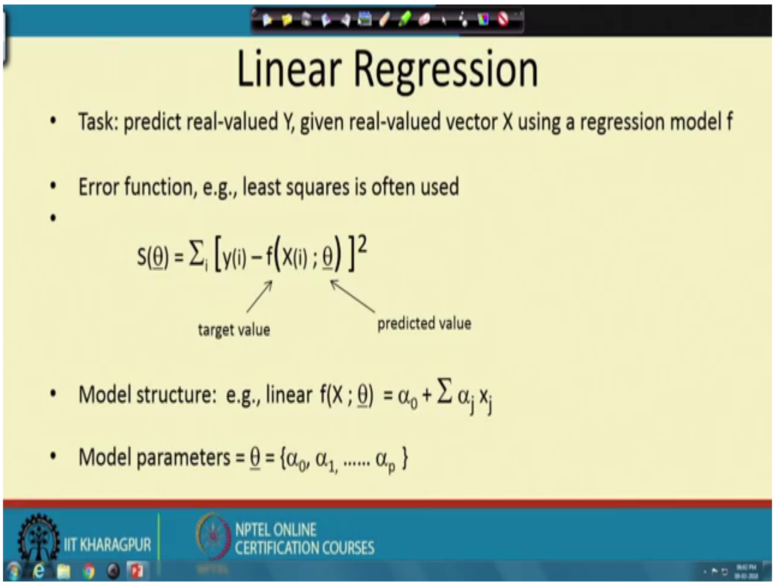
Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture - 38
Regression II

We continue the discussion on linear regression.

(Refer Slide Time: 00:23)



Linear Regression

- Task: predict real-valued Y, given real-valued vector X using a regression model f
- Error function, e.g., least squares is often used
-

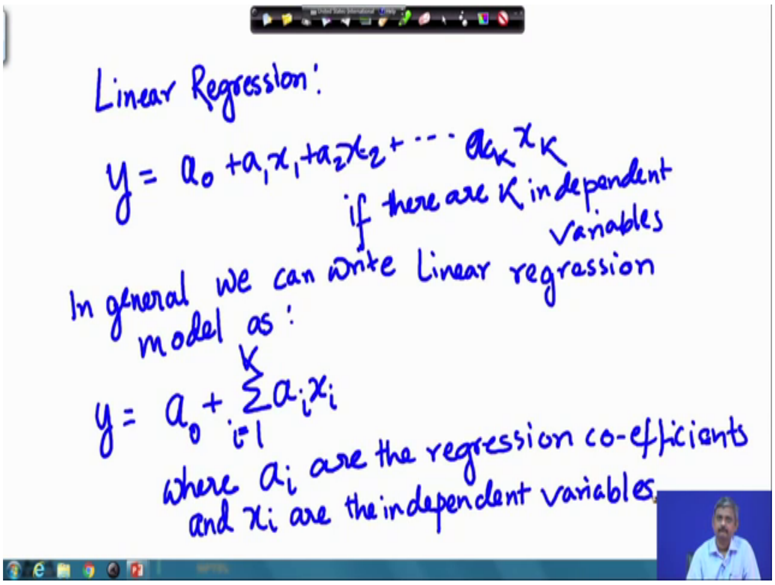
$$S(\underline{\theta}) = \Sigma_i \left[y(i) - f(X(i)\,;\underline{\theta})\right]^2$$

target value

predicted value

- Model structure: e.g., linear $f(X\,;\underline{\theta}) = \alpha_0 + \Sigma\, \alpha_j x_j$
- Model parameters = $\underline{\theta} = \{\alpha_0, \alpha_1, \ldots\ldots \alpha_p\}$

(Refer Slide Time: 00:25)



Linear Regression:

