

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

Lecture – 54

MLR

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 which is data analysis and decision making one course under NPTEL MOOC. And as you know this course total duration is 30 hours which is for 60 lectures, 12 weeks and each lecture as you know is for half an hour. And we are in the last, but 1 week which is the 11th week and this is the 54th lecture. And each week as you know there are 5 lectures each being for half an hour and after each week you have 1 assignment.

So, I am sure we are being you are being able to get a lot of help by answering these assignments and getting the things clear. Before I start of the class I know that this is a huge amount of coverage. In this lecture in this course starting from univariate statistics to different type of distribution, expected values, then different type of a discrete and continuous distribution, what we mean by PDF, PMF, then the CDF concept, the QQ plots.

Then we went into different examples of the distribution, what are the parameters, then sampling distribution, what we mean by a sample, different type of concepts of how we take a sample. Then we went into the unique 3 other distributions coming from normal case which was basically the chi square, F and T distribution and how they can be utilized to find out the interval estimation problem, hypothesis test problem.

Then we went into concept of multivariate statistics covered the concepts of multinomial distribution, then multi normal distribution, wizard distribution. Then we consider the extreme value distribution, consider the different type of Emily's methods before that obviously we consider the Emily's and the GMM methods conceptually in the univariate case also. Then we went into different concept of utility theory, decision sciences, loss functions. Then we consider the factor analysis, principle component analysis and then now here we are in the multiple linear regression.

In the multiple linear regression if you remember we have assets of X_1 to X_p independent random variables, each have n number of readings. Using this you are trying to predict our forecast for Y which is the dependent variable. And; obviously, under the ordinary square method we saw the formula could be derived by taking the sum of the squares of the errors, differentiating partially differentiating with respect to the beta 1, beta 2, total beta p putting them to 0 and finding it out.

Now, under so, you have already seen; what is the estimate for beta under the unruly square. And if you consider the LINEX loss which is the linear exponential loss function if you consider the best estimate for beta which is given by beta tilde beta tilde.

(Refer Slide Time: 03:23)

Multiple Linear Regression (contd..)

- The LINEX estimate for $\beta_{p \times 1}$ is given by

$$\tilde{\beta}_{p \times 1} = (X'_{p \times n} X_{n \times p})^{-1} X'_{p \times n} Y_{n \times 1} - a(X'_{p \times n} X_{n \times p})^{-1} \frac{\sigma^2}{2}$$

NPTEL DADM-I R.N. Sengupta, IIT Kanpur, INDIA 538

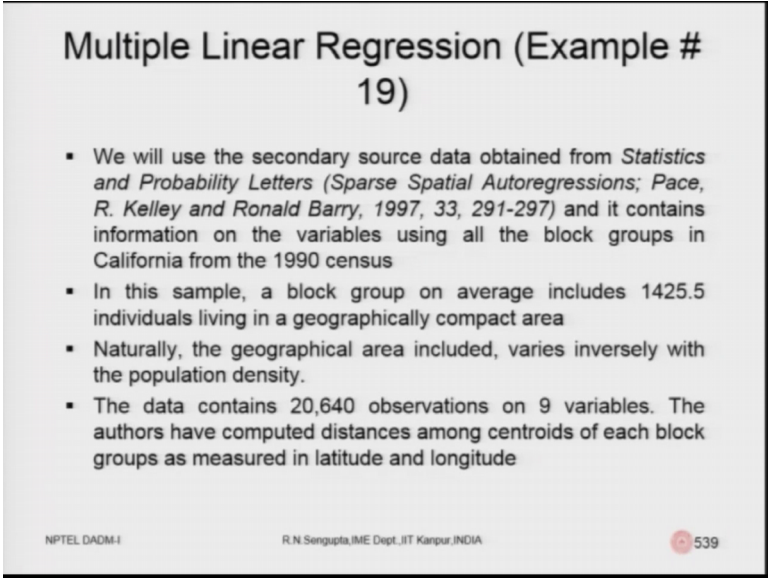
Beta tilde is given, this whole portion which I am basically highlighting now is basically beta hat as per the ordinary square which means and. So, this is actually beta hat as per the unruly square which means this is actually beta hat p cross 1 and this is true under unruly square method. But, when we have the LINEX loss you have to basically take a nuisance parameters with the minus sign.

