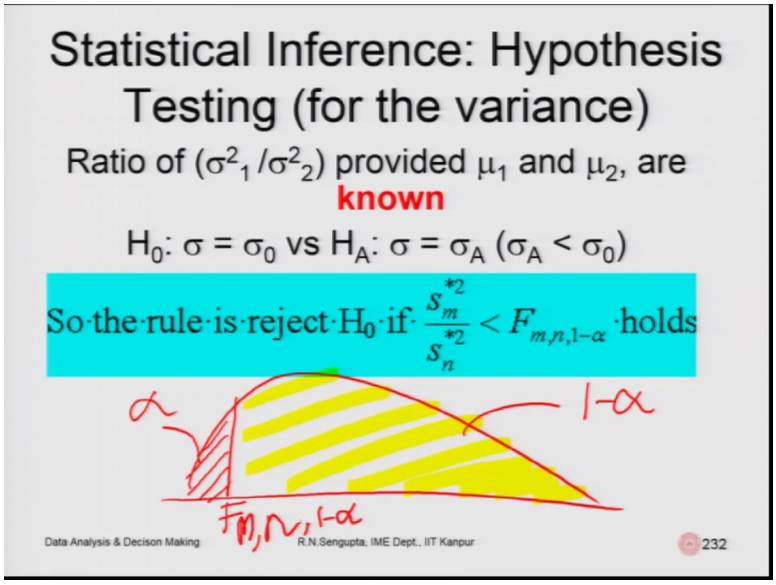
Data Analysis and Decision Making -I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 21

A warm welcome to all my dear friends and students; a very good morning, good afternoon, good evening to all of you. And this is the DADM I, lecture series which is data analysis and decision making. Lecture series under the NPTEL MOOC and we are on the 21st lecture. And if you know which I keep repeating before I start the class is basically this is a 12 weeks program, a 30 hours total, and total number of lectures would be of 16 number; each week having 5 lectures and each lecture being for half an hour. So, if this is the 21st lecture; that means, we have finished for week and we are going to start for the 5th week; 5th week first class. And I am Raghu Nadan Sengupta from IME department, IIT Kanpur.

So, if you remember we were discussing about hypothesis testing and also you do remember that at the beginning of the 20th lecture; I did not show any slide for the first 10-15 minutes. I discussed again repeated what was the plan of the hypothesis testing. So I would not repeat it again, but wherever necessary I will highlight those facts. So, for the hypothesis testing generally for the last two sets was basically where you want to find out something to do with the variance of the standard deviation of the population; provided in case 1 the mean values for both the distributions of the populations are known. And in the second case; case 2 with the mean values for both the populations are unknown. So, we will consider if you want to find out some relationship between the ratios of sigma 1 square to sigma 2 square provided case 1; mu 1 mu 2 known and case 2; mu 1 mu 2 mu 2 unknown.

(Refer Slide Time: 02:12)



So, this is the statistical inference hypothesis testing for the variance. and the ratio is basically as I said is sigma 1 square divided by sigma 2 square provided mu 1 and mu 2 1 are known. So; obviously, they can be three variants of a each case as I have been discussing. So, actually the hypothesis H naught is sigma equal to sigma naught versus the case H A which is the alternative hypothesis mu is equal to mu A when sigma is equal to sigma A and sigma is less than sigma naught or the less than type. So, first let me draw the diagram mark the nomenclature and then come back to the rule.

