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

Lecture – 57
Canonical Correlation Continued

Welcome back my dear friends and students, very good morning, good afternoon and good evening to all of you. And welcome to the DADM 1, which is Data Analysis and Decision Making 1 course under the NPTEL series. And as you know this is a course for total duration of 12 weeks, and we are in the last week. And the total duration for the course is 30 hours, which is 60 lectures. And each week as you know, there are 5 lectures, each being a for half an hour. And I am Raghu Nandan Sengupta from IME department, IIT Kanpur.

So, we are we have four classes left, four lectures left for this DADM 1. And as we know that we were considering canonical correlation method to just to recap, even though I am sure the mythology how we went about is clear. So, we considered a set of X random variables. And we want to basically find out the relationship actually which is existing between axis using so called proxies or which give the best relationship from X using Y set of random variables.

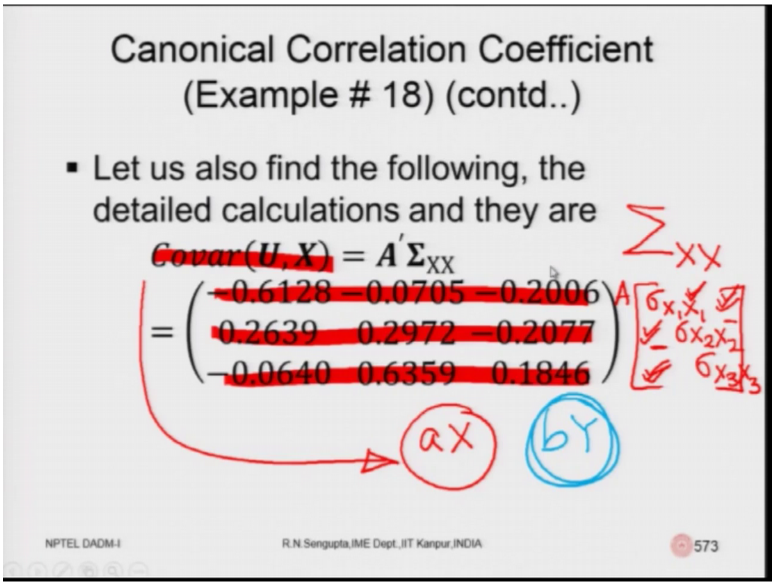
And we find out as you know that one set of vectors a and vector b, such that in reality what we try to do is that will try to find out a linear combination of um row vector a with X and b with Y, and find out the relationship of a a 1 X, a 2 X; a 1, a 2 I am mentioning they are vectors, such that the overall variance is reduced step by step. But, we try to get the maximum variability in each of the linear combination, such that we are able to give the maximum information or the relationship between the X and Y set of random variables.

Now, we also remember that in optimisation case, we will try to maximise the covariances a linear combinations the covariances of X and Y considering a and b being multiplied; a and b as I as I said they are some constants to be found out. And based on that, we try to give the relationships. Now, once we do that, we would basically find out the correlation coefficient, the test correlation coefficient, which is row star square.

And each of the correlation coefficient, which we found out, if you remember in the last class I discussed, that the correlation coefficient would be in the decreasing order that means, the first set would take out the maximum information, second set would take out the second set of maximum information so on and so forth. For the third, forth, whatever number of sets we want to, but remembering that they would be orthogonal to each other.

So, if you remember we found out in the last class, the covariances and between the X and Ys, which are basically equal to the U and V, because U and V are the linear combinations of X and Y. U being linear combination of a and X, and Y being the linear combination of b and V being the linear combinations of b and Y. So, we are going to find out the correlation coefficient of the covariances between U and X.

(Refer Slide Time: 03:43)



Canonical Correlation Coefficient (Example # 18) (contd..)

