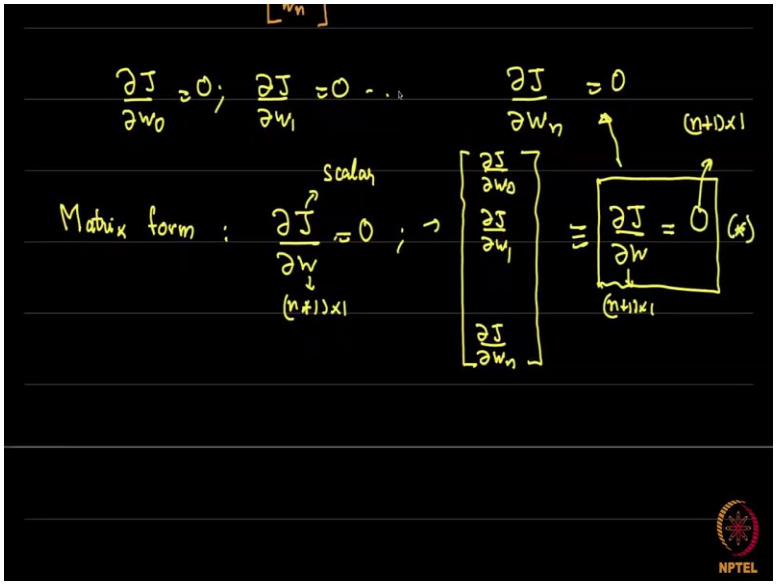
Inverse Methods in Heat Transfer
Prof: Balaji Srinivasan
Department of Mechanical Engineering
Indian Institute of Technology, Madras

Lecture No 13
Normal Equations for Linear Models (Contd.,)

(Refer Slide Time 00:21)



Welcome back. In the previous video we had seen that all linear models can be written in the form $xw = \hat{y}$, where this x was called the design Matrix and consists of all the input vectors or all the input variables. The size of this is the number of, examples are the number of sensors that you have multiplied by the number of features with which you wish to represent it, where m is the number of data points and n is the number of features and I had explained features the last time.

For example, if you have x, x itself the input variable could be made up of multiple components. For example, it could be made up of the x location y location of a specific point. more generally it could be other things also for any linear regression problem, for example, let us say you are trying to make weather prediction.

And you could have pressure and temperature as 2 components or pressure temperature humidity in case you have three features so on and so forth. W is the parameter vector and W is made up of n + 1 components. This in machine learning language is called the bias unit and

x if you remember is also augmented by a bunch of ones. So, as to multiply this and this of course is our model.

So, this problem is or this model is linear in W, that is the important point. it is linear in parameter space. It does not have to be linear in x. So, x itself could contain as we saw in the last video, it could contain x^2 or it could contain $\sin x$, it could contain $\cos x$ it really does not matter, because this is constant. it depends on data that we have given and W itself is the parameter vector and it is the parameter Vector which is the unknown, which we are trying to solve for.

We also saw that if we take this matrix formulation, we can Now write the objective function as J the objective function is equal to $W^T X^T XW$, this is a quadratic term. It is quadratic in x or it is quadratic in W sorry. These 2 are linear terms in W and this is constant and as I talked about last time to find the minimum of J, we need to minimize with respect to the variable.

The variable here is not the data points, we cannot change the location of the sensors, the location of the thermocouples or what value they have measured. But we can minimize with respect to the parameter, with respect to W, which means we are minimizing with respect to these n + 1 variables in general. Now how do we minimize. So, we minimize in the following way, we say

$$\frac{\partial J}{\partial w_0} = 0, \frac{\partial J}{\partial w_1} = 0, \ldots \frac{\partial J}{\partial w_n} = 0$$

Now this can again in the same notation, this can again be written in a form which looks like a vector. So, it can be written in the form the vector or the Matrix form is we say $\frac{\partial J}{\partial w} = 0$ notice J is a scalar, but W is a vector, $(n+1) \times 1$ vector or a matrix. So, this is the same as saying $\frac{\partial J}{\partial w}$ can be written as

$$\frac{\partial J}{\partial w} = \begin{bmatrix} \frac{\partial J}{\partial w_0} \\ \frac{\partial J}{\partial w_1} \\ \vdots \\ \frac{\partial J}{\partial w_n} \end{bmatrix}$$

