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

Lecture - 53
Loss function and MLR

Warm welcome to all my dear friends and students; a very good morning, good afternoon, good evening to all of you and welcome to this DADM which is Data Analysis and Decision Making-I course under NPTEL MOOC series. And as you know we are going to start the 53rd lecture as shown in the slide, which is in the last, but one week which is the 11th week and this course total duration is 30 hours for 60 lectures and each lecture being for half an hour. And each week as you know there are 5 lectures each being for half an hour. And my name is Raghunandan Sengupta from the IME department IIT, Kanpur.

So, if you remember we were discussing about different life type or loss functions, we gave the examples of lin lin loss functions which is linear linear loss functions; then we discussed that if the linear loss functions are weighted; weighted depending on what is the domain or the range of theta values theta is basically the parameters and the parameters can be either scalar or vector. Depending on the range of the theta the loss functions propensity or the intensity can increase or decrease. And then you generally we initially consider the squared error loss because for its nice theoretical properties, because under the unbiasedness property the t suffix n; where t is the statistic which you find out from the sample if it is the best estimate for theta; that means, the expected value of tn is equal to theta. Then the square; that means, tn minus theta whole square that expected value which we call as the risk would be equal to the variance and we try to basically minimize the variance as the case is.

So; obviously, it follows the theoretical properties and hence we are very comfortable using squared error loss. But squared error loss also it can be weighted depending on the weights you want to give, if the weights can be a fixed value like k 1 k 2 as it as it can be for the linear linear loss function also lin lin loss function. Then the weights for the squared error loss could also be weighted weights being a function of theta, which is the parameter value which you want to estimate. And then we very briefly considered these are just giving a picture.

So, we consider the zero-one loss function depending on the values how different tn is with respect to theta within a certain range where the range as per the concept of interval estimation is epsilon. So, it can be loss function can be 1 or 0 depending on whether it is out of range or inside the range. Then we went to into considering the two interesting loss functions for which Zellner has worked a lot and one is basically the linear exponential loss function which is LINEX loss function where one part is linear another part is exponential. And we considered the linear part or the exponential part are such that if you are trying to draw the along the x axis, the difference between tn and theta which is delta, and along the y x you draw the loss function which is the exponential loss function for values of a greater than 0; overestimation is more penalized than under estimation because exponential part dominates the linear part, while for on the right hand side in the first quadrant and in the second quadrant of the linear part dominates the exponential part.

Then which is a is greater than 0. If a is less than 0, then in that case just the reverse happens in the sense that underestimation under in underestimation linear part exponential part sorry exponential part dominates the linear part and for over estimation, the linear part dominates the exponential part. And we get discussed then, we discussed the balance loss function. Balance loss function was basically the concept of using goodness of fit and precision or estimation. Precision or estimation are with respect to the alphas or betas whichever you denote the parameters for the multiple linear regression you want to find out, and what is the precision of estimation for that with respect to either the squared error, with respect to the lin lin loss function, with respect to either the 0 one loss function whatever it is.

And the other part is the goodness of fit is basically once you find out those alpha and beta using their alpha hats and beta hats which are the estimates, then we use these estimates to find out y hat which is the predicted value or the forecasted value of y. And then try to find out the loss corresponding to the fact that we are not able to estimate y exactly using y, but we basically estimate or forecast y using y hat.

Now as per Zellner, the loss function or the balance loss function in the case when you both have the precision of estimation and the goodness of fit both would be squared, but we can basically have different variance of this a balance loss function which will slowly see as we discuss.

So, with this we considered I said that we will consider 3 examples for the LINEX loss function where over estimation and under estimation are unequally penalized on how the picture can be denoted. So, in the initial case, we consider the warranty problem of a new product which is being floated by a marketing company or a manufacturing company, and the company wants to find out the one (Refer Time: 05:58) with respect to its customers, and how the customers see the set of information which is there with the with the competitors also.

So, if the competitors in actual the warranty life is 6 months, then whether if you give 8 months as case 1; where you have estimating 8 months as the warranty life with respect to 4 months as the warranty life in case 2 form from your side, you will find out that in case if the loss function of the prediction is squared error then; obviously, over estimation and under estimation are equally penalized. But in the case if it is a LINEX loss function depending on the value of a, you have to find out the what ranges of a would best suit your practicality because in that sense that in case say for example, your you initially give a warranty life of 8; obviously, people will more tempted to buy a products you will gain a market share, but the probability of the products for a failing before the 8 months because the actual warranty life is 6 months would be very high then they would be a market loss for you, face value loss for you, business loss for you.

