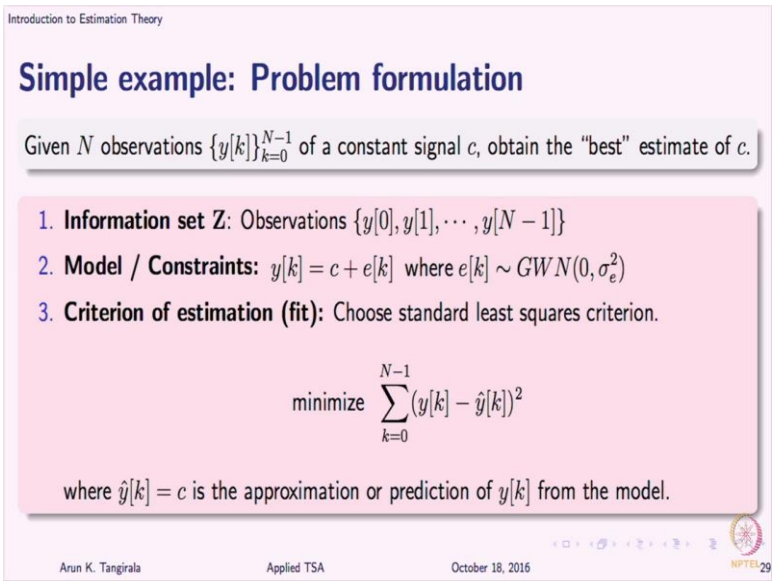
Applied Time-Series Analysis
Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture - 79
Lecture 35 B - Introduction to Estimation Theory 2

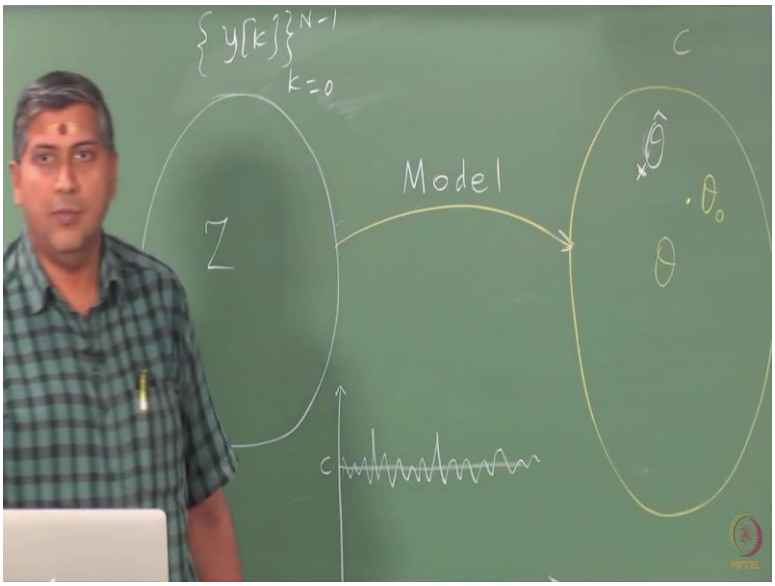
So, having now given this three perspectives let us get to the problem itself. So, the problem statement is as follows.

(Refer Slide Time: 00:20)



Given N observations, meaning N measurements of this signal C find the best estimate. Now obviously, we have resigned ourselves to saying that I will find only the best estimates, I will never be able to get the truth because of the uncertainties that I have, I will never be able to go past completely that uncertainty.

(Refer Slide Time: 01:30)

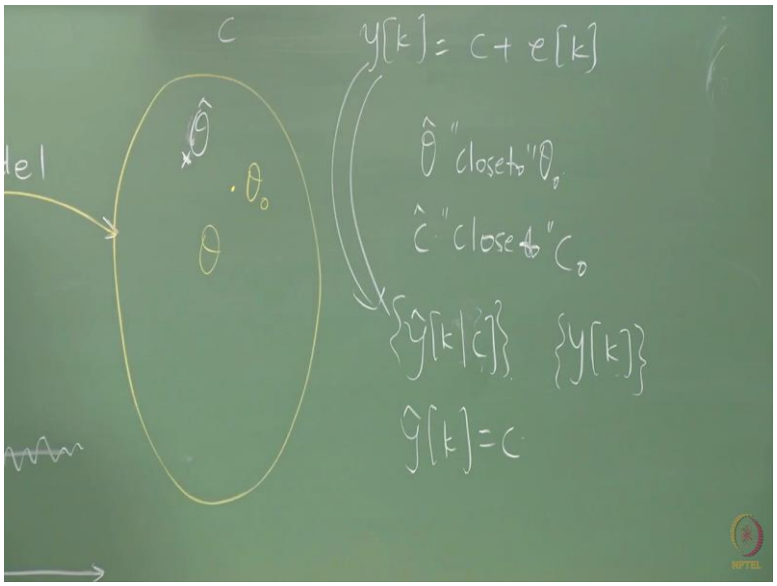


.

