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

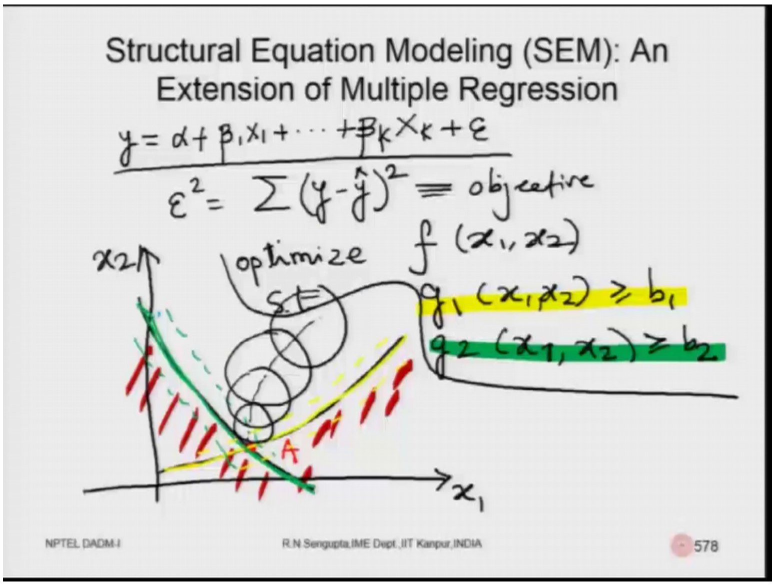
Lecture – 58
Structural Equation Modeling (SEM)

Welcome back my dear friends and students; very good morning, good afternoon, good evening to all of you. And this is as you know this is a DADM which is Data Analysis and Decision Making I, course under NPTEL MOOC. And this total course is for 12 weeks which is 60 lectures and 30 hours. Because each lecture is for half an hour, and each week you have 5 lectures, and for half an hour each and each after each we give the assignments.

So, we are in the last week can we have already completed 2 lectures for the last week and we are in the 58th lecture. And I am Raghu Nandhan Sengupta from IME department IIT Kanpur. So, we are we were discussing and as I mentioned in the last class or last lecture about the constant string probability probabilistic and how they can be considered.

So, I will first come to the diagram. So, the diagram would be discussed detail. So, I will also highlight few portions and draw of you few more diagrams as required. So, this is the diagram which we are considering.

(Refer Slide Time: 01:18)



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

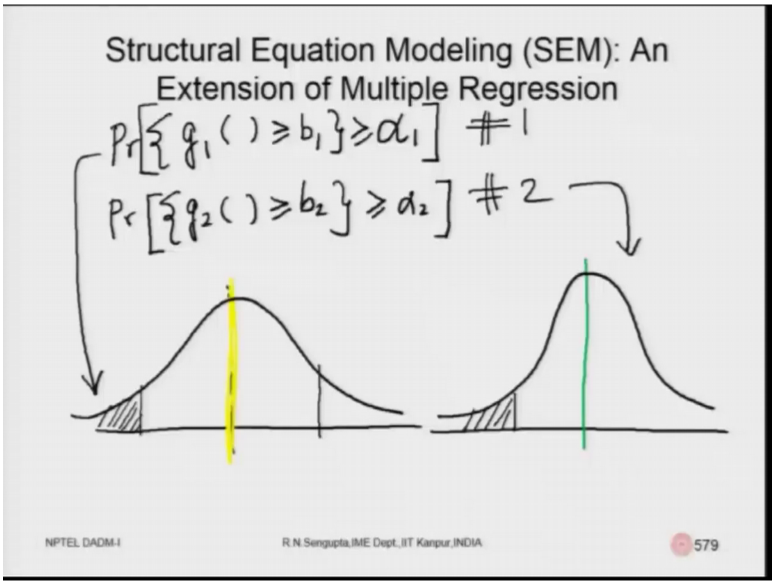
So, concentrate on the circles which you have circles means, I will not highlight, but I just show it you these the; so these are the circles which you have. Circles y, because if the if consider if they are orthogonal to each other. So, that means, x 1 x 2 orthogonal to each other and the variabilities are same. So, if the variabilities are the same and (Refer Time: 01:41) they meet.