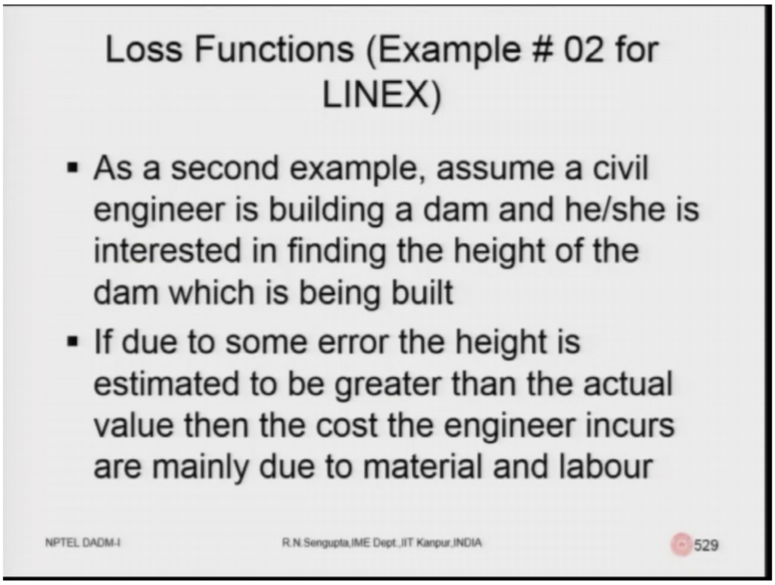
So, how you basically compensate the initial market share increase with respect to the losses which you will face from the customers as they switch their products, as they may file for a litigation all these things how you will basically balance that using either the concept of a is greater than 0 or a is less than 0 and that will depend how you see the practicality.

In the case if you give a warranty life of say for example, 4 months. So, in that same case also under estimation or over an estimation will be differently penalized because in this case initially you lose the market share because products being sold by the or competitors has a warranty level 6 months, people will be more tempted to buy that product.

But later on when if you are able to basically give products for which your products would be failing after the 4 months, because you are warranty life is 4 months and people get extra benefits of this annual maintenance contract so; obviously, there may be some positive benefit for you. So, how you balance that again will basically be a question how you see it from the practical point of view.

So, with this again I will consider two examples and come to the concept of multiple linear regression.

(Refer Slide Time: 08:14)



Loss Functions (Example # 02 for LINEX)

- As a second example, assume a civil engineer is building a dam and he/she is interested in finding the height of the dam which is being built
- If due to some error the height is estimated to be greater than the actual value then the cost the engineer incurs are mainly due to material and labour

NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 529

As a second example assume a civil engineer is building a dam. So, the dam actual height is say for example, 120 feet and you are building in the dams such that the actual value is 120 feet and in one case you make it say for example, 123 feet and in another case 1 and hypothetically another engineer would have basically given a height of say for example, 117 feet. So, in this both this cases the difference is plus 3 and minus 3. So, for 123 it is 123 minus 120 which is plus 3, and for 117 it is 117 minus 120 it is minus 3. But in main case if you want to basically estimate something to do with the dams heights or width or strength parameters, consider using a loss function and the loss function is quadratic in that case both the values where there is plus 3 or minus 3 the overall squared error value would be 9.

So, in that case over estimation and under estimation are equally paralyzed. So; obviously, theoretically it is fine, but let us see it from the practical point of view. In the case if you build a height to 123 feet. So, initial cost man hours, materials, time delays; obviously, would be there and you have to basically shelf out more amount of money for that. But consider that a flood comes our sea for example, some huge amount of rainfall happens and the water inundates the banks of the rivers and flows.

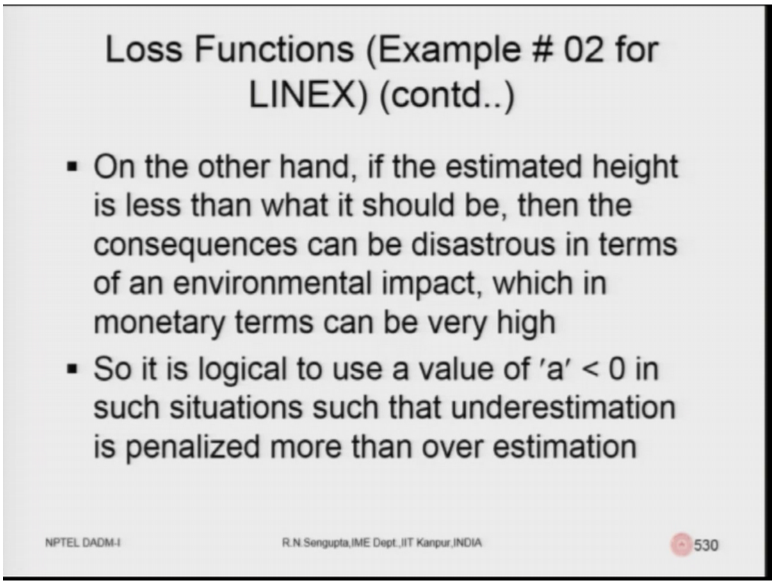
So, in that case the surge of the water breaching the dam for which the actual height should have been 120, but he have made it much higher and much stronger 123; obviously, the probability would be almost 0. In which case the loss of vegetation, loss of manpower, loss and the natural calamity would be much less. So, he (Refer Time: 10:00) hence initial cost which was a little bit high or maybe or quite high is compensated because you are able to which stand the flood.