$y = a_0 + a_1x_1 + a_2x_2 + \cdots a_K x_K$

if there are K independent variables

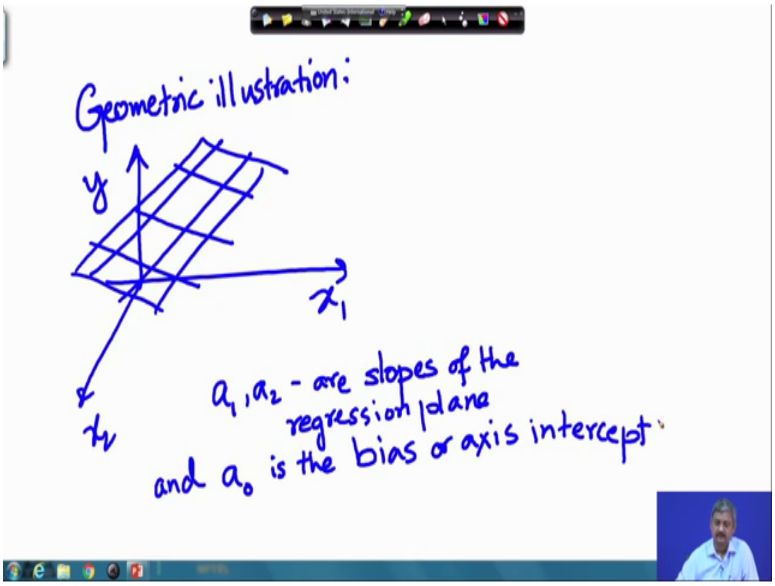
In general we can write linear regression model as:

$y = a_0 + \sum_{i=1}^{K} a_i x_i$

where a_i are the regression co-efficients and x_i are the independent variables.

As we discussed earlier we have a regression model of the following form. In general we can write as and then independent valuables.

(Refer Slide Time: 02:38)

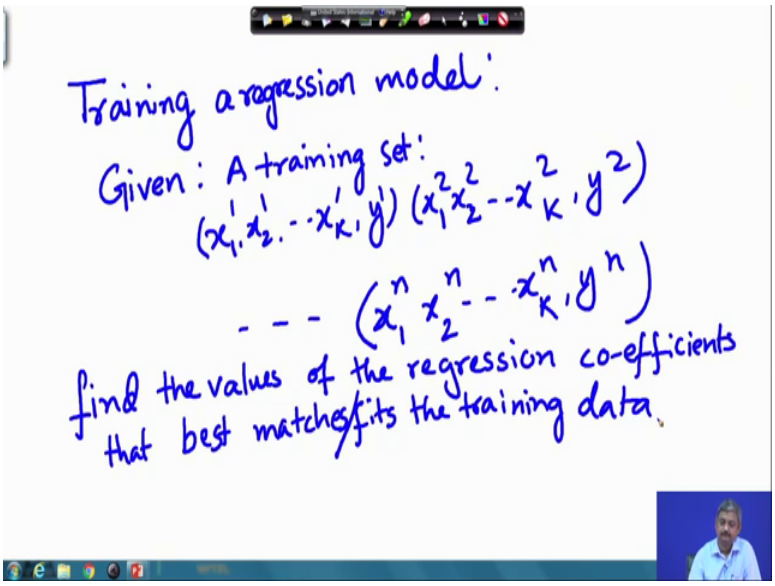


Geometric illustration:

a_1, a_2 - are slopes of the regression plane
and a_0 is the bias or axis intercept.

As (Refer Time: 03:51) so, this is the illustration.

(Refer Slide Time: 03:51)

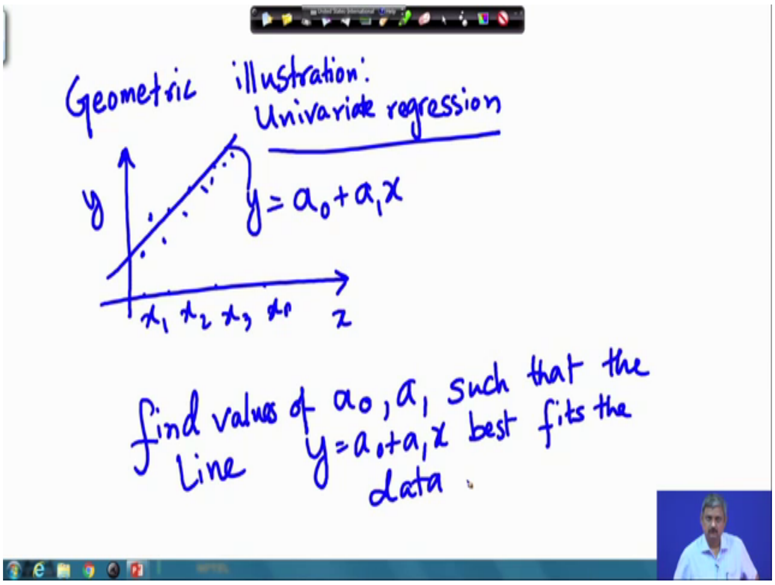


Training a regression model:

