

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

Lecture - 29
Multivariate Statistical Analysis

A warm welcome my dear friends, a very good morning, good afternoon, good evening to all of you and this is the Data Analysis and Decision Making course under the NPTEL MOOC series. And as you know this is a 12 week course for 60 lectures, 30 hours and each week we have 5 lectures, each being a duration of half an hour or 30 minutes. And we are in the 20th 29th lecture, which is basically the last but one lecture for the 6th week. And if you remember the last day we basically, spend whole lot of time trying to understand the concept of qq plots, normative plots and technically the number of slides if you consider covered was not much, only maximum 3 or 4.

I will rehearse that and continue the discussion with normally plots and other concepts of multiple linear regressions. And as I said I made a change in the analysis of the multiple linear regression I will basically do it during the multivariate statistics. So, as I said that we need to check for the normality, so the normal what we do is that I will again read it because, we have already done it once and using the excel sheet.

(Refer Slide Time: 01:37)

To check for normality of data

We need to check for the normality of X_i 's and Y

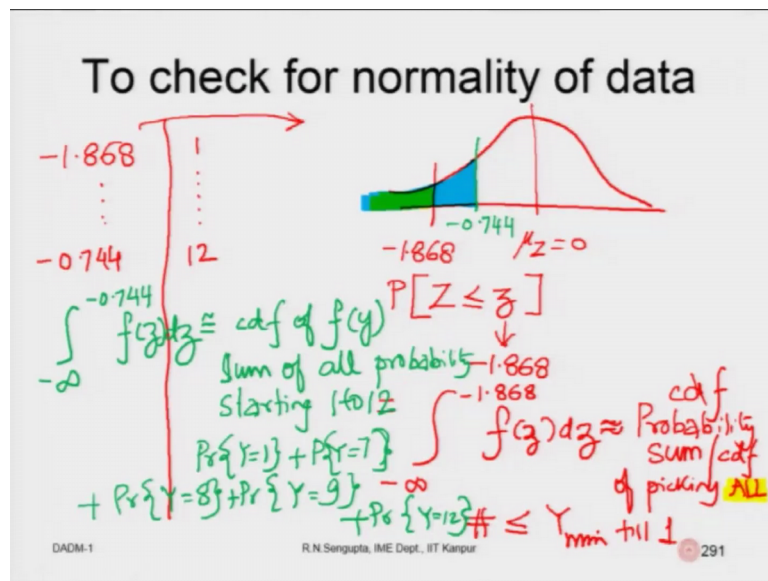
- 1) List the observation number in the column # 1, call it i.
- 2) List the data in column # 2.
- 3) Sort the data from the smallest to the largest and place in column # 3.
- 4) For each i^{th} of the n observations, calculate the corresponding tail area of the standard normal distribution (Z) as follows, $A = (i - 0.375)/(n + 0.25)$. Put the values in column # 4.
- 5) Use NORMSINV(A) function in MS-EXCEL to produce a column of normal scores. Put these values in column # 5.
- 6) Make a copy of the sorted data (be sure to use paste special and paste only the values) in column # 6.
- 7) Make a scatter plot of the data in columns # 5 and # 6.

DADM-1R.N Sengupta, IME Dept., IIT Kanpur290

So, we list the observation numbers in column 1 call it as i , then, list the data actual data for which you are going to do the $q-q$ plots and try to verify them. We list the data in column 2, then, sort the data which you have already collected which is in column 2, that will be sorted out in column 3 from the least through the highest. Now for each of the i which is basically on the first column on the i th observations we find the corresponding tail area of the standard normal distribution using the formula, which is as A is equal to i minus 0.375 divide whole thing divided by n plus 0.25 .

And then, you will basically use the norms inverse function of A to find out the column number values in column 5 and plot column 5 and the sorted out data, try to find out how the straight line is technically if it is a normal distribution with respect to the standard normal. Normal distribution I am talking about the data. Then you know it is a standard normal distribution. Now, the reason then I spend a lot of time trying to basically, give you the essence what we mean by $q-q$ plots in this I had drawn.

(Refer Slide Time: 03:05)