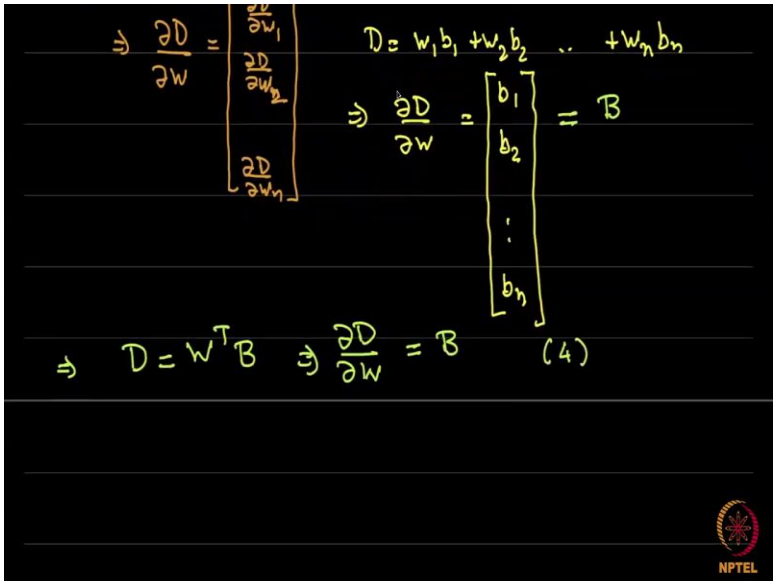
So, this is the definition of Del J by del w. So, this that we use, in order to actually set the equation.

So, we will say $\frac{\partial J}{\partial w} = 0$. This has the same meaning as the scalar system of equations. Here $\frac{\partial J}{\partial w}$ is a $(n+1) \times 1$ vector and 0 also actually is a bunch of zeros it is $(n+1) \times 1$. So, I am going to start with this equation which is $\frac{\partial J}{\partial w} = 0$. So, we have this equation, let us call this equation star and we have this original equation, let us call this equation one.

So, basically what it means is we need to find out $\frac{\partial J}{\partial w}$. Now if we look at this it is made up of three terms, the quadratic term, 2 linear terms of course $\frac{\partial J}{\partial w}$ with respect to a constant is zero. Now the way we are going to do it is first we will do the linear terms, then we will do the quadratic term, this is going to involve some amount of Matrix algebra. I hope you can follow it, even if you cannot really follow it.

Though I do have a couple of questions with this, in the exercise you should be able to understand the overall sense at least the linear terms are fairly easy to calculate. So, I hope it is a useful skill that you learn even though that is not the main point of this course. So, let us. Now differentiate these terms.

(Refer Slide Time: 07:25)



So, now consider terms of 2 forms like. So, I am going to say some Matrix E is let us say,

$$E = CW$$

and where C is a matrix and W is a matrix. So, for example C could be,

$$C = [C_1 \quad C_2 \quad . \quad . \quad . \quad C_n]$$

and W could be,

$$W = \begin{bmatrix} W_1 \\ W_2 \\ . \\ . \\ . \\ W_n \end{bmatrix}$$

It does not matter whether we start with 1 or n I am just trying to derive a general result. So, suppose I want dE_1 or $\frac{\partial E_1}{\partial W}$ which is the same as instead of calling it E1, maybe a better name we simply E just.

So, that there is no confusion I am going to call it E.

$$E = [c_1 \quad c_2 \quad . \quad . \quad . \quad c_n] \begin{bmatrix} w_1 \\ w_2 \\ . \\ . \\ . \\ w_n \end{bmatrix}$$

So,

$$\frac{\partial E}{\partial w} = \begin{bmatrix} \frac{\partial E}{\partial w_1} \\ \frac{\partial E}{\partial w_2} \\ . \\ . \\ . \\ \frac{\partial E}{\partial w_n} \end{bmatrix}$$

We are going to differentiate with respect to each one of the components. So, all we need to find out is E. Remember E is a scalar, C is a vector W is also a vector. but C is a $1 \times n$ vector or $1 \times n$ Matrix and W is a $n \times 1$ Matrix. So, E is actually a scalar, which is equal to,

$$E = c_1 w_1 + c_2 w_2 + \cdots + c_n w_n$$

So, now if we see this and we try to find out, what $\frac{\partial E}{\partial W_1}$ is? $\frac{\partial E}{\partial W_1}$ all other variables are 0 and the only term that remains is C_1. $\frac{\partial E}{\partial W_2}$ is C_2 so on and so forth until $\frac{\partial E}{\partial W_n}$ is C_n. So, this is the result of $\frac{\partial E}{\partial W}$ what is this?