So obviously, this minus sign may not really mean remain the minus sign depending on the value of a . If a is negative so, it becomes positive, if a is positive this continues to be near minus negative.

So, we overemphasize, under emphasize the beta hat value in order to basically calculate beta tilde value such that over estimation and under estimation are given due importance as per the LINEX loss which can be then used, which can be incorporated in the LINEX loss function by the values of α if you remember the parameter.

Just so this was a being positive and this α is negative you will have this graphs which you have already seen, but I am just highlighting.

(Refer Slide Time: 04:49)



Multiple Linear Regression (Example # 19)

- We will use the secondary source data obtained from *Statistics and Probability Letters (Sparse Spatial Autoregressions; Pace, R. Kelley and Ronald Barry, 1997, 33, 291-297)* and it contains information on the variables using all the block groups in California from the 1990 census
- In this sample, a block group on average includes 1425.5 individuals living in a geographically compact area
- Naturally, the geographical area included, varies inversely with the population density.
- The data contains 20,640 observations on 9 variables. The authors have computed distances among centroids of each block groups as measured in latitude and longitude

NPTEL DADM-I R.N. Sengupta, I.M.E Dept., JIT Kanpur, INDIA 539

Now, we would not solve the problem, but go through the details of problem which for which a good dataset is there. So, because the data set is very large.

So, I cannot have the luxury of downloading and working with the data. So, we will use the secondary source data and this is example number 19 in this DADM course as written there in the top. So, we will use the secondary source data obtained from the journal statistics and probability letters which the paper came out in 1997, the title of the paper is sparse spatial auto regressions by Pace Kelley and Barry in the volume 33 of that journal. And it contains the information on the variable using all the block groups in California in the 1990 census.

So, in this sample group of block would basically have about 1 1425 about individuals so, that the divided into blocks. The total number of data points is as given in the last bullet point is 20,640 observations with 9 variables. And the geographical area included

basically varies inversely with the population density. So, area being large or area being small would basically be dictated by the population density.

So, as I was reading the data consists which is the last bullet point as it mentions data comes contains 20,640 observations on 9 variables. The authors have computed the distance among the centres of each block groups as measured the latitude are not do it. This is for the information, but we would not use the consorted of latitude longitude for our calculations.

(Refer Slide Time: 06:33)

Multiple Linear Regression (Example # 19) (contd..)

- The data file contains all the variables and specifically, it contains the following which are
- Median house value (MHV)
- Median income (MI)
- Housing median age (MA)
- Total rooms (TR)
- Total bedrooms (B)
- Population (P)
- Households (H)

NPTEL DADM-I R.N. Sengupta, IIT Kanpur, INDIA 540

If you open the data file and basically see which is there in that journal. So, actually there are lot of informations, but the actual variables which will be utilized are as follows and which is given here. Here the median household value the value of the house in dollars as it is given. You have the median income of the people who are residing in that house in dollars. You have the housing median age is the house of the housing houses age, how long it has been built and how long the people have been occupying that. Total number of rooms which are there in that house, total number of bedrooms which is important that is also mentioned. The total population in their centroid and the households numbers who are there, total families which are there in technically in that.

So obviously, we will have each family having some number of people multiplied bath by both being multiplied will give you an approximate number of population in that area and then you multiplied that average number which was given as 142, 142 5. So, that

multiplied by the number of such districts on areas would give you the total population. So, technically we are considering the total population as in this problem is 20,640.

Now, as proposed by the authors in that journal statistics and probability letters.

(Refer Slide Time: 07:59)

Multiple Linear Regression (Example # 19) (contd..)

- The multiple linear regression model is given as

$$\log_e(\text{MHV}) = \beta_0 + \beta_1 \text{MI} + \beta_2 \text{MI}^2 + \beta_3 \text{MI}^3 + \beta_4 \log_e(\text{MA}) + \beta_5 \log_e(\text{TR/P}) + \beta_6 \log_e(\text{B/P}) + \beta_7 \log_e(\text{P/H}) + \beta_8 \log_e(\text{H}) + \varepsilon_1$$

Handwritten annotations on the slide include:

- A blue box around y_i and y_{20640} with an arrow pointing to the regression model.
- Red labels X_1 through X_8 above the corresponding terms in the equation.
- A green box around ε_1 with a vertical ellipsis and ε_{20640} below it.
- A red box around the coefficients β_0 through β_8 with a vertical ellipsis and β_8 below it.
- Red boxes around 20640×1 and 20640×9 at the bottom right.