Given: A training set:
$(x_1^1, x_2^1, \ldots x_K^1, y^1)$ $(x_1^2 x_2^2 \ldots x_K^2, y^2)$
$\ldots$ $(x_1^n\, x_2^n \ldots x_K^n, y^n)$

find the values of the regression co-efficients that best matches/fits the training data.

(Refer Slide Time: 05:48)

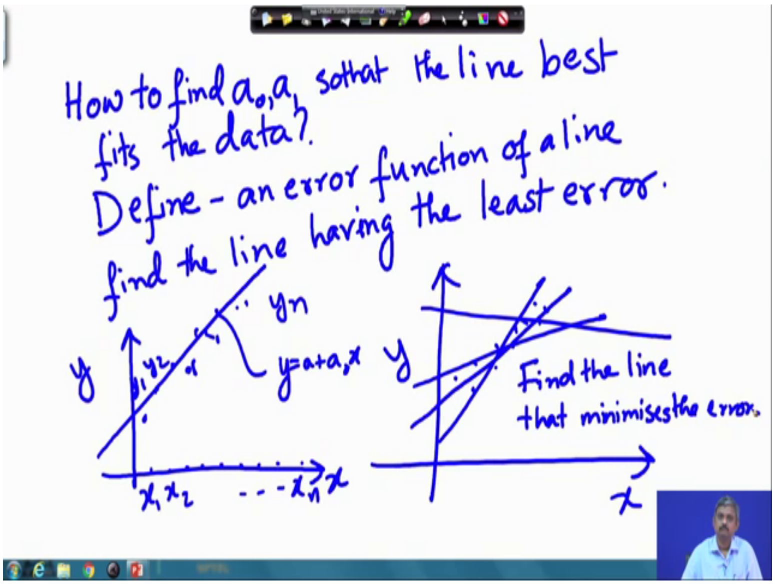


Geometric illustration: Univariate regression

find values of a_0, a_1 such that the line $y = a_0 + a_1 x$ best fits the data.

Data, find a 0 a 1.

(Refer Slide Time: 07:30)



How to find a_0, a_1 so that the line best fits the data?

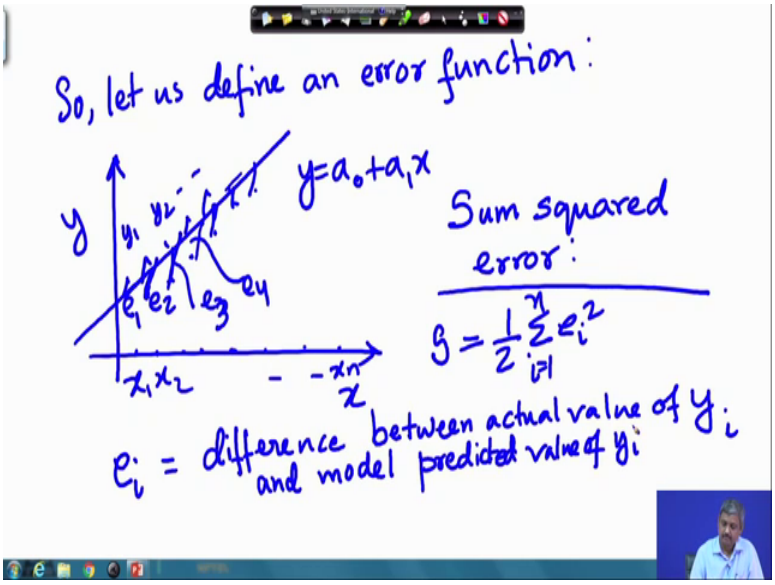
Define - an error function of a line

find the line having the least error.

y
y_1
y_2
y_n
$y = a_0 + a_1 x$
x_1 x_2 - - - x_n x
y
Find the line that minimises the error.
x

So, that it best fits the data. Ok define some error function and minimise it. There will be some error it will not exactly match the points. So, try to find the line.