Meet in the sense that if you are taking the consideration of the variabilities of both of them, they are orthogonal. So, obviously, it will be a sought of slices taking from the foot ball. Now as the variability increases for both of them, it will just basically be an expansion and contraction of the of the foot ball. So, if you take the slices again, they would be circles with increasing or decreasing diameter or radius that is point one. Point number 2 the loci or the set of points of locus of any of the circle would basically be the centre in around which you will covered that probability which is being defined by alpha 1 and alpha 2.

So, alpha 1 and alpha 2 in the case, because they are of the same the variability they would be equal. So, we will in the in standard normalized case. Why standardised normal case, I want to come to that within 2 minutes. So, they would basically be equal and the concentrics or the circles which you have not the word concentric circles which you have would be of the dimension which will basically consider the variability likewise considering alpha and alpha 2 are changing, but they are the same value. Now what you do is that, as the centre moves as the centre moves, so the first the second third fourth and it goes on. So, it will basically be the area over which it will basically be covering.

So, obviously, as the reliability changes the green line because it is it will get too complicated to draw complicated to flattered. So, this feasible point will shrink. In the sense, that it will go more towards the upper portion of the first quadrant considering that you have visible solutions are only in the first quadrant; such that you will be able to find solutions, but they would basically be more reliable. In the sense that the probability of achieving them would be higher or lower depending on the values of alpha 1 and alpha 2. So, if I come to equations and how they would be converted into standard normal considering alpha 1 alpha 2 are equal for the time being are like this.

(Refer Slide Time: 04:01)



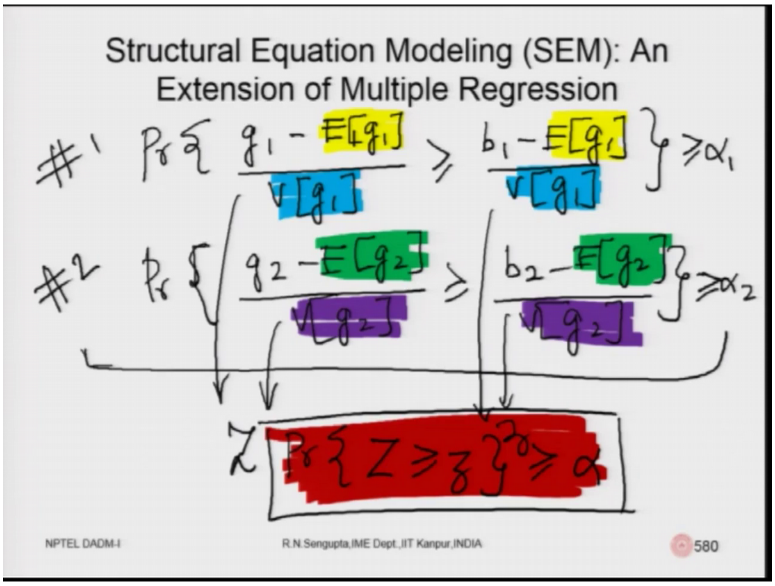
Structural Equation Modeling (SEM): An Extension of Multiple Regression

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

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

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

(Refer Slide Time: 04:02)



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

#1 $Pr\left\{\frac{g_1 - E[g_1]}{V[g_1]} \geq \frac{b_1 - E[g_1]}{V[g_1]}\right\} \geq \alpha_1$

#2 $Pr\left\{\frac{g_2 - E[g_2]}{V[g_2]} \geq \frac{b_2 - E[g_2]}{V[g_2]}\right\} \geq \alpha_2$

z $\boxed{Pr\{Z \geq z\} \geq \alpha}$

NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 580

So, the probability of this is the first equation, so is g 1 I will not write x 1 x 2 minus the expected value. So, this expected values is the line which I have drawn in the yellow and green colour of I will replace it as E 1 or E of g 1 divided by the variance of g 1 that is if you basically bring the b on to the left hand side. So, it will be basically greater than equal to b 1 minus E of g 1 by v of g 1. This is greater than or equal to alpha 1. Alpha 1 greater than equal to less than equal to it will be moderate accordingly.