- Let us also find the following, the detailed calculations and they are $Covar(U,X) = A'\Sigma_{XX}$

$$= \begin{pmatrix} -0.6128 & -0.0705 & -0.2006 \\ 0.2639 & 0.2972 & -0.2077 \\ -0.0640 & 0.6359 & 0.1846 \end{pmatrix}$$

NPTEL DADM-I

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

573

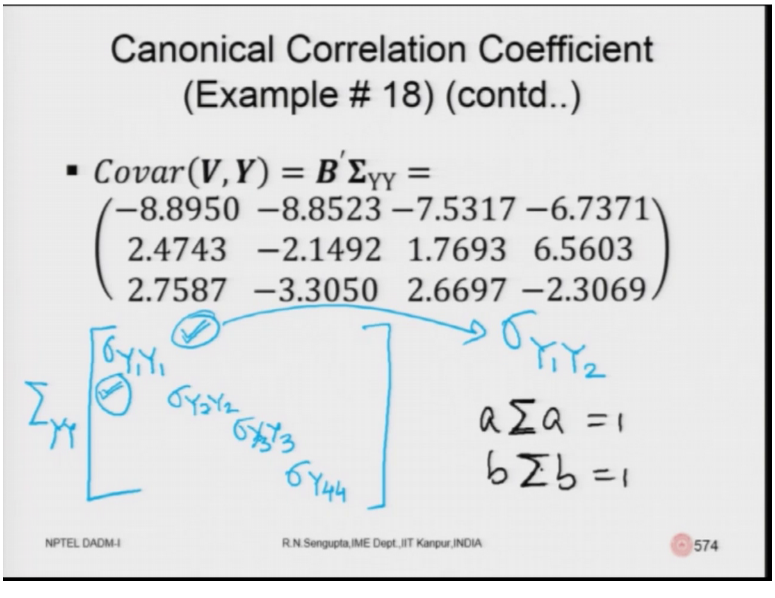
So, if you remember so we have to find out a X. So, these are vectors remember, vectors are matrices. So, I am not able to draw the board, but remember they are in board, when you write them. Similarly, you want to find out b and Y, let me use the other colour b and Y. So, the first case would be the covariances happening between a and X. So, this is what we need to do. So, the covariances of U and X would actually be the covariances of U. U is a combination a in to X, obviously to be a the covariance is being multiplied by a; covariance being for X and X.

So, once you found find it out, so this is the covariance which is existing between U and X U is the linear combination of X, so they come out to be as given. So, it can be found out. You just multiply them and find out to the covariances between them. So, we already have remember, so this would technically the so, there are three access, if you remember for the last example n cross 3, n being the sample size, which was 600.

So, the first one would be I will use this the subscripts X 1 1, X 1 2 and so on and so forth. The first one will be sigma X 1 1, and the last one would be X 3 3, and the (Refer Time: 05:41) second element will be sigma X 2 X 2. And the off the diagonal invent, which is this one or mirror image, this one is mirror image, and this one would be the mirror image. So, those values when you multiply with a is a vector, so the values come out to be minus zero point six two I am just six one.

I am just reading the first two decimal places. The second invent is along the row is minus 0.07 minus 0.20. Then this third rows are plus 0.26 plus 0.30 minus 0.21, and the third row being minus 0.06 plus 0.64 plus 0.19. Similarly, we need to find out of the covariance between V and Y, it will be the covariances, which we found find out considering b and Y, this is how we will do it.

(Refer Slide Time: 07:05)



Canonical Correlation Coefficient
(Example # 18) (contd..)

- $Covar(\boldsymbol{V},\boldsymbol{Y}) = \boldsymbol{B}'\boldsymbol{\Sigma}_{YY} =$

$$\begin{pmatrix} -8.8950 & -8.8523 & -7.5317 & -6.7371 \\ 2.4743 & -2.1492 & 1.7693 & 6.5603 \\ 2.7587 & -3.3050 & 2.6697 & -2.3069 \end{pmatrix}$$

$$\Sigma_{YY}\begin{bmatrix} \sigma_{Y_1Y_1} & ✓ & & \\ ✓ & \sigma_{Y_2Y_2} & & \\ & & \sigma_{Y_3Y_3} & \\ & & & \sigma_{Y44} \end{bmatrix} \qquad \rightarrow \sigma_{Y_1Y_2}$$