But in the other case considered the actual height by the second engineer has been proposed and 117 feet so; obviously, the initial cost of man material labour time is; obviously, low with respect to 120 feet, but what happens even the flood comes in that case the propensity of the flood to breach the dam is much higher. So, hence later on they may be calamities catastrophes loss, natural losses human life lost, cattle lost, vegetation lost and all the farmlands inundated.

So, in that case the overall loss would be much higher than in the first case where you have build the dam to 123 feet or have proposed that will to build the dam to 123 feet. So, in both the cases in the practical sense in the first case when it is 123, in the second case which is when it is 117; the overall loss would definitely be different. So, you will definitely think of trying to overestimate and build the height of the dam more than 120 feets such that the overall loss is not that much in case when the flood basically in an (Refer Time: 11:13) or bridges the dam.

So, in this case you will basically take a value of a is greater than 0. So, with this I will just read it whatever I said. As a second example assume a civil engineer is building a dam and he she interested in finding the height of the dam which is being build. If due to some error the height is estimate to be greater than actual value, then the cost the engineer incurs on mainly due to the material and the labour cost which I said.

(Refer Slide Time: 11:37)



Loss Functions (Example # 02 for LINEX) (contd..)

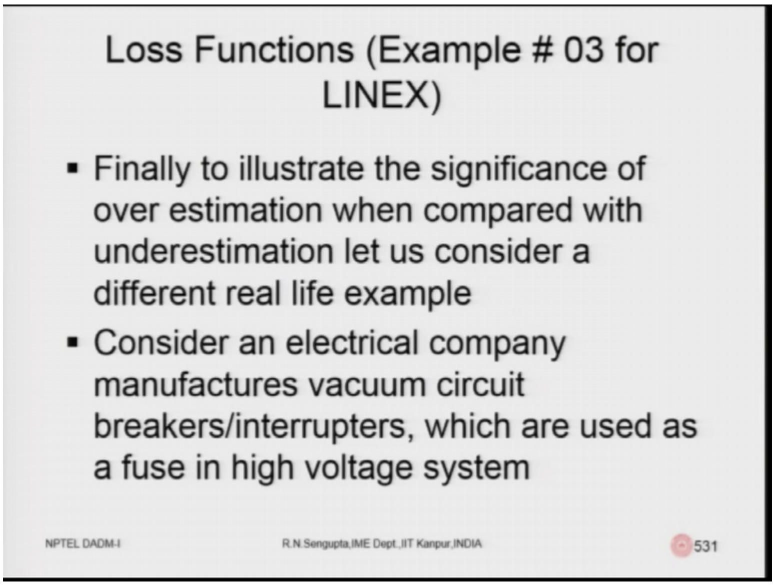
- On the other hand, if the estimated height is less than what it should be, then the consequences can be disastrous in terms of an environmental impact, which in monetary terms can be very high
- So it is logical to use a value of 'a' < 0 in such situations such that underestimation is penalized more than over estimation

NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 530

On the other hand, if the estimated height is less than what it should be, then the consequence consequences can be disastrous in terms of an environmental impact which is monetary in monetary terms can be very high so; obviously, we will think to have a situation where my mistake, in that case you will basically have an underestimation to be more penalized. I made a mistake it is not a greater than 0, it should be a less than 0.

So, it is logical to use a value of a less than 0, in such situations as that the underestimation is penalized more than the overestimation.

(Refer Slide Time: 12:12)



Loss Functions (Example # 03 for LINEX)

- Finally to illustrate the significance of over estimation when compared with underestimation let us consider a different real life example
- Consider an electrical company manufactures vacuum circuit breakers/interrupters, which are used as a fuse in high voltage system

NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 531

Let us consider a third example. So, consider that a you are making some electrical huge machines or you have purchased a huge electrical machine which is very sophisticated one it is costly; and you using that machine you basically producing some products in the factory, and the workmen work on that.

Consider that there is a warranty life for that machine where you need to basically stop the machine and change the vacuum circuit breakers. Those a trip switches or fuses say for example, and they are very essential to be changed because the warranty life if they exceeded, there is a high probability that some estimation surge of the voltage would have catastrophic loss on this machine may have may definitely affect the production and can have life threatening effects on the human beings who are working on that machine.