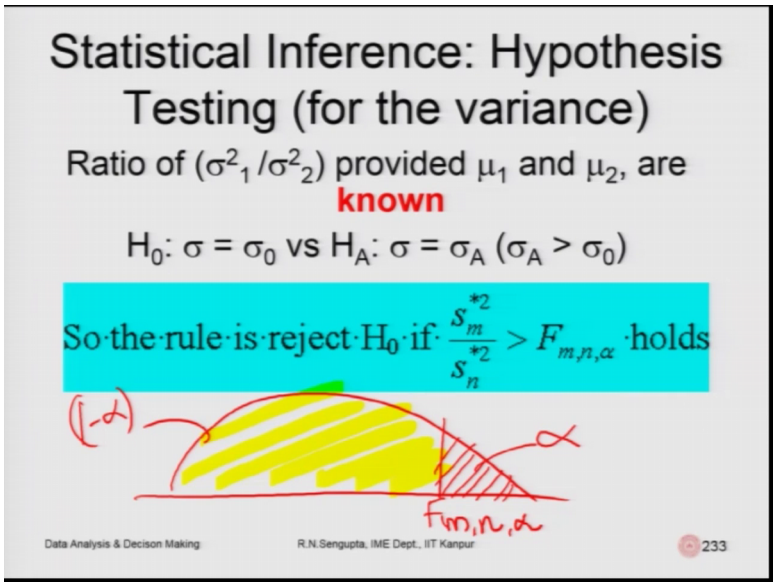
So, we will see the rules which we are stating are exactly as per the norm which have been falling time after time. So, this is say for example, they have distribution and this is the F value. Now the F value would basically have two degrees of freedom; the sample size or degrees of freedom corresponding to the first set of observations from the first population. And the other degree of freedom would basically be the set of observations the first set of observations from the second population.

So, this is m comma n and the ratio would be m would be in the numerator and n in the denominator which I discuss how the F distribution were being formulated. Now comes again the important part and this is the repetition which I keep saying time and again in order to make things clear. So, the overall area on to the left hand side is alpha, the right hand side is 1 minus alpha; so corresponding to that this would be 1 minus alpha, so the this area is this.; so exactly matches what we have said.

So, the ratio of the best estimates for sigma 1 square divided by sigma 2 square, so what is that? Sigma 1 square best estimate is S dash square suffix m because m is the sample size and this is S dash because mu 1 is known so we are not losing any degrees of freedom. So, this is the case I would not highlight, but I just hover my electronic pen over this point so it is S dash square suffix m. And in the denominator what we had actually want to find out the hypothesis testing was based on sigma 2 square so is best estimate from the population number 2 would be S dash square suffix n because n is now the sample size which we are picking up from population 2; which is again true.

Now as it is less than time because sigma A is less than sigma naught so hence it would be less than the cut particular value critical value of F; where F would have a degree of freedom of m comma n. And it would be 1 minus alpha corresponding to fact to the fact that what is the area to be covered or to the right hand side. So, the diagram we have done, the nomenclature we are following it exactly matching with the rule or vice versa whichever you discuss.

(Refer Slide Time: 05:31)

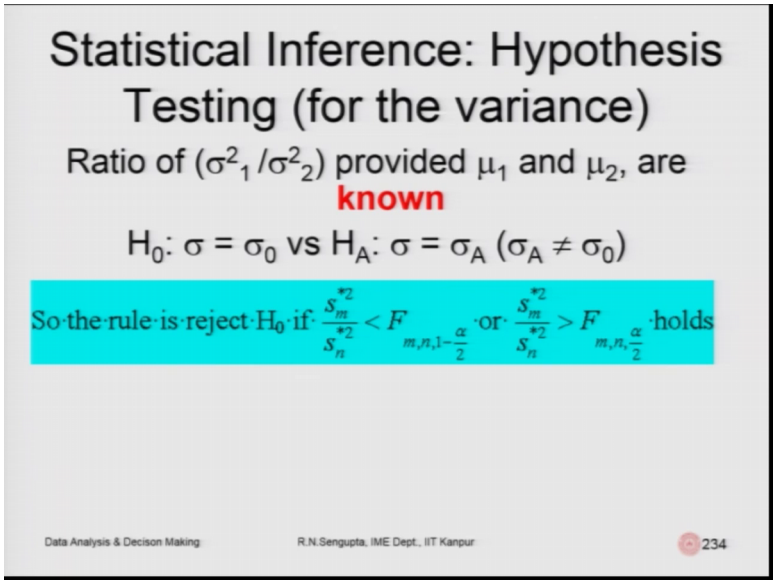


Now, let us go to the second case under this here again you want to find out something to do with the ratios. And provided mu 1 and mu 2 are known, but here the hypothesis is stated as below. H naught under H naught sigma is equal to sigma naught and under H A sigma is equal to sigma A. But in this case sigma a is greater than sigma naught; hence it

is a greater than type. This would be F distribution again m comma n the area on to the right as I mentioned is alpha.

