Applied Time-Series Analysis
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 95
Lecture 41B - Estimation Methods 1 -2

Now, we turn to least squares methods. There is no routine called MOM in r, do not search for that. You will have to go home to find mom right, but there is no routine. You have Yule Walkers methods you know for example, for autoregressive you have a r dot y w which estimates Yule Walkers ar model using Yule Walkers method, but otherwise do not expect to see a routine called MOM.

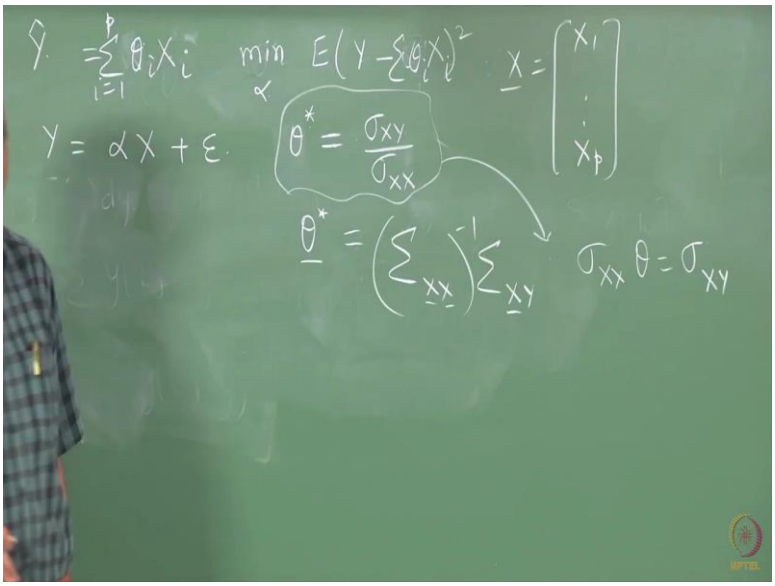
Now, we turn to the most one of the most ubiquitous method, the least squares method which we have been seeing probably from our high school and I do not have to really explain the principle behind least squares, but there are a few points that one has to remember when it comes to least square.

Now, originally if you I mean if you go to the history of least squares there is always this conflict whether Gauss proposed it or Legendre proposed it and eventually at least after both of them have live their lives. Now today mostly it is attributed to Gauss, but we never know some day Legendre may come back and in the form of some reincarnation and actually again show that know it was him who proposed the idea first and so on. So, go and read there is some nice historical masala there which will make them always history will make it more exciting and spicy.

Nevertheless, we do not worry about now who came up with this, what we should worry about is what is the method proposing; what is the principle on which it is built. If you turn to least squares; I mean the open literature that is available text books internet and so on, you will have various different presentations of the least squares method; depending on who is presenting it.

Primarily you will see this group of statisticians presenting the least squares method and then you have functional analysis people talking about least squares methods and then there are engineers talking about least squares methods and so on; each of them will give you a different presentation.

(Refer Slide Time: 02:31)



For example a statisticians would say given; I am given a random variable y and then I am given another random variable x and I would like to predict why using x or y; I would like to approximate or whatever. So, let us say I am going to write a predictor here as Y hat being alpha x plus beta will not worry about beta for now.

So, I would like to obtain an optimal estimate of alpha such that. So, find alpha such that this is minimized expectation of y minus Y hat square which means alpha x square. Many a times, you will not see this hat there, but it is extremely important to have that hat, what you are saying is you are constructing an approximation of some variable using the other variables; whether it is given in the resource or not in the reference or not at a back of your mind you should keep a cap on it alright and then you solve the problem, what is the solution to this? What is the optimal solution; the star solution?

Student: (Refer Time: 03:47)

Sorry.

Student: sigma x (Refer Time: 03:48)

Correct sigma x y by?

Student: (Refer Time: 03:53)

By?

Student: sigma x square.

