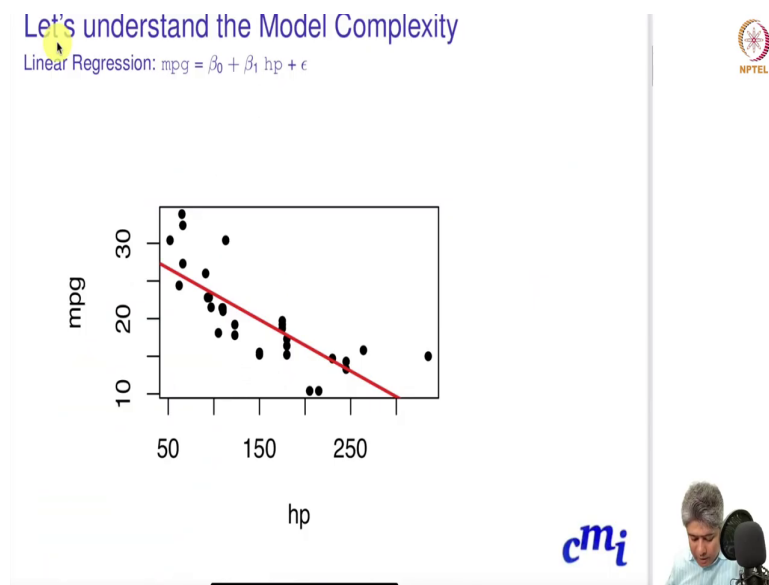


**Predictive Analytics - Regression and Classification**  
**Prof. Sourish Das**  
**Department of Mathematics**  
**Chennai Mathematical Institute**

**Lecture - 20**  
**Model Complexity, Bias and Variance Tradeoff**

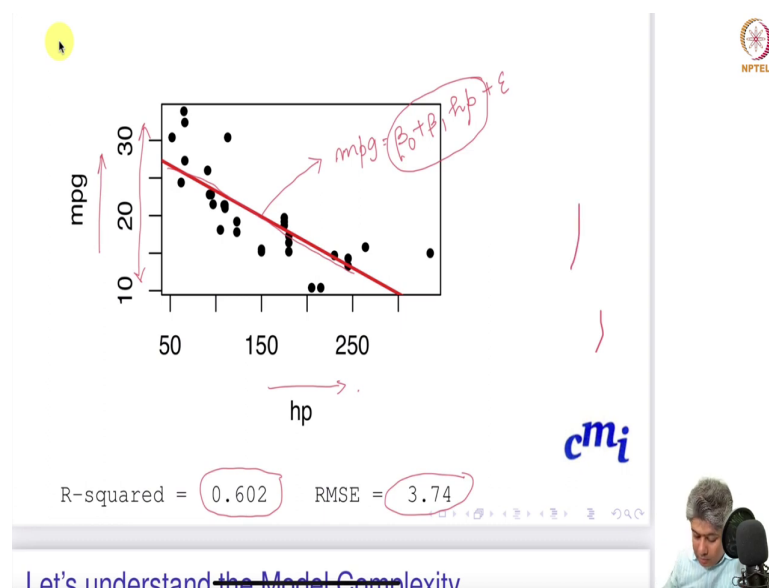
Hi all welcome to the lecture 6 part A. In this lecture we are going to discuss about Model Complexity. Now what is model complexity?

(Refer Slide Time: 00:34)



Let us start with the simple example where with our mt curve data set.

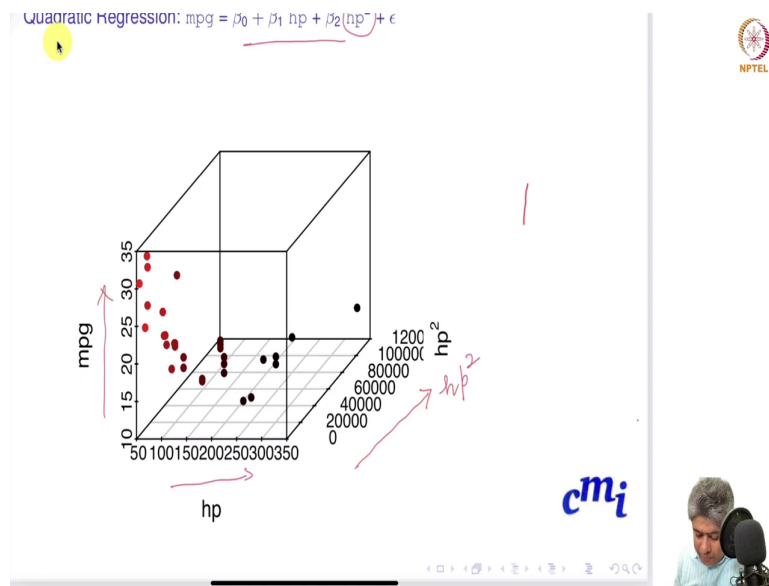
(Refer Slide Time: 00:37)