So, what do I have with me? With respect to the schematic that we just saw earlier, we have the information, what is the information that we have? The set of observations; we will denote the observation by y and we have N observations from y 0 to y n minus 1. The next that I need is a mapping, that connects the set of measurements here I have used z; z is a very generic symbol here we are using y no worries. I need now a model that connects is measurements to the unknown space. What is unknown for me here? The constant signal c.

So, I have y k observations k running from 0 to n minus 1 and the unknown parameter is c. Now I need a mapping from the known space to the unknown space then only I can recover. You can think of even a calibration, lot of you must have come across calibration. Calibration is also essentially an estimation problem. So, here you can think of it as calibration problem also, there are so many flavors that you can give to this problem.

(Refer Slide Time: 02:12)



So, the model now that we have written there is y equals C plus e k. Now there is a big assumption, of course I have stated that e k is Gaussian white noise at the moment Gaussianity is not required, but what definitely maybe required is the white noise assumption. So, that is a big assumption that I have made, no body told me I am making that assumption. Is there a way to check whether this assumption is correct?

Student: (Refer Time: 02:44).

Very good, you just look at ACF of y you will know whether this assumption is correct or not. In the simple problem life is easy. In more generic problems general problem scenario you may not be able to really verify the assumptions up front, somewhere down the line into the estimation problem you may be able to verify, something will tell you made a wrong assumption. Anyway, in this case at least assume that I have verified and I have therefore put up this model. But, why should I assume that the constant signal adds on to the white noise it could be a multiplier also or it could be some complicated model. So, there are several things that you can actually do here when it comes to the model. And the model plays as critical role as the objective function that we will shortly talk about. In the final quality of the estimates, the quality of the final estimate that you get. But to keep things simple we have chosen to work with an additive model, we are assuming that this is how the unknowns are entering the measurements.

This is some phrase that you have to get used to how the parameters enter the model. So, here the parameters are entering the model in this fashion. And let me tell you that many problems become simpler depending on how you assume the parameters enter the model. Simple example is; suppose I am estimating the parameters of AR model, it becomes easy to estimate the parameters of an AR model because of the way they enter your model. We have talked about it earlier the estimation of MA models is lot more difficult than estimation of AR models, simply because of the nature of the estimation problem. That is got to do with how the parameters or the unknowns are entering your model. Here the parameters there is only a single parameter C and it is entering the model this way.

So, you can come back and revisit this model if you do not get a good estimate for example, you may have made a wrong assumption. Then we need a criterion of a fit. So, earlier we said that we would like to drive theta hat close to theta naught, but that is an imposed expectation. On at least prim of AC I cannot expect theta hat to be driven close to theta naught just like that, why because I do not know theta naught. In fact, if I know theta naught none of this is required.

So, somehow I have to do something without knowing theta naught I have to drive theta hat as close as possible to theta naught. So, I would like to minimize the distance between theta hat and theta naught. It should be very close to theta naught, but I cannot do this because I do not know theta naught. So, what is the way out for me? The way out for me is to find something that is representative of theta hat and that is something that is representative of theta naught and drive them close with the hope that if you drive those very close to each other theta hat will also be driven close to theta naught.

Remember theta here is c. So, I would like to let us say C naught is the value I would like to drive C hat very close to C naught I cannot do that, but now I will chose some representative of C hat something that contains information about C hat, and also something else that contains information about C naught the truth. And drive them very close with the hope that if they are driven close to each other the estimates and the truth are also of the parameters are also driven very close to each other.