Correct. So, let us write it as sigma x x; just for for a reason, so in least squares you have now therefore, a variable that you are trying to predict and a variable that you are using to predict y. Although, we are not going to solve; we are going to use this problem; this problem is useless to us in practice because when I say useless, it is not useful in when I have data with me correct because solving the so called theoretical least squares problem, this is called the theoretical least squares problem. So, I have a y which is being predicted I have an x which is being used to predict or for predicting y. This x has different names depending on the context, depending on the field in which you are looking you are working. Sometimes is excess are called regresses this is called a linear regression problem for example, you are why it is called linear because the predictor equation is linear in x or linear in alpha; what is it or is it both.

Student: (Refer Time: 05:05)

So, if I fix alpha it is linear nx, if I fix x its linear in alpha it is called a linear regression problem particularly because it is linear in parameters, I can have alpha x square also that is also linear in alpha, sometimes its x is called explanatory variable because it is being used to explain for explaining y. I have some changes I notice in some variable that is a temperature or something and then I have another variable pressure. So, I am using I am arguing that temperature is changing because pressure is changing. So, I am trying to explain why I see variations in this temperature or in variable y. So, there are terminology is called regressors then there are explanatory variables, sometimes a settle difference between these two is observed.

For example, ultimately what goes into your equation are called regressors as an example; suppose I was building a model of this form then you say x is explanatory variable, x square is regressor; the variable that you are using for prediction is your explanatory variable, but the actual form of the variable that is going and sitting there is x square that is your regressors. So, you are actually regressing why onto x square, so there are certain differences that you should observe in linear regression we do not have to break our head on it.

In another terminology, another field x may be called the independent variable and y the de dependent variable and in another field; x may be called cause and y may be called

effect. So, it there are various, there are different kinds of terminologies that are used, so coming back to the point now this is the theoretical least squares problem and many a times you will also see this; you will say that the statistician would first present the model for y, you would say the y that I have is made up of two parts alpha x plus sum epsilon and then further characterization of epsilon would also be given, then epsilon would be written as here there is no notion of time, if you notice carefully, We have forgone the notion of time, we are just sitting in the outcome space right this that is why it is called a theoretical least squares problem, a statistician go on to say epsilon is a Gaussian distributed random variable and so on.

Now, you can extend this also to multiple regressors; suppose I had here instead of alpha x, I had multiple regressors. Then how would the solution change of course, optimization problem, but we now this and I have let us say P regressors and we will use theta I instead alpha I because that is the notation that will follow thetas for parameters, how would the solution change from the scalar case to the vector case. So, now I have to look at theta vector star, what would be intuitively what is the solution as a natural extension of this scalar case.

Student: (Refer Time: 08:43)

Correct very good; so inverse of the covariance matrix of what correct. So, now x is a vector you see first you have to understand that now you are regressor; it is a vector from x 1 to x P, you are using P regressors correct, so you would say now I have a covariance matrix times.

Student: (Refer Time: 09:06)

Sigma.

Student: (Refer Time: 09:11)

Right what is the size of sigma x y; that should be easy.

Student: (Refer Time: 09:25)

P cross 1 what is the size this.

Student: (Refer Time: 09:27)

P cross P, so, simple dimensionality check theta is a P by 1 notice that I am not used a hat on theta, why have I not used a hat I am only using star.

Student: (Refer Time: 09:39)