Similarly, so the expected value which you have is the central line which is yellow in colour. And the variability which you have, I will draw the variability using see for example, blue colour here, if we go back to this diagram. They would be the variability which you would have here, not on the infeasible region.

So, higher or lower variability would basically mean this dispersion the spread is more. Now come to the second equation. Probability of g 2 minus E of g 2 variance of g 2 is greater than equal to b 2 is brought to the left hand side, b 2 minus E of g 2 variance of g 2, this is greater than equal to alpha 2. Again it can be less than alpha 2 greater than alpha 2. So, here what are the expected values, let me draw it using the colours, so it makes sense. So, this would be the green line, this would be the green line. So, what are the green lines I am talking about?

So, the green lines are this. So, this is the mean value. And if I am talking about the variability, I have to use a colour separate. So, let me use the colour violet. So, this is the variability, this is the variability. And where does the variability come? So, again in the feasible region there would be variability here. So, more moving towards the feasible region and shrinking the feasible region more on to the upper part. Now why the ambient like this?

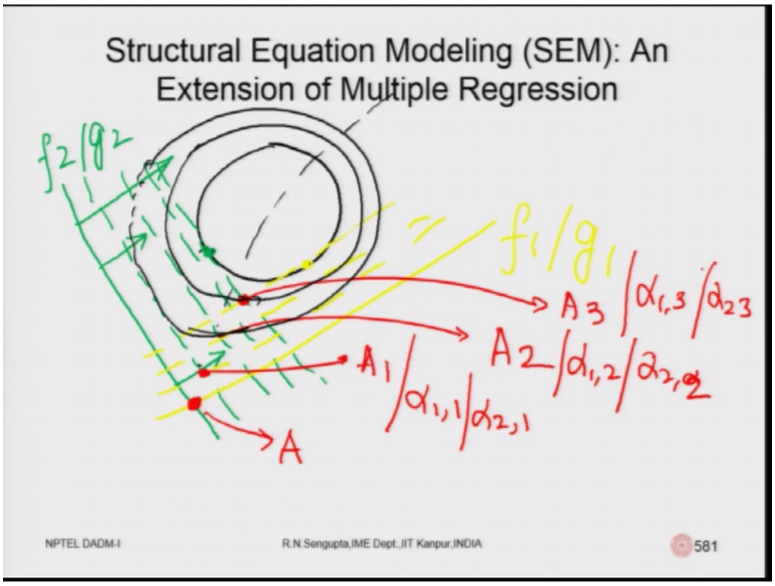
So, consider this, so I will put it equation 1 equation 2. So, this is one equation 1; and equation 2 basically means the constraint one and constraint 2. So, I am writing for both of them. So, they are basically the first part, whether you are considering equation 1 equation 2 they are technically Z, standard normal. And whether you are considering equation 1 equation 2 this is small z.

So, what you have for both the equations? Now I will remove these subscripts 1 and 2. So, it will you will understand the general equation; would hold 2 for both of them, just add subscript 1 and 2 depending on what the problem or which constraint it is. So, it will be probability of capital Z greater than equal to small z is greater than equal to alpha. So, this is your general equation, I will highlight it using the red colour.

So, they would basically be a standard standardised formulation and you can basically have it accordingly. Now the beauty of this method is that, you can work on the optimization problem or from the statistical point of view in 2 ways. Number one is either keep the circle fixed and basically expand or contract the feasible regions such that is tangent at the point where alpha values exactly equal to the area being converted into the circle, that is point one. Point number 2, keep the line fixed and contract and expand the circle.

So, the contracting and expanding the circle would basically give you the overall coverages which you are going to have in trying to basically find out the feasible region, and the part of the feasible region which will give me the best optimal so called optimal solutions in the in the non deterministic case depending on the values of alpha. So, alpha values or expected values are not with subscript now for understanding. So, it is like this.