$$\frac{\partial E}{\partial W} = \begin{bmatrix} c_1 \\ c_2 \\ . \\ . \\ . \\ c_n \end{bmatrix} = C^T$$

If C is the row matrix, this then is C^T. I think that should be clear. So, let us summarize the result if you have a matrix or if you have a scalar which is $E = CW$, this means that $\frac{\partial E}{\partial W}$ is equal to C^T.

So, let us label this result as 3. Now I am going to write another case. similarly consider another case. So, the other case is D equal to,

$$D = W^T B$$

Now what is B? B I am going to write as,

$$B = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_n \end{bmatrix}$$

So, B is a $n \times 1$, W^T of course is,

$$W^T = [w_1 \quad w_2 \quad . \quad . \quad . \quad w_n]$$

which means $\frac{\partial D}{\partial w}$ is

$$\frac{\partial D}{\partial w} = \begin{bmatrix} \frac{\partial D}{\partial w_1} \\ \frac{\partial D}{\partial w_2} \\ . \\ . \\ . \\ \frac{\partial D}{\partial w_n} \end{bmatrix}$$

And D is nothing but, you can now write it out $W^T B$.

So, which is,

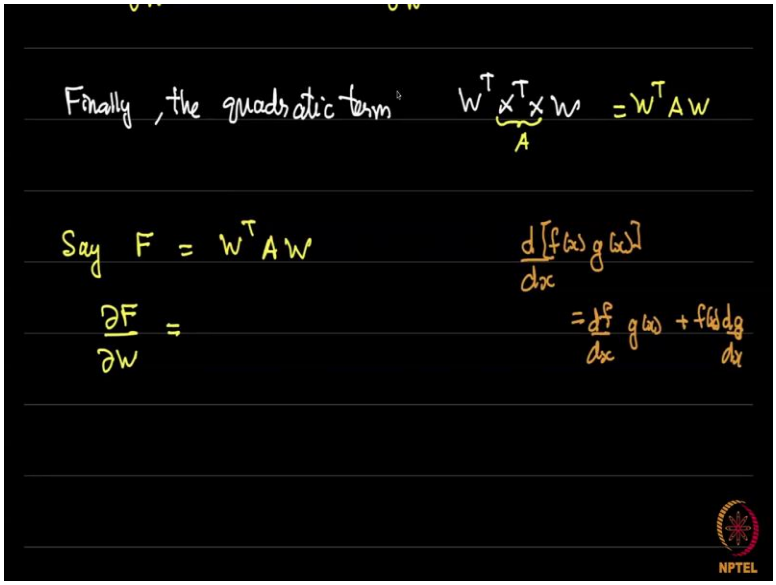
$$D = w_1 b_1 + w_2 b_2 + \cdots + w_n b_n$$

So, which means $\frac{\partial D}{\partial w}$ you can now write as differentiate with respect to w_1, it is simply,

$$\frac{\partial D}{\partial w} = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_n \end{bmatrix} = B$$

So, this tells you this is exactly the same as the original Matrix B. So, this tells you that if $D = W^T B$, then $\frac{\partial D}{\partial w}$ is simply equal to B. So, let us call this equation the previous equation was 3, this equation as 4.

(Refer Slide Time: 12:11)



So, now we are going to use 3 and 4 to solve for the linear derivatives. So, the linear derivatives are the derivatives here, these 2. So, let us look at the derivative of these 2 terms. So, the first term I am going to look at is this term here $W^T X^T Y$. So, let $E = W^T X^T Y$, I want $\frac{\partial E}{\partial W}$, I think I called it D here. So, let me keep consistent notation.

So, let this be D and let us call this term B. Now this is $W^T B$. so, we know that $\frac{\partial D}{\partial w}$ is B which is nothing but $X^T Y$. So,

$$\frac{\partial D}{\partial w} = X^T Y$$

So, this result directly leads here. So, similarly the other term, we had was the term here, this term here which is $Y^T X W$. So, let us call that E. we want $\frac{\partial E}{\partial W}$ we had the result before let us call this C.