Area on to the left is 1 minus alpha and the area covered here is 1 minus alpha. So, this would basically be alpha on to the right so let us verify. So if the ratio of S dash square suffix m divided by S dash square suffix n. So, this is the case is greater than type so; obviously, it could be on the right hand side is greater than F m comma n comma alpha. So, which you see is right so the nomenclature as stated as being followed is exactly matching the rule of rejection of H naught or vice versa.

(Refer Slide Time: 06:53)



Now, let us come to the third case for the ratios of sigma 1 square to sigma 2 square provided mu 1 and mu 2 unknown which are the population mean values for population 1 population 2 But in this case the hypothesis has been framed like under H naught sigma is equal to sigma naught under H A here, it will be under H A; obviously, it will be sigma A. But in this case rather than less than type or the greater than type we will frame it as not equal to. So, if it is not equal to there is a probability that will be both on the left hand side and the right hand side. So, first let us mark the diagram this is F distribution as usual F m comma n F m comma n.

Now the areas to be covered now are equally distributed corresponding to the fact that the overall area in between this so called upper and lower control limits would be exactly equal to 1 minus alpha which is the level of significance. So, if you are considering the

area on to the left and the right are equally divided. So, hence the area on to the right is alpha by 2 area on to the left is alpha alpha by 2. So, through the area inside is 1 minus alpha as I am mentioned it. So, the lower control limits I have not written the alpha value. So, let me write it down so it will become. So, in this case total area to be covered on to the right would be 1 minus alpha by 2 as rightly pointed out here. So, it is less than an area about on to the right is again alpha by 2 and rightly pointed out here..

So, in this case the statistic would be such that you will reject the H naught provided that ratio of the standard error square that is S dash S dash from both the case because mu 1 mu 2 are known. If that ratio is less than for F m comma n comma 1 minus alpha by 2 or greater than F m comma n comma alpha by 2. So, this m n alpha by 2 1 minus alpha by 2 all are in the suffix. In order to basically this tell that for the F distribution what are the degrees of freedom and what is the level of also called confidence level we want to put. So, here again the diagram the nomenclature matches with the rule or vice versa. So, we are basically corroborating the fact time and again that how we are going to basically mention that in our drawing in our concepts and how the rule has been framed based on that we are proceeding.

Now I have not drawn it, but I will basically highlight it accordingly. So, this is the statistical inference based on the fact that you want to find out something to do with sigma 1 square with respect to sigma 2 square provided mu 1 and mu 2 are unknown and they would be three cases. So, what are the cases I will repeat it first; in the first case the hypothesis is H naught under H naught sigma is equal to sigma naught. And the alternative hypothesis H A is sigma is equal to sigma a and sigma e is less than sigma naught that is the first case. Second one would be again same thing under H naught sigma is equal sigma naught versus the case of H a under HA you have basically sigma is equal to sigma. But in this case sigma is of greater than sigma naught.