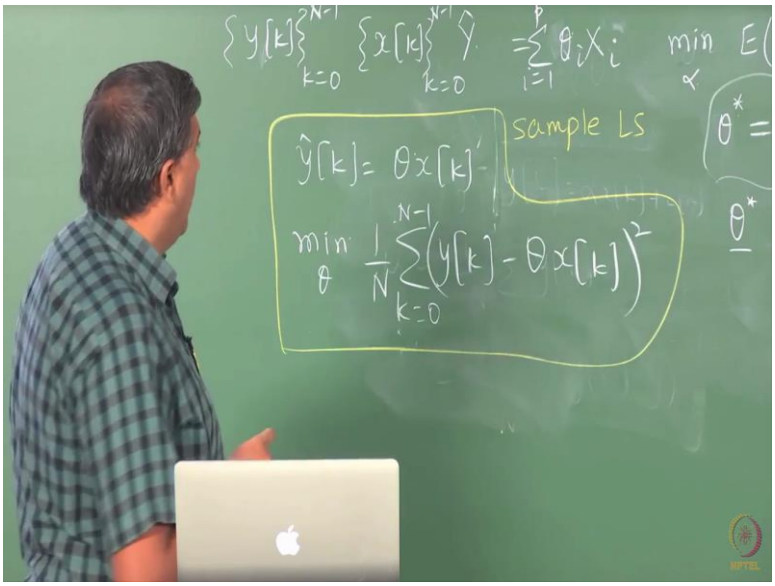
It is a theoretical optimal estimate, when I estimate this theoretical theta then I would actually put a cap on it because that is only an estimate that is why we do not have a hat, this I just a theoretical optimal solution this is called the theoretical least squares problem and this is what statistician would present for you; this is one version that statisticians would present to you, but functional analysis people would do something else and even the when it comes to setting up the least squares problem on samples of y and x.

Suppose now I were to solve the same problem, but not on the y and xs and so on, but on a observations of y and x; that means, I have data with me, I have N observations of y, N observations of x; how would I solve this problem. You may say one natural way is to use the method of moments like idea, this is not this method of moments kind of solution because this is relating if I have to rewrite this in the equation form; this equation would be sigma x x times theta equals sigma x y or you can say this itself is in the method of moments form, on the left hand side you have parameters, on the right hand side you have moments some function of moments right sigma x y is a what kind of a moment is it.

Student: second (Refer Time: 11:19)

Second order moment; very good sigma x x is also second order moment. So, it is some function of the moments and therefore, it is a method of moments approach you can say that is now from here on you can take a method of moments approach and say simply that the in practice estimate of theta would be obtained by replacing the numerator and denominator with that respective estimates. It turns outs that when you solve this so called sample least squares problem, it is not an example least square problem; sample least squares problem is the least squares problem that you would set up on the sample of y and x on the observations.

(Refer Slide Time: 12:03)



And what is that kind of a problem, that problem would be that I am given N observations of y and x, k running from 0 to N minus 1 and likewise for x. So, we are solving a very generic problem remember, so that is why I am using some generic notation.

I am given this data set and the sample least squares problem is first stating that Y hat of k is theta times x k that any observation is just a linear function the best or the prediction of any observation is simply a linear function of x, again many a times you may see this statement like this for the observations that is, you would see a statement like this y k equals alpha x k plus some epsilon and so on and then some characterization would be given on epsilon and so on.

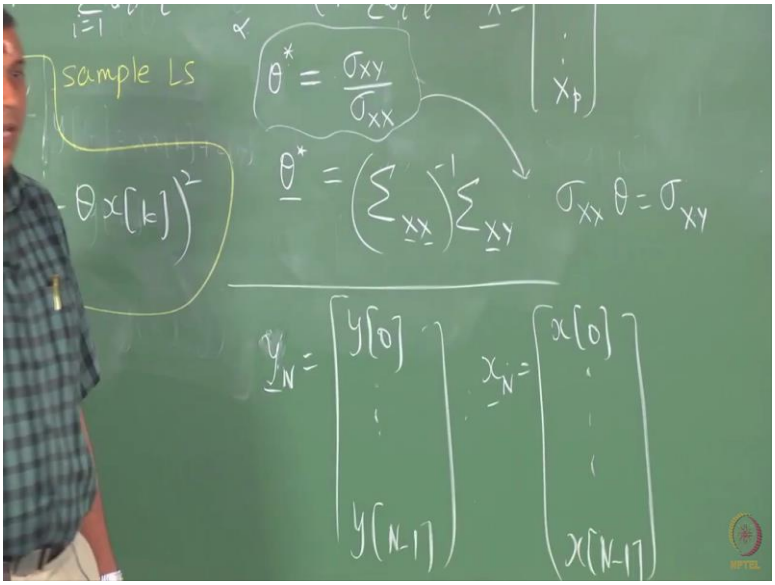
Now, at that point I would say there is absolutely no need to give any distributional characterization of epsilon. You do not have to say epsilon falls out of Gaussian white noise and so on, although many presentations will upfront impose some conditions on the distribution of epsilon and so on, that is absolutely not required as far as the derivation of the solution is concerned. That kind of a assumption is required later on when you want to make some qualifying statements about the goodness of the estimates; that is you want to comment on the bias, variance, efficiency and so on.