Where we have on the x axis we have taken horsepower and on the y axis we have taken miles per gallon. And this is the scatter plot and we have fitted this line miles per gallon equal to beta naught plus beta 1 horse power plus h hat. Now, this is the simple linear line. So, this is the line and this is our this is our this is the function that we are trying to fit, ok. This is the simple line we can everybody can understand.

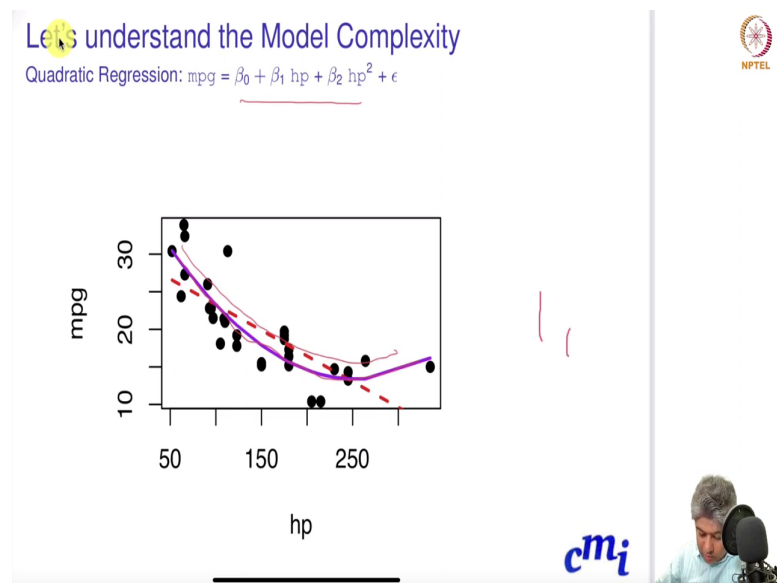
And what we are seeing that the R square is 0.602 and root mean square error is 3.74. Now R square is 0.602 means that 60 percent of the total variability of miles per gallon can be explained by this particular model, ok.

(Refer Slide Time: 01:55)



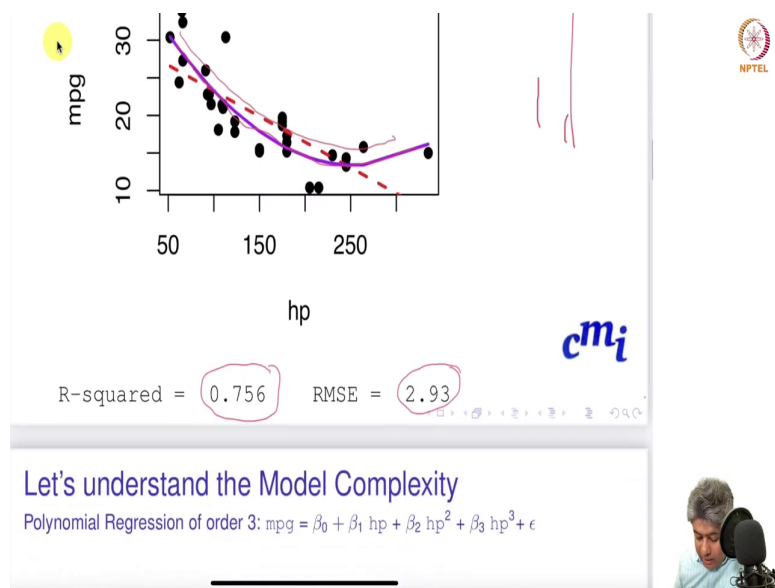
Next we have decided we thought like ok, why not we try a quadratic function. Because some kind of a horsepower square we add. When we are doing that what we what is happening is, we just put another engineered feature horsepower square. So, original data is in the 2 dimension horse power and miles per gallon. So, this was the; this was the two dimensional data the data was in 2 dimension can imagine.