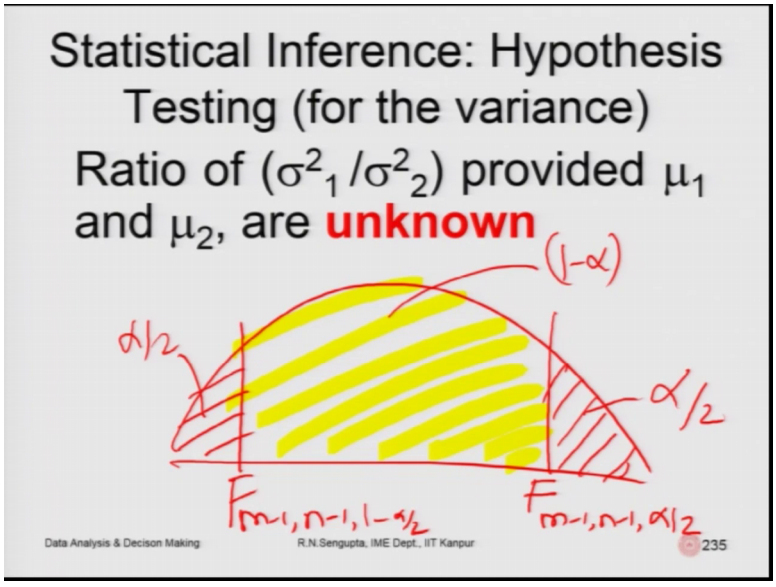
The third rule being exactly the extension have beam as we have been doing; under H naught sigma is equal to sigma naught and under H A sigma is equal to sigma A, but this sigma a is not equal to sigma naught. So, this can be of the less than type of greater at that time point 1 is taken care and I am sure we have understood. We also mentioned in this rule that mu 1 and mu 2 are unknown. So, if they are unknown what is the case we will replace them by the best estimate which is X bar m that is the sample mean for the

first population where the sample size is of m size. And X bar n is the sample mean which is the best estimate for mu 2 provided we have taken up a sample of size n.

Now, if that is the case; obviously, the corresponding standard error squares both which are the best predictors for sigma 1 square and sigma 2 square. So, hence S the S we will be using both for population 1 and population 2 as the best estimate for the variance square of the population would be S without the dash because you are losing one degrees of freedom. So, in this case it will be S square suffix m and S square suffix n. Now if you bring into the picture; obviously, that F distribution would the degrees of freedom the F distribution would also change.

So, in this case it will be F m minus 1 because you are losing 1 degree of freedom from the first population, comma n minus 1 because you are losing another degrees of freedom from the second population. And it will be the suffixes apart from m minus 1 and n minus 1, it will be 1 minus alpha by 2 and alpha by 2 depending on which side of the test are you trying to look at. So, I will basically pinpoint the F values in the diagram and you will understand.

(Refer Slide Time: 12:42)



In the first case F distribution this would be F the value, this is alpha, this is 1 minus alpha, this would be F m minus 1, first degree is being lost in the first population m n minus 1. And if you follow the nomenclature the area on to the right is 1 minus alpha

sorry it is would basically we have two values my mistake my mistake and just basically do it.

So, this would be alpha by 2 this will be alpha by 2. So, this would be F m minus 1 n minus 1; right hand side area is alpha by 2 in this case it is 1 minus alpha by 2. So, in this case we will have basically I am basically drawing the last one at together. So, now, I will basically highlight it accordingly so I jumped, but just bear with me. So, if it is of the less than type less than type where H naught as I mentioned to that for H is the important less than type. So, it will be F m minus 1 n minus 1 this area is alpha this area is 1 minus alpha.

So, this would be 1 minus alpha. So, if it is on the left hand side we will reject H naught. So, as usual and try to basically highlight the right hand side; this is the less than type 1. So let me first eraser it, then draw the diagram well the second case which is greater than type. So, this is area on the right which is alpha this is the area to the left 1 minus alpha. This will be F m minus 1 n minus 1 and the area we did not cover on to the right as per the nomenclature is alpha rule is also that. So, you will basically reject H naught provided the ratios of the best estimate of the and the variances of the population; which are now I am repeating it please excuse me is those are not the S dash they are the S without the dash, because you have lost 1 degrees of freedom from both population 1 and population 2.

So, that ratio should be on the right hand side and if that is true then; obviously, we will reject H naught. Finally, if it is not then equal to type which have already drawn, but please this is the 2 F and highlight these areas this is alpha by 2 this is alpha by 2 this area is 1 minus alpha this is F distribution with m minus 1 m minus 1 area on to the right as per the nomenclature is alpha by 2 which actually we will follow the rule. This is F and minus 1 n minus 1 area to the right is 1 minus alpha by 2. This is the area in between and basically you can give the rule that if the values of the statistics are in that mod, mod distance then you will basically reject accordingly.