All those properties depend on the distribution of epsilon, but as far as the solution is concerned, you do not have to worry about this equation at all and many a times people

are left to wonder; what has Gaussianity of epsilon got to do with my least square solution; well it has, but only when you say when you talk about the efficiency of the least squares estimate. As far as the problem solution is concerned, so you say this is Y hat and now you say minimize find theta such that this is minimized 1 over N sigma y k minus Y hat of k square right. So, here k runs from 0 to N minus 1 or 1 to N as the notation may be, this is called the sample least squares problem; this is your sample least squares which is the sample version of the theoretical problem formulation that we have looked at earlier.

This is what is of interest to us in practice; now we have written, we have come up with this equation from an observation we say that there are random variables y and x and I would like to predict one using the other and then we go on to say that I have observations and therefore, I will set up the problem directly in terms of observation and so on. So, we have come from a data analysis approach, but I can also look at this problem from a vector space approach or functional analysis approach; I can say that there is a vector y k.

(Refer Slide Time: 15:48)

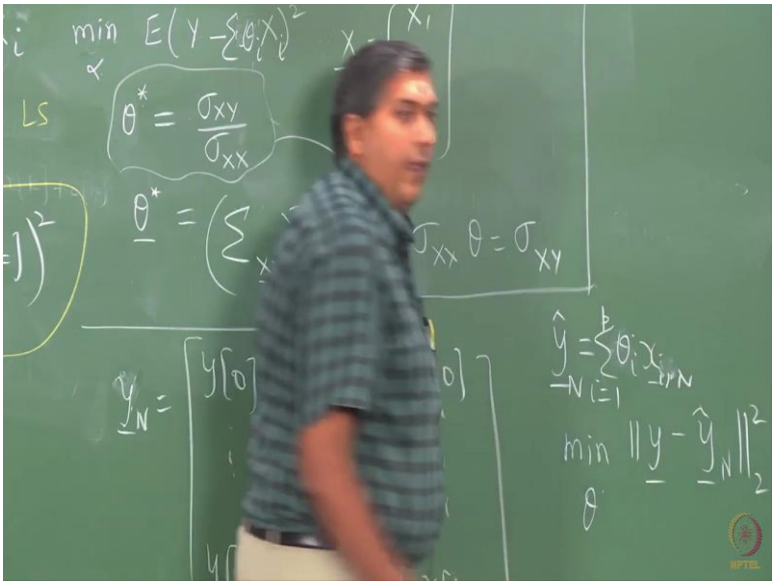


So, I can say that there is this vector Y big y N, which consist of stacked observations here and then there is another vector here x; which also is a vector of this stacked observations of x. Now I treat this y and x as vectors and I say I would like to approximate the this vector y lives in some space, vector lives x also in some space and

in that space I would like to; I know that why is made up of x and something else; that means, what we say is y lives in a higher dimensional space than x and we would like to obtain lower dimensional approximations of y in the space where x lives.

What we mean by higher dimensional space; Y contains more than what x can explain, when I say y lives in a higher dimensional space than x y is made up of x, but plus something else, I do not worry about it what it is made up of; I know for sure Y lives in a higher dimensional space, more factors are required to explain Y not just x alone that is, but now if I decide to explain Y; if I decide to approximate Y using x and the technical term in vector spaces and functional analysis is projection. So, I am going to project Y onto this lower dimensional space, what is the best projection optimal projection.

(Refer Slide Time: 17:40)



So, now I can say find that projection; assume that first your projection itself is linear. So, you say Y hat; the projection if you call; I am going to we will use N say here theta x N of course, here x is just a one dimensional, x is living in a; you can say this vector x lives in probably this N; it has N elements to it, but as far as a regressor is concerned there is only one regressor.