So, what is it that I have with me which contains information about the truth? Ultimately, I have to work with knowns only. The measurement is what I have, measurement

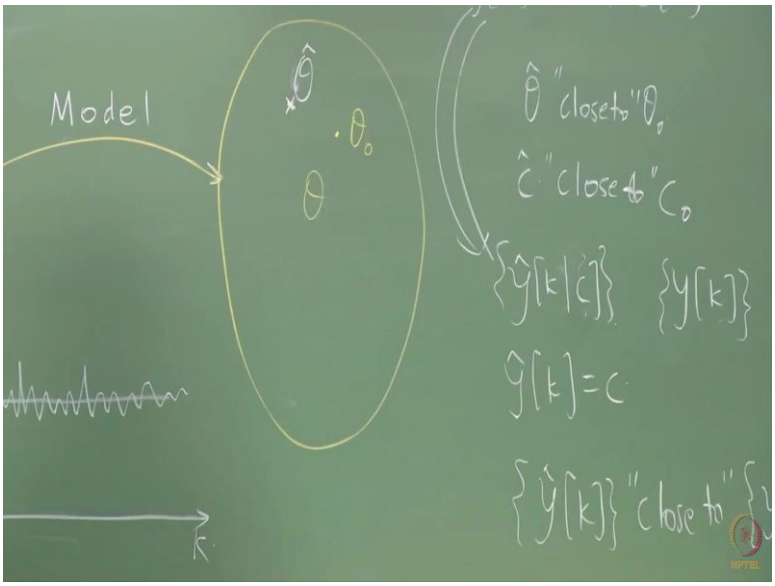
contains information about the truth. So, the representative for C naught is y k, the measurement cancels. It contains a truth in what way? In the way we have written here. We do not know what is C naught, but we know y. Now I have to find a representative for C hat. Suppose I have an estimate of C what will I do? I will actually use that to construct an estimate of y. That means, I will actually make a prediction. Ultimately why do I want to know this C? Ultimately, I want to make the prediction of next one, of the next measurement that is coming in.

So, a representative of C hat would be the prediction of y or the estimate of y; let me say which we do not have we have to construct. In fact, I would go back and say this measurement need not be bounded by this model at all, this measurement we know contains information about C naught. We are assuming that the parameter is actually affecting the measurement in this way, that is only our assumption but the reality maybe much more complicated. But the fact is the measurement contains truth, it contains information about the truth.

Now the idea is that we will drive the prediction of y or estimate of y very close to y, as close as possible to y with the hope that C hat then will be driven very close to C naught; which means I have to construct y hat from C and that is where the model plays a critical role. The use of the model is in producing y hat and that is why whatever you write down as a model has an enormous impact on the quality of the final estimate, that is how close you are to the truth because the model is going to determine what kind of estimate you are constructing of the measurement or prediction you are constructing.

So, strictly speaking this should be y hat of k given C. That is if I were to be given C how would I construct y? What would be the answer? C itself, right in this simple example C itself because that is the best prediction; so y hat of k is C. Do not think y k is C naught that is not. Y k contains information about C naught and contains something else also that I am unable to get rid of. Now you should understand the role of the model. In general the role of the model is in generating this prediction of the measurement and the idea is now to drive the prediction of the measurement as close as possible to the given measurement with the hope that C hat is also going to be driven very close to C naught, whether it actually does we will analyze later on, but the moment we will do that.

(Refer Slide Time: 10:59)



In other words, now we have changed our problem of driving theta hat close to theta naught or C hat close to C naught to driving y hat of k very close to as close as possible to y. So, you get the idea here, the original problem of driving theta hat very close to theta naught has now been converted to another problem, but this problem is now well posed because it is in terms of knowns. The earlier problem was not in terms of knowns, that is it.

