Total Quality Management - II
Prof. Raghunandan Sengupta
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur

Lecture - 36
Fitting regression Models – I

Welcome back my dear friends. Very Good Morning, Good Afternoon and Good Evening to all of you and this is the TQM II class under the NPTEL MOOC series and I am Raghunandan Sengupta from the IME department IIT Kanpur. So, if you remember we are in the last week; that means, we are going to start the last 5 lectures we will have in the sequence and if you remember in the 35th lecture because we are still left with 36th, 37th, 38th, 39th, 40.

So, in the 35th lecture fag end we are discussing about multiple linear regression and what is the concept of multiple linear regression; mainly being the parameters should be linear because that is what you want to estimate. And using that concept of estimation you want to basically put them in the future forecasting or prediction models and then try to find out how good or bad are models are.

If you remember the simple linear regression was y is equal to alpha plus beta 1 x 1 plus epsilon; where alpha was basically, intuitive feel you can have if you consider the very simple equation we learned in class ten; y is equal to m x plus c. So, alpha and c were are similar in nature. So, we basically where it cuts the y axis and the beta 1 is basically the rate of change of y with respect to x 1. That is basically the marginal rate. So, similar is that the concept of y is equal to m x plus c that m value and this epsilon is basically the error which you have.

So, that the white noise. Then we I will come to the later parts of how we do that and how we find out the betas and their beta hats, hats are the estimated values. Then we have the multiple linear regression model, where rather than having only one x you have multiple x. So, it is the equation is equal to given as y is equal to beta naught plus or alpha whatever is said. I am considering beta naught for ease of understanding. So, you basically have beta naught plus beta 1 x one plus beta 2 x 2 plus beta 3 x 3 plus dot till beta k x k, where is the number of variables which are used to predict y plus epsilon again the white noise.

Now, in this case the intuitive feel of beta 1, beta 2, beta 3, till beta k with actually means they are the marginal rates of increase of x 1, x 2, x 3, till x k considering the other variables are constant. In the sense that when I am considering beta 1, we will consider beta 2, x 2 to x k as constant, they are not changing. Hence we try to find out the partial derivative or the rate of change of y with respect to x 1.

Then when we basically go to x 2, then in that case beta 2 is basically the partial derivative or the rate of change of x 2 of y with respect x 2 keeping x 1, x 3, x 4 till x k as constant and then we do the modeling part. Now when you are considering these model, both the simple linear regression and the multiple linear regression, we always consider few very simple assumptions, which intuitive which may not be absolute practically possible, but they give you a very good feel of about the model. So, they were basically the x 1 to x k or even if there is only one x that is x 1. They are all normally distributed and hence any convex combination or normal distribution we know it is normal; hence y would also be normal point 1.

Point number 2 will consider that the errors and independent of each other; that means, the errors do not change or do not affect the errors of the next time period or one period of t period, errors is not effected by t minus 1 period. Number 2 the covariance's between the, which means basically the covariance's between the errors is 0, variance of the errors is fixed; is sigma square or one depending the simplicity on the model will consider the errors as the main value of 0. Will also considered the covariance's existing between the error and the x values x 1 or x 2; where x 3 till x k are all 0 will also considered.

