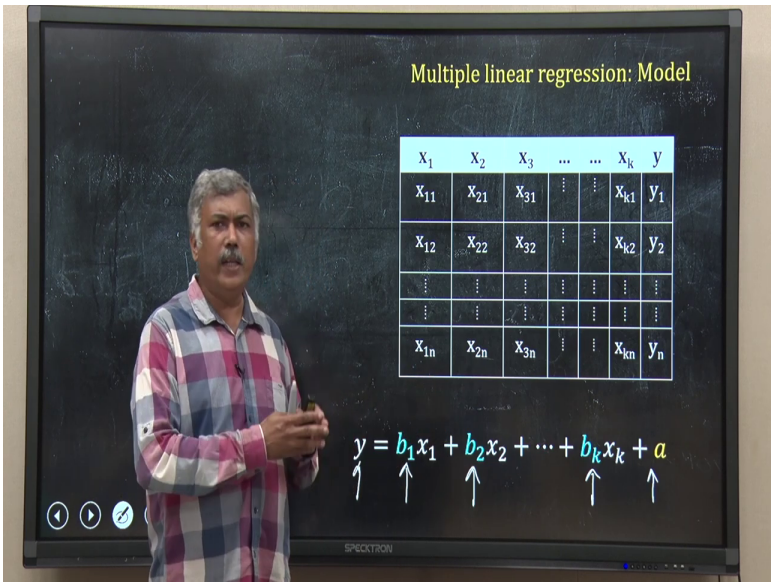
x1	x2	x3	...	xk	y
x11	x21	x31	...	xk1	y1
x12	x22	x32	...	xk2	y2
:	:	:	...	:	:
:	:	:	...	:	:
x1n	x2n	x3n	...	xkn	yn

$n > > k$

Let me, write down the data in a tabular format. So, I have a k predictor starting from x1 up to xk in the example of blood pressure, I have k equal to 4, 4 predictor, so, I am writing generalized way x1, x2, x3 up to xk and the last column is for my y which is the dependent variable dependent variable

And you must have some observation or experimental data. So, all these data suppose I have n data points, so, they are along this. So, they are rows of my table. So, I have x11 is for the first data point x12 is x1 value for the first second data point and at the nth row I have x1 n1 and such that and usually in this case if we are the number of data points should be rather must be much bigger than the number of predictors n must be much bigger than the predictor number of predictor k. So, now, if I have this data, I want to fit a linear model to these data by regression.

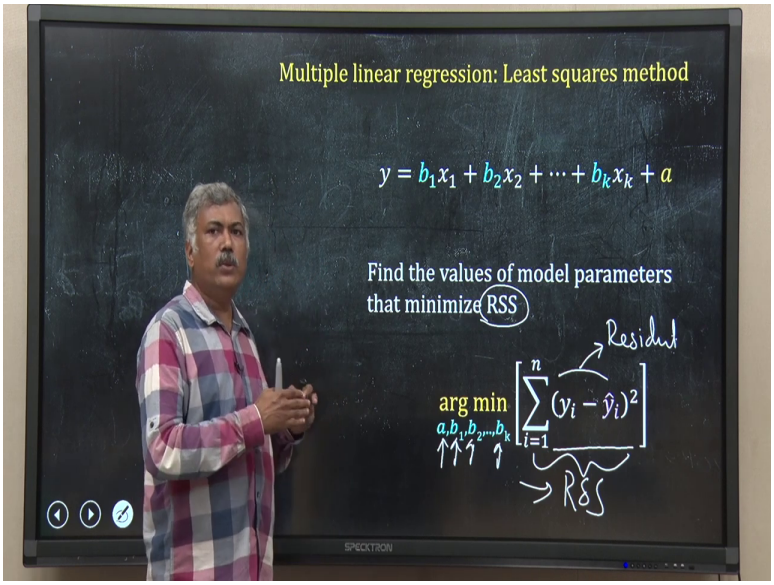
(Refer Slide Time: 04:58)



$$y = b_1x_1 + b_2x_2 + \cdots + b_kx_k + a$$

So, what is my linear model just now, I have shown y is equal to b1x1 b2x2 plus up to bkxk plus a, b1 b2 bk up to bk and a became are the parameter or coefficient of the regression and we have to determine that by fitting this model to the data by regression. How should we proceed it turns out that the mathematics or algorithms that we have used for simple linear regression where we have only one dependent variable one independent variable one predictor based system, we can use the same one exactly for this problem also.

(Refer Slide Time: 05:41)



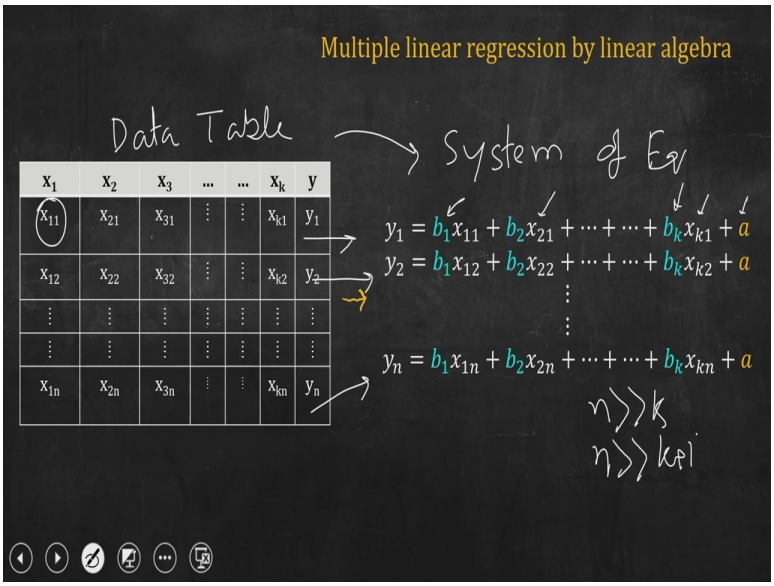
$$\arg arg\left[\sum_{i=1}^{n}\left(y_i - \hat{y_i}\right)^2\right]$$

In fact, I can use the principle of least square where I want to minimize the objective function, what was the objective function for simple linear regression it was RSS residual sum of square, the same RSS is used here, we have to minimize yi minus yi hat squared and summation of those yi minus yi hat is the residual, deviation from the data.

So, yi hat is the calculated or predicted value of y whereas yi is that value from the data. So, you are calculating the deviation squaring that and then summing for all n data point that is your objective function and that is residual sum of square. We have discussed this in when discussing simple linear regression in earlier lectures.

In this case, you are trying to minimize this RSS is by choosing correct value of a b1 b2 and bk. So, that is your goal for least squares method for multiple linear regression. In fact, for simple linear equation I discussed that we have the linear algebra base method. So, here we usually use a linear algebra base method to solve this problem.

(Refer Slide Time: 06:59)



x1	x2	x3	...	xk	y
x11	x21	x31	...	xk1	y1
x12	x22	x32	...	xk2	y2
:	:	:	...	:	:
:	:	:	...	:	:
x1n	x2n	x3n	...	xkn	yn

$$y_1 = b_1x_{11} + b_2x_{21} + \cdots + b_kx_{k1} + a$$

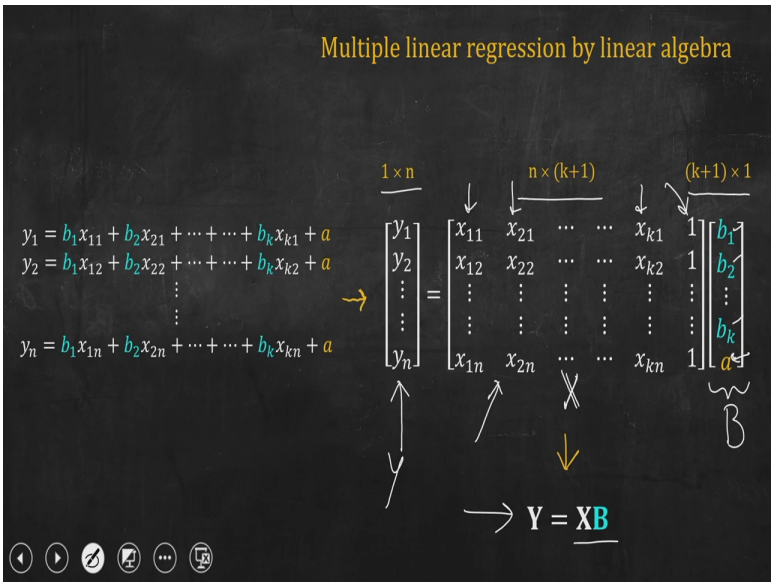
$$y_2 = b_1x_{12} + b_2x_{22} + \cdots + b_kx_{k2} + a$$

$$\vdots$$

$$y_n = b_1x_{1n} + b_2x_{2n} + \cdots + b_kx_{kn} + a$$

So, what we have to do I have the data set this is the data table I represent that as a system of equations, what do I do? I have a linear equation that I have to fit. So, for the first line of my data set, so, the first data point y1 would to be equal to b1 x1 that is the data from here plus b2 into x21 x21 up to bk into xk1 plus a. So, similarly, for the second data point I can write a equation and so, I will have n number of equations and as n is much bigger than k. So, obviously n must be also bigger than k plus 1 and you can easily see this is a over determined system and we discussed that while discussing simple linear regression earlier. So, I have this system of equations generated from the data and using the linear model that we want to fit.