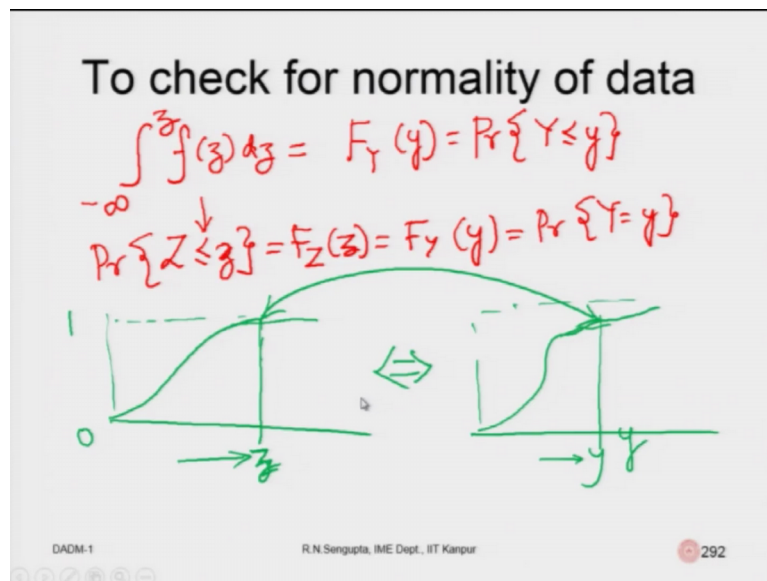
So, this is a data which I have here and basically use the highlighter as needed. So, I will just would not basically write much because or else this whole thing would be make (Refer Time: 03:18) too clutter. So, this the column when I am hovering my pen are the corresponding Z values, you have the corresponding values for the distribution and you basically do a 1 to 1 map; that means, because the cdf covered between any distribution is always 1, total c d f and corresponding to any X any Z you find out the corresponding

c d f values, equate them. If the if you are equated there been basically in mapped on to a 1 to 1 scale.

So, for each so this green and red one which I have written and exactly the calculations, but for different values you find out the c d f of f of y. And then basically, this is a discrete or a continuous whatever it is then, map it with the standard normal deviate what is the area and basically say that value of Z from the standard normal deviate and the corresponding value of Y for which the c d f values is equal to the standard normal deviate c d f, we say that they are equivalent and then basically plotted accordingly.

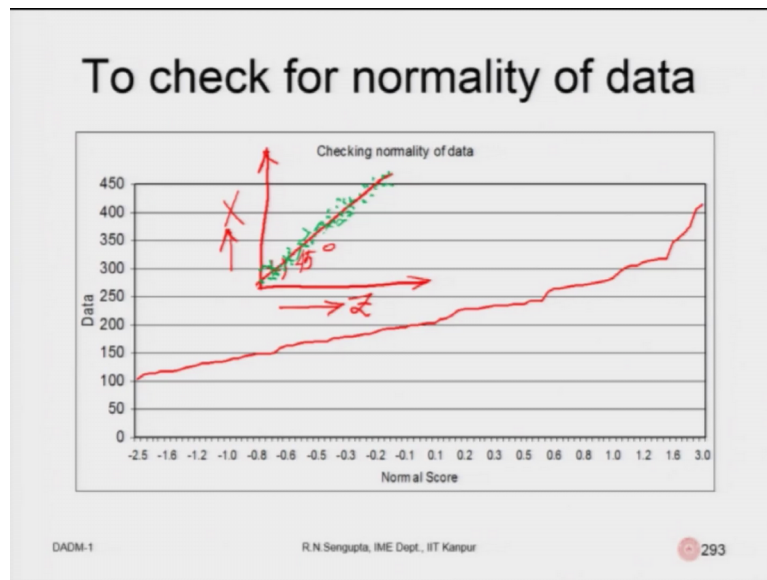
So, when we are plotting along the X axis, I will just I have a diagram in the next page, I will basically do it.

(Refer Slide Time: 04:35)



So, next to next please hang on. So, this is what we said that integrated from minus infinity to z, find out f of z and then, equate it to the c d f value for the corresponding Y distribution and calculate accordingly. And the green graphs which are there, they are basically the c d f for the Z distribution and c d f for the unknown distributions such that, we can compare.

(Refer Slide Time: 05:01)



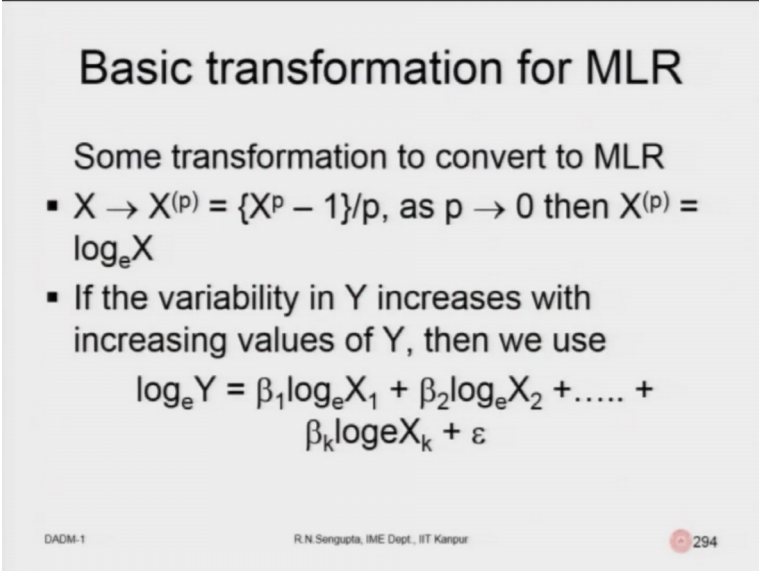
Now, when we plot this is what I wanted to say, when you plot here of the normal scores along on the x axis and the corresponding sorted out data points are along the y axis and what do we do is that, for some unit movement along x axis, you will cover basically, the z values small z values that will have a corresponding distribution corresponding c d f my apologies.