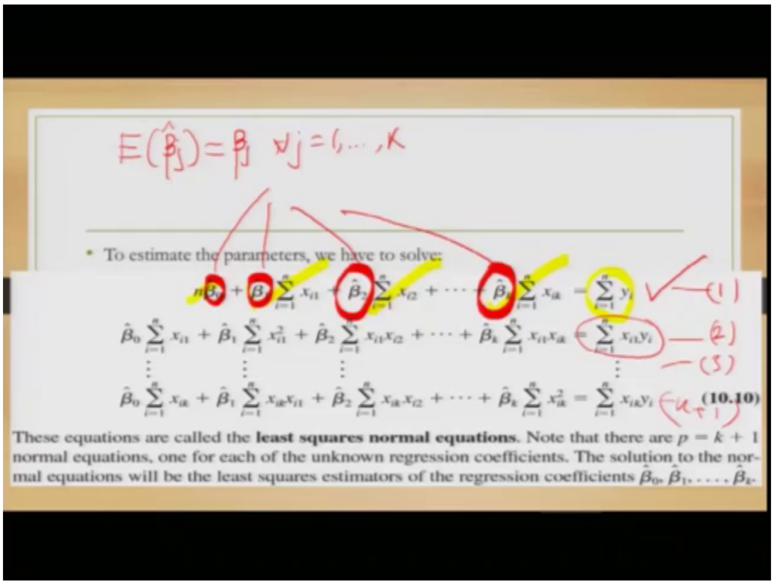
The relationship of the covariance is existing between the x is are also nonexistence; that means, they are independent. So, if they are independent it will, if you basically use the concept of matrix we will comes come to the con concept that rank of a matrix. So, the rank of the matrix depending on, there are n number of observations; that means, y 1, y 2, y 3, y 4 till y n. Hence corresponding to x 1 there were also be n number observations, x 2 also there would be n number observations, till x k also there will be n number observations. And obviously, there will be epsilon 1 till epsilon n. And when we basically denote it has a matrix multiplication, so like y is a matrix, capital X is a matrix, cap beta is a matrix, epsilon is a matrix. We are a trying to basically denote is an a matrix fast multiplication, this multiple linear regression. Will consider that the rank of x as k. So, if it is not k it will mean that one or more than one of the columns or the rows of x is can be expressed as a convex combination of the other rows; which means then it will be difficult to find out the exact values or the inwards would not exist and hence trying to find out the values of betas, which are the estimates based on which you are going to predict the future values would not be possible.

Now, whenever you are considering this we are considering that you want to basically minimize that error with respect to the unknowns; unknowns being beta naught, beta 1, beta 2 till beta k. Now the question is that how we do that? I have already discussed, but I will still repeat it. It means actually that you want to basically minimize the deviations and the deviations can be of many times. It can be minimum absolute deviations, it can be basically square of the deviations, it can be cube of the deviations, it can be only mode of the deviations, it can be some convex combination, the waited deviations.

But generally we consider is that we consider the square of the divisions, sum them up, find out the errors, totals sum of the square of the errors and then differentiate partial differentiate, remember that. Because the partial differentiation means that we can going to consider other factors are constant. What are the other factors; that means, when we cons we can be considered that the partial derivation of beta of the of the sum of the square of the errors with respect to beta 1 we consider beta 2 to beta k as constant.

When we partially different differ differentiate the sum of the square of the error with respect to beta naught, then we consider beta 1 to be k the as constant. When we basically considered partial derivative of the sum of squares of the errors with respective beta i, then will consider beta naught, beta 1 till beta i minus 1, as constant plus beta i plus 1 till beta k are also constant; that means, that is why when I mentioned that few minutes back that we that What is beta naught? What is beta 1? What is beta 2? They are the partial derivative of the rate of change of y with respect to those access. So, that is what I mean once we say that we are trying to basically partially differentiate that. So, with this will start this 36th lecture.

(Refer Slide Time: 07:55)



So, now we want to estimate the parameters we differentiate as I mentioned few seconds back with differentiate with respect to the betas. Now, if you basically do that you will find out, another thing which I have completely forgot that I did not mention as we are discussing introduction for the past 5 minutes that it is the linear equation; In the sense that it is not a square or cube with of with respect to the values of betas.

So, you can have for example, a multiple linear regression of this form; say for example, we can have y is equal to beta naught plus beta 1 x 1 plus beta 2 x 2 plus beta 1 2 x 1 x 2. So, in that case you have 3-4 parameters which you need to estimate which are beta naught, beta 1, beta 2, beta 12; suffix the 1, 2, 3, 4 whatever I am mentioning are the suffixes. And here in place of x 1, x 2 which you have been multiplied we can considered a new variable as x c and a consider beta 12 as a new estimate we want to find out which is basically beta 3.