Consider for the example the actual vacuum circuit breakers or breaker whatever it is the actual warranty life is 1 year. So, in case 1 you predict; the actually should be 1 year, but you basically due to some information set loss you have two cases in case 1 you predict a warranty life of 14 months which is plus 2 and another case you basically predict a life of 10 months which is minus 2.

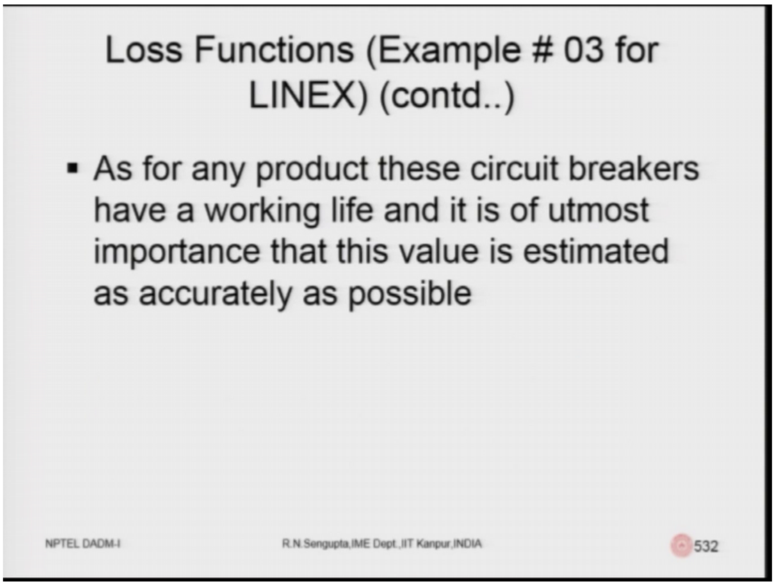
So, in the initial case if it is a squared error loss so; obviously, in that case 14 minus 2 which is plus 2 whole square is 4; and in the second case which is 10 minus 12 which is minus 2 whole square is 4. So, both the over estimation and estimation are equally penalized. But let us think what happens in the practical sense in the sense when if you are using a LINEX loss, then in the case when it is underestimated. So, what will happen? You will basically stop the production and the 10th month, replace that. So, in when you stop that and replace the vacuum circuit breaker; obviously, they would be man hours lost production lost which is fine.

But think on the other hand in case say for example, if you are not stopped that and basically continued working; then the probability of that surge of the voltage or electricity electrical circuit changes would have a catastrophic effect on the machine. It could have basically burnt your factory, maim would basically could have maimed or hurt some human beings stopped your production.

So, in this case you would and in case if we your basically overestimated it, made it at 14 so; obviously, would have stopped working on the machine, after the 14 month change the circuit breakers. Hence your protein initial production could have been high, but a in case if there is a catastrophic loss, the overall loss because the probability would be much higher because it is 14 and not 12 would have been much higher.

So, in this case you will be tempted to take the value of a in such a way that over estimation would be more penalized. So, let me read it as it is written. Finally, to illustrate the significance of overestimation when compared with the underestimation, let us consider a different life example. Consider in electrical company manufactures is vacuum [vocalised-noise] vacuum circuit breakers, interrupters which are used in as fuses in high voltage circuits or system.

(Refer Slide Time: 15:22)



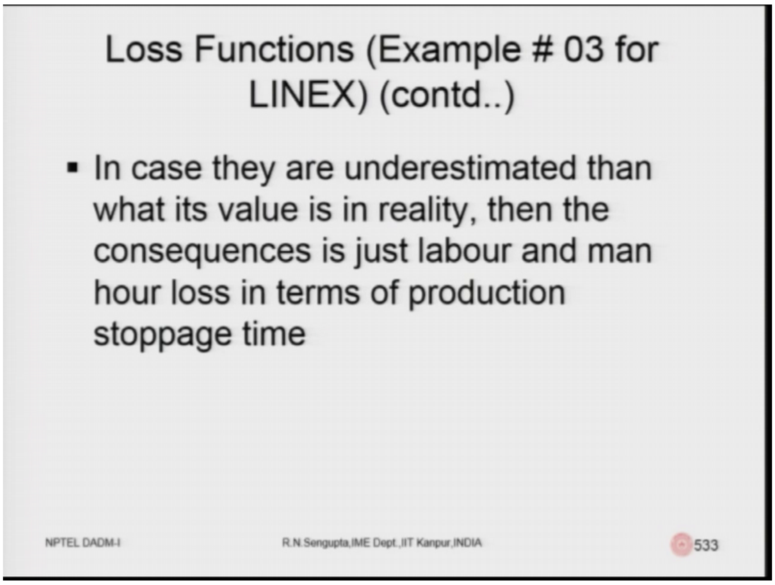
Loss Functions (Example # 03 for LINEX) (contd..)