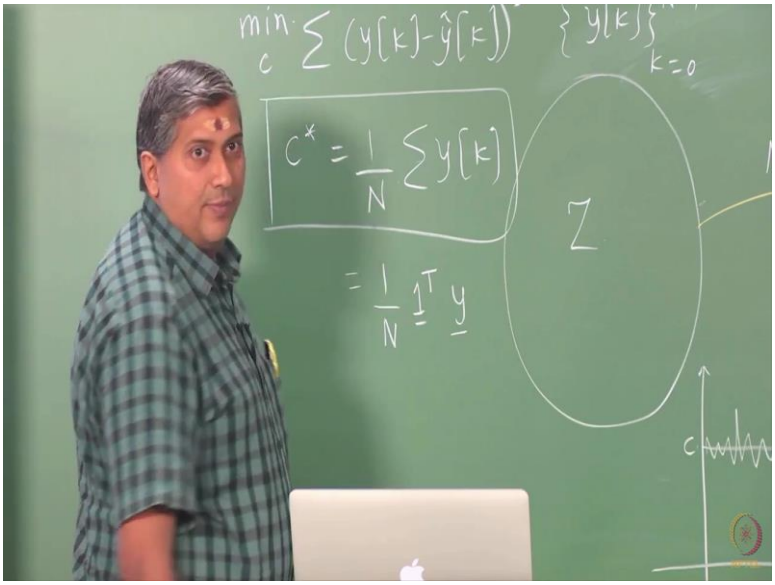
So, now I have to choose a measure of distance, and there are number of measures. We will choose a most common one the squared Euclidian distance. We can choose a Euclidian distance, but will choose a squared one because it is mathematically more friendly to optimize. So, we say now that will minimize the Euclidian distance the collective squared distance between y hat and y. What I mean by collective is, all N observations, we are not focusing on an individual observation. In fact, you should go back and ask suppose I pick just a measurement y k one single observation and I say that that is the estimate of C how bad that estimation would be.

That is I have N observations with me I close my eyes I pick just one observation and then say- here is an estimate of C, do you think it will be a bad estimate or good estimate or no problems. I just randomly pick one observation because I know that y k contains C plus some white noise uncertainty. So, a decent estimate of C would be just one observation, what could be wrong with it? Think about it, why I am a not doing it. And

why are we now using all these N observations in this manner, we will talk about it a bit later, but its good time to start thinking about it.

So, coming back to the problem we want to minimize now the distance between y k and y hat of k. This is a standard least squares problem. Can you get me the solution? Y hat of k is C, I hope you know the basics of optimization this is your objective function.

(Refer Slide Time: 13:36)



So, we want to minimize this sum y k minus y hat of k square by fine tuning c.

Student: (Refer Time: 13:43).

Sorry.

Student: (Refer Time: 13:46).

Did you actually work it out?

Student: (Refer Time: 13:51).

This is your intuition.

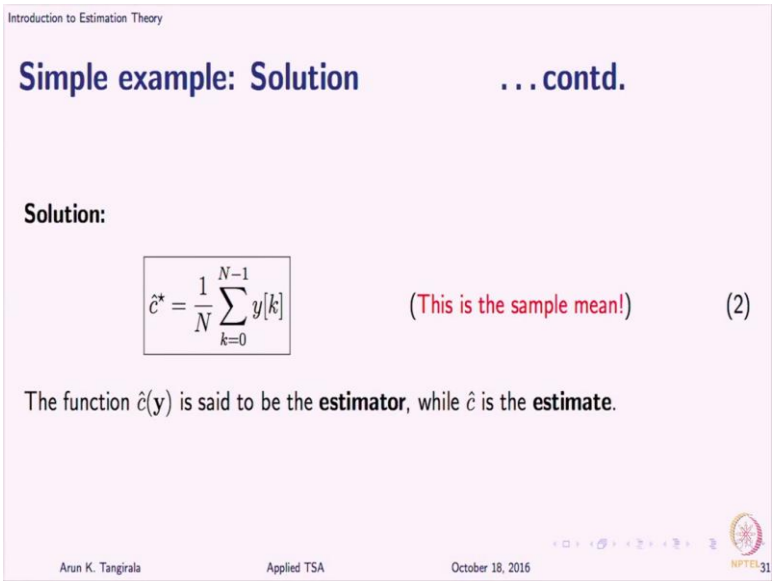
Student: (Refer Time: 13:57).

You get that, your answer is correct I just wanted to make sure that it is not through sms n, it is actually through derivations. So, if you work out the optimization C star all

optimal variables are stars denoted by stars. So, there you go the sample mean that we talk of happens to be the best estimate of c, but now we know through this problem, I mean this formulation that the sample; I mean you we could have said this early on itself. We said the statistical viewpoint of this problem is estimate the mean, you could have simply said why do not you use sample mean; yes, we could have used sample mean upfront and dismiss the problem, but we would have missed out on a number of important aspects that is point number one.

And point number two, now I can say in what sense sample mean is the best estimate. I t is the best estimate in the least square sense. What does it mean? It means that if I change the objective function, if I change the distance measure the solution is definitely likely to change. That is why is very beautiful problem, very simple problem but has all the flavors and you know all the settle points that you can observe in this problem.

