

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

**Lecture - 39**  
**Regression- III**

We will continue our discussion on the linear regression. As we discussed we had a model of the form, we had a model of the form of this type where  $y$  is a naught plus a  $1 \times 1$  plus a  $2 \times 2$  up to a  $k \times k$ ,  $x \times k$ .

(Refer Slide Time: 00:33)

**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
- $S(\theta) = \sum_i [y(i) - f(x(i); \theta)]^2$
- Model structure: e.g., linear  $f(x; \theta) = \alpha_0 + \sum \alpha_j x_j$
- Model parameters =  $\theta = \{\alpha_0, \alpha_1, \dots, \alpha_p\}$

Handwritten notes on the slide:

$$y = a_0 + a_1 x_1 + a_2 x_2 + \dots + a_k x_k$$

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

$$\theta = [a_0, a_1, \dots, a_k]$$

So, you can write it down as a particular form  $y$  equal to a naught plus sigma  $i$  equal to 1 to  $k$   $a_i x_i$ . So, this set of value, so this is the regression model or the regression equation this one is the regression equation.

Well, I have written alpha here. So, you can as well put alpha here. So, the model parameters we call it as theta would be denoted by this a naught  $a_1$  up to a  $k$  these are the these are the model parameters. So, theta is a naught  $a_1$  up to a  $k$ . What we are supposed to do is to given a set of points  $x_1 y_1$   $x_2 y_2$   $x_3 y_3$  and the corresponding  $y_1 y_2 y_3$  we have to find this parameters theta a naught  $a_1$   $a_2$   $a_3$ . So, that the error that is the actual value minus the model predicted value which is this quantity model predicted value square of that and some of that of course, it will be a function of theta because if

we change the slopes of the line here we will change this quantity sum squared error is minimum. We have to find that theta.

(Refer Slide Time: 02:34)

Estimating  $\theta$  (having least error): we can write-

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

$$= \sum_i e_i^2$$

$$= e' e$$

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

where  $e = y - X \theta$

$y = N \times 1$  vector of target values

$N \times (p+1)$  vector of input values

$(p+1) \times 1$  vector of parameter values

Handwritten notes in red ink:

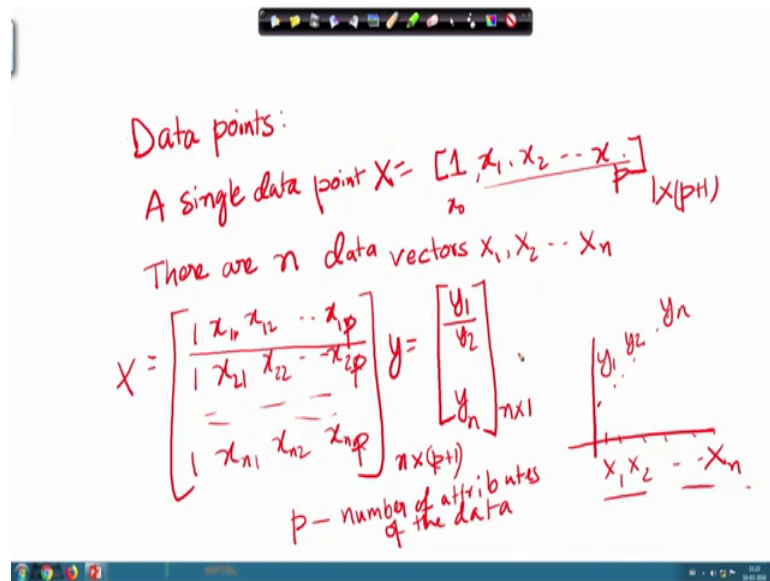
- $X = [x_1 \ x_2 \ \dots \ x_k]$
- $X = [1 \ x_1 \ \dots \ x_k]$
- $y = \sum_{i=0}^k a_i x_k, \theta_0 = 1$
- $y = a_0 \cdot 1 + a_1 x_1 + a_2 x_2 + \dots + a_k x_k$

Small video inset of a lecturer in the bottom right corner.

And what we did quickly was that and so you if you way of estimating the theta you can write down your error e as y minus summation alpha j, x j. So, what we will do actually is that, so each data point x each data point x even a multivariate class case is a collection of k or d components ok. And for the ease of the algebra what we do is that I sort of write x in a slightly different form for this a naught. So, you remember y is a naught plus a 1 x 1 I introduce a dummy component which is always 1 ok. So, then I can write y as summation i equal to 0 to k a i x k, where x 0 is always one when x 0 is always one, so just another way of writing this equation ok.

So, here x vector the multivariate x vector becomes 1 at x 0 and remaining values in the remaining positions slightly augmented. So, if we write in that form I can write my squared error e as this, squared error e as some sum this square which is summation e i square and if your e is a, for every point you have a vector. So, you have x 1 x 2 x n and you have y 1 y 2 y n and every point there is a error between these two that error vector I call that is a e vector which is nothing but y minus I will explain what is the e vector.