So, the rules are the algorithm how you basically do the problem are exactly the same for each thing less than type on to the left right greater than type on to the right hand side not equal to; obviously, on the area right or left. But only remember I am again repeating it please excuse me for doing that; when it is something to do with the mean provided

standard deviation is known then you will always use z, left or right or middle means not equal to that part. If it is something to do with the mean the standard unit is unknown that you will use the t distribution and t distribution would have 1 degrees of freedom less if it is something to do with the differences of the mean provided the standard deviation of the populations are known and you will again use the z distribution; both for the left and for the light as well not equal to; that means, in between.

if it is something to do with the differences of the mean; provided the sample variances are unknown and use the t distribution. But and obviously, in this case you will have a pooled sample variance standard error whatever it is. But remember that you would be losing 1 degrees of freedom from the first population, and 1 degrees of freedom for the second population. So, if the population if this, the sample size from population 1 is basically m 1 and the sample size from the second population is n one. So, in that case the degrees of freedom would be m plus m 1 plus n 1 minus 2 because you are losing 1 degrees from both. Now come to the fact that if it is something to do with the finding out and the actual value that hypothesis about the mean value provided the standard deviation of the population is not known.

Again with repetition I am saying you will use the t distribution with n minus 1 degrees of freedom and the standard errors of the sample of sample would be utilized at the best estimate so those are the set rules. Now when you come to basically if try to find out something to do with the variances the ratio of the variances or before that you want to find out something to do the variance from 1 population if the mean value of the population is known you will use the chi square without losing 1 degrees of freedom; for the left hand side right hand side in between. But we just assured that in place of sigma square you will use the best estimate provided mean value is known you will use S dash square and do the calculations accordingly.

Now, if it is something to do with 1 population something to do study with the variance then be rest assured that if the value of the sample mean is not known you will use the (Refer Time: 19:18) square with n minus 1 degrees of freedom that you lose 1 degrees of freedom is lost and in place of S dash you will be using S without the dash. Coming to the ratios of the which he just discussed coming back to the ratios of the 2 variances from 2 populations population 1 and population 2; obviously, as the norm is. If it is something to do with these ratios and provided the population mean mu 1 and mu 2 unknown, then

you will basically immediately use the F distribution with m comma n degrees of freedom because you are not losing any degrees of freedom because the population mean values are known.

If the population mean values are known; obviously, you will use the standard data from the sample where the standard data would be S dash with the corresponding some observation size or sample size. Similarly for the sample from population 2 you will also use as S, S dash with the corresponding sample value. In this case we are considering from the first population we are picking of a sample of size m and from the second population you are picking up a sample of size n. Now, when we come to the final stage final discussion is that something to do with you want to find out something to do with the ratio of the variances. But the ratios of variances the best estimate would be S without the dash with the dash would; obviously, it will come out as we discussed, but as you can guess it would be without the dash yes rightly..

So, because the population mean mu 1 and mu 2 are not known, if they are not known; obviously, they have to replace when you utilize that standard error to which is the best estimate for sigma square 1 or sigma square 2; that means, for population 1 and population 2. When you are using S without the dash in both the cases you lose 1 degrees of freedom. Hence the F would basically have m minus 1 and n minus 1 depending on the degrees of freedom which is there for the first in the for the numerator and the denominator.