(Refer Slide Time: 08:10)



$$\begin{vmatrix} y1 \\ y2 \\ .. \\ .. \\ yn \end{vmatrix} = \begin{vmatrix} x_{11} & x_{21} & \dots & x_{k1} & 1 \\ x_{12} & x_{22} & \dots & x_{k2} & 1 \\ .. & : & \dots & : & 1 \\ .. & : & \dots & : & 1 \\ x_{1n} & x_{2n} & \dots & x_{kn} & 1 \end{vmatrix} \times \begin{vmatrix} b_1 \\ b_2 \\ .. \\ b_k \\ a \end{vmatrix}$$

$$(1 \times n) = (n \times (k+1)) \times ((k+1) \times 1)$$

$$Y = XB$$

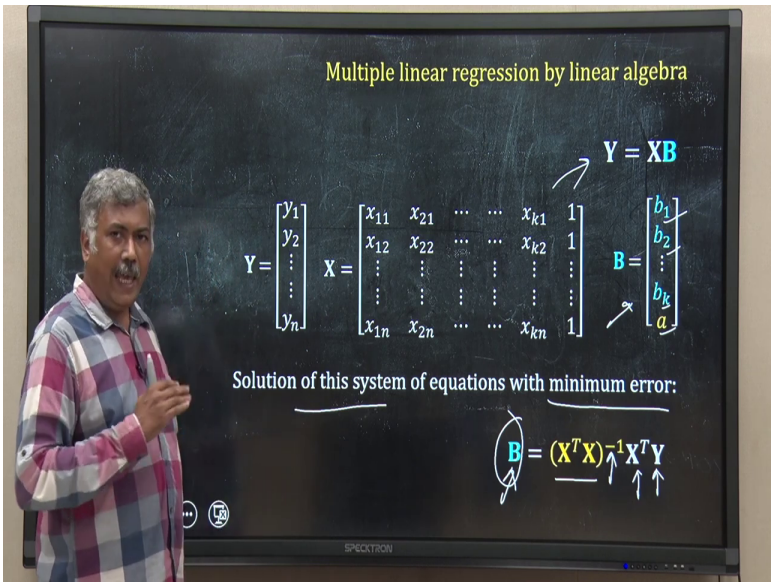
Now, I will represent the system of equation in terms of vectors and matrices the way we did for linear regression earlier. So, what will happen, I will get a vector for y, that will have y1 y2 up to yn stacked together, it is 1 into n vector. And then I will have a matrix here, which will be n into

k plus 1, it will have the x values the predictor and one column for 1 because that is associated with the parameters or the coefficient that I want to estimate from data.

And these matrix should be multiplied by my parameter vector or coefficient vector whose dimension will be k plus 1 because I have k coefficient for k predictor and one constant extra intercept, so, k plus 1 into 1. So, in short, if you remember, these will be represented at y capital Y these will be represented by capital B and these X sorry not B X and this will be represented by capital B.

So, I get a system of equations in a short form, I can write y capital Y is equal to capital X into capital B and I have to get = estimate the value of this B this B vector, I have to estimate from data. So, how should we proceed? We will proceed the exactly the same way we have done for simple linear regression earlier.

(Refer Slide Time: 09:40)



$$B = \left(X^T X\right)^{-1} X^T Y$$

So, I have the system of linear equation and from the concept of projection it can be shown just like earlier we have done then the solution the optimum solution where the error is minimum will be given by B these vector is equal to X transpose X inverse of that. So, we are taking the inverse of X transpose X into X transpose Y. I hope you can identify it from our initial lecture on linear regression, same equation we have.

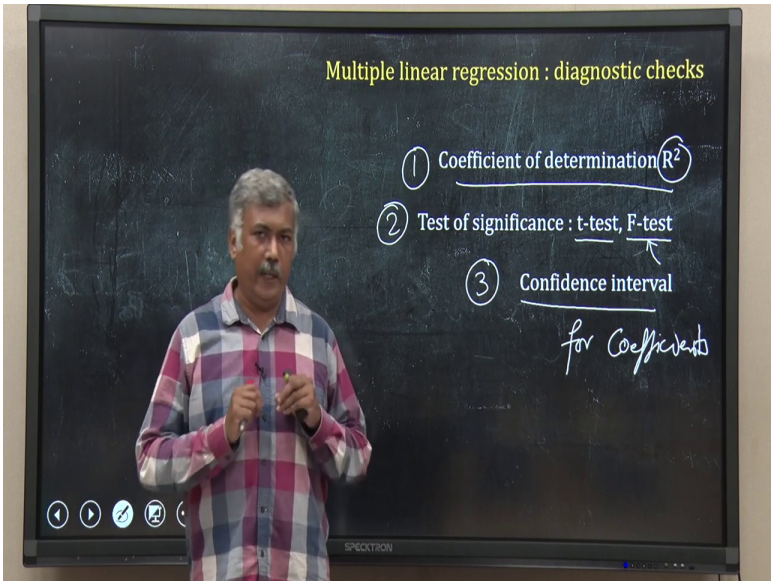
So, now we can solve this and you can manually also do that if the system is simple, otherwise, we can use any other tool like R. And in fact, the lm function the linear model function that we have used for simple linear regression in R can be used in this case exactly with a very simple tweak and it will perform the same operation and it will calculate the value of B.

So, it will give me a vector with b1 b2 bk all and a estimated. So, that is how you perform multiple linear regression, it is same as simple linear regression. The only difference is here is that the earlier linear regression we have discussed which is usually called simple linear regression, you have only 1 predictor in this case, I have more than 1 predictor up to k predictors, and but the relationship between the dependent and independent is still linear.

So, once I have done these and calculated the coefficient, what should be my next job? As we have discussed earlier for a simple linear regression, here also I have to go for diagnostic checks.

So, I have to check how good is my linear model, how good are the estimated coefficients are all these things we have to do.

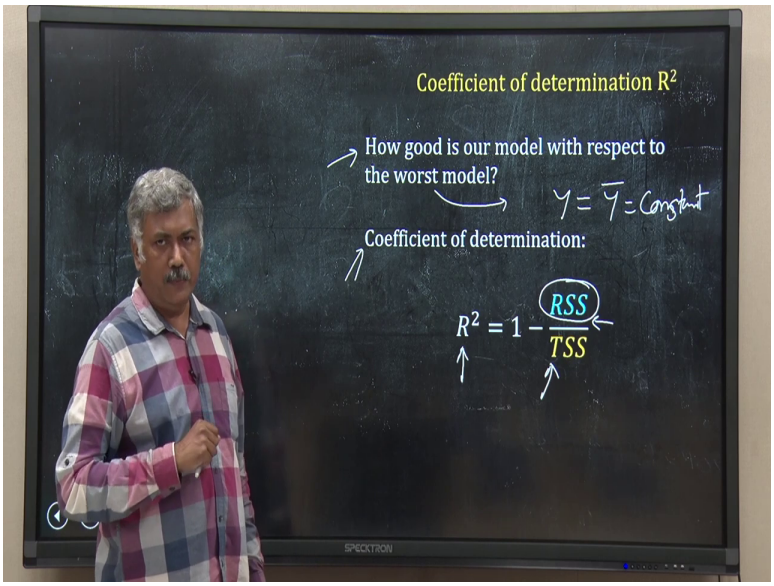
(Refer Slide Time: 11:34)



So, what I will do? The first thing I will check is I will check the coefficient of determination of R square because that compares my model with the worst model if you remember, second thing is that we should go for a test of significance and in multiple linear regression, you can perform two type of t test one is t test one is F test and the third one that you should do is you will check or calculate the confidence interval for the coefficients.

Now, confidence interval estimation and t test will remain same exactly same as you have done for simple linear regression where we have only 1 predictor, so, I will not discuss that in this lecture separately, but there will be some twist in this coefficient of determination R square. So, I will discuss that in detail. And we have not discussed F test earlier. So, I will discuss F test in this lecture..

(Refer Slide Time: 12:47)



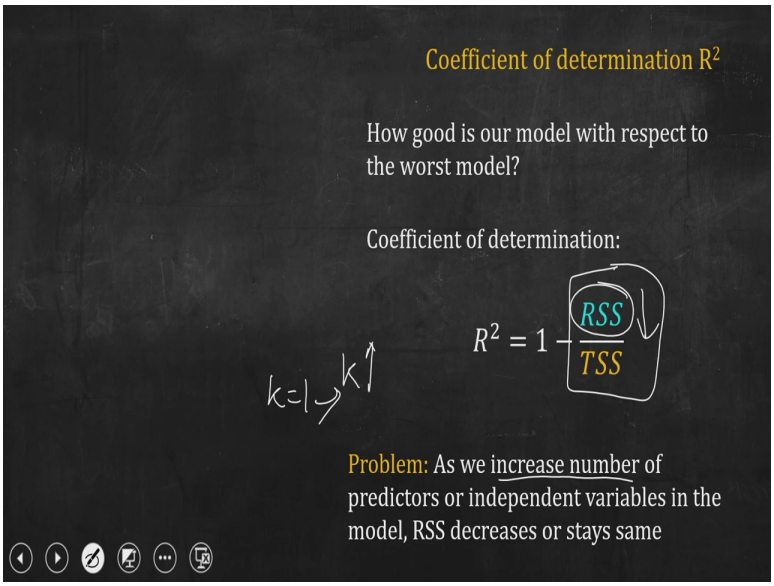
$$R^2 = 1 - \frac{RSS}{TSS}$$

So, let us start with R squared. Just to remind you, what is R square. In R square or coefficient of determination, your testing how good is your model linear model with respect to the worst model and if you remember, in the worst model, y is independent of all the predictor. So, it is equal to a constant, usually take the mean value, constant mean value of y.

This is worst model and then we compare the error between in the error in the linear model versus the error in your worst model. And that is how we define coefficient of determination which is R squared equal to RSS divided by TSS, RSS is the residual sum square error. TSS is the total sum squares of error.

Then, TSS in a way is representing the error in the worst model, whereas, RSS is the error in your regression model linear model and we try to minimize this RSS while doing the square method of regression. So, this is my r square this we have used when we have done simple linear regression and it worked beautifully. Now, but in case of multiple linear regression, we have a trouble with this R square, what is the trouble?