- As for any product these circuit breakers have a working life and it is of utmost importance that this value is estimated as accurately as possible

NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 532

As for any product these circuit breakers have a working life and it is of utmost important that this value is estimated as accurately as possible because it will be used by your customers.

(Refer Slide Time: 15:33)



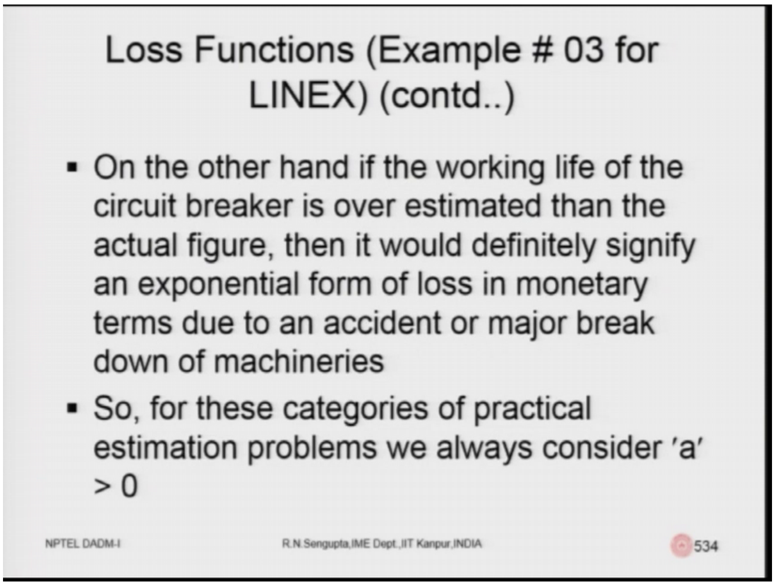
Loss Functions (Example # 03 for LINEX) (contd..)

- In case they are underestimated than what its value is in reality, then the consequences is just labour and man hour loss in terms of production stoppage time

NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 533

In case they are underestimated, then what is its value is in reality; then the consequence is just in if they are underestimated then the consequences would be just labour and man hours loss in terms of production and stop its time.

(Refer Slide Time: 15:46)



Loss Functions (Example # 03 for LINEX) (contd..)

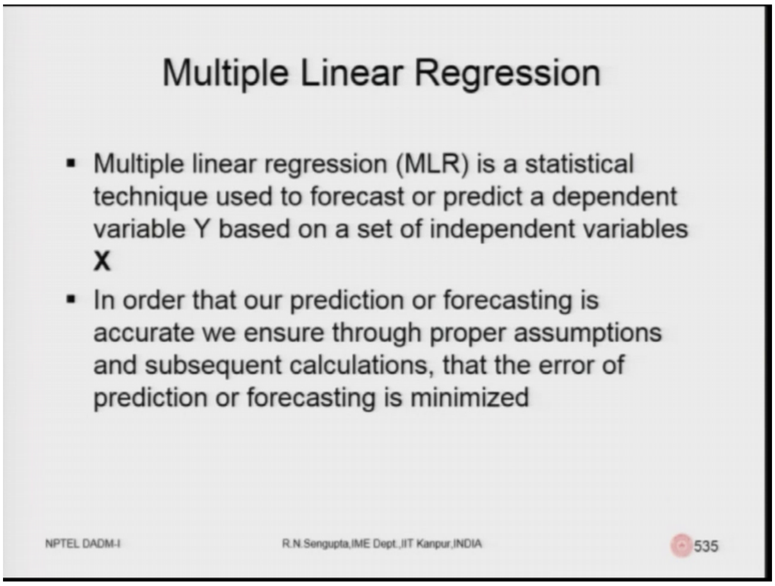
- On the other hand if the working life of the circuit breaker is over estimated than the actual figure, then it would definitely signify an exponential form of loss in monetary terms due to an accident or major break down of machineries
- So, for these categories of practical estimation problems we always consider 'a' > 0

NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 534

On the other hand if the working life of the circuit breaker is overestimated than the actual figure, then it would definitely signify an exponential form of loss in monetary terms due to accidents or major breakdown of the machineries or which could have catastrophic consequences.

So, for this case for of practicality, over estimation would be more penalized and you will basically take a value of a is greater than 0.