So, $\frac{\partial E}{\partial W}$ when E is CW is something we already know this is C^T. So, this means

$$\frac{\partial E}{\partial W} = (Y^T X)^T = X^T Y$$

$\frac{\partial E}{\partial W}$ is $(Y^T X)^T$ which as you know $(AB)^T = B^T A^T$. So, this becomes $X^T Y$ again. Now notice both these terms are the same. So, this you should know with solving quadratics. for example, if you have something like x^2, you tend to get 2x when you take a differentiation.

So, something similar is going on here, but we will go further. So, now what we get is, finally let us look at the quadratic term. This I am going to do a little bit of hand waving; I will not do it in as much detail as the previous terms I am going to appeal a little bit to your knowledge of scalar calculus or one-dimensional calculus and not vector or Matrix calculus. So, the quadratic term here is if you go back here, it is $W^T X^T X W$.

So, I am going to call this quadratic term is $W^T X^T X W$. So, let us call this term as A. So, that is $W^T A W$. Now let us say F equal to,

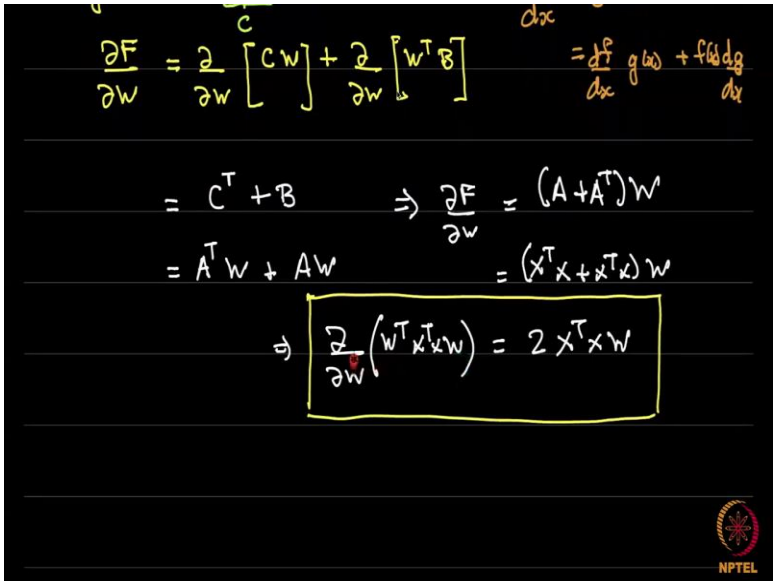
$$F = W^T A W$$

We want $\frac{\partial F}{\partial W}$. the way we will do it is using something like the chain rule. So, you know that the chain rule works this way, if you have let us say $\frac{d}{dx}[f(x)g(x)]$ the way we do it is we first differentiate this assuming g is constant.

Then we differentiate it assuming f is constant and then add the 2. So, we will say something like $\frac{df}{dx} g(x) + f(x) \frac{dg}{dx}$. So, the idea is hold differentiate this whole term assuming g is constant, then differentiate this whole term assuming f is constant, then add the 2. we are going to do the same thing here. So, we are going to treat it in the following way, we will differentiate this whole term holding this term constant.

Then they will differentiate this whole term assuming this term is constant and we will just write a sum of the 2.

(Refer Slide Time: 17:14)



So, for example I am going to call this combination, let me be consistent with my notation here. I will call this term as C and this term as B. So, if I do that this basically, becomes,

$$\frac{\partial F}{\partial w} = \frac{\partial}{\partial w}[CW] + \frac{\partial}{\partial w}[W^T B]$$

Now we know the result already when you do that. So, when you do this, you get what is derivative of C times W, we already calculated it, this C^T and what is derivative of $W^T B$ you know that that is B.

$$\frac{\partial F}{\partial w} = C^T + B$$

Now C was $W^T A$, if you take transpose of that you get $A^T W$ and B is simply AW. So, put these together you get,

$$\frac{\partial F}{\partial w} = A^T W + AW$$

$$\frac{\partial F}{\partial w} = (A + A^T)W$$

Now A itself was $X^T X$, I think that is the definition yes. So, x-transpose x, but x-transpose x, if you take a transpose of that you will again get x transverse x that is because it is symmetric.