But, so this is allowed, but what is not allowed is like this equation which would basically take us in the realm of non-linear regression, which we are not going to discuss here. It will be y is equal to beta naught plus beta 1 x one plus beta 2 x 2 plus beta 3 cube x 3. So, in that case if when you try to find out and differentiate the sum the square of the error with respect to beta 3, it would not give us a term by which you are able to find out the values of beta 3. So, this type of non-linear regressions are not allowed to be discussed here; obviously, there are the different ways you can tackle those problem, but we are going to avoid that.

So, to estimate the parameters if you differentiate with respect to the beta so, we have are the set of equation. So, if there are beta naught to beta k, you will basically have such k plus 1 betas, you will have basically k plus 1 equation; partial derivatives put to 0 based on that you can find out the values of beta naught hat, beta 1 hat, beta 2 hat till beta k hat.

So, this is what we have. So, this is the background which I gave if you differentiate. So, the equation so you have basically sum of the y values; that is basically will be equal to. So obviously, when you are adding of the sum, it will be sum of all the values of x 1 to x k. So, the left hand side so, this have been just written in the reverse direction.

So, the first variable on the right hand side of the equation was basically beta naught. So, how many times you have adding beta naught? If you add up all the n values, it will be adding n number of beta naught. That is why it is n. So, when I go to summation of x 1; so x 1 would basically x11, x12, x13 till x1k x1n sorry. So, the first one basically denotes the variable number and the second suffixes basically denotes the reading number.

Similarly for the x, it will be x 2 21, x22, x33 till x2n, for the third one it x 31 till x 3n, for the last one it will xk1 to xkn. So, this is what they are denoted. This is for the first one. I am just putting a tick mark. So, you basically have a note. This is for the second one, similarly you have 3rd, 4, 3, 6 till the kth one. Now we just note down I am just highlight use a different ink color. So, just note down these hats. So, these hats are the estimated value based on which you are trying to minimize the sum of the squares of the errors.

So, this beta 1 hat is basically the estimated values of beta 1 and obviously in the long run, as I keep repeating in many of the examples. If you toss a coin,00 times say for example, you will get for head 45 times, then you again toss it00 times you get heads 55 number of times, if you keep repeating it then find out the averages of this head. So, in the first case it will be 45 by 100, second case it will be 55 by 100, third case it will be anything whatever value. And if we keep finding on the averages of the averages in the long run for an unbiased coin it will be 0.5.

Similarly, if you do roll the die and the actual probability of getting a 1, 2, 3, 4, 5, 6 in an unbiased die, it will basically be 1-6. If you keep running it, say for example, 200 times, you get some number of once. Again you roll it for two 200 number times of the second set you will get some other number of one's. If you basically find out the radio frequencies of the day or the chances of the number of ones in the first role of 200, second role of 200, third role of 200 and basically add them up and find out the averages it, in the long run it should come out to be 1 by 6.

So, actually what we mean is that there, there are they should be unbiased; in the sense that the expected value of the beta j is j is equal to 1 to k should be exactly equal to the, so expected value of the hats sorry. So, they should be exactly equal to the values of beta j. Similarly, you have beta 2 hat and similarly for the last one will be beta k hat. So, the expected value what I mean is the expected value of beta no let me this quite difficult to use it. I will use the not a highlighter, but the pen. So, you have basically expected value of beta hat j should be equal to beta j, where for all j is equal to 1 till k. So, you want basically have this and this is what would be true for beta naught here, beta 1 here, beta 2 here till beta k. So, all this should be hold.

So, when we basically the multiply it with now when we differentiate it and then basically do the same thing for beta 1, then beta 2 till beta k and we have the k equations. Because the first one is the total number of equations are k plus 1; for the first one for beta naught, second for beta 1, third for beta 2 and the k plus 1 for the k plus 1 for the kth one beta k.

So, the first equation which where we have highlighted is basically with respect to beta naught. Similarly if you have the other equations; obviously, you will see that you will be multiplying, in the first case as with x11 multiplied by y1 one y1, then it will be x12 multiplied by y2, then x13 multiplied by y3; multiplied them sum them up and then when you basically put it in the equations you will have the equations as given which I will denote as say for example 2. So; obviously, this is the 1.