And that c d f value would also correspond to a certain value of X for the unknown distribution. So, if the quantile quantile plot values, for equal quantum of quantile on the x axis and equal quantum of quantile on the y axis, they have the same c d f value then, the unit of length moved along the x axis and the unit length of length move along the y axis to cover the c d f values are exactly the same. So the line joining would be 45 degrees aligned which means, the corresponding quantile quantile match in the corresponding distribution which we have is basically, the Z 1 and the unknown one which is X distribution.

We are basically, denoting in denoting it by the random variable X, if the values match then; obviously, it means that the quantile to quantile values are same. So, technically I am just writing here, the quantile quantile value should be exactly as 45 degrees line, along this you have Z and along this you have X, which is unknown and if, these values match so if, if the quantile let me use so, these values are along for 45 degree then

obviously, the quantile quantile plots are same. Hence it is a normal distribution for X also.

(Refer Slide Time: 07:23)



Basic transformation for MLR

Some transformation to convert to MLR

- $X \rightarrow X^{(p)} = \{X^p - 1\}/p$, as $p \rightarrow 0$ then $X^{(p)} = \log_e X$
- If the variability in Y increases with increasing values of Y, then we use

$$\log_e Y = \beta_1 \log_e X_1 + \beta_2 \log_e X_2 + \dots + \beta_k \log_e X_k + \varepsilon$$

DADM-1 R.N.Sengupta, IME Dept., IIT Kanpur 294

So, now, we will consider some transformations for the multiple linear regression model considering normalities true. So, in general, we will basically transfer to X to the power p, where p can be a value starting and can be integer non integer values. So, the transformation is given by X to the power p minus 1 by p and p can take any value. So, as p tends to 0, that transformation which is X to the power p minus 1 by p basically, becomes the Naperian log to the base e.

Now, the variability in Y increases within increasing value of Y and in that case you want to basically, bring normality in the data set such that you are able to forcefully bring normality. This is for the calculation part you are not going in a practicality. You want to bring them normality for the calculation then, what you do is that convert the values Y values or X values whatever, it into the log scale. But the problem is that in the normality scale, the values technically should be for the minus infinity to plus infinity which may not be the case when you are considering the log scale. So, these has to be basically analyze form the data analysis point of view.

(Refer Slide Time: 08:56)

Non-linear regression

- $y = (\beta + \gamma X)/(1 + \alpha X)$
- $y = \alpha(X - \beta)^\gamma$
- $y = \alpha - \beta \log_e(X + \gamma)$
- $y = \alpha - \beta \log_e(X + \gamma)$
- $y = \alpha[1 - \exp(-\beta X)]^\gamma$

NOTE: For all these and other models we minimize the sum of squares and find the parameters α , β and γ .

DADM-1 R.N.Sengupta, IIM Dept., IIT Kanpur 295

Few of examples of non-linear regressions are, if you see the first one you have basically beta and alpha in and gamma in the numerator and denominator. Hence, trying to find out the some of the square some of the errors squared up; obviously, when you are trying to differentiate with respect to alpha and partially differentiate with alpha beta and gamma, would not be basically give you simple linear equation they would may be quadratic other one may one. Hence trying to find out the betas alphas and gammas using the concept of multi-linear regression would not be possible.

Other equations which you gives you are the relationship between y and X would be the second one which is, y is equal to alpha into X in the bracket X minus beta through the bar gamma. The third and fourth one are with respect to the relationship of logarithm of X into, I am given with the relations of y, but the only problem here why it is become complicated is that because, gamma is inside the log hence trying to find out the partial differentiation fact this values and put into 0, trying to solve them, may not be straight forward.

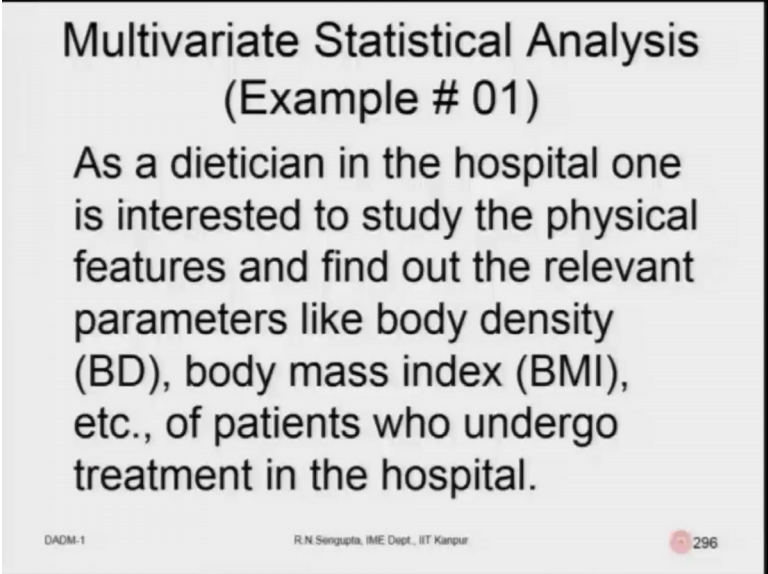
So, another the last one you where you basically, you have the relationship of X and y given the exponential function. For all these values another method we need basically, our main aim is to minimize the sum the square of those errors of errors and try to basically find out alpha beta gamma, using the alpha hat, beta hat and gamma hat. Hence