NPTEL DADM-I R.N. Sengupta, IIT Kanpur, INDIA 541

The multiple linear regression model is given as follows where you have log of median household values is equal to beta naught. So, this is the first time we will basically see if you remember the alpha value which was there.

So, this is beta naught and why I am mentioning beta naught I will come to that within few minutes, but plus beta 1 which is the first regression coefficient for the first random variable MI, which is the median income plus beta 2 into medium income square. So, you square this medium income to find out the second random variable which is technically X 2.

Then X 3 is medium income cube and that regression coefficient is beta 3. Then X 4 is basically log of medium age and the regression coefficient is MID beta 4. The next random variable is log of total room by population and the regression coefficient is beta 5. The next random variable is log of bedroom by population and the regression coefficient is beta 6 the next and I am just reading out.

So, you notice the slide you will understand it. The next random variable which is population by a household log of that and the regression coefficient is beta 7. And the

last the regression this random variable is log of household and the regression coefficient is β_8 plus the error term.

So, what actually this model applies are these. So, this is technically X_1 , I am writing over the corresponding random variables. This is X_2 , this is x_3 , this is X_4 , this is X_5 , this is X_6 , this is X_7 , this is X_8 . And correspondingly actually if you remember the beta vector which we are denoting in our model actually that is I will write it down β , β_1 , β_2 , β_3 , β_4 , β_5 , β_6 I am running out of space.

So, let me (Refer Time: 10:24) I will switch over to β_7 and β_8 . So, this is a vector of 9×1 because including β we have 9 regression coefficients, now these betas are not known to me. So, if they are not known to me what we do? We will basically use the same policy. I am again repeating it please bare with me, you find out the errors, square these errors, sum them up, differentiate with respect to partially differentiate with respect to β put it to 0.

While then partially differentiate again separately with β_1 , put it to 0, do it for all the betas, find out this betas and they those $\hat{\beta}$ would be utilized and the error terms if you know actually. So, if you take the actual data set the error term would basically be starting for the whole dataset. It will be basically starting from ϵ_1 to $\epsilon_{20,640}$ data points. Similarly, Y , I missed it, Y which is here. So, I should use a different color that will be much better for all of us to understand.

So, this will be Y , I should change the color for ϵ also let me erase it and do it again. So, this is ϵ which is given ϵ_1 to $\epsilon_{20,640}$. In case you would take all the data points; obviously, you will not take the whole data point because we consider that 20,640 as the population only based on that you will proceed.

So, you are doing a so, called assumption that that 20,640 data points is the population based on this you are trying to do the studies which would give you good results nothing wrong in that, but it just a point which I wanted to mention.

So, now, once you have this, I will just highlight in order to make things understand. So, this is Y , these are the X 's. So, X_1 , X_2 , X_3 , X_4 , X_5 , X_6 , X_7 , X_8 . So, and you find also have ϵ . So, they give you the information based on this and if I want to denote the X vector and the Y vector also let me denote. So, while I because I have been using

the blue color. So, Y would basically come here, Y 1 till Y 20,640 data points and if I use the concept for the X s, X s are should are also red in color or let me use.

So, this is size is basically as I am denoting it is 20,640 cross 9. So, these values would be 1 1. So, if this is 20,640 comma 1, this is 1 p, p means 9 and this value is 20,640 cross 9. So, this is the overall matrix which we have and based on that when you find out these values of beta tilde and beta hat beta tilde being for the LINEX loss beta hat being for the case of the ordinate least square.

So, you can check up any book to find it out, the concept remains the same.

(Refer Slide Time: 14:48)

Multiple Linear Regression (Example # 19) (contd..)