(Refer Slide Time: 09:53)



So, let us define an error function:

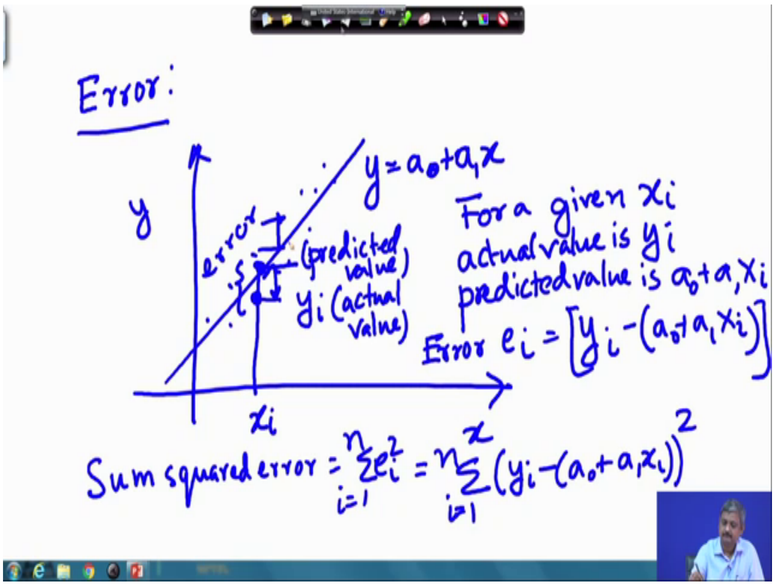
Sum squared error:

$$\mathcal{J} = \frac{1}{2}\sum_{i=1}^{n} e_i^2$$

e_i = difference between actual value of y_i and model predicted value of y_i

This line, this line or this line or this line; So, what is the error? You measure how much this point is off from the line. Ok square up all these add them up. Ok, square these are the error sum them; the difference between actual value and predicted value.

(Refer Slide Time: 12:44)



Error:

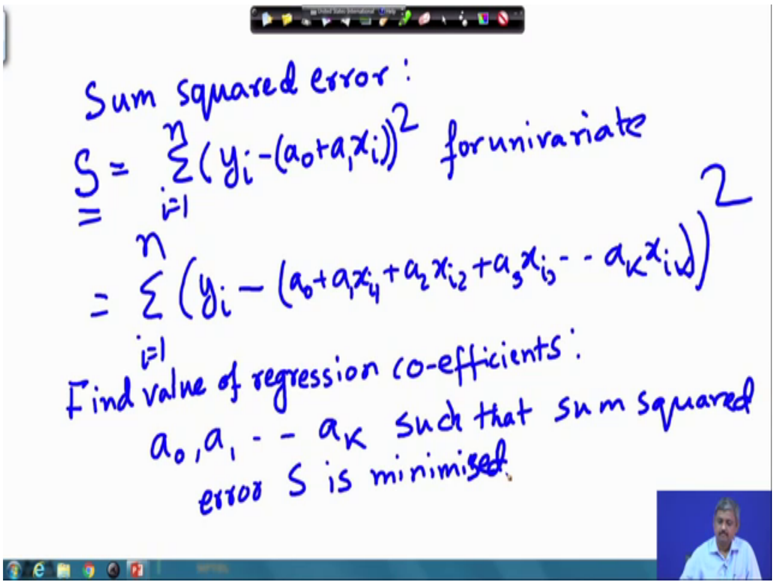
For a given x_i actual value is y_i predicted value is $a_0+a_1x_i$

Error $e_i = [y_i-(a_0+a_1x_i)]$

Sum squared error $= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i-(a_0+a_1x_i))^2$

So, I xi this is the error, this defines why I am taking square because see error can be positive as well as negative moment I take both are equally bad.