(Refer Slide Time: 02:48)



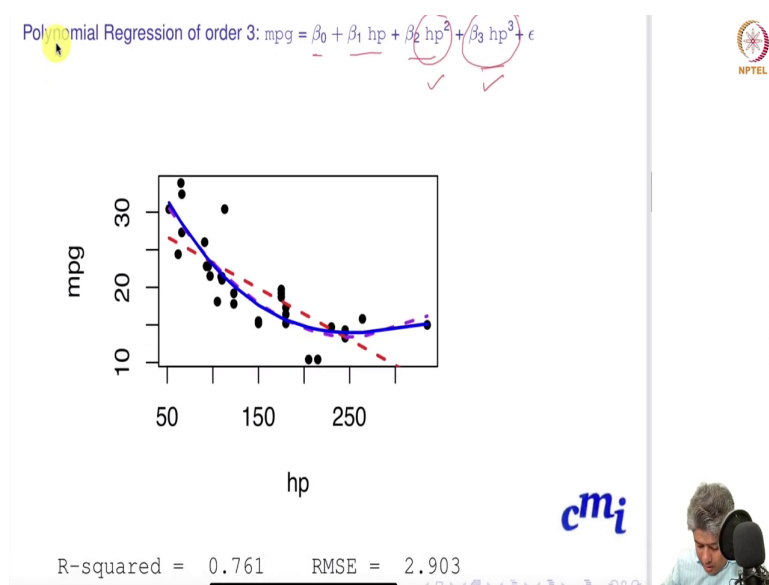
Now, we engineered a new dimension third dimension and we put the data in the cube, right. And that is what we are trying to fit. Now if you see that when we are fitting this model quadratic model this purple color curve is the quadratic model that we are fitting. This purple color curve is the one we are fitting.

(Refer Slide Time: 03:15)



Now, and what we see that R square has gone up it was previously in the linear regression it was 60 percent. But when we are using quadratic regression what we have seen? The R square has gone up to 75 percent. So, which is a good news. And then root mean square error also has gone down as expected. So, it was about three point if you go up about 3.74 now it is 2.93.

(Refer Slide Time: 03:59)



So, quite a bit of reduction in root mean square error. So, what about if we can put quadratic term? We can put also the cubic term everything else is same. So, beta naught plus beta 1 horse power beta 2 horsepower square. Now what we are doing it is the we have created other than horsepower square we have created a horsepower cube. So, now, data is in the fourth dimension we cannot draw this fourth dimension.

Since we cannot draw of in fourth dimension we can all we have only have to assume and imagine a how the fourth dimension would look like. So, the original data was in the 2 dimension. Now we are putting it into the 4 dimension we are putting the data as we are engineering new and new features, ok.

New we are engineering new features we are putting into higher dimension. And as a result we see R square is now 0.761 76 percent previously it was 75 percent in the 2 dimension when we put a quadratic model when we put a quadratic model we had 75 percent accuracy.

When we now, we are doing a cubic polynomial of order three or cubic function we have 76 percent. And now RMSE 2.903. And the model if you can see that by adding this new quadratic cubic term along with the quadratic this blue color is now a main function whereas, and almost they are same they are almost same that the quadratic curve and the cubic curve they are almost looks like same and there is no much importance.

(Refer Slide Time: 06:06)

Let's understand the Model Complexity



□ 2-dimen

linear	Model 1	R-squared = 0.602	RMSE = 3.74 ✓
quadratic	Model 2	R-squared = 0.756	RMSE = 2.93 ✓ <sub>3-d</sub>
cubic	Model 3	R-squared = 0.761	RMSE = 2.903 ✓ <sub>4-d</sub>

$$m \log = \beta_0 + \beta_1 h + \beta_2 h^2 + \dots + \beta_{10} h^{10} + \epsilon$$

$P(10)$

cmj

But you can and this is the 3 model this one is the linear right, this one is; this one was the linear this one was the quadratic and this one was a cubic. Now so remember that my data is always in 2 dimension, data is 2 dimension. So, after linear was also modelling 2 dimension,

but when we went to quadratic it was put into 3 dimension, right. It was put into 3 dimension its 3D and we went to cubic it was 4 dimension the data was then we our model we put into 4 dimension.