So, we assume that there is a theta here that will allow me to compute the projection, so these are called linear. In fact, projections are always in the linear sense only so I have here Y hat equals theta x N, Y hat is a projection of y onto x. Earlier we said regression now we are using the term projection and now I would like to compute this theta such

that this is minimized; find theta such that this squared 2 norm; what is the squared 2 norm of a vector a measure of.

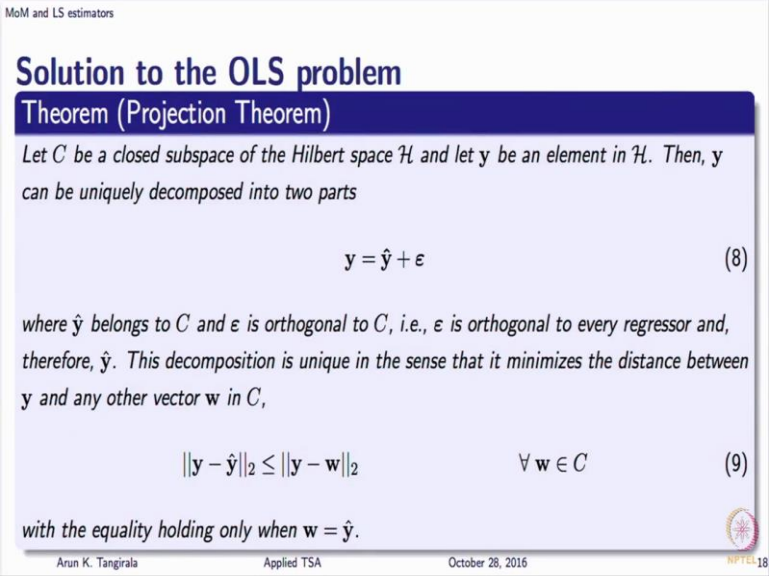
Student: (Refer Time: 18:55)

You can say it is a Euclidean distance, you can also say it is an energy or the energy contain in that vector, if you look at from an energy view point, yes if you look at norm it is essentially the distance. You can also think of it is a length measure as well correct, all in all we are minimizing the squared 2 norm of the approximation error. There is absolutely no notion of randomness at all in this entire discussion. We will not see anything on and in fact, you can see now that this is exactly the same problem as you see here, the only difference is where we started off from earlier we start from statistics and now we have started from vector spaces or functional spaces.

Now, always let me tell you this and this is something that you should always look for; whenever you have a problem formulation in statistics or in probability space, you will have expectations and you will have this terminology random variables and so on coming along and you will when you come across on optimization problem, typically the optimization problems will be in terms of either expectations like you have solved before or sample versions. You will always find the parallel between the sample versions of any optimization problem in statistics and an optimization problem in a functional analysis or linear algebra; you will always find this.

Now, we have coolly written this expression here, but you have to guaranty that these two norms exist and that is when statements like saying let Y live in a Hilbert space, what is Hilbert space, it is a vector space in which inner products are defined. When inner products are defined, norms are also defined; not all vectors need have inner products, you have to define what is this operation, this operation called squared two norm and you have to show that it exist. Rather than showing it exist; we assume that Y and x both live in in fact, Y first lives in a Hilbert space and x lives in a subset of it because we say x lives in actually lower dimensional space. So, that statement is what we will see in what is known as a projection theorem.

(Refer Slide Time: 21:39)



I am going to skip here and come back to this notation later on, but this is the theorem that we are going to solve this problem. We can solve this minimization problem in two different ways or many different ways in fact. One is to use to standard optimization trick right you can differentiate that with respect to theta and come up with the solution, that is a very straight forward thing. Of course, the general regression problem would be and I have here p parameters. So, then I would be estimating the p parameter vector that will not change and likewise here also you would say now I have p regressors right x I comma N and I runs from 1 to k. So, I can use the standard optimization tricks at the derivative to 0 get the solution or there is something called sum of squares completion, that is you break up you re rewrite the sum of squares such that you straight away see the optimal solution coming out.