$$a\Sigma a = 1$$

$$b\Sigma b = 1$$

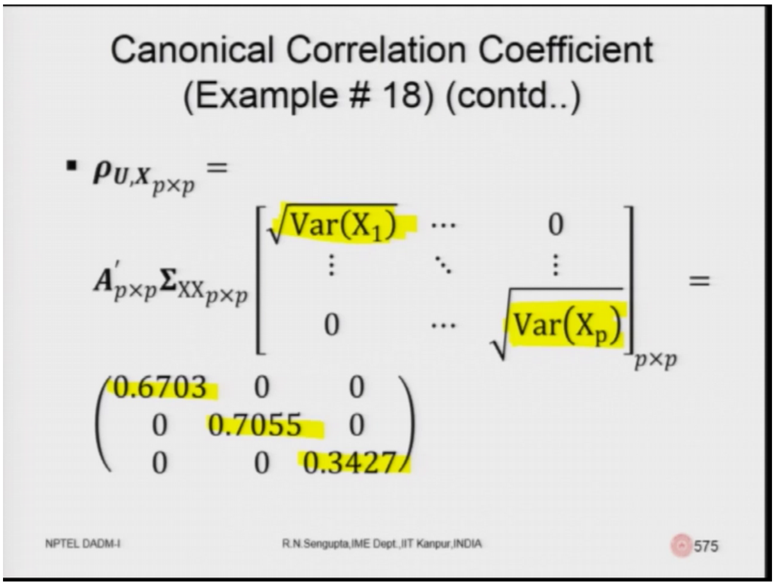
NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 574

So, again if you remember, we have, so this is Y 1 1, Y 1 1. Now, remember the size of sigma XX and sigma YY are different. In X, you are three variables X 1, X 2, X 3; and Y you had four variables, Y 1, Y 2, Y 3, Y 4, depending on the level of dependent structure we find out. So, this would be sigma Y 2 Y 2 sigma Y 3 Y 3 sigma Y 4 Y 4. So, I should basically expand. And the off the diagonal element would all be mirror images. So, these value which you write would be sigma 1 1 Y 1 Y 2. So, this would (Refer Time: 08:10) here. Similarly, you can find out all the variables.

The values come out to be as given here, the first row being. So, remember, just see watch the size, it is 4 by 4, because you at again I am repeating please do not mind, the Y had four random variable, X had three random variable. So, X was 3 by 3, Y was 4 by 4, the values being minus 8.8. I am going to read only the first decimal places minus 8.8 minus 8.8 minus 7.5 minus 6.7 and so on and so forth, for the second row, and the third row.

Now, remember one thing the values which you found out or actually the values based on which we are trying to minimise the maximise the variances of for the optimisation problem subject to the conditions we remember, we want to basically make a mu this values. So, I will use the such then conditions were as 1 as 1, this is related to the variance covariance matrix for X, and this was related to Y.

(Refer Slide Time: 09:31)



Canonical Correlation Coefficient (Example # 18) (contd..)

- $\rho_{U,X_{p\times p}} =$

$$A'_{p\times p}\boldsymbol{\Sigma}_{XX_{p\times p}}\begin{bmatrix}\sqrt{\mathrm{Var}(X_1)} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{\mathrm{Var}(X_p)}\end{bmatrix}_{p\times p} =$$

$$\begin{pmatrix}0.6703 & 0 & 0\\ 0 & 0.7055 & 0\\ 0 & 0 & 0.3427\end{pmatrix}$$

NPTEL DADM-I

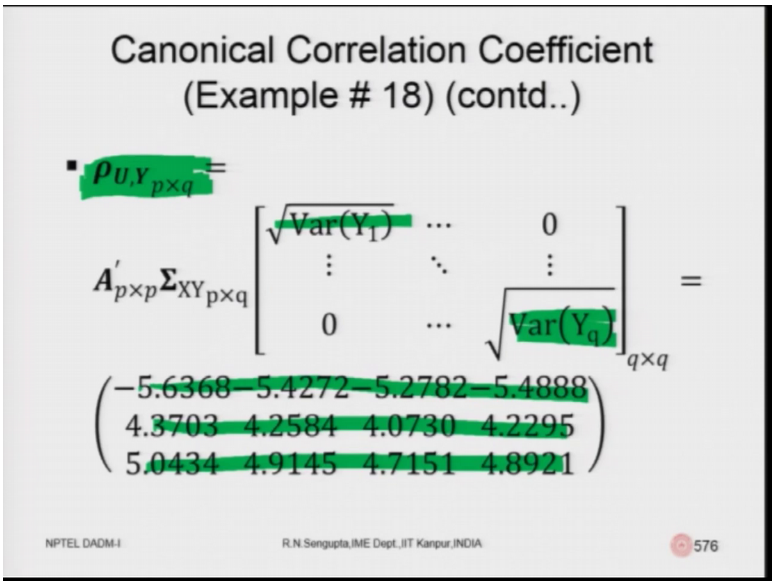
R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA

575

Now, we will go to find out the correlation coefficients. And the correlation coefficients, obviously they are remember, we are trying to find out the orthogonality, so obviously, the off the diagonal element would be 0, and the principle diagonal would be the all the variances, which we have. So, once we find out the actual vector of a multiplied by row XX, so the values are this is the standard deviation of X 1, similarly X 2, and similarly the last one, because there are three value values. So, the corresponding values would be

0.67, 0.70, 0.34; so obviously the others as noted down, normally 0. You can find it out very easily by multiplying a with the corresponding variance covariance of X. Only remember the size. Again I am repeating p or for X it is 3, for Y it is 4.