Then we do it for the third equations. So, in third equations you will basically have which is not return here. I am just reading, it will be beta naught hat summation of I is equal to 1 to n because there are n number of readings. It will be x11 and then it will be a corresponding to that it will be x12 then beta12, then you will basically have a summations i i is equal to 1 to n and correspondingly we will basically have the second term, third term. So, it will be x11 square, then the third term, fourth term till the last term.

So, will basically have the third equation and the last equation if we note down, it will be exactly the same just replace; in the first case x 1was being multiplied for all terms. In the second equation x 2 were basically multiply for all the terms, in the last equation it will be x k be multiplied for all the terms as it is. So, it will be x 1k multiplied by beta naught hat, then x 1k into x1 i1; so i i11, 12, 13, 14 so on and so forth. So, the third the third term would basically be beta naught hat and x ik into x i2.

So, it will be basically if we sum them up, it will be x the kth variables first term multiplied by the second variables first term. Then the kth variables second term multiplied by second term of the second variable and it will go on like this. So, you will basically have the k plus 1th equation.

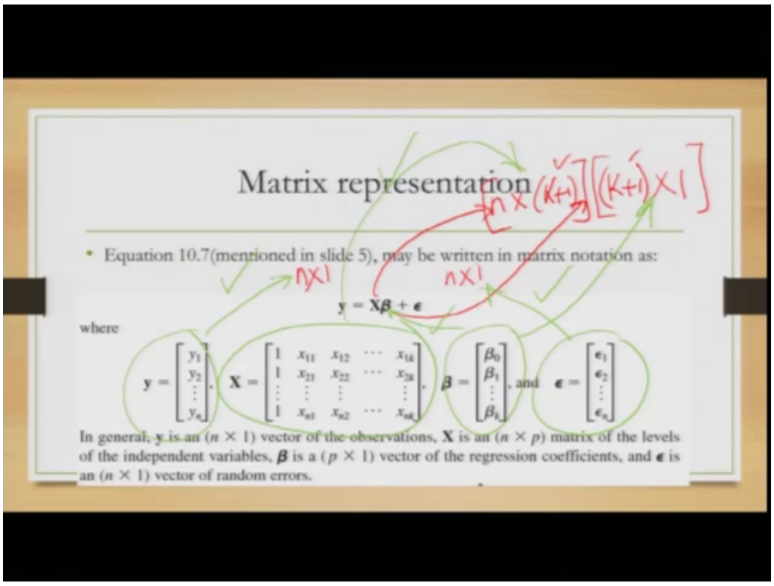
So based on that, you can basically find out the beta naught to the beta ks; these equations are called the least square normal equations. Why least square because, you are trying to minimize the sum of the squares and normal equation because you have normalize them in such a way that, you want to find out the beta naught hat till beta k hat. Note that there are yes, there are p is equal to k plus 1 equations as I have been mentioning, one for each of the unknown regression, regression coefficient. So how many regression coefficients they are? Beta 1 to beta k, for all the axis and the first term being beta naught.

The solutions to the normal equations will be the least squares estimator of the regression coefficients as I mentioned; beta naught hat, beta 1 hat till beta k hat, such that when you find out the nth plus 1 value of y which is the estimated value of y which will be given by y hat suffix n plus 1 and you already have the actual value of y n. So, the difference would be the error. So, you write down the error for the nth plus 1, then again you use this beta naught hats; remember the hats are being used. Then you find out the y hat for the nth plus 2, write it down, find out the difference between y n plus 2 and y hat n plus 2 which is the errors n plus 2. Keep doing it for the third time; that means, n plus 3 reading. Then you do it for the n plus 4 reading, n plus 5 reading so and so forth.

Say for example, you collect 100 such errors. Now technically if errors means n plus 1 error, n plus 2 error, n plus 3 error till the n plus 100 error. Now, if you basically go back to the assumptions which I did mention time and again, then the expected value of the errors in the long run should be 0; which means the if you add up the sum of this error for the for the nth plus 1 term, nth plus 2 term till the nth plus 100 term, then it should technically be 0.