(Refer Slide Time: 14:17)

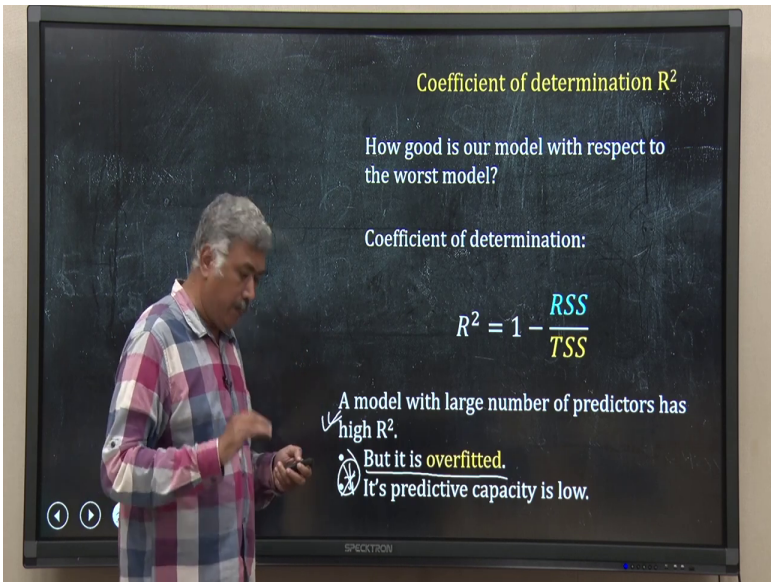


$$R^2 = 1 - \frac{RSS}{TSS}$$

As you increase the number of predictors that is k is increasing, in simple linear regression k is equal to 1, but in multiple linear regression k is more than 1. It can be 2, 3, 4, n up to 100 or something bigger than that also, depends so k is increasing. So, as the predictors is increasing, what will happen RSS, the residual sum of square error will either stay sane or it will keep on decreasing.

That is because we are trying our optimization method of least square method is actually trying to reduce RSS. So, if I add some extra predictor either RSS should decrease or it should stay same as when we do not have those predictors. So, as I keep on increasing the number of predictor in my model, the RSS either stays same or it keeps on dropping. So, if RSS keeps on dropping what will happen, this whole thing will become smaller. So, this thing R square will become bigger and bigger. And as we have learn earlier, the bigger R square is a better R square that means a better model.

(Refer Slide Time: 15:34)

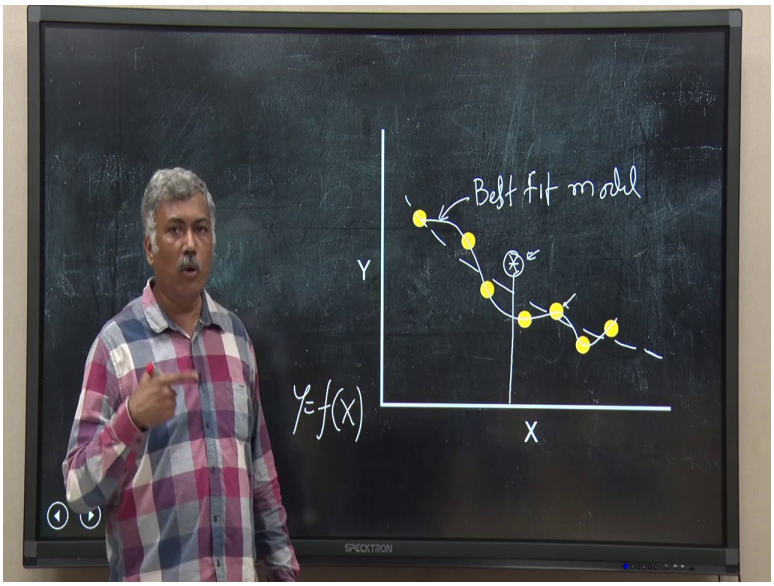


$$R^2 = 1 - \frac{RSS}{TSS}$$

So, this will lead us to a conclusion that as R square is high, we are getting a very good model. But this is a trap, a very high R square model, maybe actually over fitted model that is what I have written. These models may have a very high R square, but it may be over fitted one and the over fitted at model is bad for its predictive capacity. And this is not a problem for multiple linear regression alone.

Anytime, we try to create some sort of model based on observational data, we always get me get trapped into this trap as we increase the number of independent variables are predictors in our model, irrespective of that is linear or nonlinear in all of these things, as you increase the number of predictor or independent variable your model the model may get fitted better to the data, but that is actually overfitting and that you cannot use for generalized purpose. Let me explain this general concept a bit before we move into this particular case and how to sort it out.

(Refer Slide Time: 16:47)



So, suppose this is my data, and suppose, there is another data point which I have not taken into consideration only the yellow dots I am taking into consideration while calculating the model creating the model. So, what can be the best fit model? If you looking to this whole data set including this one, you may imagine that maybe the origins of the real population level model is something like this. Let us guess it maybe something like this.

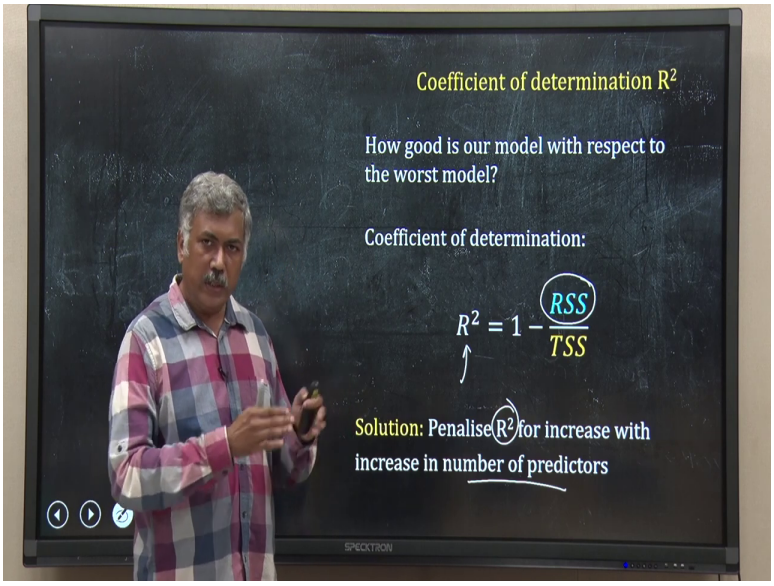
We do not know it, I am just assuming that that may be actually the right model at population level, and our data are dispersed around it, and this yellow data point I have seen. So, now, if I want to forcefully fit a model, some function linear nonlinear, I am not talking about a some function I want to fit where Y is a function of X, I want to fit that in such a way that the RSS should be smallest as small as possible, what can be the smallest value of RSS 0. So, if I fit a curve like this, I take there I start from this one go to the second third one fourth one fifth one sixth one this.

So, this undulating curve has gone exactly through all the data point. So, it is RSS should be as smallest 0. So, then you may think okay, this is a best fit model. So, this is the best fit model. But looking at these you can easily say without knowing what is X and Y that possibly this is a wrong model, although it has fitted very good with respect to the data, but maybe the original model is something like this dotted line or a straight slant line.

This curved line may not be actually realistic model and in fact, may not be most cases possibly this will not be a right model what we have done, we have over fitted that model to the data. So, we have chosen a model which fits very well, but actually may not represent the reality. And in this type of model these over fitted model, if I take any value outside these yellow one for example, if I take this 1, my prediction will go completely wrong.

So, this as I said, this is generalized problem of overfitting when we fit models to data and we keep on increasing model parameters, or rather in a predictor in our model. So, let me go back to the problem of multiple linear regression and where we say that as we keep on increasing predictor R square become high and that gives me over fitted model, which is actually not good, which is not much useful for us.

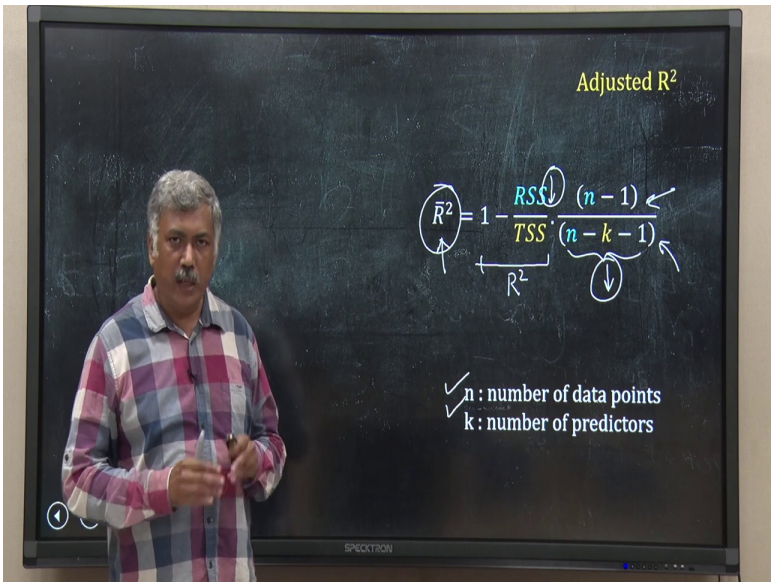
(Refer Slide Time: 19:54)



$$R^2 = 1 - \frac{RSS}{TSS}$$

So, how can I avoid this problem? We have to solve this. So, what do you do. We actually try to modify this R square in such a way that as RSS will drop my R square will get some penalty. That means, as we increase the number of predictor, R square is get penalized. So, that is what we have to do. And that is where the adjusted R squared comes, I will not go into detail derivation how they are getting that.