So, what is it mean? It means we are increasing the complexity of the model that the model is becoming complex more complex. Now you can tell me that sir why I need to stop at 3 I can go up to as many as possible, I can do I can develop a model which is polynomial about a 10. Yeah, you can do that, right.

I can develop a model like  $\beta_1$  horsepower plus  $\beta_2$  horsepower square plus dot dot dot plus  $\beta_{10}$  horsepower to the power 10 plus epsilon I am I can easily fit a polynomial of order 10 nobody can stop me. But the point is what is the interpretation? What is the interpretation?

What is the interpretation that horse power to the power 10 has what kind of effect on miles per gallon, can we explain that, can we explain this term in your model? Yes, if you fit a more and more complex model with increasing complexity your predictive accuracy we can see this your predictive accuracy will go up, your root mean square error will go down, you can see that it will keep going down.

But as you increase the model complexity eventually you will see that you will not be able to explain the model.



(Refer Slide Time: 08:47)

7 of 14

## Model Complexity

**M1 Regression Line**  
$$\text{mpg} = \beta_0 + \beta_1 \text{hp} + \epsilon$$

**M2 Regression Plane**  
$$\text{mpg} = \beta_0 + \beta_1 \text{hp} + \beta_2 \text{hp}^2 + \epsilon$$

**M3 Regression 3-dimension hyper plane**  
$$\text{mpg} = \beta_0 + \beta_1 \text{hp} + \beta_2 \text{hp}^2 + \beta_3 \text{hp}^3 + \epsilon$$

**M3' Regression 3-dimension hyper plane**  
$$\text{mpg} = \beta_0 + \beta_1 \text{hp} + \beta_2 \text{wt} + \beta_3 \text{disp} + \epsilon$$

NPTEL

m.

And that is where all problems lies. So, you can see that model complexity the initially we considered this model that a simple linear model then quadratic model, then we put it into 3 dimension the cubic model 2 sorry, 4 dimension. The another type of thing you this is also 3 dimension hyper plane that instead of horsepower and that engineered feature you put more and more new predictors weight, displacement if you have more and more features you will have a more and more predictive power.

So, if you put more and more feature or more and more engineered feature the model complexity will go up. As the model complexity go up the predictive accuracy goes up and typically the you will see root mean square goes down. So, which is a good news, but at what cost? The cost is you are paying the explain ability you cannot explain the model many times you this model is becomes almost like a black blocks.

Now, if you are working in a say in a banking and finance kind of environment you will see that regulator will always ask question why you need this particular term or why you need this complex model. Can you do the same predictive accuracy with a simpler model or simonies model, right?

Similarly, if you working in the say pharmaceutical company explain ability of the model is extremely important if you are using very complex model upon which your drug approval is depending upon and you cannot explain the model what is happening then there are lot of ethical issue comes into the picture. So, therefore, people prefer a model which is less complex, but good accuracy good predictive accuracy is being achieved.

(Refer Slide Time: 11:05)

The slide is titled "Higher order Regression with High Model Complexity". It contains two main sections:

- 1 Terms for curvature in linear regression**  
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_p x_i^p + \epsilon_i$$

is a polynomial of order  $p$
- 2 sine cosine functions of increasing frequencies**  
$$y_i = \beta_1 + \beta_2 \sin(\omega x_i) + \beta_3 \cos(\omega x_i) + \beta_4 \sin(2\omega x_i) + \dots + \epsilon_i$$

Such model is also known as Fourier model

The slide also features a hand-drawn sine wave, a small inset image of a person speaking into a microphone, and logos for NPTEL and cmj.

So, with higher order regression high model complexity you can fit any model of order  $p$  as by as we have discussed. So, any polynomial of order  $p$ . Similarly, you can put sin cosine

functions of increasing frequency say  $\beta_1 \beta_2 \sin \omega x_i \cos 3 \omega x_i \sin 2 \omega x_i$ . If you feel like your data has lot of sinusoidal behaviour then this kind of Fourier models Fourier with Fourier terms helps very a lot.