$$\frac{\partial F}{\partial w} = (X^T X + X^T X)W$$

$$\frac{\partial F}{\partial w} = 2X^T XW$$

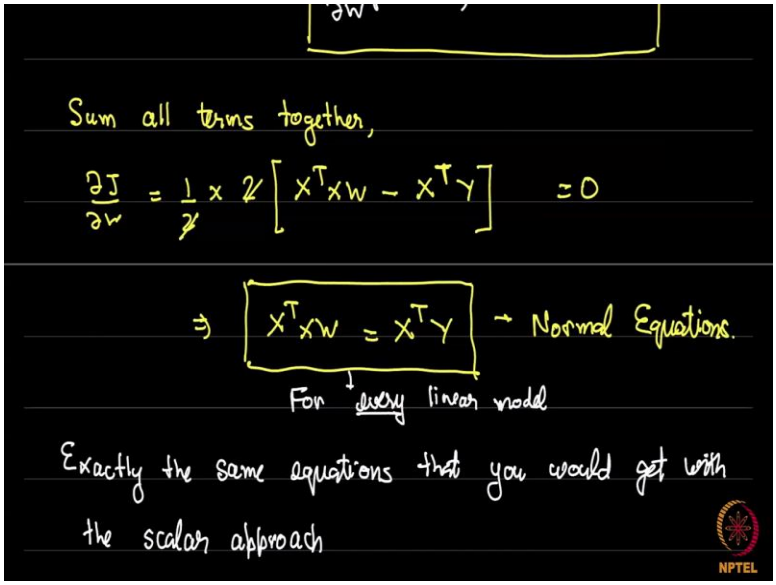
So, you get $2X^T XW$. So, this means,

$$\frac{\partial}{\partial w}(W^T X^T XW) = 2X^T XW$$

So, again like I said this is kind of a hand waving derivation, you can do a more formal derivation. but nonetheless none of what I said here was strictly speaking wrong in any way. All you need to remember again is to treat, if you treat this as a scalar. let us assume W is a scalar then it will look like $\frac{\partial}{\partial w}(W^T X^T XW) = 2X^T XW$.

So, it works in a scalar and Matrix you have to be a little bit careful, in order to ensure that all the matrices match.

(Refer Slide Time: 19:55)



So, all put together, if we now some all terms together, you will get $\frac{\partial J}{\partial w}$. Now you notice this half actually is very useful for us, this half up front. you will get a 2 from the differentiation of this term, you will also get a 2 from the sum of the differentiations of these terms. So, just look at the results,

$$\frac{\partial J}{\partial w} = \frac{1}{2} \times 2[X^T XW - X^T Y] = 0$$

once you have this 2 times $X^T XW$ you also have an $X^T Y$ and another $X^T Y$ from here.

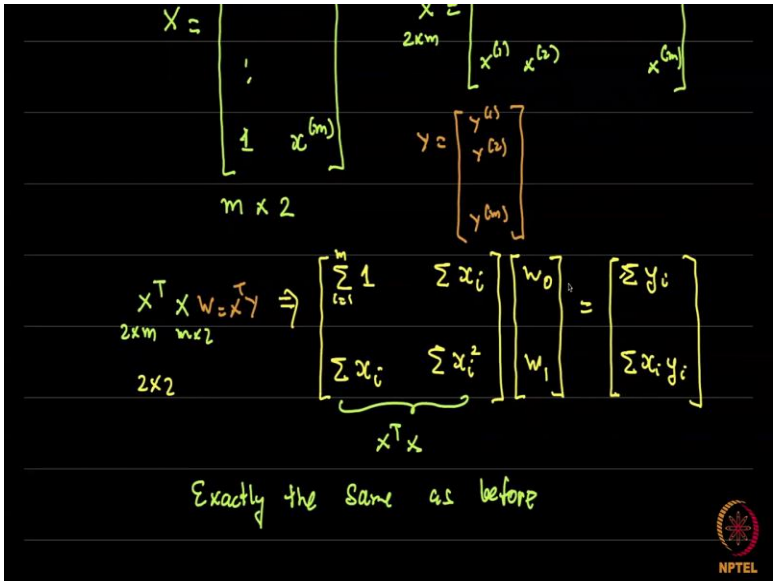
So, when you add all that. So, you will get off times that would be at 2 then $X^T XW - X^T Y$. If this is not entirely clear on how this came, I request you to just go back and look at each individual term. So, these 2s of course cancel out and this is equal to zero. So, this tells us that,

$$X^T XW = X^T Y$$

This set of equations are called the normal equations.