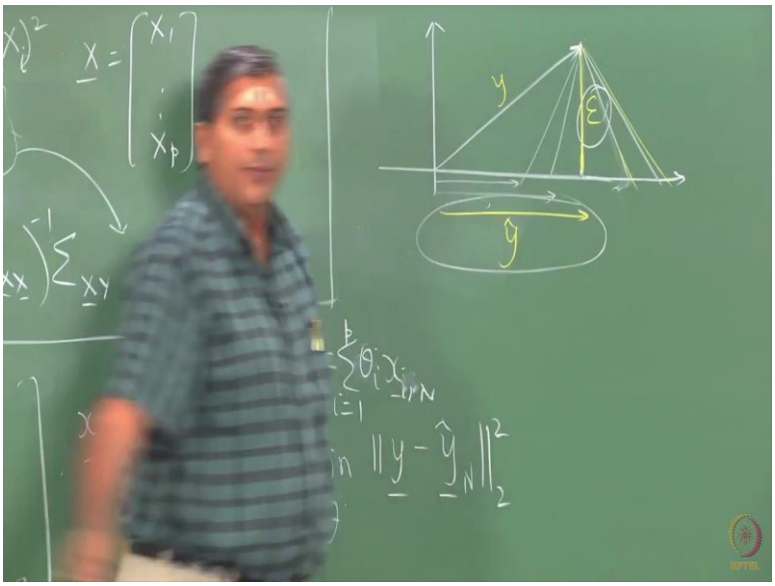
And the third approach is to use what is known as a projection theorem, this projection theorem that we have talking about now has a parallel in statistics called the decomposition theorem, that is what projection theorem says is looking at is the optimal projection of Y onto over lower dimensional subspace assuming that we are in the Hilbert space and that is what the statement says let us see be a close subspace of the Hilbert space h and let y be an element in h, then it says y can be uniquely decomposed into Y hat and some epsilon; epsilon is your pro approximation error or projection error, such that Y hat belongs to C because h hat is a projection of y onto a subset and epsilon is orthogonal to C, that is you can say that epsilon is orthogonal to every regressor that

makes up the C; that is what we mean these are 2 orthogonal subspaces how do you define orthogonality of vectors the dot product is 0 right dot product is a special case of inner product.

Now, what it says is that when you choose epsilon, when you decompose y this way; it is claiming two things; one it is claiming uniqueness; that means, there is only one solution and says that this solution has a property as given in equation 9; that is the two norm. I can decompose Y into Y hat and epsilon in infinitely different ways. If you decompose in this way such that epsilon is orthogonal to every element that makes up Y hat then the it has a beautiful property which is that it has the least two projection error in the 2 norm sense.

You can choose any other projection, you can choose any other way of de breaking up y into Y hat and epsilon. Among all such projections this one is optimal in the sense that it has the minimal length, what has the minimal length the projection error has the minimal length; among all possible projections in that is what essentially nine says. If w is another projection of Y hat y onto x is the generic projection then this particular projection Y hat that is generated in which way in such a way that epsilon is orthogonal the residual we call that epsilon as a residual, the residual is orthogonal to the projection.

(Refer Slide Time: 25:40)



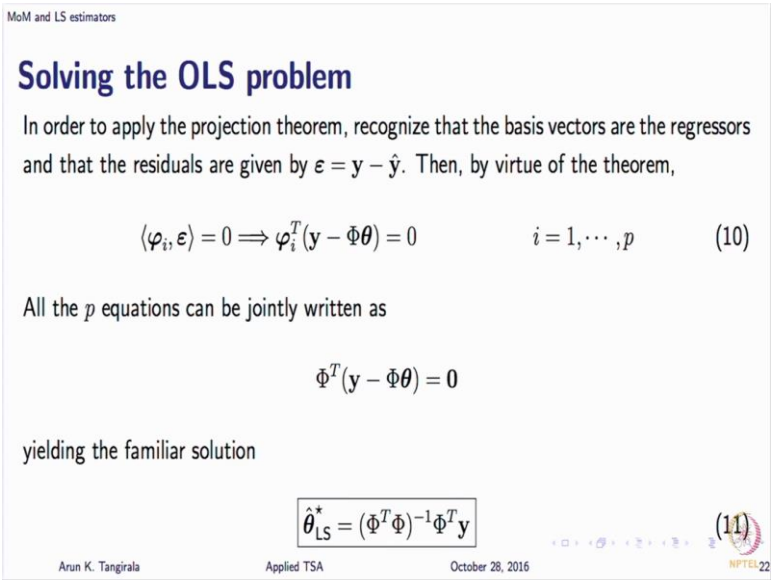
So, I use this standard example; suppose I am looking at y living in some two dimensional space. So, let us say there is a vector here and let us say I am trying to