corresponding to the fact that you are basically differentiate partially with respect to this parameters.

So, now we will start basically little bit go the quantum of understanding would change. So, we will basically, consider the multi various statistical analysis and whatever you have studied in almost 30 lectures, we will try to basically utilize that in the remaining 30. So, we I thought that it is good that I go into details about simple uni-various statistics and give you thus information's are that you are able to handle the multivariate case in much better way.

So, considering these example, we it will be more example based because, that will basically bring out the flavor of the discussion.

(Refer Slide Time: 11:23)



Multivariate Statistical Analysis
(Example # 01)

As a dietician in the hospital one is interested to study the physical features and find out the relevant parameters like body density (BD), body mass index (BMI), etc., of patients who undergo treatment in the hospital.

DAAM-1 R.N.Singupta, IME Dept., IIT Kanpur 296

As a dietician in the hospital, you are considered one of them. A very big hospital say for example, Jaslok, (Refer Time: 11:34) or Calcutta medical college, whatever it is or aims. So, you work there and in the hospital one is you are basically, interested to study the physical features of the patients, who are admitted or who come for cure or medications whatever.

Anyone to find out the relevant parameters because, the dietician wants to find out the relevant parameters because, he or she will take tell the patient what should be the diet, what should be the amount of carbohydrate the person should take, what is amount of

protein one should take, if he or she the patient is a diabetic one. So, what should be the rules and regulations for food whether, the person should take fruits or not and all these things. If, somebody is allergic to milk, which is lactose intolerance, so then, how would you basically analyze the food habits such that, you need the parameters. Basically, you need to know some values of the human being as such the patients whom you are going to treat.

So, things relevant for the studies are body density, what is the body mass index. So, in many of the cases you try to find out the ratio of the height to the weight, then what is the fat content in the body and all these things are there. You basically get or the dietician gets all these from the patients who undergo treatment in the hospital.

(Refer Slide Time: 13:05)

Multivariate Statistical Analysis
(Example # 01 contd..)

Dietician's job is to decide on the right diet plan based on the data/information like percent body fat, age (years), weight (kg), height (cms), etc., of the patients

DADM-1 R.N.Sengupta, IME Dept., IIT Kanpur 297

The dietician job is to basically decide the right diet plan because, as I said diet plan whether amount of food calorific should be high low or maintain. May be some people have been told due to their restrictions that, the amount of water has also be to be maintained at a certain level. So, it basically, decide the right diet plan based on the data information like percentage body fat, what is the age of the patient, what is the weight of the patient, what is the height of the patient etcetera is important to be noted now.

(Refer Slide Time: 13:45)

Multivariate Statistical Analysis
(Example # 01 contd..)

Let us use the data
<[http://wiki.stat.ucla.edu/socr/index.php/SOCR Data BMI Regression](http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_BMI_Regression)>
which consists of a sample set of 252 patients, as this enables you to do a detailed study/analysis of the different characteristics, like body fat index, height and weight using 3D scatter plot

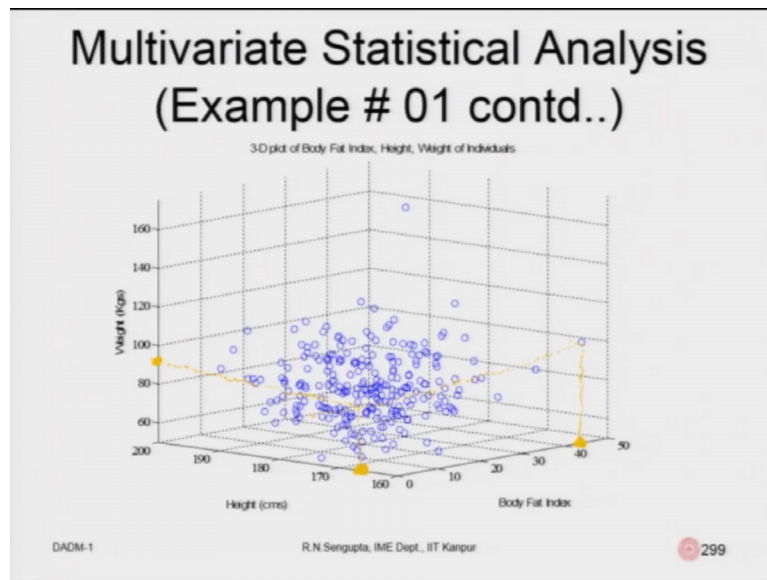
DADM-1 R.N.Sengupta, IIM Dept., IIT Kanpur 298

So, for this let us use the data, I am trying to basically use data in the real sense what is available in public domain and which is used by different people, in order to study these different concept, based on which the data has been uploaded or being discussed.

