

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

**Lecture - 11**  
**Normal Equations**

Welcome back to the Predictive Analytics Regression and Classification course, this is lecture 3 part a.

(Refer Slide Time: 00:25)

Regression Model

# samples      # predictors

▶ Given a vector of inputs  $\mathbf{X}_{n \times p} = ((X_{ij}))$ , we predict the output  $\mathbf{y}$  via model

$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\epsilon}_{n \times 1}$ 

 $X_{ij}$ :  $i$ th row  $j$ th column  
 $X_{ij}$ :  $i$ th pred of  $j$ th sample


$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1}$

$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}_{n \times p}$

$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n \times 1}$ 
 $X_{i1} = 1$

▶ It is convenient to include the constant variable 1 in  $\mathbf{X}$ , to include the intercept.

▶ How can we estimate  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ ?






In this lecture we will focus on understanding the normal equations and how different issues of normal equations. So, we will start with simple regression model which has say  $n$  rows and  $p$  columns. Now, what I mean by  $n$  rows  $p$  columns?  $n$  is the number of samples; so,  $n$  is the number of samples and  $p$  is the number of predictors, number of predictors ok.

So,  $X_{ij}$  is the  $i$ th row and  $j$ th column; so, this is  $i$ th row and  $j$ th column. So, basically  $X_{ij}$  stands for  $j$  predictor of the  $i$ th sample,  $X_{ij}$  is the  $j$ th predictor of  $i$ th sample. Now,  $y$  is  $n$  rows; so,  $n$  responses that we have here  $n$  responses, this is the design matrix or  $X$  matrix and these are the residual errors vector.

(Refer Slide Time: 02:05)





- ▶ Many different methods, most popular is *least squares*.
- ▶ minimize the residual sum of squares ✓

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

$$= \sum_{i=1}^n (y_i - x_i^T \beta)^2 = \sum_{i=1}^n \epsilon_i^2$$

$$y = X\beta + \epsilon \Leftrightarrow \epsilon = (y - X\beta)$$

$$\epsilon^T \epsilon = (y - X\beta)^T (y - X\beta)$$



Now, many different models you can fit, but before that; so, it is convenient to include a constant variable 1s in  $X$ . Say suppose you have these are the predictors and then what you can do? You can make the first predictor as all 1s. If you do that then automatically this will be your constant predictor ok, it will be your like intercept it will behave as a intercept ok. If you just make each of like basically  $X_{i1}$  equal to 1, then it will behave as a first column will behave as a intercept, you take each of them as a 1.

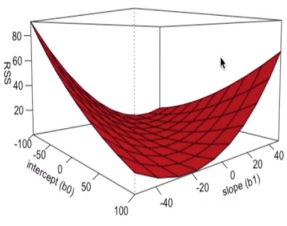
So, in this you know you all you have the data that you have is  $y$  and  $X$ . So,  $y$  responses and  $X$  predictors and we want to estimate the coefficient values; so, this is our main goal that how to estimate the betas that is what we want to estimate. So, there are many different methods to estimate the beta or the regression coefficients, the most popular is least square method.

Now, minimizing how the way least square method works is by minimizing the least square or the residual sum of square. So, why it is called residual sum of square? So, you see your model is  $y$  equal to  $X$  beta plus epsilon. So, you can write it this model as epsilon equal to  $y$  minus  $X$  beta: so, essentially this epsilon is  $y$  minus  $X$  beta.

So, if you take epsilon transpose epsilon that will be basically the residual  $y$  minus residual sum of squares  $X$  minus  $X$  transpose  $y$  minus  $X$  beta that is what you are getting ok. Now, this is same as you can write it as  $i$  equal to 1 to  $n$   $y_i$  minus  $x_i$  transpose beta is epsilon  $i$  square. So, I do not need this thing yeah epsilon  $i$  square; so, our objective is to minimize the sum of squares of error that is our objective.