- The solution for the multiple linear regression model is given as

$\hat{\beta}_0 = 11.4939$	$\hat{\beta}_1 = 0.4790$
$\hat{\beta}_2 = -0.0166$	$\hat{\beta}_3 = -0.0002$
$\hat{\beta}_4 = 0.1570$	$\hat{\beta}_5 = -0.8582$
$\hat{\beta}_6 = 0.8043$	$\hat{\beta}_7 = -0.4077$
$\hat{\beta}_8 = 0.0477$	

NPTEL DADM-I R.N Sengupta, IIME Dept., IIT Kanpur, INDIA 542

So, based on that if you do the calculations here are the values. Now the solution for the multiple linear regression model as being used from the orderly square would be these values. So, beta naught is 11.49 I will only read till the second places a decimal 11.49, beta 2 is minus 0.02 basically 6 is being converted.

So, I will take with the 2 places of decimal and then normalized, beta 1 is basically 0.48, beta 3 is equal to 0.00. So, if you consider the 4 pacer decimals it becomes minus 0.0002, beta 4 is 0.16, beta 5 is minus 0.86, beta 6 equal to plus 0.8 0 beta 7 is minus 0.41 and beta 8 is 0 plus 0.05.

Now, the issue is that what do we next? So, we have to find we have found out this value. So, I have written remember one thing, I have not put a hat value here. So, the hat

value which actually I have not put them here. Reason is that I am considering these are the population, if you remember I did mention it few minutes back. So, these are the populations based on that we will try to basically compare the hat values which we calculate. So, technically I should remove these hats.

(Refer Slide Time: 16:27)

Multiple Linear Regression (Example # 19) (contd..)

- Taking the a set of sample points from the data say say n and the estimate $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ and β_8 , For this use the formulae for $\hat{\beta}_{9 \times 1} = (X_{9 \times n}^T X_{n \times 9})^{-1} X_{9 \times n}^T Y_{n \times 1}$, here n=100, 150, 200, 250, 300, 350, 400, 450, 500
- Also remember to take such number of n in each case 10 number of times
- Plot the histograms of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ and β_8

NPTEL DADM-IR.N.Sengupta, IIT Kanpur, INDIA543

Now, let us discuss the problem, how you solve it. Let us take the values; so taking a set of sample points what do you consider by the sample points. So, let me make the blank slides and explain.

Now, what it says is taking the us taking the set of sample points from the data say n any number and the estimated values we want to find out for beta naught to beta 8; so, what do you do?

(Refer Slide Time: 17:09)

Multiple Linear Regression (Example # 19) (contd..)

$$Y = X\beta + \epsilon$$

$y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \epsilon_{n \times 1}$
 $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$
 $E(\hat{\beta}) = \beta$
 $\epsilon_{n+1} = (y_{n+1} - \hat{y}_{n+1})$

NPTEL DADM-I R.N. Sengupta, IIT Kanpur, INDIA 544

So, let me use the plant. So, consider this is the equation and this is of size 20,640. So, this is why I am not going to add the dimension because it will become 2 cluttered. This is X, this is beta and this is epsilon.

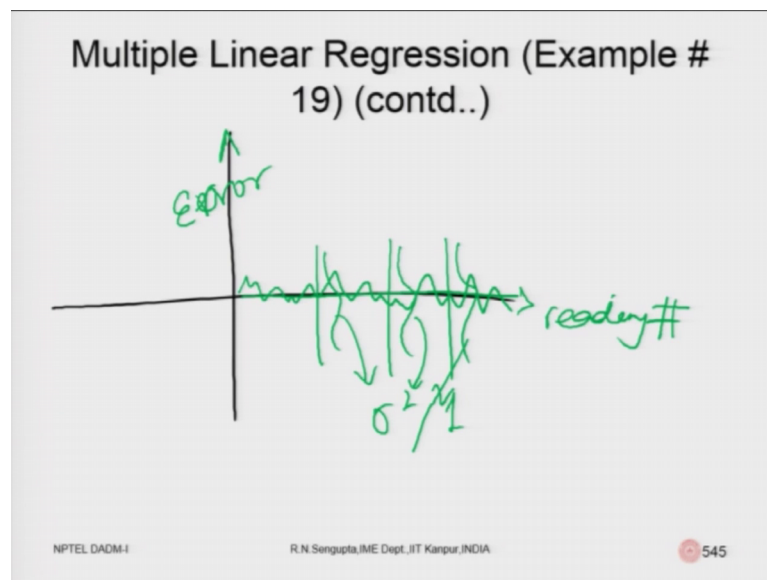
Now, what do we do? I will use the red color first. So, what we do is that we take a small sample size here which is of size n. Similarly, you will have a size n for X also. So, Y will take a size n cross 1, X will take basically n cross p, p remains same because that is 9, because 0, 1, 2, 3, 4, 5, 6, 7, 8, beta will consider again p cross 1 and epsilon we will consider which is basically n cross 1.