(Refer Slide Time: 20:29)

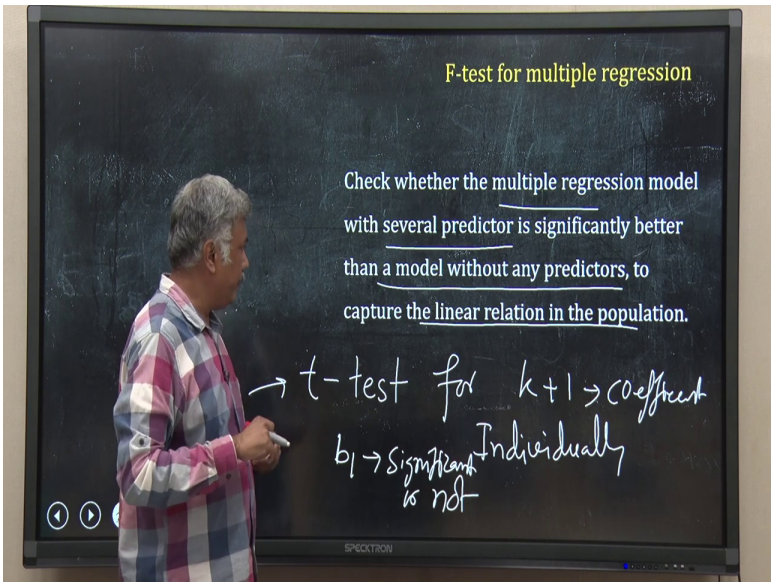


$$R^2 = 1 - \frac{RSS}{TSS} \cdot \frac{(n-1)}{(n-k-1)}$$

Just let look into the definition of adjusted R square and in R you can easily calculate it. So, adjusted R square usually written in R bar squared is 1 minus RSS by TSS, this part is same as the R that R squared we have known earlier into n minus 1 divided by n minus k minus 1, n is the number of data points k is the number a predictor.

So, notice what will happen if I keep on increasing k as k increases, this part decreases, at the same time RSS also decreases. So, now, if this get decreased, and these also get decreased, so I have decreased both in the denominator and numerator. So, in a way, they cancel each other effect. So, I get an adjusted value of R square and we call it adjusted R square. So, if you are doing multiple linear regression, look for adjusted R squared not just simple R square value.

(Refer Slide Time: 21:34)



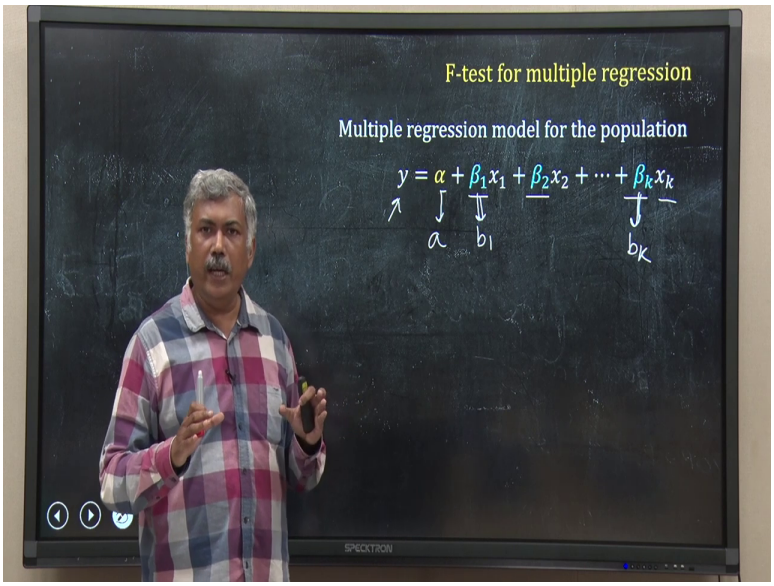
Now, come the question of statistical test. We have discussed in a separate lecture, what is statistical test of significance and then we have discussed t test in simple linear regression, what were we doing in t test for simple linear regression? We have fitted the data or model to a sample data of a sample. And now we want to know whether this model with one predictor and one constant term y equal to nx plus c for example, is a valid representation is a correct representation of the model in the population level that is why you have to do a statistical test.

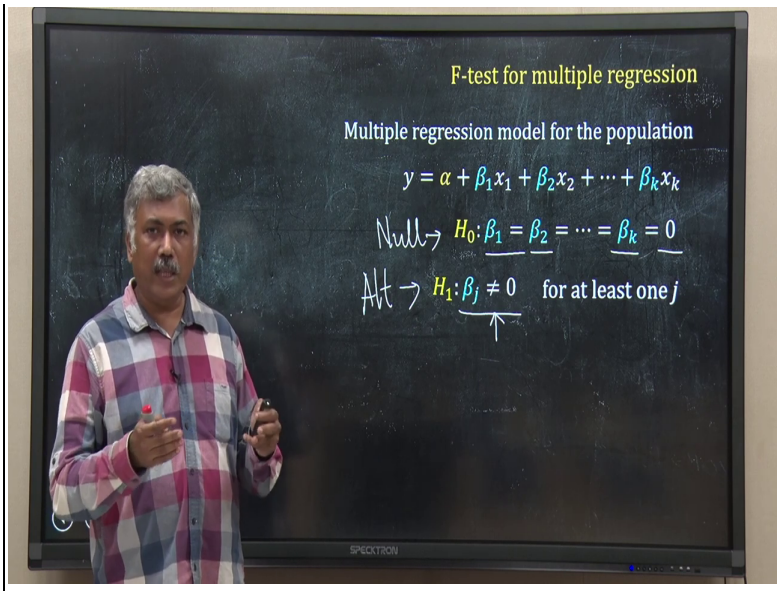
And in that case, you do t test and you actually check the significance of each one not each predictor, because simple linear regression has only one predictor in that case, what you do, you are trying to check the significance of one predictor and its coefficient. So, you have y equal to nx plus c you calculate whether n significant or it is equal to 0, that is what you do by t test. Now, in case of multiple linear regression, you have k coefficient for k predictors and I have a additional a term for the intercept constant. So, for each of these coefficient each of these k plus 1 coefficient I can do t test. So, I can do t test for all those k plus 1 coefficient individually.

So, I can check whether our suppose the first coefficient b1 is significant or not. Similarly, I can check for the second coefficient third b and all up to bk. So, that is t test, it will remain same as you have done for simple linear regression. But in this lecture, I will discuss something which is call F test, and why are you doing F test?

In F test, what you are trying to check is that whether the multiple regression model with several predictor is significantly better than a model without any predictors that means, there is no x term in the model to capture the linear relation in the population. We can remember it is not the issue with respect to sample data, it is respect to the what is the real model in the population. The population that you have not seen, but just you have a sample data. So, just like any other statistical test of significance will take the same steps. What will be the first one?

(Refer Slide Time: 24:38)





$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

$$H_{\circ}: \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$H_1: \beta_j \neq 0 \; for \; at \; least \; one \; j$$

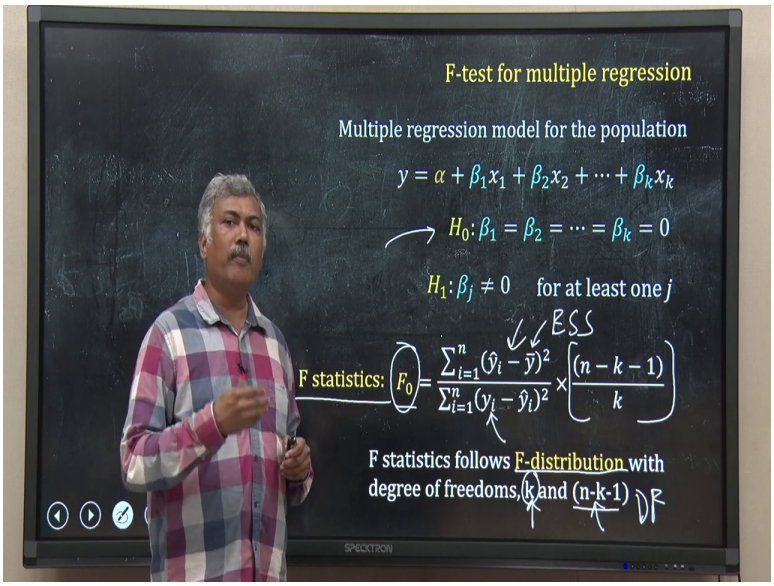
The first one is following the model that we have fitted to our sample data, the data that I have collected, I create a linear model for the population. So, that is equal to y equal to alpha plus beta 1 x1, beta 2 x2 plus upto Bk x beta 2 beta k xk… What is beta 1? Beta 1 is equivalent to b1 in multiple linear regression model. Alpha is equivalent to a in my model, same way, beta k is equivalent to bk in my linear regression model. The only difference is alpha beta these are for the population level model. So, there is a population parameter whereas, a b1 bk b2 these are my samples parameter sample model parameters.

Now, once you have this model, you have to create null hypothesis and alternate hypothesis. So, what is the null hypothesis in this case? Remember in this case you are creating comparing your linear model with respect to your model where is no predictor. So, that means, my null hypothesis says bu beta 1 beta 2 beta k all of them are equal to 0 that is my null hypothesis H naught.

What is my alternative hypothesis? This is my alternate hypothesis H1. In that case, I believe at least one of those beta 1 up to beta k, at least one of those is not equal to 0. That means, my model has some predictor it is it is not without any predictor. So, I have these 2 model null model and the alternate model and I want to, I believe, actually the alternate model is correct, but I am checking it respect to the null model.

So, this is the second step of any statistical test of significance. The third step for statistical tests of significance is to create test statistics, in t test, you create the t value or t statistics you define and calculate that for null hypothesis. Here in F test, what I will do I will define F statistics and I will calculate that F statistics or F value for the null hypothesis.