(Refer Slide Time: 15:38)



Sum squared error:

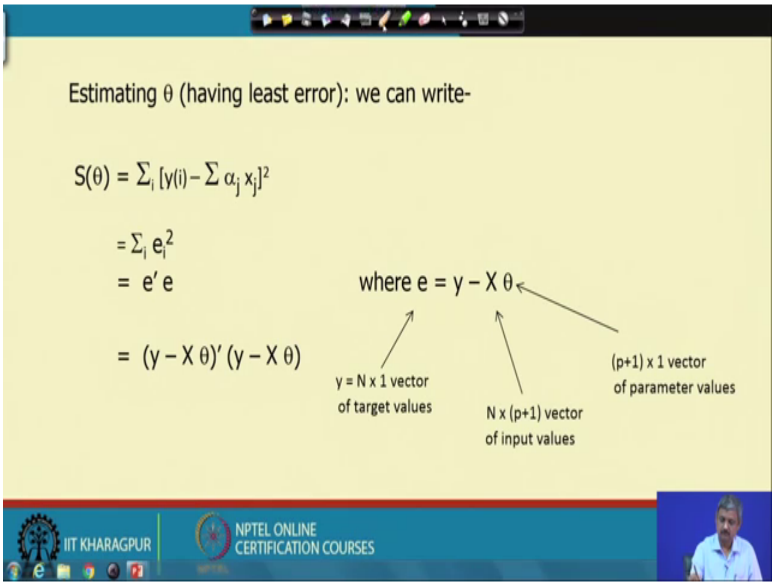
$$S = \sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i))^2 \text{ for univariate}$$

$$= \sum_{i=1}^{n} \left(y_i - (a_0 + a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3} - - a_K x_{iK})\right)^2$$

Find value of regression co-efficients: $a_0, a_1, - - a_K$ such that sum squared error S is minimised.

So, moment I make it positive a square it is becoming positive, such that minimised ok. So, find a 0 a 1 a 2 so that this quantity is smallest for a given training set ok; So, here I have used a slightly different methodology this is alpha and these coefficients are called the model parameters; So, the squared.

(Refer Slide Time: 18:03)



Estimating θ (having least error): we can write-

$$S(\theta) = \Sigma_i [y(i) - \Sigma \alpha_j x_j]^2$$

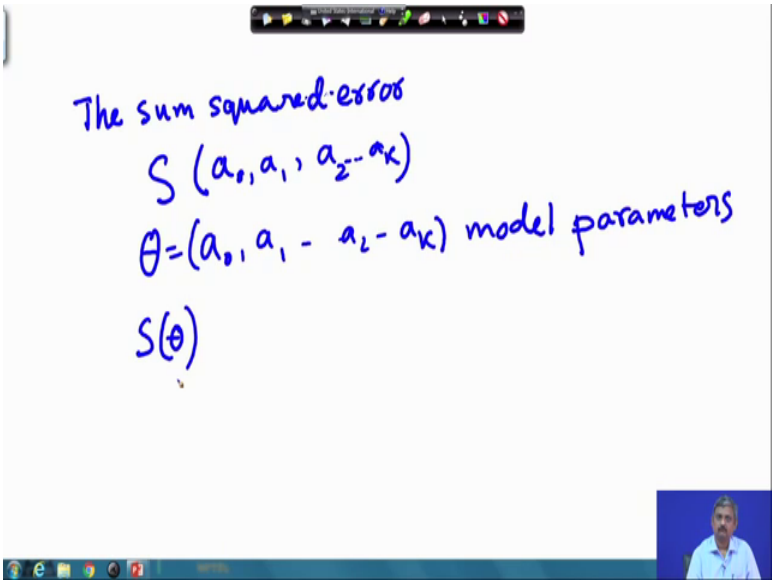
$$= \Sigma_i e_i^2$$

$$= e' e \qquad \text{where } e = y - X\theta$$

$$= (y - X\theta)'(y - X\theta)$$

y = N x 1 vector of target values
N x (p+1) vector of input values
(p+1) x 1 vector of parameter values
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

(Refer Slide Time: 18:28)