So, let us consider the data which is there in university California Los Angeles, which any consist of a sample set of 252 patients. So, all the different things are given for patients. Please have a look at the data you will understand. So, here is I am more of our interest making session, where you will basically, understand why I am sighting this examples and what is the relevants of them.

So, the data consist of a sample set of 252 patients, as this enables you he is the dietician to do a detailed study or analysis of the different characteristics like body fat index height and weight using 3 diamond dimension scatter plot. So, I will basically use these 3 different variables of parameters in order to understand how the multivariate in set of informations can basically be gleaned such that, you at one go with the diaganical understand how it basically is.

(Refer Slide Time: 15:15)



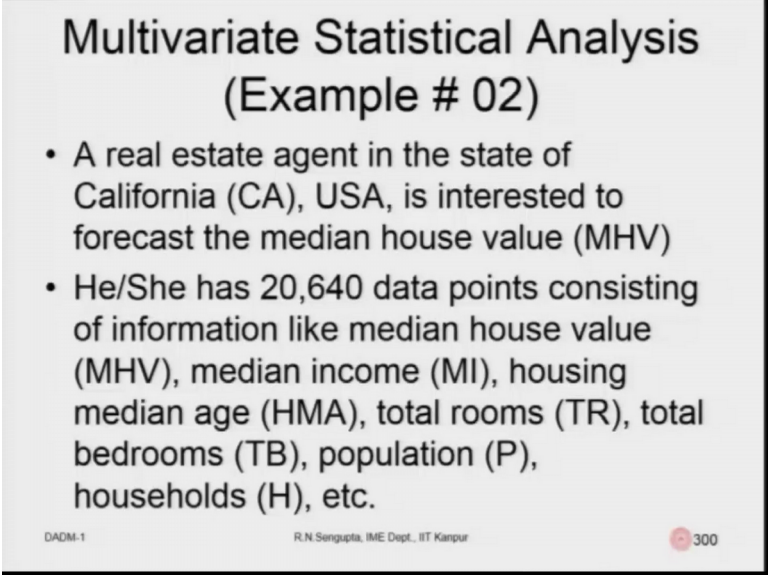
So, in the next slide you are basically, the diagram which gives the 3D diagram along the X axis you have body fat index which is here then, along the Y axis you have the height and along the Z axis which you have the weight. So, height and weight are basically, measured in centimeters and kilo grams So, all the scatter plots, the blue circle points which you see in the diagram, they give basically give you so, what is the corresponding different 252 values, which are there. So, if I consider see for example, arbitrary, let me consider this value this person So, in that case, the person body fat index if you drop a vertical, the body fat index would be about 42 and just arbitrary given, if you go in this way; that means so basically, I go and I want to touch the Y axis this is the plain here, which basically is the height.

So, basically, it would be somewhere it is in between so, 170 and 60. So, consider it will be about 100 and 65. So, basically, drop vertical here, so this is 165, where this 42 and basically, I want to find out that what is the weight of that integral. Again I go to in orthogonal direction and basically, I had the weight 1. So, weight 1 comes out to, this I am trying to do (at crude method trying to make an understand.

So, this is the values of the person is and the weight which, would about 96 97. This is the height which is about 167 and this is about 142 only 142. So, this person basically has the values of height weight and body fat index as we have. So, you can basically find out from the 3D where the particular persons stands and obviously if you have a

threshold, threshold can be either for the weight either for the height or the body mass index, based on that you can say yes, these people basically need to be careful about that height. Hence you will basically take actions accordingly. Consider the next data set it is basically, a again it is a different example.

(Refer Slide Time: 18:00)



Multivariate Statistical Analysis
(Example # 02)

- A real estate agent in the state of California (CA), USA, is interested to forecast the median house value (MHV)
- He/She has 20,640 data points consisting of information like median house value (MHV), median income (MI), housing median age (HMA), total rooms (TR), total bedrooms (TB), population (P), households (H), etc.

DADM-1 R.N.Sengupta, IIM Dept., IIT Kanpur 300

A real estate agent in the state of California in USA is interested to forecast the median household value, not the mean median.

So, based on that, the real estate if you knows the median household so, an obviously, he can incase if, he is buying the house is not duped and basically, knows that, what actually the price of the house should be and when he is also selling the house, he should also give the right information to the customer and not basically, go for a loss or not basically, see to that business proposition based on which is doing is business really is profitable.