(Refer Slide Time: 05:11)

Residual Sum of Square : Surface



- ▶  $RSS(\beta)$  is a quadratic function of the parameters
- ▶ Its minimum always exists, but may not be unique.

cmj



Now, what happens with the how the sum of squares of errors behave? So, if you look into the sum of squares of errors; so, this sum of squares of errors my  $y_i$ 's and  $x_i$ 's are essentially known to me. It is from the data nothing to change it is exactly constant effectively it is behave like a constant. Now, only thing is unknown to is beta if, I change a beta little bit; so, naturally this  $\epsilon_i^2$  will change.

Now, if a  $\epsilon_i^2$  also change; so, RSS of beta is a function of beta. RSS residual sum of square turns out to be function of beta. Because, if I change beta a little bit; so, residual sum of squares of beta will change completely. So, if you look into the you know the surface, if you can consider this is like you know intercept different values of intercept ok and different values of slope.

For different values of slopes and intercept, you plug it in there and then you calculate the residual sum of squares and what you get is a surface. Now; obviously, I would prefer that slope and somewhere here maybe let me try a different color maybe blue. So, I will probably try to reach here by maybe this value here and somewhere this value; so, this is the these are the values where we will get a most you know minimum value.

So, wherever the where the valley is minimum we will try to get there ok. Now, its minimum always exists turns out, we can show that it minimum will always exist, but may not be the unique and we will discuss why what I really mean by this.

(Refer Slide Time: 07:31)

▶ minimize the residual sum of squares

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

$$= \sum_{i=1}^n (y_i - x_i^T \beta)^2$$

$$\frac{\partial}{\partial \beta} RSS(\beta) \stackrel{\text{set}}{=} 0$$

$y_{n \times 1}$   
 $X_{n \times p}$   
 $X^T_{p \times n}$   
 $(X^T X)_{p \times p}$   
 $(X^T y)_{p \times 1}$



$$\frac{\partial}{\partial \beta} (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

$$\underbrace{(X^T X)}_{p \times p} \underbrace{\beta}_{p \times 1} = \underbrace{(X^T y)}_{p \times 1}$$

$(X^T y)_{p \times 1}$

$cm_i$

▶ p linear equations with p unknowns

Minimizing residual sum of square of error; so, essentially how can we do that? We can do that by just choosing differentiating residual sum of squares of beta with respect to beta. So, and set it equal to 0; so, we will differentiate and set it equal to 0. If you then what we will

have is essentially just you will set up a equation and that equation will give us the solution for if we solve that equation we will solve and that will give us the solution for beta.

Now, if we differentiate with respect to beta and set it equal to 0, what we get is typically known as the normal equation. Now, if you look into this normal equation what we have is X transpose X; now, X is a n cross p; so, X transpose is a p cross n matrix; so, X transpose X is a p cross p matrix. Similarly, so, this is X transpose X is a p cross p matrix and beta is a vector of length p.

So, here on the left side we are getting p equations and on the right side we have X transpose y. So, X transpose is p cross n and y is n cross 1 correct; so, X transpose y p cross n, n cross 1 its giving us p cross 1. So, we setting up p linear equations with p unknowns, these are the p unknowns with p linear equations.

(Refer Slide Time: 09:48)

The slide content includes the following elements:

- Normal Equations:** 
$$\underbrace{(X^T X)}_A \beta_x = \underbrace{(X^T y)}_b$$
- Re-expressed as:** 
$$Ax = b$$
- Properties:**
  - $A_{p \times p}$  is known matrix
  - $b_{p \times 1}$  is known vector
  - $x_{p \times 1}$  is unknown coefficients

The slide also features the NPTEL logo in the top right corner, the CMJ logo in the bottom right, and a small image of a person in the bottom right corner.

So, the normal equations turns out to be this is my normal equation. Now, what I am doing, I am bringing the typical popular notation of linear algebra; let us call  $X$  transpose  $X$  as  $A$  and  $X$  transpose  $y$  as  $B$ . Then we can re expressed our normal equation as  $Ax$  equal to  $b$ ; where,  $A$  is a  $p$  cross  $p$  known matrix,  $b$  is a  $p$  cross one known vector and  $x$  is actually my beta. So,  $x$   $p$  cross 1 is unknown coefficient and this  $x$  we want to solve.