(Refer Slide Time: 10:50)



Canonical Correlation Coefficient (Example # 18) (contd..)

- $\rho_{U,Y_{p\times q}} =$

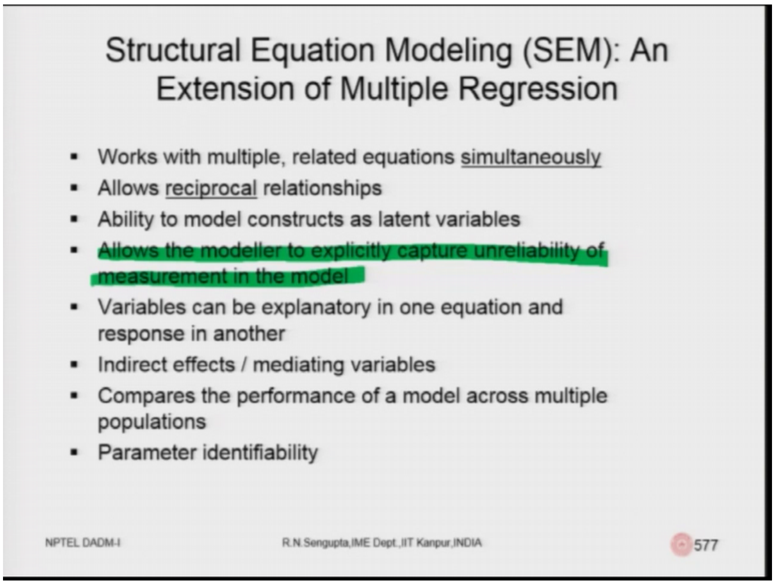
$$A'_{p\times p}\Sigma_{XY_{p\times q}}\begin{bmatrix}\sqrt{Var(Y_1)} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{Var(Y_q)}\end{bmatrix}_{q\times q} =$$

$$\begin{pmatrix}-5.6368 & -5.4272 & -5.2782 & -5.4888\\ 4.3703 & 4.2584 & 4.0730 & 4.2295\\ 5.0434 & 4.9145 & 4.7151 & 4.8921\end{pmatrix}$$

NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 576

Similarly, when we go to find out the correlation coefficient between V and Y, which is here correlation coefficient between this is this is this should be V with Y. The values would be again the principal diagonals are right way around the standard deviations of Y 1, standard deviation Y 2 till the standard deviation of Y 4. And the off the diagonal elements you can find out, so the values come out to be as given here.

(Refer Slide Time: 11:22)



Structural Equation Modeling (SEM): An Extension of Multiple Regression

- Works with multiple, related equations simultaneously
- Allows reciprocal relationships
- Ability to model constructs as latent variables
- Allows the modeller to explicitly capture unreliability of measurement in the model
- Variables can be explanatory in one equation and response in another
- Indirect effects / mediating variables
- Compares the performance of a model across multiple populations
- Parameter identifiability

NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 577

Now, we will go into little bit different topic, which is structural equation modelling. So, I will go through the theoretical basis a little bit slowly, we will consider structural equation modelling, where it works somewhat like the in the principles of multiple linear regression. And I will basically go slowly with the assumptions, and then go through the theoretical build up of the standard standardized structural equation modelling.

So, structural equation modelling is a methodology, which is basically a sort of extension of the multiple linear regression, and it is quite heavily used in social sciences, where the relationship which is to be found out between the variables are considered, where the quality factors can also be consider in a in a in a different way. So, what are the general bullets points or structural equation modelling are as follows.