(Refer Slide Time: 27:05)



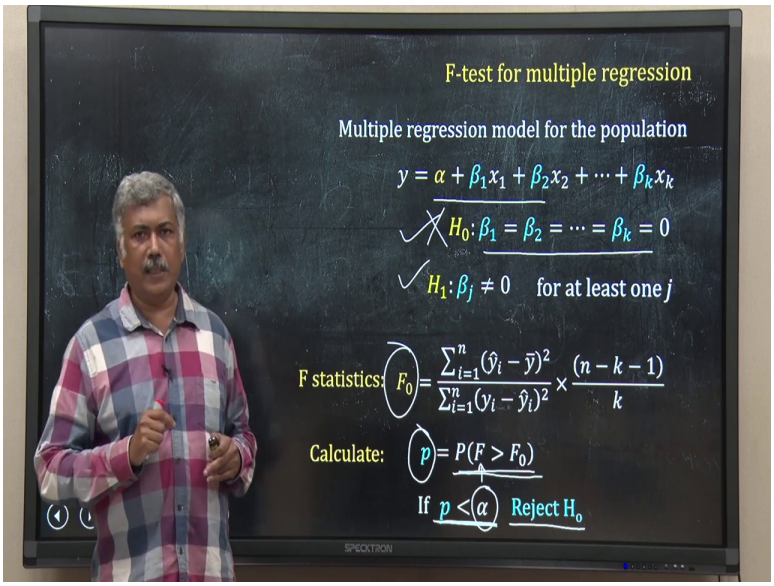
$$F\ statistics:\ F_{\circ} = \frac{\sum_{i=1}^{n}\left(\hat{y_i} - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \hat{y_i}\right)^2} \times \frac{(n-k-1)}{k}$$

So, for the null hypothesis that I have considered here the F statistics F naught not is actually defined I will not go in detail into that is defined as the ratio of 2 errors, see, this is also error that is yi hat is the value predicted for yi and y bar is the mean value of y in your data. So, that is actually error deviation of the predicted value from the mean square of that and some of it and if you remember this is actually explained sum of square ESS. And what you have in the denominator that is also error that is yi minus yi hat whole square.

So, essentially it is a residual sum of squares is not it that is what we have here. Apart from that this ratio is multiplied by some values n minus k minus 1 divided by k these are associated with degree of freedom of certain thing that we are dealing here in this model, will not go in detail of that. So, we have this definition of F naught the F value for or F statistics for my null hypothesis.

Now, it can be shown that these F statistics actually follow a probability distribution function called F distribution is a 2 parameter distribution, and those 2 parameters are called degree of freedom and in this case, the degree of freedom DF will be k and n minus k minus 1. So, I have defined a F statistic for the null hypothesis and it follow F distribution.

(Refer Slide Time: 28:51)



$$F\ statistics:\ F_{\circ} = \frac{\sum_{i=1}^{n}\left(\hat{y_i} - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \hat{y_i}\right)^2} \times \frac{(n-k-1)}{k}$$

$$Calculate:\ p = P\left(F \geq F_o\right)$$

$$If\ p < \alpha \Longrightarrow Reject\ H_{\circ}$$

Now, come my fourth step. So, I have the experimental data, I have treated the model and I have the null hypothesis alternate hypothesis, I calculate a numerical value of F naught, The F statistic have statistics for null hypothesis, and then I looking to the F distribution. I know the PDF of that, and from that I calculate what is the probability of getting an F value bigger than F naught. Probability of getting F value F statistics bigger than F naught and that is my P value.

If the P value is less than a threshold, remember, if the P value is very small, I will reject the null hypothesis, but how small is small to decide that I need a threshold, that threshold come from the level of significance. So, that is alpha. So, if P is less than that threshold, I will reject null hypothesis. So, I will say H naught is null hypothesis is wrong. If null hypothesis is wrong that means they are not equal to 0 at least one of them is important here in this equation that means at least 1 of the predictor is required in the population level model, you cannot remove all of them

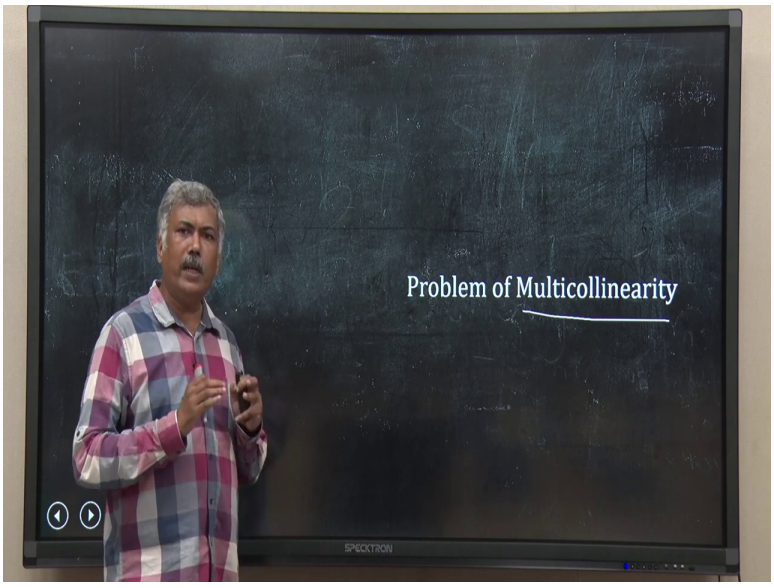
So, what you are doing here? In t test you can actually test the significance of the each of the predictor and its associated coefficient. If a predictor is not important is not significant then the t

test for its coefficient will fail. So, in that case what will happen you will reject we will simply removed that predictor.

So, in that case, you are actually testing for each of the predictor. Whereas, for F test you are comparing the whole model with respect to a model null model you are comparing all the predictors together with respect to a model where there is no predictor at all. So, you are comparing the as the whole behaviour as a whole effect of all the predictor to the predicted one or the dependent variable.

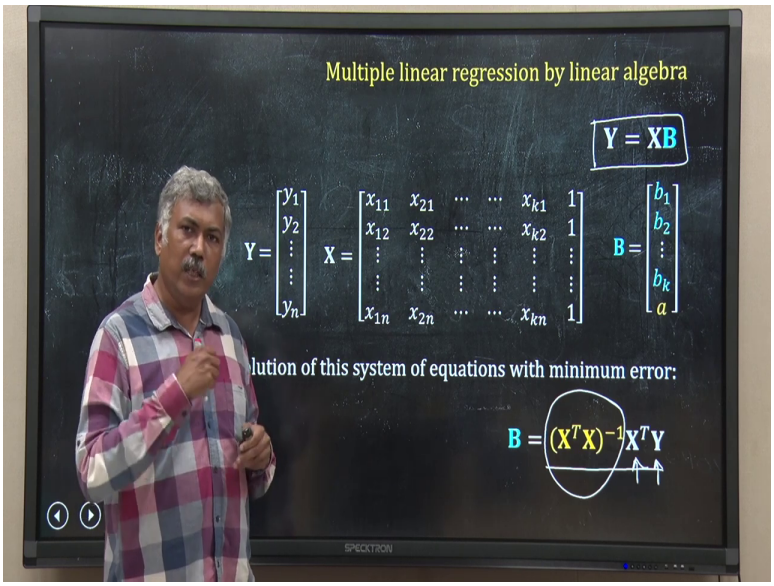
So, this is F test. So, what we have learned till now, we have learned about how to do multiple linear regression. And then I have discussed that you can do multiple linear regression the way we do simple linear regression using the same function in R. Once you have done that, you do some diagnostic check like this F test, adjusted R square t test and confidence intervals. So, that is what completes usually your multiple linear regression problem. But there is a unique problem associated with multiple linear regression that we have not faced in case of simple linear regression. We have to be very careful about that problem.

(Refer Slide Time: 31:52)



And let me discuss that this problem is called the problem of multicollinearity. What is that?

(Refer Slide Time: 32:02)



$$Y = \begin{vmatrix} y1 \\ y2 \\ .. \\ .. \\ yn \end{vmatrix}$$

$$X = \begin{vmatrix} x_{11} & x_{21} & \dots & x_{k1} & 1 \\ x_{12} & x_{22} & \dots & x_{k2} & 1 \\ .. & : & \dots & : & 1 \\ .. & : & \dots & : & 1 \\ x_{1n} & x_{2n} & \dots & x_{kn} & 1 \end{vmatrix}$$

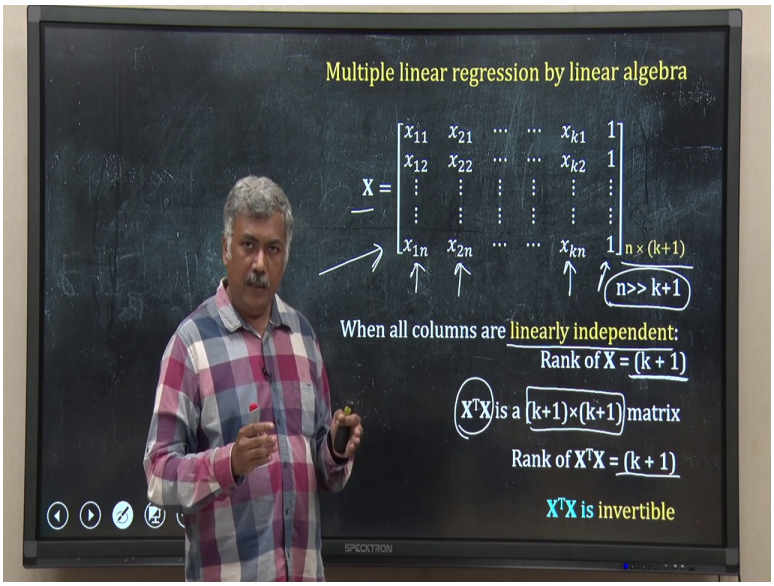
$$B = \begin{vmatrix} b_1 \\ b_2 \\ .. \\ b_k \\ a \end{vmatrix}$$

$$B = \left(X^T X\right)^{-1} X^T Y$$

To understand the problem of multicollinearity let us look back what we are doing in case of multiple linear regression. I have represented the problem as a system of equations using these vector matrix notation Y equal to X into B, and I have to calculate the B, and how I am calculating the B? B is obtained by this relationship a inverse of X transpose X into X transpose Y.