(Refer Slide Time: 05:10)



Let me write down the notations a single data point is a if we have or let me write it as p the notation. So, suppose there are p components. So, this is 1 cross p plus 1 p of this and 1 for x naught y n. So, I can write a big vector x like this. So, this is first vector first component first vector, second component, first vector k component ok. So, this is all these vectors I have write as a matrix. Let me call the capital l a a x matrix.

So, what is the dimension of this matrix? It is it is n cross naught k p, let me write p is the number of components of the multivariate input number of attributes of the data. So, we have p plus 1 and n such vectors. So, this is x. What is y? So, each of this x have a desired y, so y is a n cross one vector ok. So, what is e? Let me clear (Refer Time: 08:32) x n p at each of this point what is the y.

(Refer Slide Time: 08:30)

Handwritten mathematical notes on a whiteboard:

- Matrix  $X$  is defined as  $X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$  with dimensions  $n \times (p+1)$ .
- Vector  $Y$  is defined as  $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$  with dimensions  $n \times 1$ .
- Model vector  $\theta$  is defined as  $\theta = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{bmatrix}$  with dimensions  $1 \times (p+1)$ .
- The error vector  $e$  is defined as  $e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$ .
- The error  $e_i$  is given by  $e_i = y_i - (a_0 x_{i0} + a_1 x_{i1} + \dots + a_p x_{ip})$ .
- The sum of squared errors  $S$  is given by  $S = e^T e$ .
- A small diagram shows a point  $x_i$  on the x-axis and a corresponding value  $y_i$  on the y-axis, with a regression line  $a_0 + a_1 x_i + a_2 x_i^2$  passing through it.

I have a model vector theta which is the coefficient of the regression. So, I have a naught or alpha naught whatever we want to write a p there will be p coefficients for a p dimensional p attributes plus 1 extra coefficient. So, it is 1 cross p plus 1. So, what is the error vector? Error vector is y minus if you take each of these small x's, x i let me call it. So, y i, e i is y i minus x i into theta because see they sorry x i transpose x i into theta transpose I can I can write because wait not x i into theta transpose x i transpose into theta transpose into ok, theta also let me make it the other way round. So, theta is actually let me write it as this vector.

So, e i is y i minus sorry a 0 x i 0 which is one a 1 this is my e i error at point i, error at the I eth point if you remember the diagram error at the eth point is take x i find out a 0 1, a 1, x i 1, a 2, x i 2 and so on. Find out this and find out the actual y i the difference between that 1 ok. So, this e i, so I can also write a e vector as each of the point have an error ok. So, I take y i I take x i transpose multiplied by theta that is this quantity subtract that from y i, I will get my e i I will write it down e i as a vector. So, my sum squared error is you can imagine is sum of e i square e 1 square e 2 square e 3 square. So, I can write it as e transpose e this transpose e it will be this square plus this square if you do the vector multiplication ok. So, that is my error that is my error e transpose e if you match the dimensions you will get it.

(Refer Slide Time: 12:43)

Estimating  $\theta$  (having least error): we can write-

$$S(\theta) = \sum_i [y(i) - \sum \alpha_j x_j]^2$$
$$= \sum_i e_i^2$$
$$= e' e$$
$$= (y - X \theta)' (y - X \theta)$$

where  $e = y - X \theta$

$y = N \times 1$  vector of target values

$X = N \times (p+1)$  vector of input values

$\theta = (p+1) \times 1$  vector of parameter values

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, then we did all the minimizations; that means, take the derivative with respect to theta.

(Refer Slide Time: 12:55)

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

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

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

Set this to 0 to find the minimum of  $S$  as a function of  $\theta$ .

$\theta = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{bmatrix}$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Note that, note, note that theta is this a p this vector ok. So, you take derivative partial derivative with respect to each of this theta you will get this and if we equate that to 0 you get if you equate that to 0 you get this thing you get theta equal to, you get theta equal to this quantity.

(Refer Slide Time: 13:42)

Set to 0 to find the minimum of  $S$  as a function of  $\theta$  ...

$$\Rightarrow -2 X' y + 2 X' X \theta = 0$$
$$\Rightarrow X' X \theta = X' y \quad (\text{known in statistics as the Normal Equations})$$

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

$$\theta = C^{-1} b = (X' X)^{-1} X' y$$

Handwritten in red ink:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & \dots & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & \dots & x_{np} \end{bmatrix}$$
$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$
$$\theta = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{bmatrix}$$

Logos: IIT KHARAGPUR, NPTEL ONLINE CERTIFICATION COURSES

So, so this is defining this. So, the theta is equal to this quantity. So, basically you note the dimensions again. So,  $x$  as I told is  $1 \times 11$ ,  $x_{12}$ , up to  $x_{1p}$ ;  $1 \times 21$ , up to  $x_{2p}$ ;  $1 \times n1$  up to  $x_{np}$ . So,  $x$  is this  $y$  is this and this quantity theta is this. So, theta's dimension is  $p + 1$  cross  $1$  this dimension is  $n$  cross  $1$  this dimension is  $n$  cross  $p + 1$  ok. So, if you just do this multiplication matrix do the matrix inversion multiply you can figure out that they will be the same if you multiply their dimensions will match you get your parameters of the equation theta ok.