(Refer Slide Time: 16:11)



Multiple Linear Regression

- Multiple linear regression (MLR) is a statistical technique used to forecast or predict a dependent variable Y based on a set of independent variables **X**
- In order that our prediction or forecasting is accurate we ensure through proper assumptions and subsequent calculations, that the error of prediction or forecasting is minimized

NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 535

Now, you will switch our discussion to multiple linear regression. So, in multiple linear regression is basically a statistical techniques used to forecast a predictor dependent variable Y, based on a set of independent variables X. So, X is basically a vector. So, if you remember I did discuss this point, I will try to basically go more in details not for the proofs, but more of the solutions. So, consider you have you want to predict the temperature and for the temperature you think that humidity, then pressure, then height at which place you are measuring are the 3 variables which are very important and consider the wind speed it is also important.

So, he will basically have x 1, x 2, x 3, x 4 where p is equal to 4 are the 4 independent variables which are basically considered in order basically predict the temperature. Now obviously, if you remember we are discussed many of the assumptions based on which will do the work assumptions being x 1 to x p, where p is the number of random variables independent random variables which are there for x, which are independent, they are normally distributed, they have a mean value of mu 1, mu 2, mu 3 till mu p plus a variance 0.1. Point number 2 there is; obviously, there would be an error as I mentioned that error would basically have a normal distribution in the very simplest case with the mean value of 0 and a variance of 1. And the covariance existing between the between x

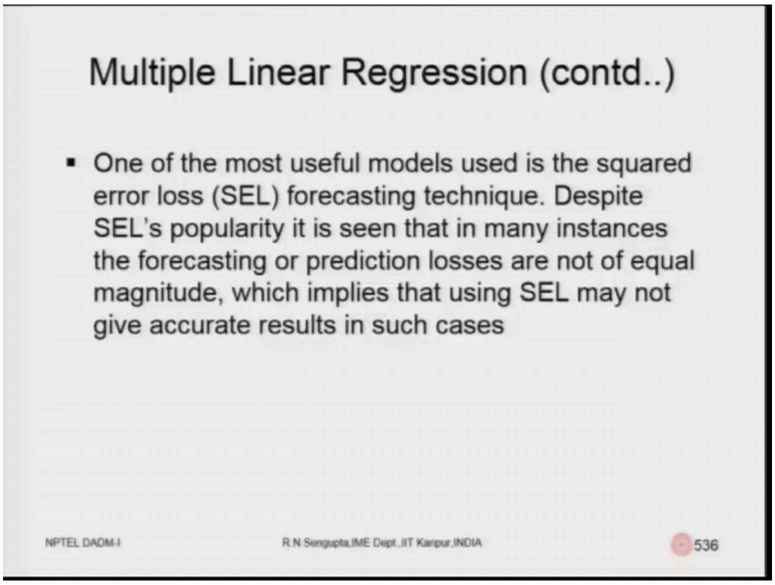
1 to x p and with respect to individually with epsilon would be basically 0, which means the errors and the random variables which are the independent ones are independent to each other. We will also consider the random variables are independent to each other in the sense one error does not affect the other, and we will we will consider that to be true even though practicality that may not be true, but will consider that to be true in order to basically solve a problems is accordingly.

So, if you remember those different types of assumptions which you have assumed. So, in order that our prediction of our forecasting in accurate, we ensure that proper assumptions and subsequent calculations trend that the errors of prediction the forecasting are minimized. So obviously, we will see if you remember what we did was, we considered the errors error was basically y minus y hat; y was the actual value y hat is the predicted value.

Now your main aim was basically to estimate the values of the parameters corresponding to x 1 to x p which one basically will denote as beta 1, beta 2, beta 3 till beta p. So, our main task was basically to estimate the those parameters because if the population values was not known our best way would be to take a sample, estimate those values of beta 1 to beta p and then, try to utilize those hat values that is beta 1 hat, beta 2 hat till beta p hat which is the estimated values use them to basically predict the future values of y, and then compare the errors which are there with respect to the actual value of y and the predicted value of the forecasted value of y.

Now whenever you are trying to do; obviously, you will always ensure that the values of betas which you find out in the long run the estimated values of beta hats would be exactly equal to beta, and also the variances what we will consider would also be calculated in order to find out the variances of y, which will depend basically on the individual variances of x as well as the variances of the error.