So, from that you use the formulas which is basically beta this is bold, I am not able to denote the bold, but this is bold this is the beta hat. We get beta naught hat till beta 1 hat for some n. Using that we basically find out the nth plus 1 value; so y n i k plus 1 value which is already there which is the next value.

You will find out the estimated value using nth plus 1 value because, that what you will do is that you multiply these betas with the X n-th plus 1 values. So, n a X na plus values they would basically be such values correspond to. So, this should be p, I am not writing in 9. So, it is p and based on that the nth plus 1 value from the matrix X we calculate the error. So, this is the error which is for the nth plus 1. Similarly, find out the error for nth plus 2 and nth plus 3 and so on and so forth.

Now, if we add up the errors what did what is actually should be. If you remember we had mentioned that the expected value of the errors is 0, and the variance is basically fixed as 1 or sigma square or whatever it is. So, the error terms technically would look like this.

(Refer Slide Time: 19:47)



So, this is the graph and the errors if I use the green color the errors would be like this. So, the expected value would be 0 and the variance would basically be fixed considering is not dependent on time. This value of their variances would be equal to sigma square or 1 depending on the problems which you have. So, here you are basically measuring the error and this is the reading number.

So, you can find it accordingly, let me go back again. So, I am sorry; I am really flip flopping going forward and backward for this slides, please bear with me. Now, for this we will use the formulas as given which we already discuss many times this is the formula. So, this T is the transpose. So, we and hence say for example, n is equal to 100 you do many of the calculation. So, what you do is that whole data set you have different type of sets of 100.

So, the first 100 you find out beta hats, the next 100 you find out the beta hats, next 100 you find out the beta hats, you continue doing it may say for example, take 30 number of times, 40 number of times, 50 number of times, find out the averages of these averages. Because, for the first 100 you will basically have 1 set of betas, then the second 100 and

you will again have a second set of betas, find out the averages of all the beta naught's from this 30 such sub samples you have taken.

Again, find out the average for all the beta beta ones for this 30 sub samples you have. Once you find out keep repeating it in the long run the averages of these averaged should x actually be the values of the beta 1, beta 2, beta 3 or beta naught which you have found out using the 20,640 points. So, larger the number of samples is or larger the sample is and you keep repeating it then obviously the difference between the estimated value and the beta value which you have taken in the long run should be 0.

Similarly, there are larger number of values which you take for to estimate the betas and more number of such idioms estimation which you do hence, the errors which you want to find out using the concept of betas beta hats and the X is the. And errors with respect to the difference between the actual value of Y and the estimated arrival is Y in the long run should be exactly and equal to 0 as per the assumption.

So, here what we are saying is that take n is equal to 100, 150, 200, 250, 300, 350, 400, 450 and 500 keep repeating it as the many number of times, considering n is fixed. For 100 you will have 1 1 set of readings, for 150 you will have one set of readings.

Similarly, for 200, 250, 300, 350, 400, 450, and 500, but remember one thing for this number of small sub samples if you have the same number of repetitions then for the largest samples which is 500 there the errors in the on the average which is the difference between the actual value of Y and the predicted values Y would basically be turn out to be 0 faster point one. Point number 2 the estimation values of the betas would be much better. That means the difference between the beta hat and the actual value of beta which you have already found out using the 20,640 points would be as low as possible.

So, what we are saying is that also remember to take such number of n's in each case as 10 as in number I have said 30, it can be 10, 15, 20 whatever it. But if you do such number of repetitions considering sample size is also large, number already 2 position is also large, the value should definitely be as close as possible to the actually values. If you plot the histograms of beta 1 to be beta naught beta 1, beta 2 till beta 8 what is interesting is that the histograms would basically slowly come out to a normal distribution.

And the expected value of this histogram of the normal distribution would be in the actual values of beta naught to beta 8 which you have found out using the 20,640 data points which you have used for our studies. So, this is what so, if you take second n. So, what I am saying is that you take another n, another n, another n, keep repeating it. So, this is the first n, second n, third n, fourth n, do it for these values and you will get the values of beta hats.