Now here if you know the omega if you know the value of omega then it is a then you can fit this model in using simple OLS method. Because if you know the value of omega then you know what will be the omega xi you know what is the sin omega xi. If you know the omega then you know omega xi you know what is cos omega xi. Similarly, all these terms are known to you, right. And now as you know all these functions.

(Refer Slide Time: 12:22)

The slide is titled "Higher order Regression with High Model Complexity". It contains two numbered points:

- 1 Terms for curvature in linear regression  
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_p x_i^p + \epsilon_i$$

is a polynomial of order  $p$
- 2 sine cosine functions of increasing frequencies  
$$y_i = \beta_1 + \beta_2 \sin(\omega x_i) + \beta_3 \cos(\omega x_i) + \beta_4 \sin(2\omega x_i) + \dots + \epsilon_i$$

Such model is also known as Fourier model

A blue box with the text "Linear in Parameter" is overlaid on the slide. The NPTEL logo is in the top right corner, and a small video inset of a person is in the bottom right corner.

So, that means, if the model is linear in parameter this is very important. Your model has to be linear in parameter and feature part has to be completely known. If your engineered feature

these are all the engineered feature or transformed variable we can call it transformed variable. The transformed variable part should not include any unknown component.

Your engineering feature engineered part should not include any unknown component. If you if your engineered feature is completely known then the model is linear in parameter and you can put use simple ordinary least square method to fit the model, train the model, do the prediction you can do everything.

But the problem is as you put more and more engineered feature the model become highly complex, it is you lose the interpretability and explain ability of the model.

(Refer Slide Time: 13:25)

Remarks

$X = \begin{bmatrix} \end{bmatrix}_{n \times p}$

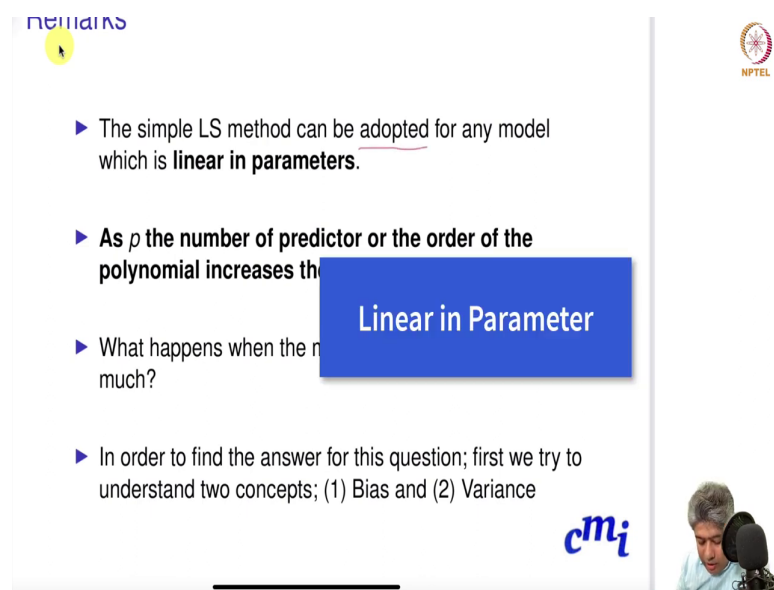
- ▶ The simple LS method can be adopted for any model which is **linear in parameters**.
- ▶ As  **$p$**  the number of predictor or the order of the polynomial increases the model complexity increases.
- ▶ What happens when the model complexity increases too much?
- ▶ In order to find the answer for this question; first we try to understand two concepts; (1) Bias and (2) Variance

NPTEL

cm<sub>i</sub>

So, some remarks.

(Refer Slide Time: 13:35)




REMARKS

- ▶ The simple LS method can be adopted for any model which is **linear in parameters**.
- ▶ As  $p$  the number of predictor or the order of the polynomial increases the
- ▶ What happens when the  $n$  much?
- ▶ In order to find the answer for this question; first we try to understand two concepts; (1) Bias and (2) Variance

Linear in Parameter

cmj

NPTEL



The simple least square method can be adopted for any model which is linear in parameter this is very important. The model has to be linear in parameter, ok. As  $p$  the number of feature or number of predictors or the order of the polynomial increases model complexity increases,  $p$  is the number of feature or the number of column in your design matrix, ok. So, typically it is  $n$  cross  $p$ . So,  $p$  is the number of predictors or the number of feature in your design matrix as it increases your model complexity increases.