So, these are called the normal equations and these are exactly the same equations, that you would get with the scalar approach, that is the approach with which we derived our x-bar formula etcetera on the last time. that is exactly what you would get with this approach also except it is compact and it is one general formula for every single type of model, true for every linear model.

(Refer Slide Time: 22:28)



A few other things about this model, I would like to say here. The first is this that let us take the case of a simple linear model which we did. So, the linear model that we had was $\hat{y}^{(i)} = w_0 + w_1 x^{(i)}$. So, if we look at the design Matrix capital X in this case, this is going to be,

$$X = \begin{bmatrix} 1 & x^{(1)} \\ 1 & x^{(2)} \\ . & . \\ . & . \\ . & . \\ 1 & x^{(m)} \end{bmatrix}$$

Look at the size of x, x is an $m \times 2$ matrix what is X^T,

$$X^T = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ x^{(1)} & x^{(2)} & . & . & . & x^{(m)} \end{bmatrix}$$

Now suppose we want $X^T X$. So, this of course X^T is a $2 \times m$ Matrix. So, $2 \times m$ multiplied by $m \times 2$ is going to be sum 2×2 Matrix. Now what is that? So, let us see X^T, the first row into first column it is just a bunch of ones. I am going to write it as 1 + 1 + 1 of course it is just $\sum_{i=1}^{m} 1$, but I will leave it as $\sum 1$.

Next this location will be first row into second column you can see $1 \times x_1 + 1 \times x_2$. So, this is simply $\sum x$ or $\sum x_i$ this symmetric. So, this one is also going to be $\sum x_i$ the final ${x_1}^2 + {x_2}^2 + \cdots + {x_m}^2$. So, this is $\sum {x_i}^2$.

$$X^T X = \begin{bmatrix} \sum 1 & \sum x_i \\ \sum x_i & \sum {x_i}^2 \end{bmatrix}$$

Now the equation was

$$X^T XW = X^T Y$$

$$X^T XW = \begin{bmatrix} \sum 1 & \sum x_i \\ \sum x_i & \sum {x_i}^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

So, this is the $X^T XW$. Now we next have to do $X^T Y$, Y we already saw is simply,

$$Y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ . \\ . \\ . \\ y^{(m)} \end{bmatrix}$$

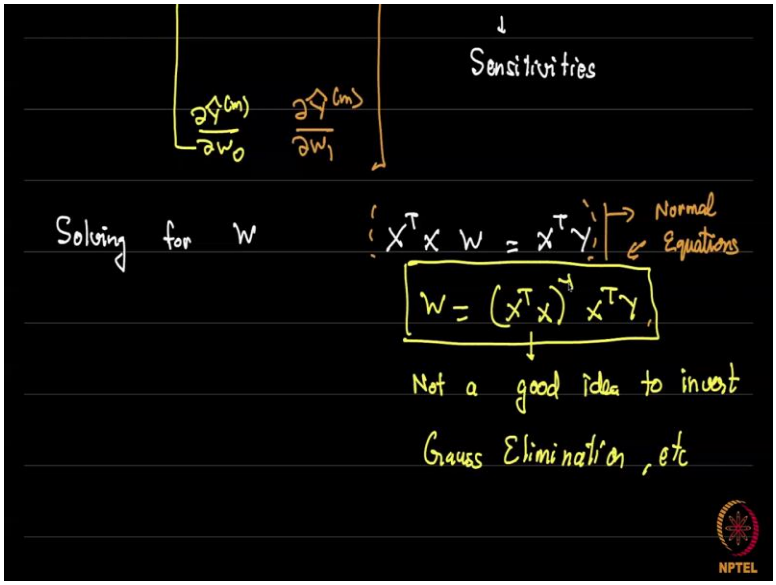
Now what is $X^T Y$ it has to be a 2×1 matrix. first row into First Column is simply $\sum y$ and second row into first column is $x_1 y_1 + x_2 y_2 + \cdots$ So, this is $\sum x_i y_i$. Now notice what has happened this is the $X^T X$ Matrix and all you got was a 2×2 you summed up across all these things and you got a 2×2.

$$\begin{bmatrix} \sum 1 & \sum x_i \\ \sum x_i & \sum {x_i}^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

In case of a quadratic, this will become a 3×3. Now notice if you go back to your notes from the last week, this is exactly the same as before.