(Refer Slide Time: 19:41)



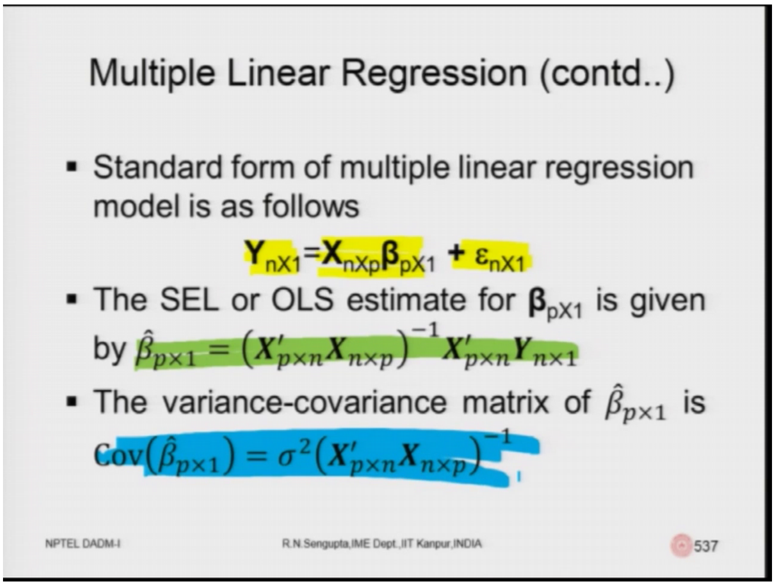
Multiple Linear Regression (contd..)

- One of the most useful models used is the squared error loss (SEL) forecasting technique. Despite SEL's popularity it is seen that in many instances the forecasting or prediction losses are not of equal magnitude, which implies that using SEL may not give accurate results in such cases

NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 536

So, our model which will have would be like this will come to that. One of the most useful models used in the square is the squared error loss or ordinary least square forecasting method, despite squared error loss popularity it is seen that in many instances the forecasting or prediction losses are not of equal magnitude there is as I discussed. In the LINEX loss function which we implies that using SEL that the squared error loss me not give accurate result, but still will basically solve and understand the problem from this ordinary least squared or the squared error last point of view.

(Refer Slide Time: 20:10)



Multiple Linear Regression (contd..)

- Standard form of multiple linear regression model is as follows

$$\mathbf{Y}_{nX1} = \mathbf{X}_{nXp}\boldsymbol{\beta}_{pX1} + \varepsilon_{nX1}$$

- The SEL or OLS estimate for $\boldsymbol{\beta}_{pX1}$ is given by $\hat{\beta}_{p\times 1} = \left(\boldsymbol{X}'_{p\times n}\boldsymbol{X}_{n\times p}\right)^{-1}\boldsymbol{X}'_{p\times n}\boldsymbol{Y}_{n\times 1}$
- The variance-covariance matrix of $\hat{\beta}_{p\times 1}$ is $\mathrm{Cov}\left(\hat{\beta}_{p\times 1}\right) = \sigma^2\left(\boldsymbol{X}'_{p\times n}\boldsymbol{X}_{n\times p}\right)^{-1}$

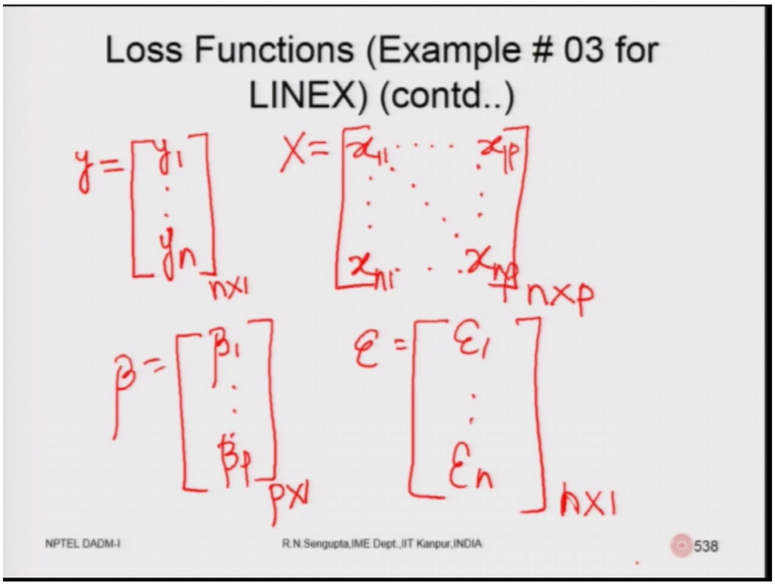
NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 537

Now, the general standard form of the multiple linear regression model is like this.

So, this is Y. So, you want to basically predict or forecast Y; Y is basically for the sample is n cross 1 matrix. So, first let me make a blank ppt and then add these values. So, I will add 3 blank slides and then continue. So, the standard formula multiple linear regressions is this Y which is the independent variable is dependent on X and then error term.

So, if you just note down which I would like to highlight; note down the size of the vector of the matrix. Y is the size n cross 1, X is the size of n cross p, beta is the size of p cross 1 and epsilon is basically size n cross 1. So, n cross 1 actually equals to n cross p into p cross 1 which is again n cross 1 and plus n cross 1 so; obviously, the dimensionality is maintained.