It works with multiple related equation simultaneously. So, you this simultaneous so called equations, technically should be linear, but will consider them to be non-linear as an when required. This allows for reciprocal relationship. Now, reciprocal relationship with say for example if X and Y are related that means, X is related to Y, and the relationship is say for example 0.23, then the reciprocal relationship that means, how Y would be effecting X need not be like the covariance variance matrices, which is standardised and along the principle diagonal across the principle diagonal, it is a mirror image, it would not be the case depending on the problem structure.

So, it is able to model constructs and for the latent variables also. Latent variables can be need not be always in qualitative and quantitative in nature, they can be qualitative in nature depending on how the effect of the latent variable can be considered accordingly. So, it allows the modeller to explicitly capture unreliability of measurement in the models. So, unreliability of the measurement in the model, I will briefly go through it with the diagram, it is not a part of the structural equation modelling concept. But, I will just consider that form of very simple part of view, how you consider reliability or say for example, stochasticity on non-deterministic in nature in the data.

So, this fourth point I will come to that, using the concept of which is a simple diagram. These variables can be explanatory in one equation and similarly they can be response in another equation. So, if you consider in very simple case, when you considering a multiple linear regression; in multiple linear regression what we have, we have Y which is dependent linearly on X 1 X 2 X 3 plus a white noise.

And we did considered that the X is X values which are the X random variables, which are they are they are independent of each other that is why if you remember, I did mention that when you are trying to find of the rank of the X matrix, it should be exactly equal to the number of rows and number of columns depending on number of independent variables which they are that means, any of the rows are any of the columns cannot be expressed as a linear combination of the other rows and the other columns.

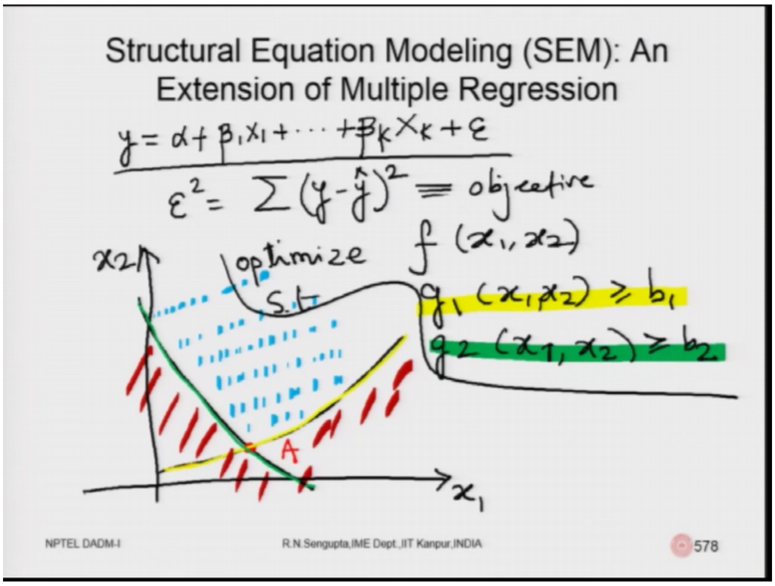
We also considered in multiple linear regression, the relationship between the random variables Xs and Ys and the errors epsilon should not be there, then we will also we did consider the relationship between errors from time to time is independent. And I will draw diagram there also for the multiple linear regression. I did draw it, but I will still highlight it as required. So, the variables in depending on their construct can be explanatory one set of equations can be the dependent or independent variables depending on the modelling frame work.

So, we will have indirect effects on mediating variables, which would be there, and which will basically have an effect how they can have a relationship, which where random variables X 1 affects X 2, and then again X 2 affects X 3, and again X 3 may affect X 1. So, this looping method can be considered. In structural equation modelling, it compares the performance of model across multiple populations. So, depending on the what samples you are taking from which sets, they will basically be compared in order to give us the best fit model.

And identifiability of the parameters should be there, such that we are able to find out set of parameters, which really affects the so called simultaneous linear or non-linear equations of the case may be. So, before I go into structural modelling, I will first go into the concept of reliability or stochasticity and then from a very theoretical point of view give you an idea and then go into the structural equation modelling.

So, I will try to make blank slides as I do and then I explain it accordingly. So, let me create some blank slide for our discussion. So, this is the general concept which I wanted to discuss about simple stochasticity and we will consider a set of distributions initially to be normal and then slowly relax it.