So, this way is just a compact way. compact way means it will look like a long way. but it is compact way in the sense that you can do this one equation one normal equation derivation for every single case on earth for linear regression. So, this also is exactly the same as before. So, notice this. So, in general you are going to have an $(n + 1) \times (n + 1)$, if you have n features. So, in this case this is a 2×2 Matrix.

(Refer Slide Time: 26:30)



Now another thing about this, is x itself as an interpretation, which we will use later on when we move to non-linear regression. So, let us take the x matrix, I am going to write this again. x is

$$X = \begin{bmatrix} 1 & x^{(1)} \\ 1 & x^{(2)} \\ . & . \\ . & . \\ . & . \\ 1 & x^{(m)} \end{bmatrix}$$

$$\hat{Y} = \begin{bmatrix} w_0 + w_1 x^{(1)} \\ w_0 + w_1 x^{(2)} \\ . \\ . \\ . \\ w_0 + w_1 x^{(m)} \end{bmatrix}$$

This is w_1 should not look like w. Now if you look at the relationship between the $\hat{Y}$ and X, it is not immediately obvious but if you stare at it for a little bit longer you notice this.

That the first term here one is nothing but $\frac{\partial \hat{y}^{(1)}}{\partial w_0}$. how am I saying this notice this, this is $\hat{y}^{(1)}$. So, if I differentiate this term with respect to w_0, I get one of course if I differentiate this term with respect to $\hat{y}^{(2)}$sorry with respect to w_0, I get 1, 2. So, we I am going to write it this way the first term is $\frac{\partial \hat{y}^{(1)}}{\partial w_0}$, the second term is $\frac{\partial \hat{y}^{(2)}}{\partial w_0}$ and the last term is $\frac{\partial \hat{y}^{(m)}}{\partial w_0}$.

The next term here the next column this column you can. Now notice is x_1 is nothing but the derivative of the first term with respect to ∂w_1 you notice the w_1 coefficient is x_1. So, this is $\partial \hat{y}^{(1)}$ with respect to ∂w_1. Similarly, $\frac{\partial \hat{y}^{(2)}}{\partial w_1}$ and the last term is $\frac{\partial \hat{y}^{(m)}}{\partial w_1}$.

$$\frac{\partial \hat{Y}}{\partial w} = \begin{bmatrix} \frac{\partial \hat{y}^{(1)}}{\partial w_0} & \frac{\partial \hat{y}^{(1)}}{\partial w_1} \\ \frac{\partial \hat{y}^{(2)}}{\partial w_0} & \frac{\partial \hat{y}^{(2)}}{\partial w_1} \\ \vdots & \vdots \\ \frac{\partial \hat{y}^{(m)}}{\partial w_0} & \frac{\partial \hat{y}^{(m)}}{\partial w_1} \end{bmatrix}$$

Overall, at least in this case, we saw that $\frac{\partial \hat{Y}}{\partial w}$ with a little bit of abuse of notation this this kind of this Jacobian, you should have a few transposes here but we are going to this equals x.

$$\frac{\partial \hat{Y}}{\partial w} = X$$

It looks at least like x, whether you want to write it as $\frac{\partial \hat{Y}}{\partial w}$ we can debate but you can see that the differentiation of the $\vec{\hat{y}}$ with respect to the $\vec{w}$ is related to the $\vec{x}$. So, these quantities have physical meanings and these are called sensitivities and we will use them and we are also going to use them for non-linear extensions in the next week. The final point when we solve for W.

So, remember the equation is,

$$X^T X W = X^T Y$$

The way we solve for W is of course we can write the equation W is,

$$W = (X^T X)^{-1} X^T Y$$

So, generally not a good idea to invert. In practice we use Gauss elimination etcetera. There are other methods too. but my point is you should not build the Matrix and kind of invert it, that is generally not such a great idea we will also look at alternative iterative methods later on.

I will also show you a code later on this week, using partially this idea I am going to use some internal routines in MATLAB to do this. So, within this video we did a derivation of this some people call this also the normal equation, but it is preferable to call this equation the normal

equation. But either way you take this gigantic Matrix as long as you have a linear model you can always come up with this.

There are some other interesting things that are going on here, which I will come to when we come to the non-linear models. Now later on this week we are going to move on to certain variants and where all this one simple model is useful, you will see that in the future videos, Thank you.