So, I have an inversion of a matrix X transpose X is a square matrix and you are inverting it and that is where the problem of multicollinearity comes this X transpose X has to be invertible. All four matrices are not invertible. If this is not invertible, then you cannot solve this problem. So, let us look into that situation when we will have this problem that X transpose X is not invertible.

(Refer Slide Time: 32:56)



$$X = \begin{vmatrix} x_{11} & x_{21} & \dots & x_{k1} & 1 \\ x_{12} & x_{22} & \dots & x_{k2} & 1 \\ .. & : & \dots & : & 1 \\ .. & : & \dots & : & 1 \\ x_{1n} & x_{2n} & \dots & x_{kn} & 1 \end{vmatrix}$$

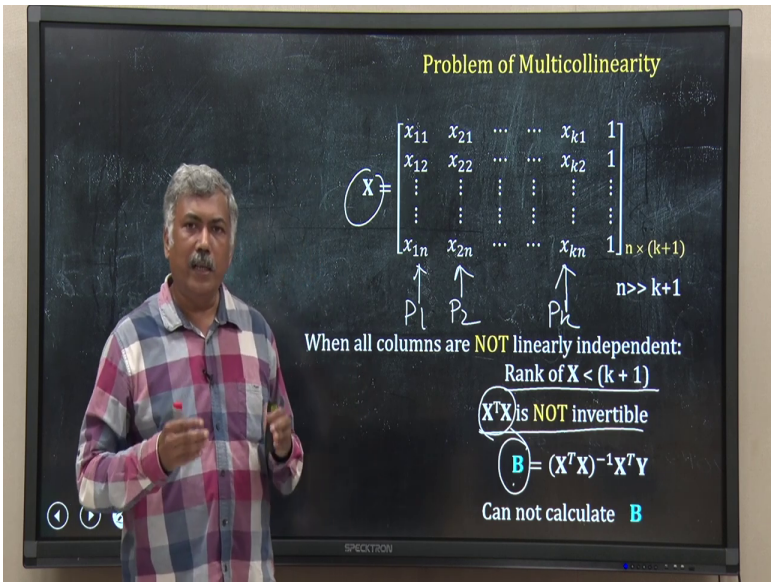
So, let us this is my X data set represented as a matrix. So, I have capital X in my system of equation is the column for the first predictor column for the second predictor column for my kth predictor, and 1 extra column of 1 for the coefficient. This is my X matrix what is the dimension is n into n by k plus 1. Usually, as we said to do linear regression we need n much bigger than k plus 1.

Now, if this k plus 1 columns of this matrix X are linearly independent, then what is the rank of this matrix? The rank of this matrix will be equal to number of column that is k plus 1. So, when these columns of this matrix are linearly independent that means, I cannot represent any of this column by combination of any other column 1 or more columns.

So, in that case the rank of this matrix will be k plus 1. If the rank of X is k plus 1 then you can easily check the rank of X transpose X will be also k plus 1. So, it is a full rank matrix number of columns and rows in the X transpose X matrix is k plus 1 by k plus 1 and his rank is also k plus

1. So, it is a full rank matrix that means, it is invertible and that means, I can calculate B. no problem.

(Refer Slide Time: 34:40)



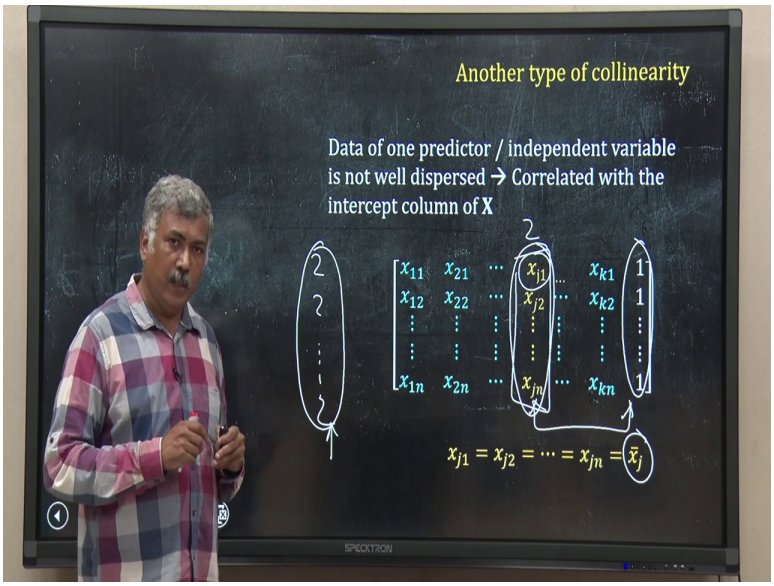
$$B = \left(X^T X\right)^{-1} X^T Y$$

But suppose somehow any of these columns in my X matrix are not is not linearly independent. That means, suppose there is at least 1 column which I can represent as a combination of any other columns or as a multiple of some other column in this matrix that means, this column is not independent it depends on other column and then the rank of X will be less than k plus 1 number of column in the matrix and in that case, X transpose X also have less rank is not full rank anymore and it is not invertible.

So, when these columns in my X matrix are not linearly independent that means, these predictor P1 P2 Pk , these are predictor represented by x1 x2 xk. When these predictors are not linearly independent, then I have a problem my X transpose X is not invertible and I cannot calculate B. So, this is the problem of multicollinearity.

You may have design experiment and have collected data where your predictors that you have chosen to create the model may have dependency among themselves linear dependency among themselves. In that case, your X will have columns, which are not linearly independent there is a dependence there and I will not be able to calculate B. So, this is the called the problem of multicollinearity.

(Refer Slide Time: 36:21)



$$\begin{vmatrix} x11 & x21 & \dots & xj1 & \dots & xk1 & 1 \\ x12 & x22 & \dots & xj2 & \dots & xk2 & 1 \\ : & : & \dots & : & \dots & : & 1 \\ : & : & \dots & : & \dots & : & 1 \\ x1n & x2n & \dots & xjn & \dots & xkn & 1 \end{vmatrix}$$

$$x_{j1} = x_{jn} = \cdots = x_{jn} = \bar{x}_j$$

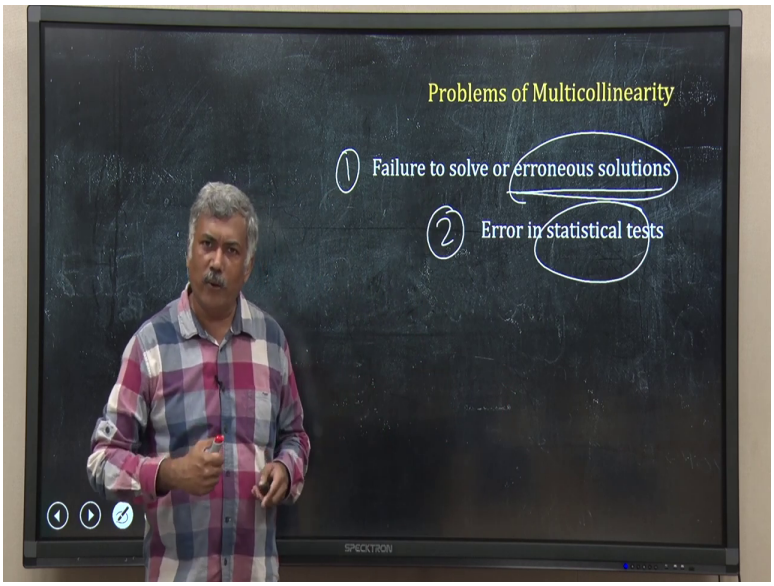
Multicollinearity can happen by another means also. Suppose, I have chosen the data point in such a way for this predictor, the data value xj1 xj2 the j predictor are very close to each other, or suppose exactly same. Suppose all of them is 2 2 2. That is how you have collected the data. That is how you design that experiment. And that is how you collected data. Now, if all of them is equal to 2 for example, I have shown it as xj bar, then see this column the column of 2 can be represented as a multiple of this column. Because that last column has 1.

So, I can multiply with any scalar value 2 for example, and I can get the other column. So, if any column in this matrix for any predictor your data is not well dispersed, it is almost same or exactly same for all the data points, then again, you will face this problem of multicollinearity because then these 2 are dependent on each other, they are not independent and then the X transpose X will not be invertible and you will not be able to calculate B.

So, now, this is actually an extreme case where I have what we call exact correlation or exact linear dependence between 1 or more columns in your matrix X. You may not have exact relation

you may not have exact depend linear dependency, but the dependency can be very strong also, even then, I will face trouble.

(Refer Slide Time: 38:02)



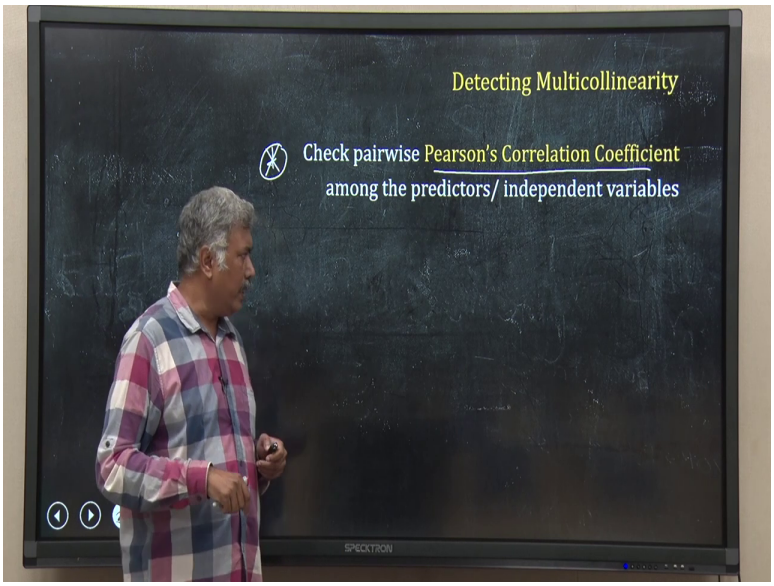
Actually, you will face 2 type of trouble. The trouble first trouble is either you will fail to solve because you have X transpose X is not invertible or it is inevitable, but there is no exact multicollinearity but there is strong multicollinearity and that leads to error in your calculation. So, you will have error in your solutions.