Now in this case, why we take the median one because rather than mean median would basically, give me a better indication how the distribution is, so it basically, device the whole distribution into half and half. So, the real state he or she has about 20640 data points, which is a huge data set, which consist of information along for the data say for any point of the data set. You will basically have the median household value; the median income for the people who are there, what is the housing median age.

So, obviously different houses have different ages, so you have to find the median age. What is the total number of rooms in that house, what is the total number of bedrooms, which is house, what is the population, what is the house numbers so and so forth. So, they are basically denoted by MHV, which is the median household, median income MI, housing median age MH HMA, total rooms TR, total number of bedrooms TB, population P and household H. So, obviously they were other set of information also if you when I come to the data source about the longitude and latitude of the place, for each and half of the houses like those 2640 data points had their location also.

But that is not necessary for our calculations. So, I have not mentioned it, but if somebody is interested, he or she can definitely have a look at the source.

(Refer Slide Time: 20:13)

Multivariate Statistical Analysis
(Example # 02 contd..)

- Information about the data-set can be obtained in the paper titled Sparse Spatial Autoregressions, Pace R. Kelley and Ronald Barry, *Statistics and Probability Letters*, 1997, **33**, 291-297
- Let us fit the multiple linear regression (MLR) model (as used by the authors) to this data, then one obtains the ordinary least square (OLS) regression coefficient vector

DADM-1 R.N.Sengupta, IIM Dept., IIT Kanpur 301

This information about the data set, can be obtained in the paper titled Sparse Spatial Autoregression by Kelley and Barry in the *Statistics and Probability Letters*, which came out in 1997. So, first let us fit the, we will basically fit the multiple linear regression. The models everything is there, I do not want to go it to details. I will come to that later on when, we will be doing the multiple linear regression.

So, let us fit the multiple linear regression, MLR model as used by the authors to this data and we will basically, use the exact formula which has been used by the authors. Then will obtain the ordinary least square regression coefficients obviously, this regression coefficient of the beta if you remember will consider them that, they are the

sample one. Even if the data size 20640 is huge, but still we will consider the data size based on which, we are doing that study, is just a sample, so they would basically with the estimate.

(Refer Slide Time: 21:13)

Multivariate Statistical Analysis (Example # 02 contd..)

- $\beta_e = (\beta_{0,e} = 11.4939, \beta_{1,e} = 0.4790, \beta_{2,e} = -0.0166, \beta_{3,e} = -0.0002, \beta_{4,e} = 0.1570, \beta_{5,e} = -0.8582, \beta_{6,e} = 0.8043, \beta_{7,e} = -0.4077, \beta_{8,e} = 0.0477)$, here the suffix e is the estimated value
- Thus if one wants to forecast the 20621th reading which is 11.5129255 then the forecasted value is 12.3302108 which results in an error of 0.8172853

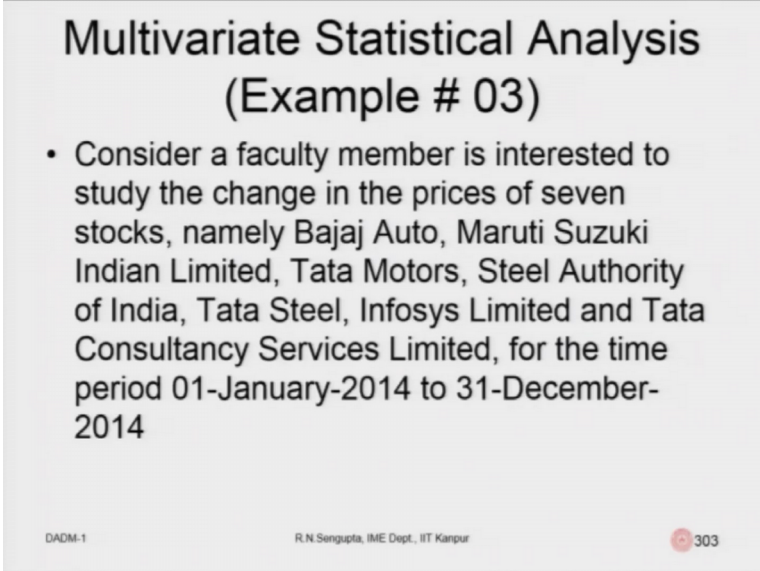
DADM-1 R.N.Sengupta, IIM Dept., IIT Kanpur 302

So, the estimate we denote by suffix e and the bold beta is basically the vector. So, in this model we consider there are 9 beta, starting from beta naught. So, it will be beta naught, beta 1, beta 2, beta 3 and each would basically, have their own significance. So, beta naught would be basically be the one for which is the coordinate at which, the straight line linear regression best with cuts the Y axis. So, obviously, the multidimensional case it would be plain not a line, which is basically hyper plain, which is easy to understand in the 3 dimensional case, may be a little bit difficult to understand in the higher dimension.