In the exercises I will in the assignments I will give you a problem where you have to actually work this out and find out the value of the theta.

(Refer Slide Time: 15:55)

**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

*easy to design.*  
*- baseline to compare other models.*

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, people solve it by matrix inversion or by some numerical methods.

(Refer Slide Time: 16:03)

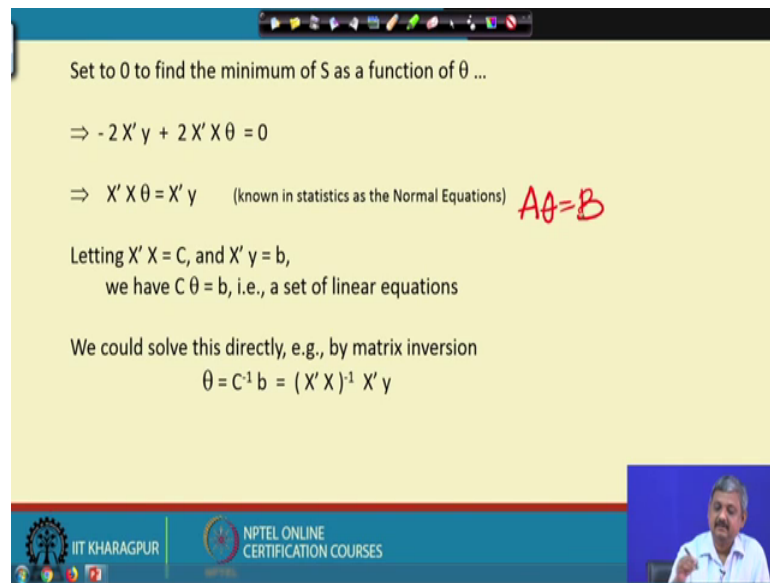
**Solving for the  $\theta$ 's**

- 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

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Basically it is like solving a set of linear simultaneous equations of the form if you. So, this equation is like solving something of the form  $A\theta = B$  ok. So,  $\theta$  is a vector  $\theta_1 \theta_2$  (Refer Time: 16:42).

(Refer Slide Time: 16:25)



Set to 0 to find the minimum of S as a function of  $\theta$  ...

$$\Rightarrow -2 X' y + 2 X' X \theta = 0$$
$$\Rightarrow X' X \theta = X' y \quad (\text{known in statistics as the Normal Equations}) \quad A\theta = B$$

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

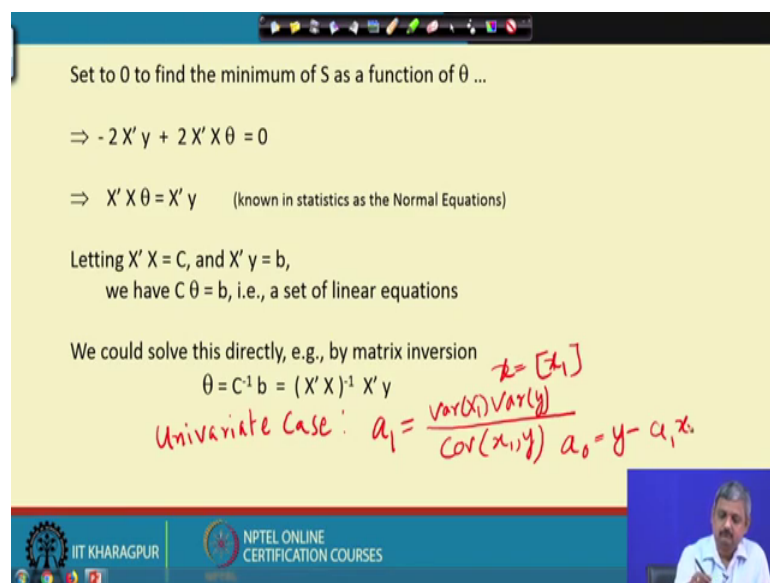
We could solve this directly, e.g., by matrix inversion  
 $\theta = C^{-1} b = (X' X)^{-1} X' y$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

This is like a linear simultaneous equation. So, that people solve by matrix inversion or someone other method ok.