And if you find out the variance of that you should also be whatever you assumed is before starting of the solution, it can be either 1 of sigma square. But obviously, this sigma square is not changing with respect to time which I did mention as an important assumption in multiple linear regression.

(Refer Slide Time: 19:22)



So, the equation now if you remember I did mentioned that it can be written as a matrix 1. I did not want to write it down there because it will be coming up soon as I thought and here it is. So, the equation actually can be written as y is equal to. So, these are all bolds. So, I am not as shown in the slide I am not mentioning them. So, they are bolds so, y is equal to x into beta plus epsilon. So, what if we want to basically find out whether is a matrix and whether matrix multiplication concept is valid. So, this y is basically of size n cross 1.

This epsilon I am I am just for the time being skipping x into beta. I will come back within few seconds. Epsilon is of size n cross 1, so obviously, if you have n cross 1 on the left hand side and n cross 1 on the right hand side also. So, this x into beta should also be of the size n cross 1. Now in this case we already have beta as so I am basically writing beta. It is of the size k plus 1 because there is k plus 1 terms; beta naught till beta k cross 1. And the value of x, obviously would be n cross k plus 1. So, n in cross k plus 1 into k plus 1 cross 1; so obviously, this and this would vanish so; obviously, and in the end you have n plus 1 for the terms x into beta also.

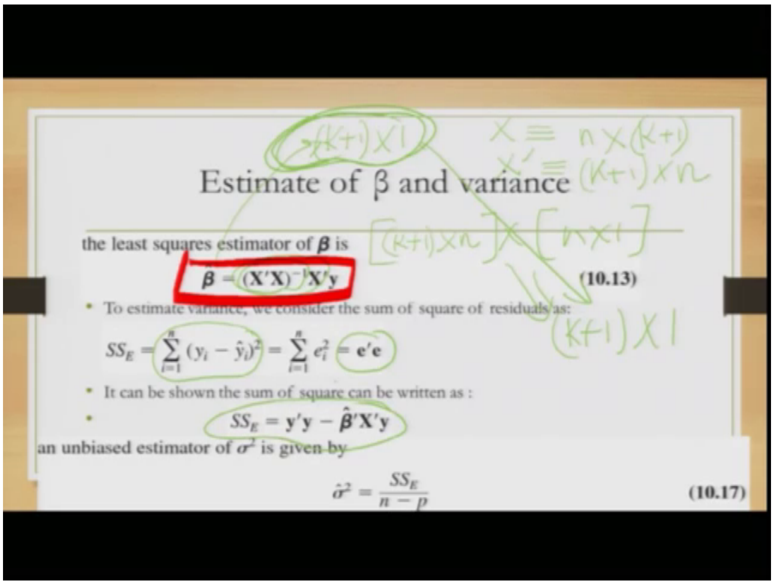
Now, let us see whether it matches. I have not seen in going to the details of the matrices look like. So, here this y will let me use a different color to highlight. So, this y is n plus 1. Yes it is matching. So, let me put a tick mark with the concept which I am mentioned. Let me go to the epsilon first and then come back to x beta. So, this value of epsilon is n cross 1 it is matching well good. Now the value of beta which I mentioned was k plus 1 cross 1. Now, let me see what is there? It is matching.

Why? Because there is beta naught is a 0th one or the first one. Then you have 1, 2, 3, 4 till k. So obviously, be k plus 1. So, this is also matching and let me put mark here also so, this is matching. And when I come to the x matrix, so obviously, I say that it will be n cross k plus 1. So, let this find out how many such rows and columns are there. So, this is the first row second row third row till n number rows how many columns. So, this is the first one, this is the second one, third one so on and so if the k plus 1. So, this is also matches here.

So, on our discussions are in line with our with our concept which we are basically discussing. In general y is n cross 1 vector observations as mentioned few minutes back; n is n cross p. So, here remember p is the size k plus 1. So, n is a n x is a n cross p matrix of the levels of the independent variables as it is rightly so. And in case by the way in case is beta naught is not there, the first column which you have for x would vanish. So obviously, in that case beta would be of size k cross 1 and x would be size of n cross k because that that the first beta naught is not there.