(Refer Slide Time: 09:35)



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

f2/g2
f1/g1
A3/$\alpha_{1,3}$/α_{23}
A2/$\alpha_{1,2}$/$\alpha_{2,2}$
A1/$\alpha_{1,1}$/$\alpha_{2,1}$
A

NPTEL DADM-I

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

581

So, in the first case yes so, you have the so called feasible region and there is a circle. Circle is basically the circle depending on the values alpha. So, what is happening is that, you are slowly moving depending on the value of alpha, you are slowly moving the shifting the deterministic one going towards such that it will be tangent at that point.

So, that tangent value where it will be tangent for both the values will give you the optimum solution in the non deterministic, how? So, this is the movement of function 1 or which is rewrite to be g 1. Similarly, you will have function 2 which is g 2 and let me I should I have to use let me use that different colours. So, it will be easier for all of us to understand. This was in green colour that was function 1,sorry. So, this would be the actual point, now it is moving.

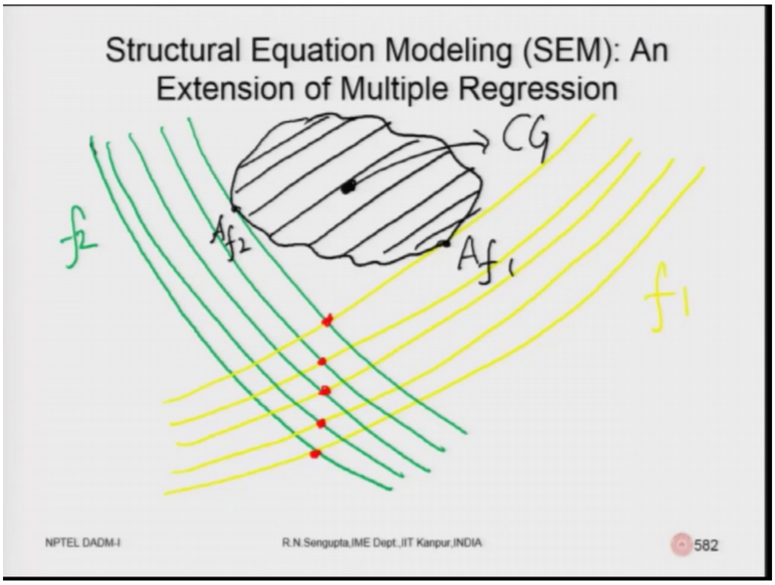
So, this is f 2 or g 2 and this is moving depending on the values of alpha. Alpha I am using same value which is technically alpha 2. Now when I come to the equation 1, this colours would change here there yellow in colour remember. So, this was the point so they would now be moving. So, consider this is again tangent.

So, this was tangent, this was tangent. So, initially you had the deterministic solution. So, these what basically f 1 g 1. So, it is yellow coloured is a little bit light, plus please be here with me. So, technically this was the point a for the determistic one. So, as you move out this becomes A 1 as a shrinking the combinations these becomes A 2, and the final solution based on this basically be A 3. So, obviously, in this case alpha is not applicable. In this case it will you will basically have alpha 1 alpha 2 with a certain value. So, let me denoted better with suffixes. So, this will be alpha 1 coma the first value for the first level alpha 2, for the first level then you have alpha 1 2 alpha 2, 2. And finally, you will have alpha 1 3 and alpha 2 3.

So, these 3 values 1 2 3 would dependent on the on the region. So, technically you would have circles like this. So, in this case it will be a circle going like this. So, I am trying my level best to draw the circles. So, they would basically the locus would be moving in this direction. Now what happens if it is a non normal distribution which is more important for us to understand, and we will try to draw it I am solution methodology would be little bit difficult to concentrate and find it out in this course, but we will try to do something in the DADM course DADM 2 course which is to do with reliability optimisation all this things, but these are very interesting topic.

And if you do any type of problems of robust optimisation reliability programming, stochastic programming they would give you some ideas or it can be done accordingly. So, let me go to us blank slide related to non-normal and any arbitrary distribution. So, let me draw it first.