The sum squared error

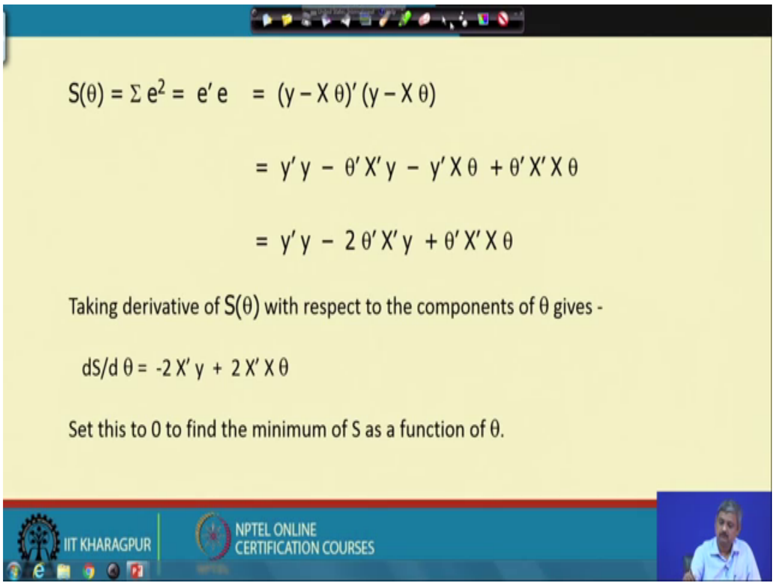
$S(a_0, a_1, a_2 .. a_K)$

$\theta = (a_0, a_1 - a_2 - a_K)$ model parameters

$S(\theta)$

What S? It is a function of the parameters what if you choose certain parameters sum error you will get. So, this is my error, where like this. So, I can write this like this e transpose into e is this thing there is this thing. So, if we expand you get this.

(Refer Slide Time: 20:35)



$S(\theta) = \Sigma e^2 = e'e = (y - X\theta)'(y - X\theta)$

$= y'y - \theta'X'y - y'X\theta + \theta'X'X\theta$

$= y'y - 2\theta'X'y + \theta'X'X\theta$

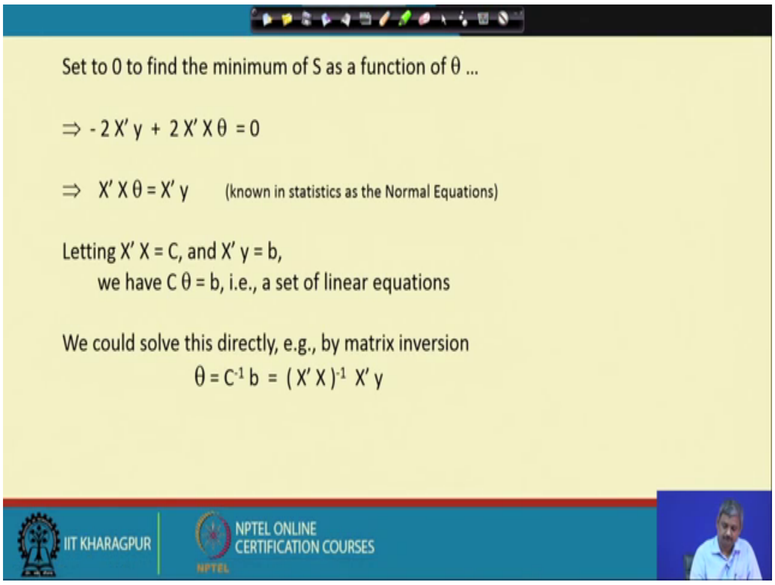
Taking derivative of $S(\theta)$ with respect to the components of θ gives -

$dS/d\theta = -2X'y + 2X'X\theta$

Set this to 0 to find the minimum of S as a function of θ.

I am just writing it down so finally, you will get theta is the spectre. Now, at minima this will be equal to 0 the minima will be derivative; minima this thing ok.