(Refer Slide Time: 15:31)



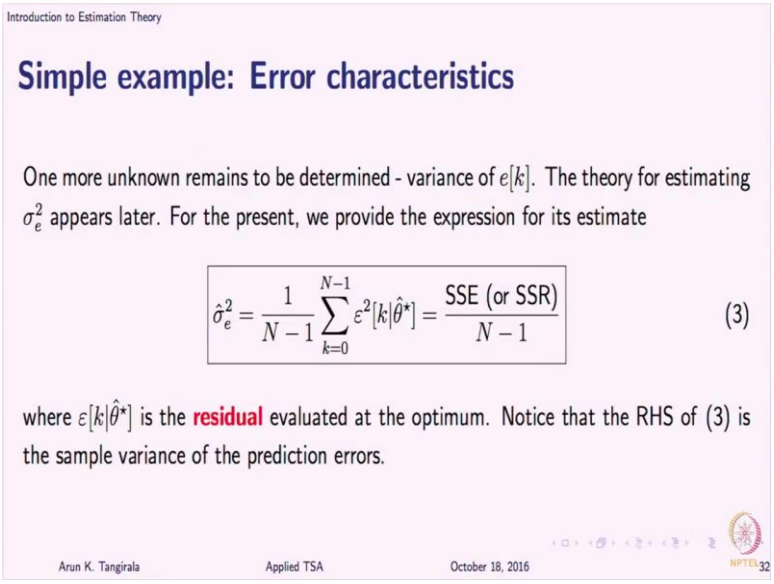
Anyway, it turns out that; now this is the optimal solution and it is the; this is called the estimator now for you this is the formula. So, in goes data out comes C hat, C hat stuff. So, it estimator has taken in the data and also taken in the model, the objective function everything and produced this estimate for you. We call this formula here as an estimator. Is this a linear function of the observations or non-linear function?

Students: (Refer Time: 16:03).

Of the observations, correct. Remember that you can write this as a nice vectorized form you can say it is 1 over n times is 1 vector transpose times y vector. So, you can actually write this as 1 vector transpose; 1 vector is the vector of one's of length n always vectors are column vectors by convention. And then you have y vector which is the vector of observations. From this you can say it is a linear estimator.

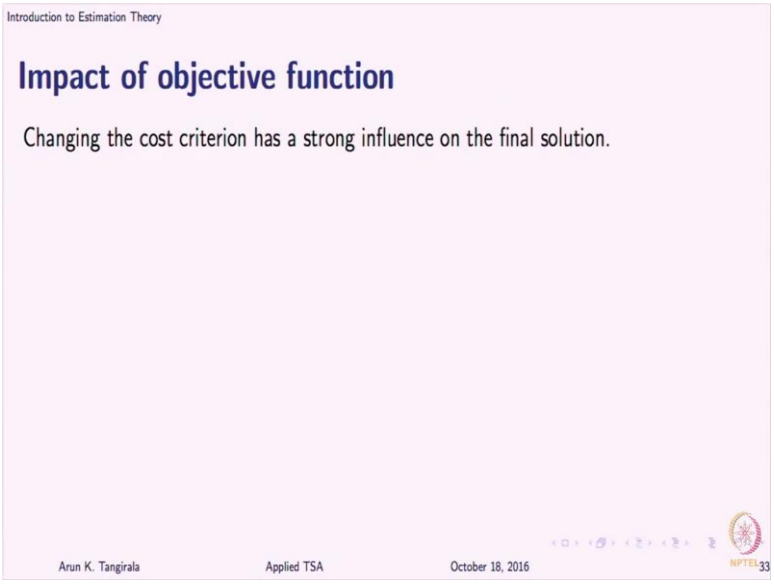
There are several advantages of working with the linear estimator. We will discover that our as we go along, but one of the most obvious advantages is implementation. If I were to implement this online it becomes very easy; coding in any hard device also it is very easy.