project this vector onto the single dimension then I can project what I mean by project is approximating this vector which is in two dimensional space with some vector here of certain length right. I can project this is one projection then I could have another projection, this is another projection and so on.

See projection the other name for projection is shadow, as I always say when we are walking in the day light or sometimes even in full moon; we are three dimensional bodies there are people who are trying to be two dimensional, but that is what is called some 0 something I do not know, but you project, if you look at the shadow it is a projection of our body onto the two dimensional road; that is essentially projection, it does not carry the entire information, but it carry some information.

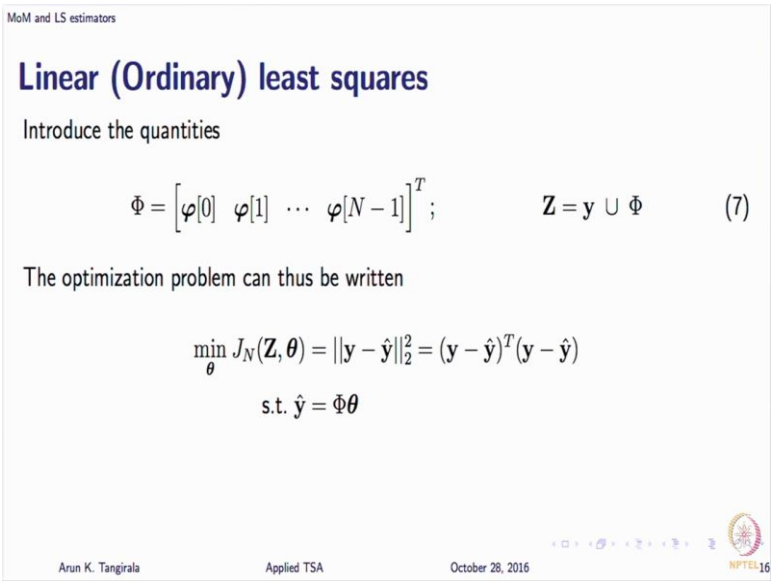
So, your Y hat does carry some information about y and there is something left out which is epsilon. So, these are all the different projections depending on the angle that you are looking at, how the light is focusing on you have different projections. It says that projection is optimal in the two norm sense which generates; so this one here, this projection here is optimal in the sense because in the two norm sense because this epsilon now this is Y hat, this is Y; Y has been broken up into Y hat and epsilon, this epsilon here has the minimal length if you can see you can also project this way you can do that, but in all such projections the epsilon has more length and what least squares is speaking up is this solution that is all.

(Refer Slide Time: 27:54)



So, we use this property to derive the solution when we meet next week, I will show you that ultimately the solution turns out to be this theta hat least squares; that is now we are using a hat because you are working with observations is phi transpose phi inverse times phi transpose y. What is this matrix phi that is where I will take you back very quickly and then we will adjourn.

(Refer Slide Time: 28:13)



So, this phi here is nothing, but I am sorry just missed it. So, phi here is instead of x I am using psi here that is only difference, this phi is just a collection of your vector of regressors. Remember you have regressors, you have observations of P regressors at every instant in time. This phi is just a collection of those regressors, so the psi that you see here is instead of x the psi vector. So as a result of which your phi matrix N by p, it has p columns and N rows and theta is a p by 1 and your y vector is an N by 1 that is all.