So, technically the expected values of the beta hats individually in the long run should be equal to the beta. So, I am using the vector notation again, if you delete now I am saying I am delete; I am going to come to that later.

(Refer Slide Time: 25:16)

Multiple Linear Regression (Example # 19) (contd..)

$$Y = X\beta + \varepsilon$$

$\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ $E(\hat{\beta}) = \beta$

NPTEL DADM-I R.N.Sengupta, IIT Kanpur, INDIA 544

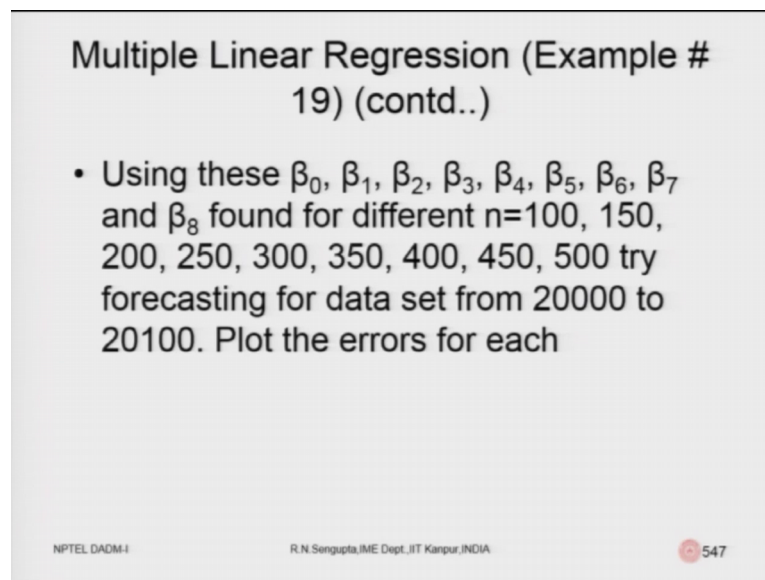
Now, if I change n as I said put a different colour because n is not changing again some n, n, n, n, n, in that case the ns were if you remember 100, 115, 200, 215, 300, 315. So, these are the number of repetitions so consider them as n 1 so, they are being repeated. So, again you will find out Y would be a size n 1 cross 1, X 2 the size of n 1 cross t, p is 9 beta will be a size of p cross 1 epsilon would be a size n (Refer time: 25:49) cross 1 cross 1 sorry.

Again you find out with the hats. So, this beta hat is bold the way I am hovering the pen. So, you mean the expected values of beta hats again these are bold and if you keep repeating it for different ns you will find out. So, technically as it will mean the p use the black colour now as n tends to increase and as if you do more repetitions then the (Refer

Time: 26:29) city in that the rate at which beta will converge to his actual value of beta which was found out from 20,640 data points would be much faster and the errors would be much better. Errors much much better in the sense that the difference of the errors would be basically be very small.

I expected value of the errors would be 0 as fast as possible. So, the errors I mean is the difference between Y and Y hat based one which you find out.

(Refer Slide Time: 26:57)



Multiple Linear Regression (Example # 19) (contd..)

- Using these $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ and β_8 found for different $n=100, 150, 200, 250, 300, 350, 400, 450, 500$ try forecasting for data set from 20000 to 20100. Plot the errors for each

NPTEL DADM-I R.N. Sengupta, I.M.E Dept., IIT Kanpur, INDIA 547

So, using this beta naught to beta 8 we find out the errors plot them and if you plot them you are certain to get the errors tending towards 0 as fast as possible. And this can be used in a very simple it will take time, but only you spent about 1 hour and spend in using the excel sheet for a good regression model, and whatever I repeated would definitely come out to be true for all the calculations.

Now, the question is that you may be asking would these answer results be also true for the case when we use the LINEX loss? Yes, with some modifications and considering beta tilde you can also prove the same for the LINEX loss and any other loss function estimates which we have.

With this I will close this last part one lecture in the eleventh week which is the 54th one and in 55th will do something to the LINEX loss estimation problems and continue for the discussion about that. Have a nice day.

Thank you very much.