(Refer Slide Time: 17:05)



(Refer Slide Time: 17:06)

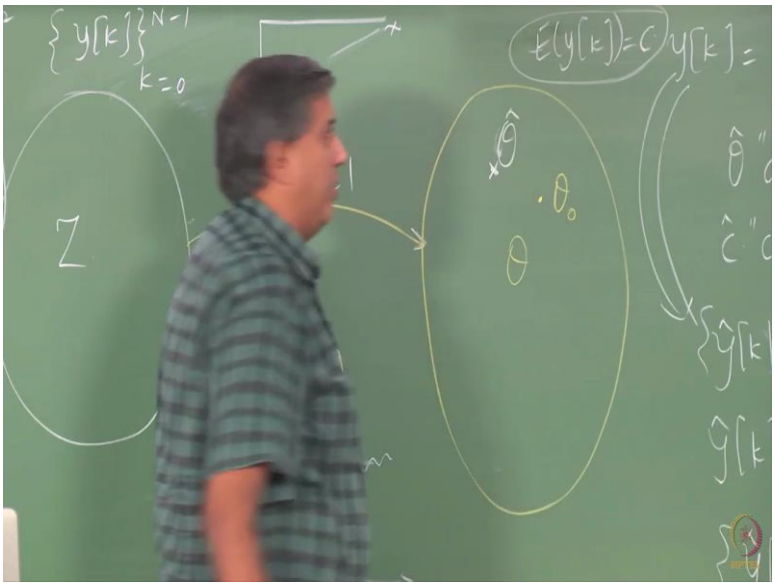
.



So, now let us we will come back to sigma square e that is also another thing that we have to estimate, but we will come back to that. Let us ask what happens if I change the objective function? What happens if I change my metric of measure of distance from squared Euclidean to 1 norm. We are moving from square 2 norm to 1 norm. This 1 norm distance is also known as, there is absolute but there is another name to it; it is called the taxi driver's distance.

There is still a name for auto drivers distance people are not able to figure out, because auto drivers thinking is lot more squid. And it depends if you are in Chennai maybe you can come up with a norm we are yet to find a mathematical function that can reflect the auto drivers distance. Taxi drivers are supposed to be a lot more straight forward, at least in New York they say and that is why it is also known as Manhattan distance.

(Refer Slide Time: 18:11)



So, this when you are going from point a to point b from here to here this is your straight distance, but then maybe this path is closed so you have to go this way or maybe go if you are in Bangalore you go all over the place and come because its one way. So, these distances are called Manhattan distance or the cab drivers or the taxi driver's distance and so on it is a 1 norm measure. Suppose I choose to minimize the 1 norm distance instead of the squared Euclidean distance, what is the solution to this problem?

Student: (Refer Time: 18:43).

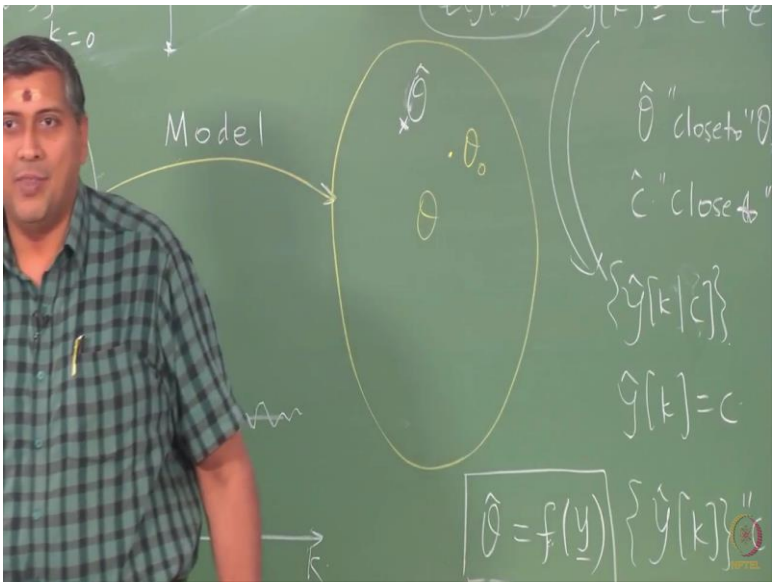
How do you know? It is a 1 norm minimization you cannot use your straight forward it is convex, but it is not possible for you at least on paper write now to go through the standard setting the derivative of the objective function and so on. But yes, it is correct the answer is median, when I say median here sample median. You can prove that the solution to this 1 norm minimization is sample median. How do you calculate sample median? Simply sort the data and pick the middle value.

Now look at the beauty, with the change of objective function we are able to generate a completely different estimate. And the nice part about this is now I know sample median is optimal in the 1 norm sense. Earlier we discovered that sample mean is optimal in the least squares in the 2 norm sense, now I know that sample median is the best estimate of the average. Remember, under this model what is the average of y C, and this is what we are planning to estimate I mean we are being intending to estimate. So, I can use sample

mean, I can use sample median to estimate the average. This is what I have being saying that truth is one, but estimators are many. And at the estimated depends on what objective you have what criterion of estimation you have. So, with the change of objective function I am able to change the estimate.