So, now, continuing the reading. So, beta is a size of p cross 1 which is an k plus 1 cross 1 vectors of regression coefficients and epsilon is the error of size n cross 1 vectors of random errors and all obviously, all the assumptions which have been discussing holds.

(Refer Slide Time: 23:27)



Now, the least square estimates. Now what we do you mean by least square estimates? We have differentiate the sum of the squares with respect to beta naught, put it to 0, then differentiate the sum of squares with respect to beta 1 put it 0. So, we have basically k plus 1 equations, we have k plus 1 unknowns find them and once we find them the values of beta naughts or beta 1, beta 2, beta 3; all in the hat values. Then because there estimated values would basically give us the best estimates of the beta values with respect to the concept of trying to minimize the sum of squares. So, when we find it out the actual matrix notation is like this. So, this is the matrix notation. So, how do you find out the values of beta hat? Beta hat would be you basically multiply.

So, now remember what is the actual size of beta hat? If there is beta naught it is of size k plus 1 cross 1. So, let are let why I mentioning that let me basically give the concept in a in to deform. So, it will make sense to you. So, this is of size if beta naught is there, k plus 1 cross 1. So, let is let us see with this balances on the left hand side; on the right hand side. Now we have already considered the value of x. So, what was X size? X size was n cross k plus 1 and y. So, so X transpose would be of size k plus 1 cross n.

Now, let us go by one by one. For the first term which is X transpose into X. So, X transpose into X would be of size k plus 1 cross n into n cross k plus 1. So obviously, n n n is not there it will be of size k plus 1 cross k plus 1. So, at an inverse of that would also be of size k plus 1 cross k plus 1. So, the first step is done now you have again going to multiply with X transpose.

So, what is this term the size of X transpose is k k plus 1 cross n. So, k plus 1 cross k plus 1 into k plus 1 cross n would basically be of size k plus 1 cross n. So, this is also taken care. So, this is now k plus 1 cross n. Now what is y? y we know is basically of size n plus 1. So, n cross 1, so obviously, it will be cross n cross 1. So obviously, in the end you will have the output as k plus 1 cross 1.

Now, let us see where is balances left hand side? You have already said beta where I am highlighting with my highlighter is size k plus 1 cross 1 and it basically matches this one. So obviously, the dimensions are maintained and the way of we have we have multiplied the concept of betas are right. To estimate the variance is consider the sum of squares of the residuals as I mentioned that we find out y whatever the value of nth plus 1, nth plus 2, nth plus 3; that actual value of y minus the estimated value of y. Find out the difference. It can be positive-negative.

Then square them up, sum of the errors. To estimate the variance we considered the sum of the squares of the residuals as mentioned here. This is the actual value, this is estimated value; we square them up sum them up. So, this is a square term which is basically a vector multiplied by transpose. So obviously, the size would be of dimension 1 cross 1. It can be shown that the sum of the square of the errors can be written as, as I have been explained you can find out the sum of the square of the errors and you can find out the sigma squares. Sigma squares would be sum the square or errors divided by the degrees of freedom.

Now, degrees of freedom like to spend one minute here total number readings is n, so for each trying to find out the beta naught, you do is 1 degrees of freedom. So, the total loss of degrees of freedom would be p which is k plus 1 because you have you lose 1 degree of freedom to find out beta naught, 1 degree of freedom basically to find out beta 1 so on and so forth. So, the total loss would be k plus 1. So, hence it will be divided by n minus in the bracket k plus 1 or n minus p as you have been denote it.

So, with this I will end the 36 lecture and try basically continue with the discussions on the multiple linear regression in more details. And basically see that the discussions we had till now how to find out the sum those squares divided by the degrees of freedom, find out the f values, they can be also utilize in ANOVA table for the regression models and you can pass very nice judgment that whether the variables we are considering or whether they are actually required to find out the overall level of fractional factorial models which are there for the design of experiments and the quality control part.

Thank you very much and have a nice day.