(Refer Slide Time: 10:51)

▶ Suppose that for a known matrix  $A_{p \times p}$  and vector  $b_{p \times 1}$ , we wish to find a vector  $x_{p \times 1}$  such that

$$Ax = b$$

▶ The standard approach is ordinary least squares linear regression.

$$\text{minimize}_x \|Ax - b\|^2$$

where  $\|\cdot\|$  is the Euclidean norm.

▶ Solution for  $x$  is

$$\hat{x} = A^{-1}b$$

▶ What happened  $A$  is not invertible? ?

NPTEL


cmj

Now, suppose that for known matrix  $A$   $p$  cross  $p$  and vector of  $b$   $p$  cross 1 a  $p$ , vector of length  $p$ ; we wish to find a vector  $p$  such that this is the system of the equation system of equation. So, that the standard approach is ordinary least square linear regression is just minimize this square, in under the Euclidean norm and the solution will be  $A$  inverse  $b$ .



Then the question is what happened A is not invertible, what happened then? This is a big question. If A is invertible, then we are very good shape we have a; we have a solution for the system, but if it is not invertible then what we should do.

(Refer Slide Time: 11:50)

### Solution to System of Equation



- 1 ▶ If  $\text{rank}(A|b) > \text{rank}(A)$  then solution does not exist.
- 2 ▶ If  $\text{rank}(A|b) = \text{rank}(A)$  then at least one solution exists.
- 3 ▶ If  $\text{rank}(A|b) = \text{rank}(A) = p$ , that is A is a full-rank matrix, then  $A^{-1}$  uniquely exists and the solution  $\hat{x} = A^{-1}b$  is unique.
- 4 ▶ If  $\text{rank}(A|b) = \text{rank}(A) < p$ , that is A is a less than full-rank matrix, then x has infinitely many solutions. This is considered as ill-posed problem. Which solution to choose and how to choose?



Now, we have to figure out when the solution what happens if under what condition solution does not exist, under what condition solution does exist and all this situation. The first thing that happens is if the rank of A augmented b is strictly greater than rank of A then solution does not exist at all for the system of equation, for that particular system solution will not exist; so, this is completely out of the window.

Second situation is second condition is if rank of A augmented b is equal to rank of A, then you can ensure at least one solution exists there could be more than one solution, but at least one solution do exist. Third situation is if rank of A augmented b is equal to rank of A and



that is equal to  $b$  that is a full rank matrix that is very important. If it is a full rank matrix, the  $A$  inverse uniquely exists; so that means, the solution  $A^{-1}b$  is unique; so, that is very important.

So; that means, if  $A$  is a full rank matrix, then the solution we have a unique solution. Now, what and the fourth situation is if rank of  $A$  augmented  $b$  equal to rank of  $A$ , but it is less than  $p$ . So; that means,  $A$  is less than full rank matrix, then  $x$  has infinitely many solution and this is considered as a ill posed problem.

If you have this kind of situation this is considered as ill posed problems; means the solution you may have a solution, but it is you have infinitely many solution that is very difficult. That means, effectively you are saying that every possible value on the real line could be a solution, if you have infinitely many solution and that is not really helpful. So, this kind of kind of situation are being called ill posed problem and we will often face this ill posed problems.

(Refer Slide Time: 14:40)

**Theorem**  
For normal equations,  $X^T X \beta = X^T y$

$\text{rank}(X^T X | X^T y) = \text{rank}(X^T X)$

- ▶ Whatever may be your data, irrespective of that, normal equations guarantees at least one solution.
- ▶ At least one solution always exists - if you adopt least squares method.
- ▶ If  $X^T X$  is nonsingular, i.e.,  $\text{rank}(X^T X) = p$ , then the unique solution is given by

$\hat{\beta} = (X^T X)^{-1} X^T y$

NPTEL

cmj

Now, we have a theorem for normal equation one can show that rank of  $X$  transpose  $X$  is equal to rank of augmented  $X$  transpose  $y$  equal to rank of  $X$  transpose  $X$ , this is you can show. So, what does it mean? So, if you have a normal equation like  $X$  transpose  $X$   $\beta$  equal to  $X$  transpose  $y$ . So, all you have to do if you have a data all you have to do, you have to set up the normal equation and this theorem guarantees that you have at least one solution.