What happens? Now here is a big question. What happens when model complexity increases too much? Let us try to understand what happens when model complexity increases too much. Obviously, one loss is interpretability and explain ability. But in order to answer this question first try to understand two concept. One is bias another is variance.

(Refer Slide Time: 14:49)

**Bias and Variance**

*Statistics*

▶ **What is Bias?** If  $\theta$  is an unknown parameter and  $\hat{\theta}$  is an estimator of  $\theta$ , then

$$\mathbb{E}(\hat{\theta}) = \theta + b,$$

where  $b$  is known as **bias**. If  $b = 0$ , then  $\hat{\theta}$  is known as **unbiased estimator** of  $\theta$ . *bias = 0*




▶ **What is Variance?** The variance of  $\hat{\theta}$  is

$$\text{Var}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2,$$

▶ Consider the prediction of the new response at input  $x_0$

$$y_0 = f(x_0) + \epsilon_0.$$

$\hat{f}(x_0) = x_0^T \hat{\beta}$  is an estimator of  $f(x_0) = x_0^T \beta$ .



So, what is bias? If  $\theta$  is unknown, if  $\theta$  is unknown and  $\hat{\theta}$  is an estimator of  $\theta$  then if expectation of  $\hat{\theta}$  is equal to  $\theta + b$  then  $b$  is known as bias. So,  $b$  is the bias, ok. And  $\hat{\theta}$  is and if  $b$  is 0 if your bias is 0 then  $\hat{\theta}$  is known as unbiased estimator, ok.

Now, what is variance? The variance of  $\hat{\theta}$  is expected value of  $\hat{\theta}$  minus expected value of  $\hat{\theta}$  whole square. So, this is typically what you will see the definition of bias of estimator and variance of estimator in statistics in statistics. Now in machine learning or statistical machine learning it is slightly different, ok.

(Refer Slide Time: 16:11)

5

▶ **What is Variance?** The variance of  $\hat{\theta}$  is

$$\text{Var}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2,$$

▶ Consider the prediction of the new response at input  $x_0$

$y_0 = f(x_0) + \epsilon_0$  (new response)

$\hat{f}(x_0) = x_0^T \hat{\beta}$  is an estimator of  $f(x_0) = x_0^T \beta$ .

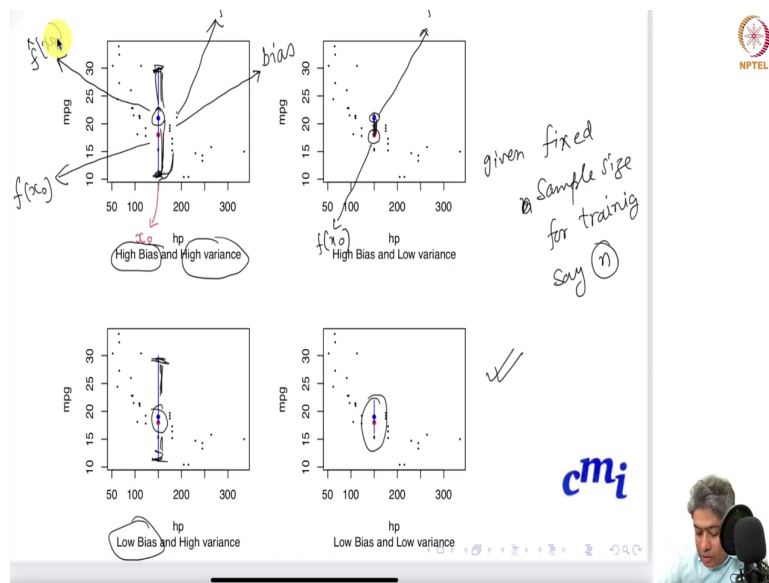
**cmj**

### Bias and Variance

Let us try to understand. I mean it is the same concept slightly context is slightly different. Let us do it little slowly, ok. If you want you can do it slowly or you can repeat the video this part you because this concept is subtly different same concept, but subtly different. Consider the prediction of new response at input  $x$  naught. So,  $x$  naught is a new response, ok. New response and  $y$  naught is the  $f$  of  $x$  naught plus epsilon hat epsilon naught.