(Refer Slide Time: 21:42)



Set to 0 to find the minimum of S as a function of θ …

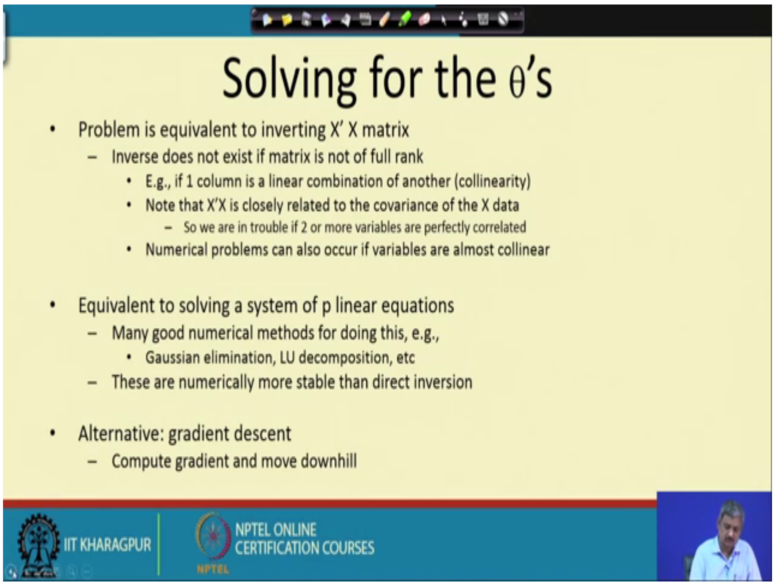
$\Rightarrow -2X'y + 2X'X\theta = 0$

$\Rightarrow X'X\theta = X'y$ (known in statistics as the Normal Equations)

Letting $X'X = C$, and $X'y = b$,
we have $C\theta = b$, i.e., a set of linear equations

We could solve this directly, e.g., by matrix inversion
$\theta = C^{-1}b = (X'X)^{-1}X'y$

(Refer Slide Time: 22:40)

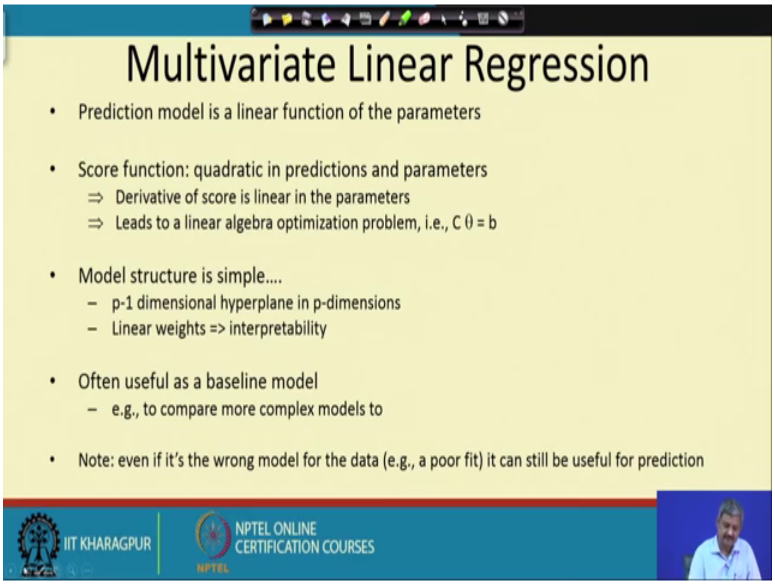


Solving for the θ's

- Problem is equivalent to inverting X' X matrix
  - Inverse does not exist if matrix is not of full rank
    - E.g., if 1 column is a linear combination of another (collinearity)
    - Note that X'X is closely related to the covariance of the X data
      - So we are in trouble if 2 or more variables are perfectly correlated
    - Numerical problems can also occur if variables are almost collinear
- Equivalent to solving a system of p linear equations
  - Many good numerical methods for doing this, e.g.,
    - Gaussian elimination, LU decomposition, etc
  - These are numerically more stable than direct inversion
- Alternative: gradient descent
  - Compute gradient and move downhill

So, there are different methods, get theta there are different methods ok.