(Refer Slide Time: 17:08)



Structural Equation Modeling (SEM): An Extension of Multiple Regression

$y = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$

$\varepsilon^2 = \sum (y - \hat{y})^2 \equiv$ objective

optimize $f(x_1, x_2)$

s.t. $g_1(x_1, x_2) \geq b_1$

$g_2(x_1, x_2) \geq b_2$

NPTEL DADM-I

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

578

So, consider that we have I will I will just slowly and very simply state the let us consider the multiple linear regression first. So, you have Y is equal to alpha plus beta 1 X 1 plus beta k X k plus epsilon. So, we have considered that Xs and Ys are all normal; similarly epsilon is also normal with the certain mean and a certain standard division. And for epsilon we considered the mean value was 0 and variance was 1 or sigma square.

Now, consider that on the other hand consider a simple optimization model also very simply. So, I am trying to basically find out the so called the errors and trying to minimise the errors. So, you have the errors as given. This is the actual value; this is the predicted of the forecasted value as square them up and then I try to minimise. So, consider the problem is this is the objective function in general. So, I will basically consider a simple objective function with constraints.

So, consider it is single objective minimum maximum does not matter, only concentrate how the constraints are developed. So, I will basically consider optimise some f of x is a function is can also be epsilon square. So, f of x 1, x 2 for the time being considered that the random variable which we are going to consider in order to minimise or optimise consists of two random variables because it will be easier for me to draw that is why that is the only reason.

So, in the higher dimension it can be done accordingly subject to constraints again some g 1 X 1 X 2 is say for example, greater than equal to b 1 these are just constant consider b 1 is constant and similarly consider g 2 x 1, x 2 being greater than equal to b 2. So, this greater than sign less than sign can be tailored according to your problem and there are other constant.

Now, let us draw the diagram how it can be done. So, let me draw a two-dimension one. If you remember I mention that I am going to consider x 1, x 2 only considering the easy for me to explain point 1. Point number 2 let us also consider x 1 and x 2 are only positive; that means need not be, but let us consider them to be positive. So, g m this function g 1, x 1, x 2, and g 2, x 1, x 2 are non-linear. So, they are greater than b 1, b 2, so that can be handled.

So, let me draw them to be non-linear, I am just arbitrary drawing a diagram. So, it will make you understand. So, now consider this optimization model that we segregated, so my drawing is easier. So, let us go one by step by step. Consider the optimum solution for this optimise f x 1, x 2 given g 1 and given g 2 is in the deterministic case is at the boundary, obviously it would be in the boundary.

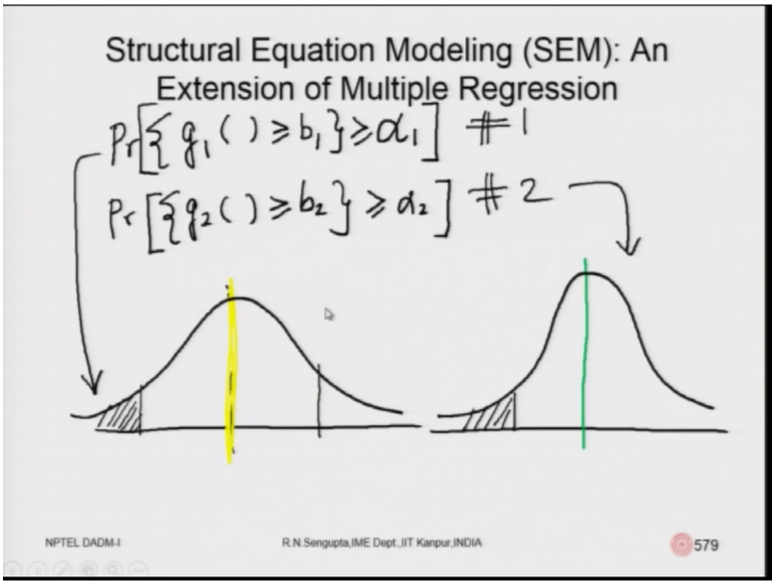
So, once you and the x 1, x 2 for simplicity, I am considering x 1 along x axis, x 2 along y axis. So, this deterministic point is say for example A. But, now let us go one step forward. Consider the function let me highlighted, so with different colours. So, this is probabilistic. And similarly, this is also probabilistic. So, how do we analyse that? So, this the function first one g 1, which is yellow in colour, we will consider this is the points.