So, if you go back here it says that the second is true for whatever be the data whatever be the data second is true; that means, at least one solution exist; so, this is very strong. So, whatever may be your data irrespective of that you have at least one solution, the normal equation guarantees at least one solution; whatever may be your data.

Irrespective of that normal equations guarantees that is a very strong result we have at least one solution. So, I am not going into the proof of the theorem, I am leaving it for you to figure

it out it is very simple to prove this theorem, but what I am going to do? I am going to explain you the implication of this theorem.




So, at least one solution always exists if you adopt at least square method. If you adopt any other method we do not know whether the solution exist or not for that method, but for if you adopt a least square solution or least square method then at least one solution is guaranteed. Now, if  $X^T X$  is non singular that is rank of  $X^T X$  is  $p$ , then unique solution is given by just simply  $X^T X^{-1} X^T y$ .

So, this is the solution that we have in for normal equation; so, all you have to do that once you have the you know data you have to figure out set up the normal equation.

(Refer Slide Time: 17:49)

Ask yourself: Can we use Mean Absolute Deviation?

▶ What about **mean absolute deviation**?

$$\Delta(\beta) = \sum_{i=1}^n \|y_i - x_i^T \beta\|$$
$$RSS(\beta) = \sum_{i=1}^n (y_i - x_i^T \beta)^2$$


And then this is the analytical solution analytical solution for the normal equations. Next you can ask yourself can we use mean absolute deviation; so, we are using least square method. In the least square method what we are doing is we are trying to minimizing RSS of beta which is essentially trying to minimize  $y_i - x_i^T \beta$  whole square  $i$  runs from 1 to  $n$ .

It is trying to minimize this residual sum of square; now, here question is can you use can you try to minimize this absolute deviation. I would recommend that you pause the video, think about this problem for about five minutes and then come back and let us discuss. I hope you got the solution, let us discuss what can we do.

(Refer Slide Time: 19:08)

▶ What about **mean absolute deviation**?

$$\Delta(\beta) = \sum_{i=1}^n \|y_i - x_i^T \beta\|$$

▶ Conceptually no problem - certainly you can do that.

▶ we can try to develop some numerical solution

▶ But we do not have any guarantees that for sure we will have at least one solutions.

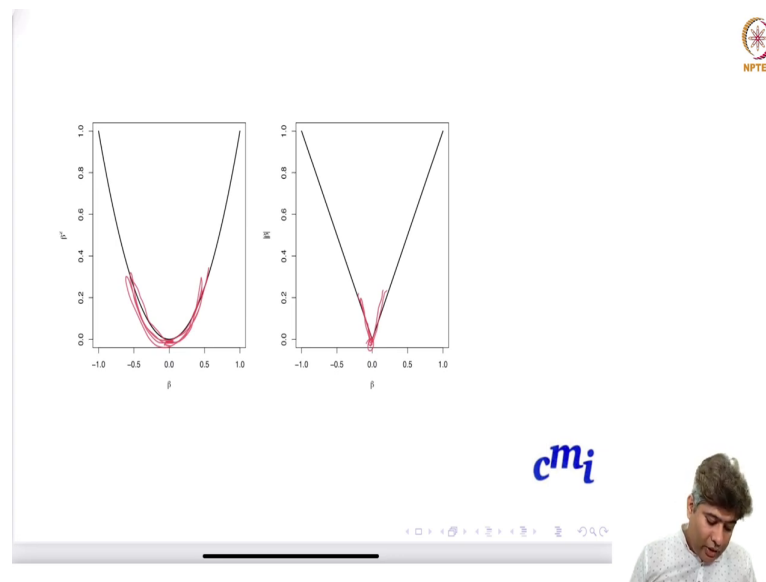
NPTEL

cmj

So, conceptually there is no problem certainly you can use this mean absolute deviation, you can do that we can try to develop some numerical solution. Some numerical solution certainly

can be developed, but we do not have any guarantees any theoretical guarantees that for sure we will have at least one solution.