(Refer Slide Time: 21:20)



Loss Functions (Example # 03 for LINEX) (contd..)

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \qquad X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}_{n\times p}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{p\times 1} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

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

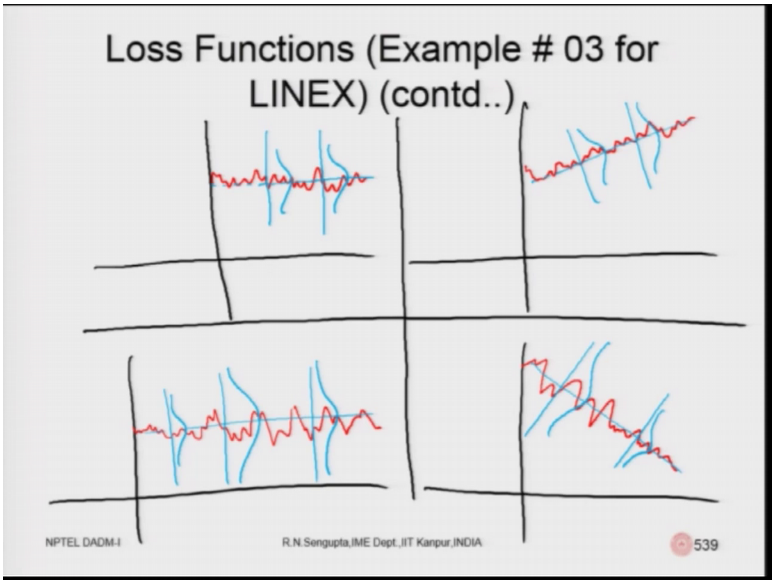
So, let me. So, your I am we have done that, but I have still I will I will repeat it. So, Y this is capital Y is equal to Y 1 till Y n; X is equal to this is n cross 1, n cross. I will write down the values. So, for the first variable you will basically have the data’s given here. So, for the first data point so; obviously, they will be x 1, x 2, x 3.

So, corresponding to that this one where I am highlighting the one comma one element is basically the first element with the first random variable, and in this case where which is basically the last element in the first row is basically the first element for the p th random variable.

Similarly the extreme left most left down value which I am now highlighting which is x n 1 is basically the value of the first random variable its n th value, similarly the last one which is the cell n comma p is basically the n th value for the p th random variable.

So, these are the betas beta 1 to beta p, epsilon 1 the errors which is epsilon 1 to epsilon n. Now based on that will basically try to find out the values of beta hats which will also be your size let me write the size here p cross 1 (Refer Time: 23:52) n cross 1. Now when we I will just move back a little bit in the sense in the discussion, if when I mentioned that dependent structure of the errors and dependent structure of the x s are independent actually in the diagrammatic form they would look like this and draw it.

(Refer Slide Time: 24:14)



Loss Functions (Example # 03 for LINEX) (contd..)

NPTEL DADM-I

R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA

539

So, I will draw up panel of 4 diagrams. So, in case if the random variables has the mean value fixed with respect to not changing with respect to time, and the variance also not changing with respect to time it would be like this. Mean values here and the variances are also are also like this. (Refer Time: 25:04) the mean value is changing increasing or decreasing whatever it is, but the variance is fixed consider is increasing, but the variance is fixed. Look like this, in case if the mean value is fixed, but the variance is changing.

So, something like this. In case mean value is increasing or decreasing, consider in decreasing, variance is also decreasing. So, you will basically have these type of graphs with respect to the random variable, which is fluctuating. So, let me come go back to the initial slide.

So, the squared error loss of the order least square estimate of beta; so, I am just giving you the formulas. So, this would be very easy for you to find out any book and read it. So, what you do is that, you find out the errors, square them up, some them up, differentiate with respect to the betas put them to individual to 0, differentiation is done with as partial differentiation, as you put them to 0 find out beta 1 hat, beta 2 hat, beta 3 hat all these things in the vector form it is as given as a violet.

Let me change the colour I think it will be much easier for you to appreciate, and the corresponding variance covariance for the matrix for the betas which are the errors, the principal diagonal will be the variance of first with respect to first. 2 comma 2 would be the variance of beta 2 with respect to beta 2 and similarly the last column n comma p comma p would be the variance of beta p with respect to beta p and all of the diagonal elements are the covariances.

So, this is again values you can basically. These are just calculations can be found out in any simple book, will try to utilize this concept in the later class. So, with this I will end the this lecture and considering that we have another two lectures for this 11th week, I will try to wrap up regression as you point and just discuss a simple problem for that have a nice day and.

Thank you very much.