(Refer Slide Time: 23:49)

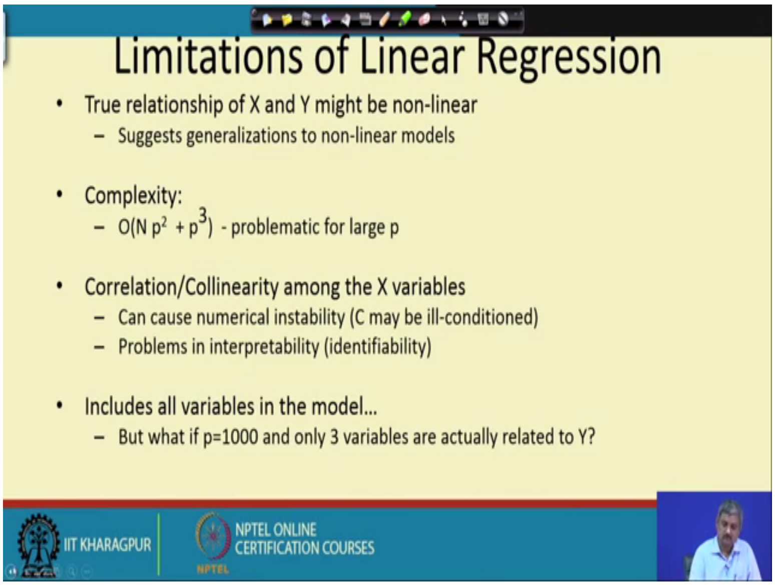


Multivariate Linear Regression

- Prediction model is a linear function of the parameters
- Score function: quadratic in predictions and parameters
  - ⇒ Derivative of score is linear in the parameters
  - ⇒ Leads to a linear algebra optimization problem, i.e., $C\,\theta = b$
- Model structure is simple....
  - p-1 dimensional hyperplane in p-dimensions
  - Linear weights => interpretability
- Often useful as a baseline model
  - e.g., to compare more complex models to
- Note: even if it's the wrong model for the data (e.g., a poor fit) it can still be useful for prediction

So, this is like solving simultaneous equations ok. So, you can extend it to multivariate case also.

(Refer Slide Time: 23:57)

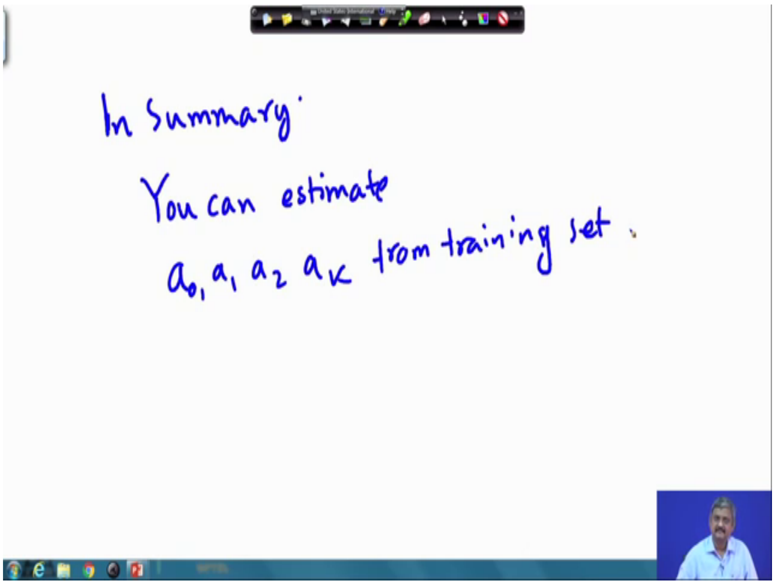


Limitations of Linear Regression

- True relationship of X and Y might be non-linear
  - Suggests generalizations to non-linear models
- Complexity:
  - $O(N\,p^2 + p^3)$ - problematic for large p
- Correlation/Collinearity among the X variables
  - Can cause numerical instability (C may be ill-conditioned)
  - Problems in interpretability (identifiability)
- Includes all variables in the model...
  - But what if p=1000 and only 3 variables are actually related to Y?

Alright so, that is how ok.

(Refer Slide Time: 24:13)



In Summary:

You can estimate

$a_0, a_1\ a_2\ a_K$ from training set.

So, in the next lecture I stop here, in the next lecture I will explain how to extend it to non-linear cases.

Thank you.