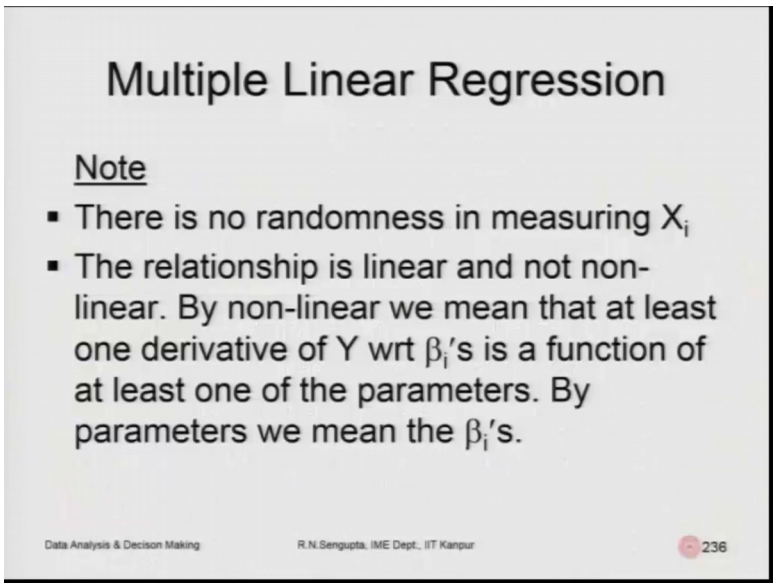
And in case when this is done; obviously, in this would be true that if you are using for all these cases I just missed it I will just recap as at 1 another millet point. Then if it is the less than time then you will use the right hand side of the area of 1 minus alpha and left on the side of the areas alpha. If it is greater than time the area on to the right hand side would be alpha and the left hand side would be of 1 minus alpha. If it is not equal to basically we will have then, then three g regions over the upper control limit it will be alpha by 2 below the lower control limit it will be alpha by 2 and the in between area; obviously, would be 1 minus alpha which is the level of significance.

Now we will consider the concept of linear regression and so consider a background for the problem. So, consider that in the in the case of linear equation you have the you have some set of variables which are to be utilized for prediction. And there is one variable

consider that as why which is to be predicted or forecasted whatever. And we will assume that the, these variables which are to be used for forecasting have some properties. I will consider those properties in details further on. So, our main concept in regression we will want to basically regress on the past data so the past data would can be depending on. Say for example, I want to find out the oil price, so oil price would definitely depend on the GDP, the GNP what is the dollar to rupee rate I am finding the oil price with relevant to India.

So, it would be dollar to rupee rate, then euro to rupee rate, what is the inflation rate and all these things. So; obviously, we will have all this variables which will be utilized to predict the in the oil price would comes be considered as independent variables and the one which you want to predict or forecast will be called the dependent variables. So, and; obviously, the relationship between the independent and independent variables will continue and we will discuss that soon..

(Refer Slide Time: 23:35)



So, in multiple linear regression there is no that they would not be any randomness in measuring x. So, what are these X xs are those variables which would be utilized and they are the independent variables. So, we will consider there is no randomness in measuring X is. So, consider there are X 1 X 2 X 3 X 4 till X k number of such random variables which are there. The relationship is linear and not non-linear so; obviously, in this case we will consider the as the word is multiple linear regression we will consider

the relationship between the Xs or X 1 X 2 X 3 X 4 whatever they are. They are the independent variables and they are dif and their effect on the dependent variable which will be mentioned as y is basically linear in structure. By non-linear we mean that at least 1 derivative of y with respect to the beta. So, what are the betas let me explain.

So, consider that you have a regression where y is dependent on some alpha which is a fixed value. So, it will this fixed value will have some significance I am going to come to that later on. Alpha plus beta 1 X 1 plus beta 2 X two so and so forth till beta k X k considering there a k number of random variables plus some epsilon which is an error. So, this technically betas beta 1 beta 2 beta three till beta k very simply would mean; that keeping all the other variables constant, if I am only concentrating on the first variable then beta 1 would basically give me the rate of change or the rate at which the function y would be changing for 1 unit increase in X 1 keeping other things fixed.