And the second problem is that when this situation arise, actually the statistical tests start falling failing. I will not go in detail of that just to tell you briefly what happens is that for the statistical tests for t test or F test, you have to calculate standard error for each of the coefficients a, b1 b2 bk, you have to calculate the standard errors for those that means, you have to calculate the variance of those.

And when you have these multicollinearity these variances are there is a huge error in these variance calculations. So, if you have error in your variance calculation, then there is error in your standard error calculation and then obviously, your statistical test goes wrong. So, when you have multicollinearity problem, you will face these two problems.

Now, we know this problem is unique for multiple linear regression, but how do I detect that my model has this problem, because if you are using R, R may have done the solution and you may have got a solution which is erroneous and it has done all the statistical tests that is also wrong suppose, but how will we know that these are wrong? Because you do not know the real model.

(Refer Slide Time: 39:37)

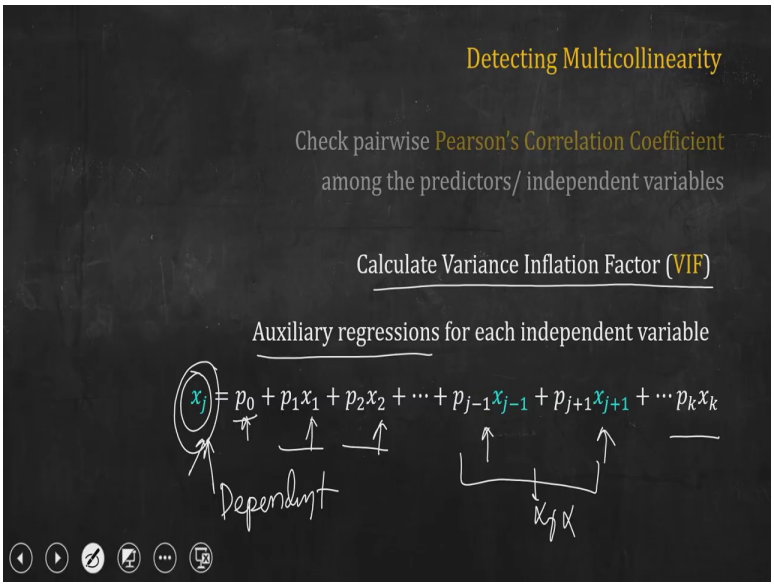


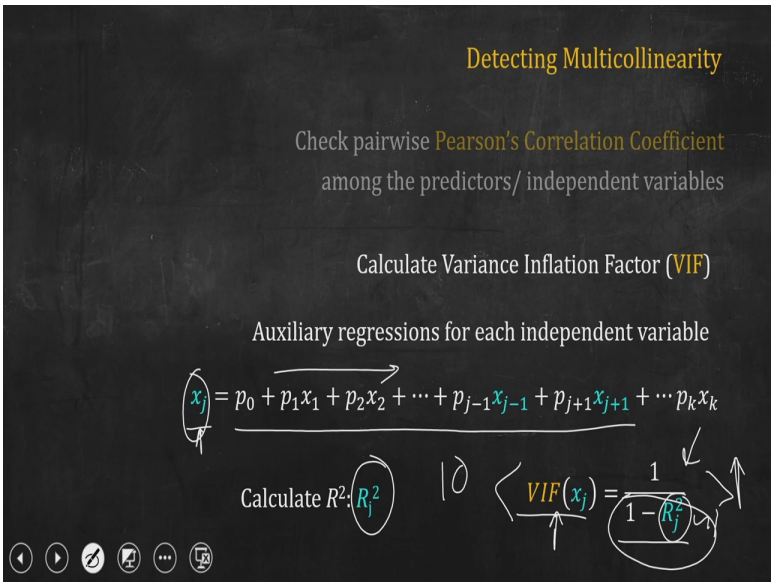
Now, to dictate the multicollinearity problem in your model, what we can do? We can do some diagnostic checks. The first one the simplest one you can check actually, you can look into the data for each of those predictor the x values and may calculate the Pearson correlation coefficient between any two pair.

So, you take x1 data of x1 and data of x2 and calculate the Pearson correlation coefficient. If the correlation coefficient is very high that means x1 and x2 has some sort of linear relation. So, in this way you calculate the correlation coefficient for all possible pairs of these predictors or independent variables.

If any of these pair have high value of correlation coefficient, then you can assume that okay they may have some for linear dependence and you can look into it and you may sort out the problem. But this is always pairwise. But the linear dependency between columns and x may not be pairwise 1 column may depend upon a combination of multiple other columns.

(Refer Slide Time: 40:43)





$$x_j = p_o + p_1x_1 + p_2x_2 + \cdots + p_{j-1}x_{j-1} + p_{j+1}x_{j+1} + \cdots + b_kx_k$$

$$VIF(x_j) = \frac{1}{1-R_j^2}$$

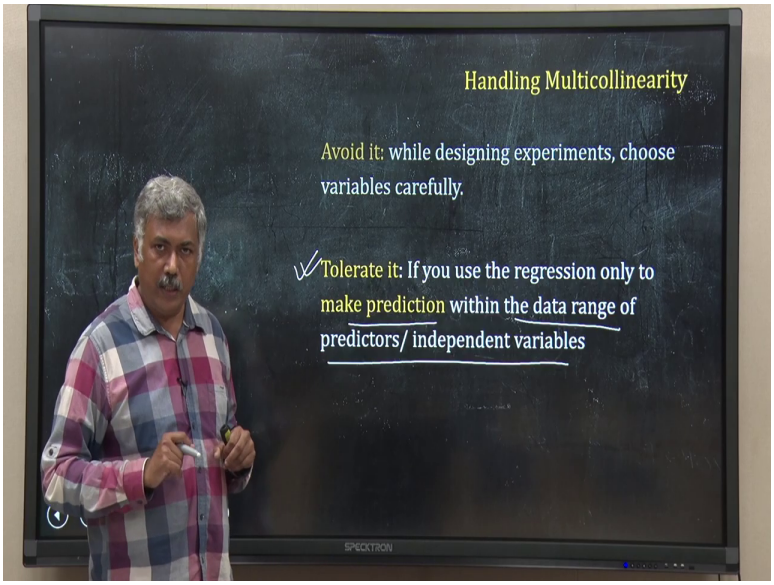
So, in that case actually VIF a particular diagnostic check comes very handy. So, we will calculate the variance inflation factor, what is variance inflation factor? See in multiple linear regression what we are doing we have a independent dependent variable y and we have set of x which are predictors.

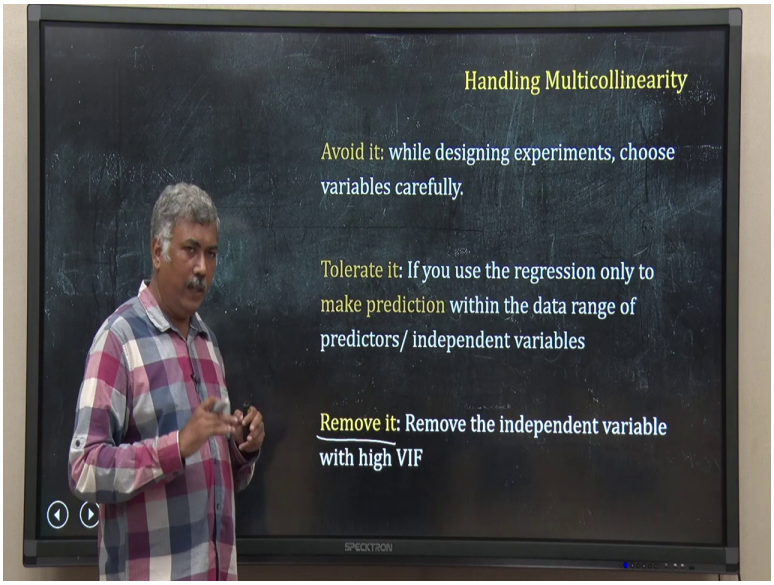
Now, you do multiple linear regression, but you remove y you take one of those predictor suppose xj and leave rest of the predictor and create a linear model. So, what do we have, xj equal to p naught a constant plus p1 x1 plus p2 x2 up to pk xk, but notice here I do not have xj here, because I have taken xj on this side. So, that mean in this case xj is a dependent variable and rest of the predictor or independent variable.

So, what do you do you perform multiple regression on this. So, for all the predictors starting from x1 to xk, you create this type of model and perform independent multiple linear regression these are called auxiliary regressions. And for each of this regression, you calculate the R square for example, for xj a predictor xj, maybe I have done the regression and I have got the R squared value that is called Rj square. Now, VIF Variance Inflation Factor for xj is defined as 1 divided by 1 minus Rj square. So, notice this if xj has high linear dependence on some of the some of the other predictor on the side of my equation Rj square will be high close to 1.

So, if it is close to 1, if this one is close to 1, then what will happen this denominator will become a reach towards 0 that means, VIF will grow. So, if xj a predictor depend upon one or more predictor, then it is VIF will be very high, high mean how much high? Usually the thumb rule is, if VIF is greater than 10, then I consider usually that xj has linear dependence on one or more other independent predictors in my model. In this way, you can calculate VIF for all the predictor and identify the linear relationship linear dependency between any of these predictor with other predictors.