So, the values corresponding to beta naught to beta 8 are given as 11.49 then, 0.47, I am only reading 2 places are decimal correspondingly. Then 0.01, 0 point it is almost 0, so beta 3 would be 0 it is estimated value, beta 4 is 0.15, beta 5 is equal to 0.8 5, beta 6 is 0.80, beta 7 is minus 0.40, beta 8 is equal to 0.04. So, here the suffix is estimated value, so you have found out the betas. And what is needed to be done is see for example, thus if you want to know the forecast after a 2600 and 21st reading which is given in the data as, I have this I should mention, I am considering because this such a huge data, I can take random sorting and when sort them out and basically, find out any value which is the hiberately consider from the same huge sample of 20640.

And then basically that value is 11.512, this is 2620, so obviously, I randomly take the 20621 value and then using those betas I forecast them, the value comes out to be 12.33. I am again reading only two places of decimal and this results in a error of about 0.81.

(Refer Slide Time: 23:37)



Multivariate Statistical Analysis
(Example # 03)

- Consider a faculty member is interested to study the change in the prices of seven stocks, namely Bajaj Auto, Maruti Suzuki Indian Limited, Tata Motors, Steel Authority of India, Tata Steel, Infosys Limited and Tata Consultancy Services Limited, for the time period 01-January-2014 to 31-December-2014

DADM-1 R.N.Sengupta, IME Dept., IIT Kanpur 303

Let us consider the third example, a faculty member in any of the IM's so or good management institutes or an engineering institute whatever. He or she interested to study the change in the prices of 7 stocks, which are being sold and what in the stock exchange with the NSC or BSC. And the stocks are being generally, Bajaj Auto, Maruti Suzuki, Tata Motors, Steel Authority of India, Tata Steel, Infosys Limited and Tata Consultancy Services, which is TCS. And the data points are starting from 1st January 2014 to 31st January 2014. So we have, say for example, technically the total number of digit 364 or 365, but the trading number of days would be on an average 240.

So, we are basically using 240 number of data points basically, to find that out. So, now, we want to basically find out, in which directions the prices, not a prices in what clusters this different type of stocks would be. So, whether they are the same clusters, with their different clusters, whether does it mean that, Bajaj Auto and Steel Authority of India would be in the same cluster, so you have basically do it.

(Refer Slide Time: 24:53)

Multivariate Statistical Analysis
(Example # 03 contd..)

- The faculty member utilizes the prices of these seven stocks from National Stock Exchange (NSE), which is available at
<<http://in.finance.yahoo.com>> or
<<http://www.nse-india.com>>

DADM-1 R.N Sengupta, IME Dept., IIT Kanpur 304

The faculty member uses the prices of the 7 stocks for National Stock Exchange as I mentioned which is given the second URL is the National Stock Exchange, but you can get the information from yahoo finance also, so, this URL's are given.

(Refer Slide Time: 25:09)

Multivariate Statistical Analysis
(Example # 03 contd..)

- A closer look convinces the faculty member that the price movement for the first three scripts, namely (Bajaj Auto, Maruti Suzuki Indian Limited and Tata Motors), the next two (Steel Authority of India and Tata Steel) and the last two (Infosys Limited and Tata Consultancy Services Limited) move in tandem as separate groups as they are from the automobile, steel and information technology sectors respectively.
- The surmise, is valid as the companies that are in the same sector tend to vary together as economic conditions change and this fact is also substantiated by the factor analysis

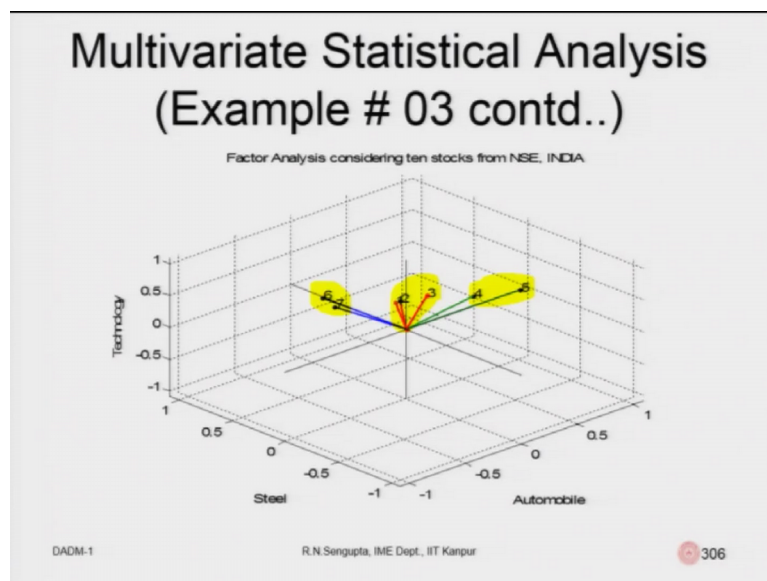
DADM-1 R.N Sengupta, IME Dept., IIT Kanpur 305

Now before he or she does the analysis, he looks he or she looks at the data, a closer look convinces the faculty member that, the price movement for the first 3 scripts which is basically, Bajaj Auto, Maruti and Tata Motors, the next 2 which Steel Authority and Tata Steel and the last 2 which is Infosys and Tata Consultancy Services move intend them.