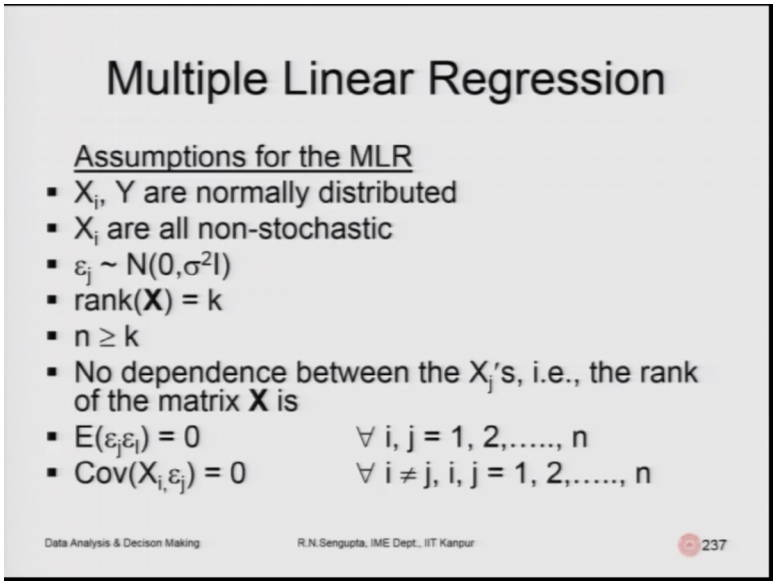
In the similar way when I considering beta 2 or beta 3 or beta 4 considering that I am considering beta i, where i is equal to 1 to k we will consider that. In that case beta i would be the rate of change of the y function with respect to X i provided others are fixed. Now in this case the alpha which I mentioned I will I will draw this diagram in later in details. When I am considering this alpha basically means the coordinate at the value at which it cuts the y axis. So, if you in school if you remember we have done the equation y is equal to mx plus c, where m is basically the slope, X is the variable based on which you want to predict y, and the c value is basically where the straight line is basically connecting the value of y axis where the X value is 0..

Now whenever you are trying to basically consider that; obviously, that epsilon which I mentioned is basically the white noise. So, obviously, white noise is there randomness is near and this randomness would basically have some distribution. In general we consider the randomness has a distribution which is basically normally distributed with a certain mean and a certain variance in the simplistic case. We consider the randomness of the white noise of the error to be there which has a 0 mean value and once standard deviation of variance which is basically the standard normal deviate.

But; obviously, you can change the variances accordingly, but technically we will always consider the expected value to be 0. So, let me continue reading it there is no randomness in measuring Xi's. The relationship in linear and non-linear is not it is linear and not non-

linear. But non-linear we mean at least 1 derivative of y with respect to beta is a function of at least 1 of the parameters by parameters we mean basically the betas.

(Refer Slide Time: 26:56)



So, this slide I will consider in more detail because consists considering the 30 minutes for this class are almost over. I will discuss that, but again in the next class also I will start from the same slide. So, do not because in general whatever slide I take up I generally wrap it up before the end of the class, but this would we will need more discussions. So, we will consider that in the next class also.

So, assumptions for the multiple linear regression model. So, here we will have X 1 X 1 2 3 4 whatever; y is a normally distributed. I am just reading it I will come to the significance later on Xi's are non stochastic as I mentioned. The error term which I mentioned is basically normal 0 mean and this is given as sigma square i, but we will I will consider very simply that they have a variance of 1. The rank of this matrix is k; that means, rank if you remember is basically the linear dependence of this of the rows and columns which is there for the matrix, where it is coming I will go to that. But k remember is the number of independent variables which are there.

And; obviously, n the number of readings based on which you will try to basically predict for the nth plus reading i is greater than equal to k. And this has also another significance because if you are solving some linear equation it would mean then the number of equations should be if there is the all the equations are independent of each

other then the number of equations should be there should be more than the number of variables to be found out because if it is not that case then can they can be redundancy in the calculations. And; obviously, we know from simple linear equations which you have studied.

So, there would be no dependence structure between Xs that is the rank of the matrix X is basically as given. And obviously, to mean that the dependence structure not being there been Xs it would mean that they are independent of each other. The expected value of the covariance's between the errors is zero; that means, they are not dependent on each other. And the covariance's of the independent variables with the error would also be zero; that means, the white noise is its by itself or stand on 1 case it is not going to basically affect the readings of either X 1 X 2 X 3 or X k.

I will basically close it here; continuing the discussion more about the assumptions and basically draw the equation in order to basically. And highlight how it looks like and what how what uses the multiple linear regression can have. With this I will close this task and have a nice day and.

Thank you, very much for your attention.