So, now when  $x$  naught transpose beta hat is we can call it  $f$  hat  $x$  naught is an estimator of  $f$  of  $x$  naught which is  $x$  naught transpose beta. So,  $x$  naught transpose beta hat is an estimator of  $x$  naught transpose beta, ok.

(Refer Slide Time: 17:25)



Now what is how the bias variance behaves? Now suppose this is the point for which you want to; you want to do the estimation. So, true  $f$  of  $x$  naught is suppose here this red point, ok. This red point is true value  $f$  of  $x$  naught this  $f$  of this is true  $f$  of  $x$  naught. Now this blue point this blue point is  $\hat{f}$   $x$  naught. So, naturally you have this difference is can be views as a bias this difference is view as bias. And this red line can be viewed as variance of  $\hat{f}$   $x$  naught.

Now, what is this blue line? Sorry, this blue line previously I said red line its blue line. This blue line gives you that ok, this could be the prediction. So, let me just little bit erase this part, ok. So, this could be the prediction, but this is the actual value a confidence interval actual prediction could be anywhere in this range. So, the variance of the confidence in the



prediction is very less. So, that is the variability possible variability is very high. So, this is a high bias and high variance situation.

Now, I am talking about high bias, but low variance. So, similar thing you can see this this is a true  $f$  of  $x$  naught this is  $f$  naught  $f$  hat this is estimated  $f$  hat  $x$  naught this is true  $f$   $x$  naught and if you see this the variance variability is so less than very tight and lot of bias.

Now, what is happening you can see that bias is almost bias is almost null there is almost no bias. This is the low bias situation, but very high variability my confidence in my prediction is very very less the variability of my possible the my possible prediction could be anywhere in this region. Whereas, this is the low variance low bias situation.

We obviously, would like to arrive somewhere in this situation this is probably the best situation where we have low bias and low variance situation. However, here is the bad news that I am bringing to you. The bad news is that we cannot given a fixed sample size, given fixed sample size sample size for training say  $n$  training or fitting whatever you call it.

Given this fixed sample size if you try to reduce bias too much your variance will shoot up, if you try to reduce variance too much your bias will shoot up this is typically the case.

(Refer Slide Time: 21:35)

**Bias-Variance Tradeoff**

▶ Consider the prediction of the new response at input  $x_0$

$$y_0 = f(x_0) + \epsilon_0.$$



$\hat{f}(x_0) = x_0^T \hat{\beta}$  is an estimator of  $f(x_0) = x_0^T \beta$ .

▶ The MSE of  $\hat{f}(x_0) = x_0^T \hat{\beta}$  is

*Result*

$$\begin{aligned} \text{MSE}(\hat{f}(x_0)) &= \mathbb{E}(\hat{y}_0 - f(x_0))^2 \\ &= \mathbb{E}(\hat{y}_0 - \mathbb{E}(\hat{y}_0))^2 + \mathbb{E}(\mathbb{E}(\hat{y}_0) - f(x_0))^2 \\ &= \text{Var}(\hat{y}_0) + \text{Bias}^2(\hat{y}_0) \\ &= \text{Var}(x_0^T \hat{\beta}) + \text{Bias}^2(x_0^T \hat{\beta}) \end{aligned}$$