So, now I am going to basically highlight the equation as yellow for the first function, equation 2, which is for the second function. Consider, there are now they are normal. So, if it is normal, the mean value of that constrain actually would be the lines, which I have drawn. So, the mean value for the mean value means, the expected value. So, as you repeat try to find out different values of x 1, x 2, it will be a random variable, but it will have an expected value.

So, the expected value of that line of the constraint, I am considering to with the line which is green and yellow in colour as shown in the slide here. Now, they are some probabilities. So, the mean values means if they are probabilities, they would basically be shifting either on to upper side or on the lower side depending on what is the overall randomness, which you are trying to bring.

Also consider the space, which we have for x 1, x 2, this is the feasible region. So, I am just highlighting it. So, this is the feasible region and these are the infeasible region, which is here. So, these are the in feasible region, where the solution cannot take place. Now, how it happens, so let me I will come I will be switching the slides, so please bare with me. So, generally the first equation just notice on the equation is g 1 x 1, x 2 is greater than b 1. And g 2 x 1, x 2 is greater than b 2.

(Refer Slide Time: 23:44)



Structural Equation Modeling (SEM): An Extension of Multiple Regression

$Pr[\{g_1(\ ) \geq b_1\} \geq \alpha_1]$ #1

$Pr[\{g_2(\ ) \geq b_2\} \geq \alpha_2]$ #2

NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 579

So, let me write using the probabilistic one. So, it will probabilistic one means, the first equation basically is probability of g 1, I am I will only write g 1. So, without mentioning x 1, x 2, which is we can understand is greater than equal to beta 1, but this is true for a value of say for example, alpha 1 as per the problems. So, alpha 1 is basically if you remember, this level of significance which you considered it is something to do with this. So, this is equation 1.

Similarly, the second one is probability of g 2 greater than equal to b 2 is greater than equal to alpha 2, again alpha 2 is basically a level of significance, which we have. So, now let us basically highlight these two into the diagrams. So, the first part is this is the mean value. So, mean value is if you remember this mean value, which I have drawn, so let me use at colour for the first equation. This is actually the yellow line, which you have drawn. So, similarly I will drawn the green one meter also.

So, the distribution is like this. So, they would be some alpha value alpha 1, alpha 2, values. So, g 1 is greater than b 1 by value of alpha 1. So, this would basically be given as alpha 1 depending on the level of significance. So, this is for equation 1, this is equation 2. So, I will draw it here also. So, the mean value, obviously would be different and the variance can also be different. So, how I am going to include that, let us go. So, I am drawing the mean value was green in colour. What was green in colour, but I am trying is this one, the mean value here. We will come to the red point, which is a later on. So, it cause (Refer Time: 26:23). So, this is alpha 2 remember the variances are different.

Now, if there is a dispersion for both equation and equation 2, how do we basically consider them. So, let me, so at this internal part the area, which is marked in blue is the feasible region. So, I will just remove the blue colour in now to draw it more clearly. So, this is how the lines would be. So, they are fluctuating. So, they would be mean value over and below. So, the below parties in infeasible; obviously, it would not be considered. And for this one the again this above and below, below part would not be considered because in feasible.

Now, what you actually have is I will use the red one, they would be a circle because white circle will consider and the circle would be would be true if both the variances are are true. So, I am going to draw the circle depending on the way the equation varies. So, this will be the loci of the midpoint. So, consider the white circle consider you are take you are looking at a football from the top. You are taking slices. If you are taking slices, it means the dimension in one direction and orthogonal directions are of the same radius which means the variability in both x 1 and x 2 directions are same that means, the variability which is cannot consider for both equation 1, equation 2 are same.

Now, considered that you have a rugby ball or the so called American football. So, if you plays the American football either on a horizontal space or in the vertical space in the sense you are placing in front of you then the variability along the x and x 1 and x 2 direction would be different. So, in that case the variability which is therefore, equation 1 and equation 2, which is of the constant 1 constant 2 would be different. So, we will draw that later on, and obviously, as the distribution change will consider little later. Now, consider for the time being the variability is are equal, and you are trying to draw the locus of this centre of the variability in such a way that has alpha 1, alpha 2 change you are able to trace it out.

So, with this I will just close this 37th lecture and continue the flow and trying to explain how we can consider the concept of stochasticity in the model then try to go into the structural equation modelling in details. So, this slide for the first time I will keep this slide and use this slide also for 30th lecture as required. So, I will end it here.

And have a nice day and thank you very much.