(Refer Slide Time: 14:29)



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

NPTEL DADM-I

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

582

So, these are produce the colours scheme same. So, there is no confusion, the first value the second value and I will draw the mean value they are moving. So now, they are all bold in order to make you understand. So, this is the mean values moving depending on the reliability level. So, they are parallel. And this is the mean values are moving.

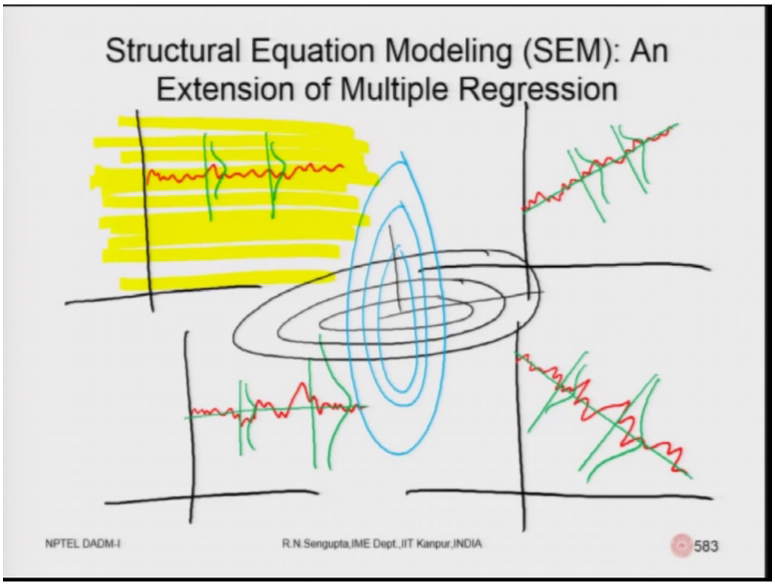
Now the points are these. So, 1 1, so it would be 2 2 3 3 2 1 2 3 4 and the 1 2 3 4; 1 2 3 4 this is the fourth one, and they would be another line. So, this would be the fifth one, this is how it is moving. Now concentrate I will only draw one diagram in order to make you understand. Now consider it is non normal. And obviously, it is any arbitrary distribution both the f 1 so let we write down also. So, this is f 2 or g 2 this is f 1 which is g 1.

So, now the overall area would be a distribution where the centre of gravity would basically with the point which will give me the loci of how the area would be covered such that the value of alpha and 1 for 2 whatever is to be equal depending on the of level of significance would be the overall area which is being covered inside that so called non normal distributions. Now, trying to find out the tangent points so, tangent points being this or this, based on which you are able to find out the centre of gravity. So, tangent points let me denote as see for example, A for f 2, A for f 1 trying to find out would be difficult.

So, we will basically use different type of simulation techniques to find it accordingly. Now if you remember that I did mention that the level of the in the significance which we consider alpha 1 and alpha 2 would be concentrated on the overall area of covered inside that non normal distribution.

We I also mention that if see for example, the errors are dependent on each other, and that would basically have an effect and what is the shift of these yellow and green lines, because that would be dictated by whether the errors are independent point 1. Among them self-point 2 is that whether the x value which is in this case x 1 and x 2 are dependent among themselves or whether they are affected by the error terms also. So, let me draw or take another blank slide.

(Refer Slide Time: 18:01)



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

NPTEL DADM-I

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

583

So, for the error terms, so what will I will draw I have drawn it, but I will just draw it again. So, let me draw the diagram 4 panels. In the first case the mean value and the standard deviations are fixed, in the second case mean values are not fixed. Some deviations are changing. In the third case mean values are fixed and deviations are changing, and the 4th case both of them are changing. So, I will basically use the red colour to draw it. In this case mean values is increasing consider.

So, I will use the same colour scheme everywhere. So, variance is same, mean value is same, but the variance is increasing first. And in this case the mean value is also decreasing, and the variance is also increasing. Now how would they look, and that obviously that would have an; we have done this I would I will explain that how it will have a consequence of the calculation.