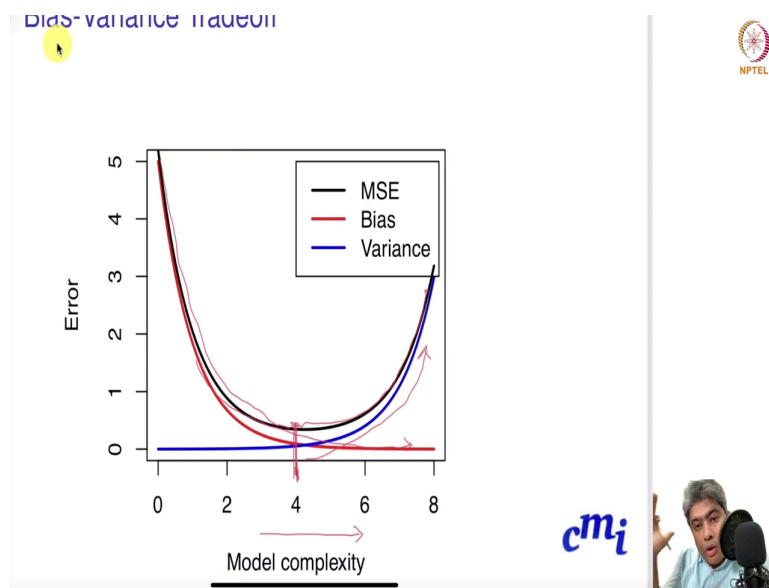
*cmj*



So, what people do try to do? So, what people try to do is they first try to define bias and variance with respect to mean square error. So, this is result this result mean square error is a function of variance of  $x_0^T \hat{\beta}$  and bias square. So, how it is started definition of a mean squared error of  $\hat{f}(x_0)$  is  $\hat{y}_0 - f(x_0)$  whole square and then with some algebra with a little bit of algebra you can show that this is equal to variance of  $\hat{y}_0$  plus bias square, ok.

And I am not going into the derivation of this mathematics I am just going to use this result, ok. I am going to use this result. If you I think in the Tibshirani's book of elements of statistical learning you can find the derivation. If you are interested you can have a look into the textbook.

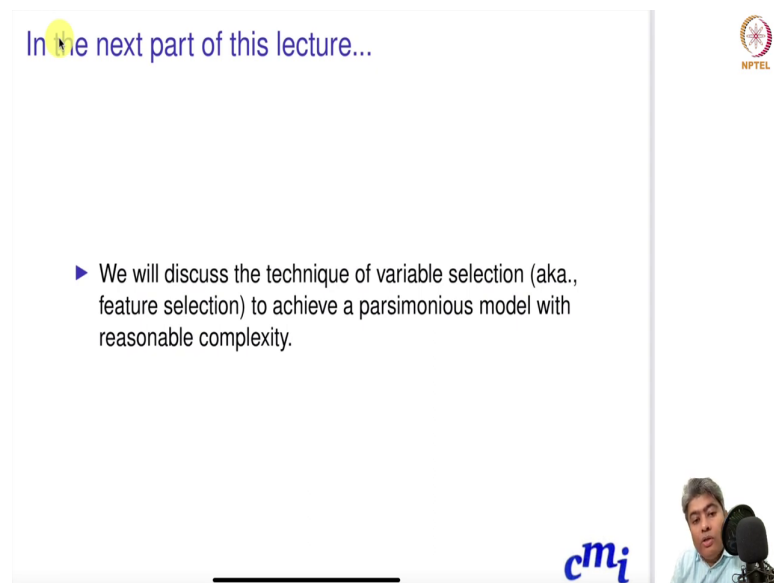
(Refer Slide Time: 23:01)



Now what I am planning is to have a look how typically it behaves. Now if you see if you consider a model whose model complexity increases what happens that the variance with the increasing model complexity the bias typically goes down and variance typically goes up and because mean square error is a function of variance plus bias square.

So, what happens it first goes down comes to some kind of minimum and then it starts going up. So, instead of trying to minimizing bias and variance together what we try to do we try to minimize the mean square error, we try to minimize the mean square error. And perhaps this is a Parsimonious model where we want to stable we want to stop we do not want to increase the model complexity any further, ok.

(Refer Slide Time: 24:15)



In the next part of this lecture...

- ▶ We will discuss the technique of variable selection (aka., feature selection) to achieve a parsimonious model with reasonable complexity.

cmj

NPTEL

This is typically the how the model training happens. I will stop here in the next video we will discuss the technique of variable selection also known as feature selection to achieve a Parsimonious model with reasonable complexity. So, we will talk about it. That as instead of using all the features and all the possible engineered features.

We will drop some of them and try to come up with some number of reasonable number of features in the model and try to arrive some kind of a Parsimonious level somewhere here where mean square error is minimum and bias and variance is somewhere in the in a controlled level and not too high not too low, ok.

Thank you very much see you in the next video.