So, what are the, so you can usually people do it for multivariate there are more than one attributes by the way if you if you go back to this definition if it is for a univariate so x is just 1, then your slope of the line a 1 becomes variance x 1 variance y y slope of the line and a naught you can find by y minus a 1 x.

(Refer Slide Time: 17:18)



Set to 0 to find the minimum of S as a function of  $\theta$  ...

$$\Rightarrow -2 X' y + 2 X' X \theta = 0$$
$$\Rightarrow X' X \theta = X' y \quad (\text{known in statistics as the Normal Equations})$$

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  
 $\theta = C^{-1} b = (X' X)^{-1} X' y$

Univariate Case:  $a_1 = \frac{\text{var}(x_1)\text{var}(y)}{\text{cov}(x_1, y)}$   $a_0 = y - a_1 x_1$   $x = [x_1]$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, these are some of the advantages. It is linear, it is because it is linear it is simple and interpretable and it is it is easy to design ok. So, often it is used as a base line (Refer Time: 19:10).

What are some of the limitations?

(Refer Slide Time: 19:30)

The slide is titled "Limitations of Linear Regression" and lists four main points with handwritten red annotations:

- 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? || Dimensionality

*attribute selection*

The slide footer includes the IIT KHARAGPUR logo and the text "NPTEL ONLINE CERTIFICATION COURSES". A small video inset in the bottom right corner shows a man speaking.

The actual laser might not be linear within y and x. For large dimensional problem linear regression is a problem, if x and if the  $x_1, x_2, x_3$  they are correlated there can be numerical instability and the final problem which we will discuss in our future thing is the problem of dimensionality. And you have to do something called a which attribute to use for equation. For example, if you are detecting whether some person how much loan amount he will repay the name of the person is irrelevant attribute. So, you have to do an attribute selection.

(Refer Slide Time: 20:50)

**Non-linear Regression**

- We can generalize further to models that are nonlinear:  
$$f(\underline{x}; \underline{\theta}) = \alpha_0 + \sum \alpha_k \beta_{k0} + \sum \beta_{kj} x_j$$
  
where the  $g$ 's are *non-linear functions*. ✓
- In statistics this is referred to as a generalized linear regression
- Closed form (analytical) solutions are rare.
- We have a multivariate non-linear optimization problem (which may be quite difficult!)

Handwritten notes:  $y = g(x_1, x_2, \dots, x_n)$  with an arrow pointing to "non-linear". A graph shows a curve  $y = g(x)$  on a coordinate system with axes  $x_1, x_2, \dots, x_n$ .

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, people extended this methodology to the to something called a non-linear regression ok. So, earlier we had only this much  $\alpha_k \times j$  now you can have some function  $g$  on this quantity which is a non-linear function. So, you try to find a  $y$  as some  $g$  of, where  $g$  is in the linear case  $g$  was  $\alpha_1 \times 1 + \alpha_0 \times 1 + \alpha_1 \times 0 + \alpha_2 \times 2$  and so on; so even though that when the actual data is non-linear, you would like to go for this generalized linear regression. Problem is the solutions of finding the optimal parameters of  $g$  is not so simple as in linear equation ok.

(Refer Slide Time: 22:35)

**Optimization in the Non-Linear Case**

- We seek the minimum of a function in  $d$  dimensions, where  $d$  is the number of parameters ( $d$  could be large!)
- There are a multitude of heuristic search techniques
  - Steepest descent (follow the gradient)
  - Newton methods (use 2<sup>nd</sup> derivative information)
  - Conjugate gradient
  - Line search
  - Stochastic search
  - Genetic algorithms
- Two cases:
  - Convex (nice -> means a single global optimum)
  - Non-convex (multiple local optima => need multiple starts)

Handwritten notes: Two vertical red lines.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if you want to find the best parameters of  $g$  which will fit the data you have to use more complex methods like some of these techniques, some of these techniques.

(Refer Slide Time: 23:03)

**Other non-linear models**

- Splines
  - “patch” together different low-order polynomials over different parts of the x-space
  - Works well in 1 dimension, less well in higher dimensions
- Memory-based models
  - $y' = \sum w_{(x',x)} y$ , where  $y$ 's are from the training data
  - $w_{(x',x)}$  = function of distance of  $x$  from  $x'$
- Local linear regression
  - $y' = \alpha_0 + \sum \alpha_j x_j$ , where the alpha's are fit at prediction time just to the  $(y,x)$  pairs that are close to  $x'$

The slide features a graph on the right side showing a piecewise linear function (splines) and a smooth curve (local linear regression). A video inset in the bottom right corner shows a man speaking.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

There are some other non-linear methods which instead of a single function fit piecewise linear functions ok. So, like local linear or splines they do that. The methods for finding them are even more complex, all right.

So, this is in general the methods of regression. There are some more extensions I will discuss in the next lecture, but it is a very powerful technique.

Thank you.