Now the question again is, is this sample median a linear estimator or a non-linear estimator? Linear with respect to observations; what makes it non-linear? I mean that is a partly right answer, but I want a much more technically better answer. How do you determine if a function is linear or not? See ultimately your theta hat is a function of the data.

(Refer Slide Time: 21:08)



Remember, theta hat is a function of y. It is this function that we are out to determine, we have already determined two such different functions: one happens to be the sample mean and the other happens to be sample median. How do you determine whether a function is linear or not? Super position principle; so you should apply the idea to sample median and see if sample median qualifies to be a linear function or not.

Somebody said non-linear, but then after that fell silent. Intuitively you may have a feel, but I would like the technical answer to come out.

Student: Linear.

Linear, answer [FL] even then I am going to ask why.

Student: (Refer Time: 22:05).

Does it satisfy.

Student: (Refer Time: 22:12).

No why do you think that median satisfies the supervision principle.

Student: (Refer Time: 22:18).

Not even, what is involved in calculating median we just discussed about it.

Student: (Refer Time: 22:32) always the same (Refer Time: 22:40).

Ok, but can we put it in more simple terms?

Student: So, the medium of the sum I thought is the median of individual.

What that the heart of calculation of median?

Student: Sorting.

Sorting is one of the standard algorithms that is used. Suppose someone asks you sorting is a linear operation or not, what is your answer? Yes or no? If you have two vectors let us say and you add them up, you have two vectors: y 1 and y 2 you are going to add them up and sort. Is it going to be the same necessarily as the sum of the sorted y 1 and sorted y 2? That is it, it is as straight forward as that we do not have to really worry about anything else. That is what makes it non-linear.

So, the sample median is a non-linear function, is a non-linear operator, non-linear function whatever you want to call it as. So, observed that with it change of objective function we have moved from a linear estimator to a non-linear estimator. And that kind of gives you an idea of what objective functions in a estimations can do to your estimation problem. And why are non-linear estimators painful? Yes, they are. Are they bad? No, they are good. They need not bad, but in painful in what sense? Well in some sense online implementation, but more so in analyzing how good this estimates are.

We have not yet talked about the goodness of estimates, when we will shortly mention those. But to theoretically analyze how good these estimates are, linear estimators are a

lot more amenable and friendly as against non-linear estimators. But if you have to live with the non-linear estimator then so be it. You cannot say no theoretically they are very difficult unless I will not use it, you cannot do that.

Now that is one of the demerits of working with a non-linear estimator. But in this example, suppose you were looking at robustness of on estimator; do you understand what is robustness of an estimator robustness with respect to out layers. Some data takes on some observation point takes on some unusually high value or low value extreme values which among the sample mean and sample median is robust to it.

Student: Median.

Median, right. So, sample median is that teacher in the class who can tolerate even three or four highly mysterious students, whereas sample mean wipes like crazy. Just with one student can make it cry make this sample mean cry. In robust statistics there is a terminology called break down point, very right fully so. This sample mean simply breaks down looking at one mischievous student. One out layer can actually take it out of the stadium, whereas sample median stays where it is. Like this short balls that people talk of in cricket right, one short ball some cricketer is out, but there are other cricketers who can take any number of short balls.

So, there are different properties of an estimator. Now you can see we are looking at different aspects online implementation, ease of analysis, robustness, it turns out that and then there are a few more very important properties like consistency, efficiency and so on; it turns out that a single estimator cannot have all the desirable properties. In estimation theory it is impossible to have an estimator that is efficient, consistent, robust, easy to implement and gives you optimal estimate in every sense that you can think of impossible you have to sacrifice some properties for the other.

And that is what also you get to learn from the simple example. Sample mean is very sweet, its very cute very, simple addition man. What more complicated things can get I mean it this is the simplest you can think of. But this poor thing actually breaks down at the sight of even one mysterious data point, just wipes like crazy. But sample median is very robust it says does not matter, I will not budge from my estimate let there be some 2 or 3 even 50 mysterious elements I am not going to be budging from my place. But it is

non-linear, so that is the price that you pay. So, there is a price that one has to pay, you have to choose what is important to you.