That means, the price movement for Bajaj, Maruti and Tata Motors move in the same direction, the same manner not exactly same, but in the same manner, then Steel Authority and Tata Steel move in the same manner, but obviously, that is a different with respect to the first group. And the last 2 which is Infosys and TCS also moves in the same manner, which is different from the first and the second one. So, we want to be find out this groups and name them is automobile sector, steel sector, information sector and just check whether the proposition which you have making that, they move in with respect to the sector is right or wrong.

The guess is valid and you all will see that later on. That has the companies that are in the same sector tend to vary together, as economic conditions change and this factor is also substantiated by the factor analysis diagram, which you have given. I am just mentioning factor analysis I will come to that later on.

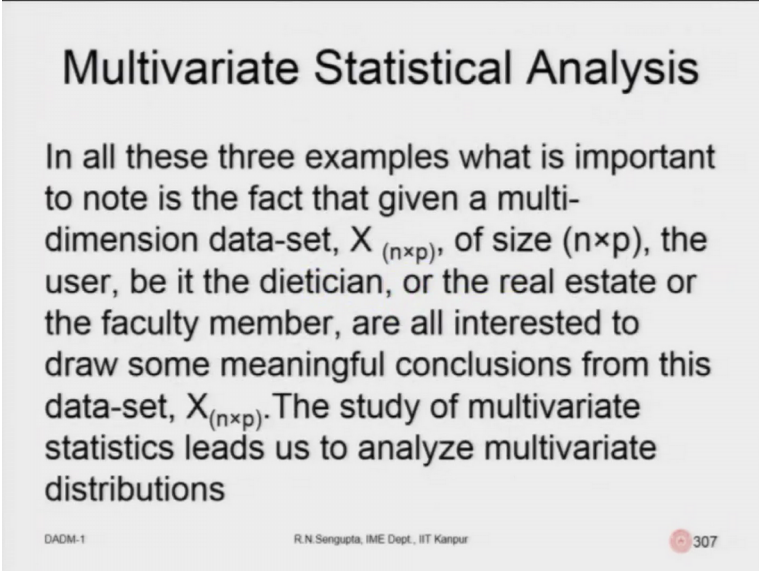
(Refer Slide Time: 26:33)



So, this is the factor analysis, which I do, so let me highlight. So, 1 2 3 which is basically, for the auto sector they are moving in time name in this direction, while 4 5 which is basically the steel sector, they are moving in this direction, the angles would be important later on, so basically they are not orthogonal. And the last 2 stocks which we have considered as TCS and the other company Infosys they move in this z direction, so they are not orthogonal but obviously, their co movements are similar.

So, you 6 7 compared amongst them, or you consider 1 2 3 compared against them, you should consider 4 5 compare against them, so they move in the same direction.

(Refer Slide Time: 27:27)



Multivariate Statistical Analysis

In all these three examples what is important to note is the fact that given a multi-dimension data-set, $X_{(n \times p)}$, of size $(n \times p)$, the user, be it the dietician, or the real estate or the faculty member, are all interested to draw some meaningful conclusions from this data-set, $X_{(n \times p)}$. The study of multivariate statistics leads us to analyze multivariate distributions

DADM-1 R.N.Sengupta, IME Dept., IIT Kanpur 307

Now, in all these 3 examples, what is important to note is the fact that given a multi dimensional data set of X and this suffix n cross p is basically the dimension, n number of rows, p number of columns. It could have been basically, k also n into k , but I am basically, stick if you remember I did mention k at one point time, but we will consider p . And the size of the matrix whether it is p cross n or n cross p that is only the nomenclature concept, but generally the dimensionality should be checked for each and every matrix, multiplications and addition.

This is of size n cross p , the user the whether she or he or she is in dietician, the real estate agent or the faculty member all of them are interested to draw some meaningful conclusion from the data using a set, which has p number of variables. What are the characteristics I am not going to come that to them immediately, an n number of data points. And the size of the study is basically is known as multivariate statistical analysis. And if, you remember I did mentioned some of the properties of the multiple linear regression that, the covariance between x has being independent. Then the covariance between the error and axis is independent, covariance of the errors themselves is 0 and there is a mean values for the x 's, mean value for the y 's, so all these and then there is the

errors would be distributed with mean value 0 and some standard deviation for the simple case we consider 1 and so and so forth.

So, we will consider all these things in much detail, this is just a warm up one and will go through the characteristics in multivariate statistics slowly and I am sure you will understand it but having said that their concept which we are considering in the univariate case would be repeatedly used as we proceed in the multivariate statistical analysis. With this I will close this 29th lecture and.

Thank you very much for your attention, have a nice day.