(Refer Slide Time: 19:39)



Also one of the reason that it is difficult to have a analytical solution, because if you use residual sum of square it is it behaves like you know sort of a nice continuous curve the elliptical curve. So, it is differentiable at 0 it is continuous and it is differentiable at 0, but the problem with this mean absolute deviation is it is continuous, but it is not differential at 0, it is not differentiable at 0.

So, you cannot really set up a take a derivative and set it at 0 and it is because at 0 it is differentiable. And try to find the solution; so, that is the main problem with mean absolute deviation.


(Refer Slide Time: 20:42)


Example:

Suppose there are four boys and four girls and their heights are recorded as follows:

Heights	Gender
160	Boy
147	Boy
173	Boy
177	Boy
160	Girl
148	Girl
166	Girl
178	Girl

*Handwritten notes:*  
✓ → Categorical or string  
4 boys  
4 girls








Now, we will consider a problem and exercise, suppose there are four boys and four girls and their heights recorded as follows. So, some there are 4 boys and 4 girls and these are their heights.

(Refer Slide Time: 21:10)

▶ We want to fit the model

$$\text{heights} = \beta_0 + \beta_1 \text{boys}$$
$$y = \begin{pmatrix} 160 \\ 147 \\ 173 \\ 177 \\ 160 \\ 148 \\ 166 \\ 178 \end{pmatrix}, \quad X = \begin{bmatrix} \text{Intercept} & \text{boys} \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

▶ We want to estimate  $\beta = (\beta_0, \beta_1)$



So, we want to model height as a function of gender boys or girls; so, we said that ok beta naught plus beta 1 times boys. And then we are saying that if it is boys indeed then we will have a we will do a one hot encoding or will create a dummy variable for boys. Like, 1 1 1 1 first four values are 1 and then next four values are 0s and first all set of values are all 1 its create a intercept; then so, our model is height equal to beta naught plus beta 1 times boys; so, we want to estimate the beta naught and beta 1.

(Refer Slide Time: 21:51)




▶ Our model:  $\text{heights} = \beta_0 + \beta_1 \text{boys}$

▶ We want to estimate  $\beta = (\beta_0, \beta_1)$

▶ 
$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} 8 & 4 \\ 4 & 4 \end{pmatrix}$$

▶ 
$$(\mathbf{X}^T \mathbf{X})^{-1} = \begin{pmatrix} 1/4 & -1/4 \\ -1/4 & 1/2 \end{pmatrix}$$


▶ 
$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} 163 \\ 1.25 \end{pmatrix}$$





So, what we do? We compute first X transpose X which gives me 8 4 4 and 4 then X transpose X inverse is this. And then beta hat is X transpose X inverse X transpose y we solve this and this is the this is our beta naught hat and this is our beta 1 hat.



(Refer Slide Time: 22:30)



- ▶ Our model:  $heights = \beta_0 + \beta_1 \text{boys}$  ✓
- ▶ Our fitted/trained model: fitted/trained model:  
 $heights = 163 + 1.25 \text{ boys}$
- ▶ If a new test case is a boy then expected height is:  
 $heights = 163 + 1.25 * 1 = 164.25$
- ▶ If a new test case is a girl then expected height is:  
 $heights = 163 + 1.25 * 0 = 163$  ✓



Navigation icons: back, forward, search, etc.

Now, what we have here; so, this is our model; now, our fitted trained model is height equal to 163 plus 1.25 times boys. If a new test point is a boy then expected height is 163 plus 1.25 times 1, because for boys we will take 1. So, that gives us expected height to be 164.25 and if a new test case is a girl then expected height is 163 plus 1.25 times 0 which is 163.

So, that is how we can handle the dummy variable also or the categorical variable also. So, in the data set the gender is a categorical variable or gender is a categorical or string variable categorical or string variable ok. So, whenever you got a categorical variable or string variable, you create a one hot encoding or create a dummy variable for a label and then like 1 1 for boys in this case we created.

You can do it for the girls also, it will be nice exercise to do it for the girls and check how if the intercepts what the intercept values and what is the beta 1 values you are getting. And if

you are finally, getting the same answers or not check it out and we will get back in the next video.