(Refer Slide Time: 43:34)





Now, suppose you have done that, what we will do? You have created a model and you have found that there are one or more predictor features linear dependency on other that means, you have multicollinearity problem, what will you do? Honestly speaking, the best way is to avoid it from the very beginning, much before you do linear multiple linear regression, when you are designing your experiment you are creating your hypothesis design the experiment in such a way that you know, the independent variables or the predictors that you are taking and measuring are not dependent upon each other.

For example, in some cases as we have discussed earlier in the lung cancer example, it has been observed that the people who smoke they also drink more many a time. So, there is a strong

correlation or dependency between these two variables. So, if you have in some experiment, you have already taken smokers, there is a smoking habit in consideration as a independent variable, you may leave alcoholism from that, because they may already have some correlation, they may have some linear dependence already.

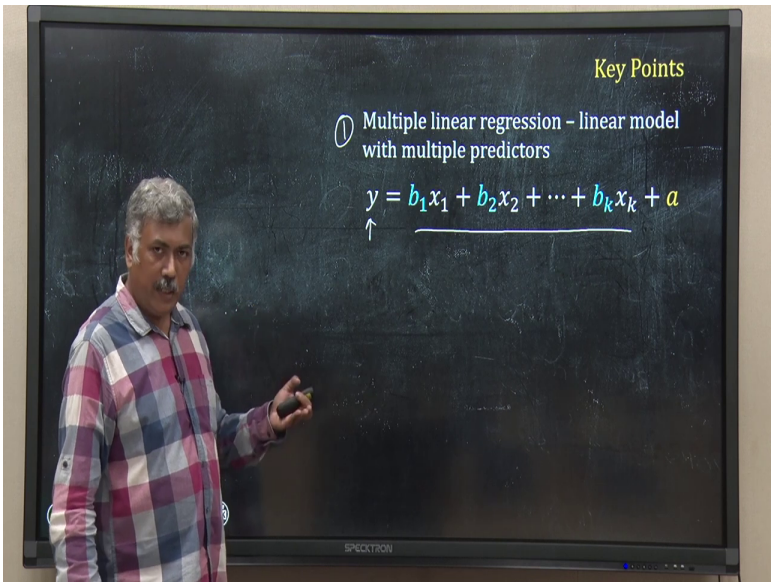
So, this is I am just giving a rough example. So, for your problem, you have to think through before you go for data collection and do the experiment design the experiment in such a way that you know, the predictors that you have taken does not have any dependency among themselves. This is the best way to handle this problem, you will not face the problem at all when doing the regression. But suppose you have already done the experiment what you will do, you cannot change it anymore.

So, in that case, you have to think over again what is the utility of your model. In many cases, actually, you can simply tolerate these multicollinearity. For example, you are using this model this equation just to make some prediction. It is fine you can in most cases tolerate it, but suppose you are creating this model and then you want to decipher some causal relation. So, you have 10 predictors in your model and a disease outcome is connected to that predictor as per your linear model. And now, you want to explain that okay, this predictor actually causes the disease.

So, you are not actually making any numerical prediction here you want to make some causal relation between these two. So, in that case, if you have taken a situation where you have linear dependency between these predictors, you may make wrong conclusion. So, if that is the case, you cannot tolerate it, but in general, if you are just making a prediction within the data range, which you have used to calculate the do the regression, you may safely use it. So, in that case, tolerate these multicollinearity.

The other one, which is the extreme 1, that okay, you cannot tolerate it. And you may have to make some causal inference from the model, then what you have to do, if you have identified some variable, the predictor as dependent on other, you should simply remove them. That is how you handle multicollinearity which is unique for multiple linear regression problem not present in simple linear regression. So, that is all for this lecture.

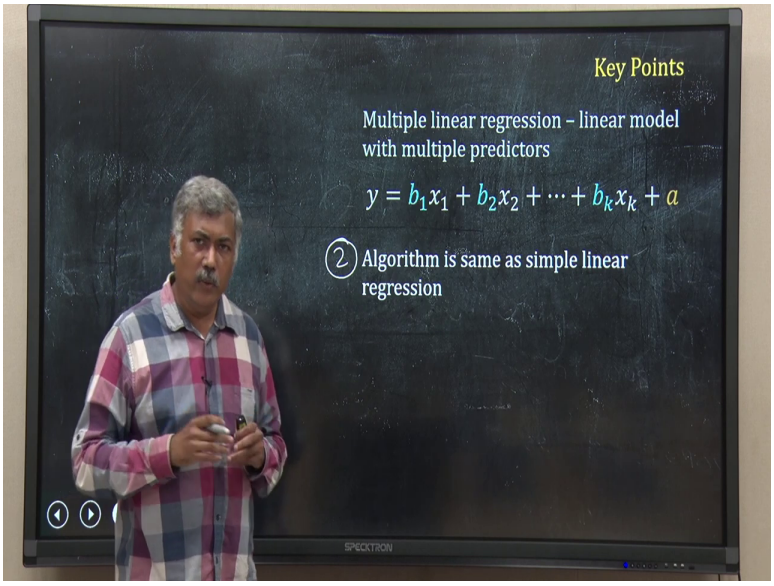
(Refer Slide Time: 47:01)



$$y = b_1x_1 + b_2x_2^2 + \cdots + b_kx_k^k + a$$

Let me jot down what we have learned. In this lecture, we have learned a multiple linear regression where you have more than one predictor or independent variable and we have one dependent variable, which has a linear relation with these predictors and you have a data set you want to fit this model to the data set and you use the linear regression for that.

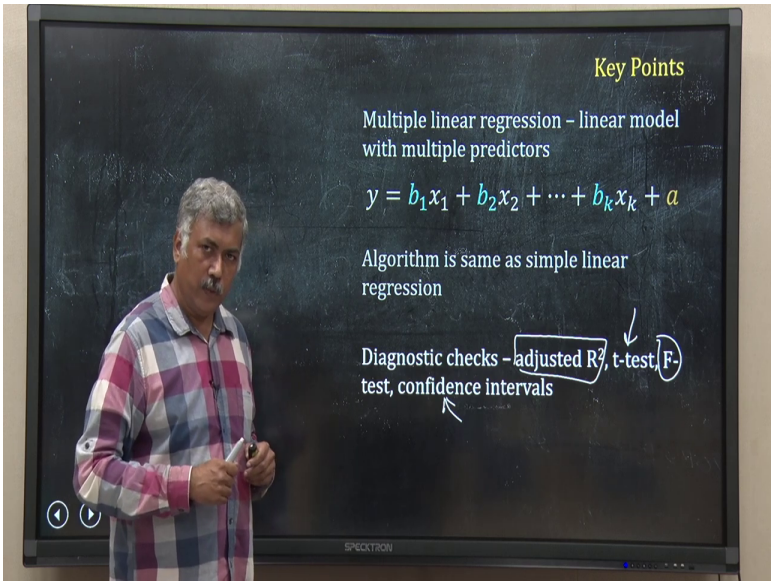
(Refer Slide Time: 47:24)



$$y = b_1x_1 + b_2x_2^2 + \cdots + b_kx_k^k + a$$

Now, the algorithm of least squares method that we have used for simple linear regression, we can use that as it is for our multiple linear regression also.

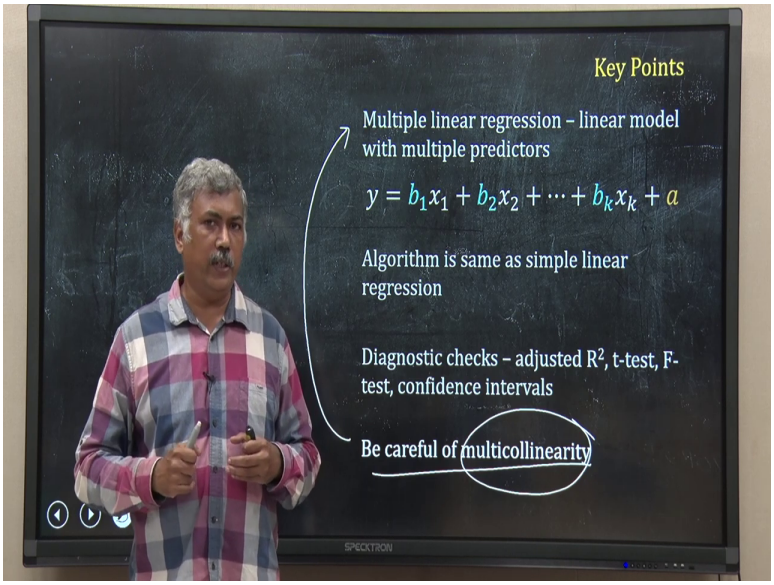
(Refer Slide Time: 47:40)



$$y = b_1x_1 + b_2x_2^2 + \cdots + b_kx_k^k + a$$

Third, who you heard you have discussed we have discussed about the diagnostic checks. For example, you we have discussed adjusted R square, we have discussed F test, t test and confidence interval we have discussed for simple linear regression, and those are applicable for multiple linear regression also.

(Refer Slide Time: 47:57)



$$y = b_1x_1 + b_2x_2^2 + \cdots + b_kx_k^k + a$$

And at the end, I have discussed about the problem of multicollinearity. Why does it arise, and what it can affect how it can affect our model and how to handle it. But just to remind you, actually, we should always think of it before the first point, before you design a experiment where you have a hunch that you may go for multiple linear regression in future, think through the experiment, identify all the predictors or independent variables which may have a relation among themselves, separate them out, you will never face the problem of multicollinearity. That is all for this lecture. In the next lecture, I will discuss how to use R to perform multiple linear regression for a data set. So, see you in this lecture. Till then happy learning.