So, in this case the mean values and the standard deviations are same, so there is no change. In this case the mean value is increasing the standard deviations are same, no change for the standard deviation. In this case mean values are same and the standard deviations are changing. In this case the mean values are decreasing increasing decreasing does not matter. So, whatever and the standard deviations are also change. So, if you have this type of diagrams, so obviously they and the issue is that the reliability of the reliabilities. So, for optimisation concept of the diagrams which I have shown to you is only applicable for this case.

So, obviously, if they are changing it will have an effect accordingly. For mean values changing obviously, it will mean the shifting of the boundary. For the variability changing it will be a little bit more difficult to understand; because if you remember variability both being same, a normal distribution you basically take slices which will be basically the case for or like taking the slices of football, and then expanding and contracting the loci would basically be the centre of a circle. Now in case if it is variance are different, mean values are same or not same, let us not consider it.

Now then, obviously, they would be as I said the slices of rugby ball or American football ball, and you can take the slices depending on whether the major axis minor axis are; it is like you will basically get ellipses, major axis minor axis being along the x axis and y axis depending on where the variability is high. So, if the variability for x 1 is high along the x axis. So, you will have the concentric values would be like this. So, it will be more along the x axis, let us along the y axis. Now in case if the variabilities are just reversed, we have more variability along the y axis less along the x axis, it will look like this.

So, obviously, you are loci of the central gravity of the centre of the ellipse would moving such a way, at the overall area covered in that ellipse or in the circle would basically be the one corresponding to alpha 1 alpha 2 which you have find out. So, alpha 1 alpha 2 in this case would be different, because alpha 1 alpha 2 are same when you had the concentric circles as guys you are taking slices of the football. Now in order to solve this there are different type of methods from the optimization case.

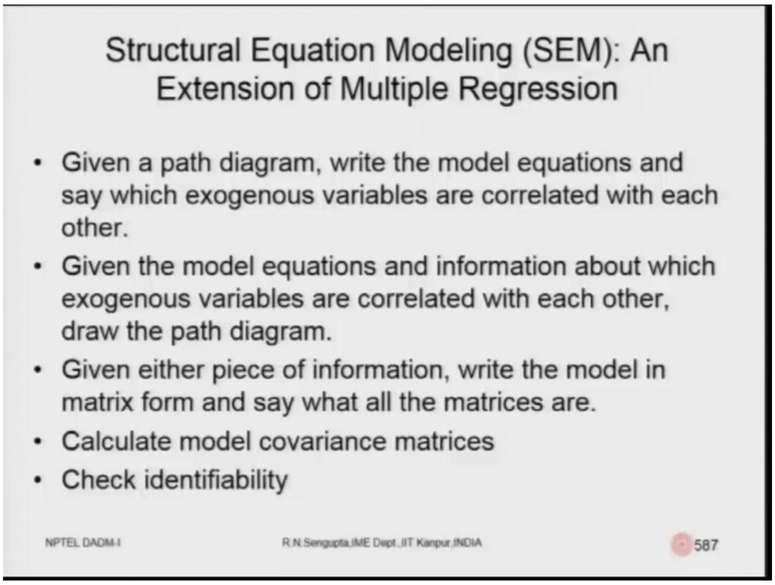
So, generally it is done in a way that you basically simulate. Simulate means the number of times it is applicable, and based on that you basically find out in the case that how many such values either are above or below this point of b b 1 or b 2 depending on which concentrate constraint you are considering, and based on that you find out the level of significance in alpha 1 and alpha 2 and solve the problems accordingly. We will we have different type of methodologies to solve.

So, generally the methodology which we consider is 2 folds. First we see what type of structure of the problem it is. It if it is a simple optimization problem or either the integer programming, mixed integer program whatever it is, these are also very heavily used in statistics that is why I am mentioning it. So, you basically solve the problem using the optimization method which you have, and then use the reliability problem to basically shuts the space to find out see for example, capital A 1 capital A 2 capital A 3. So, this capital A 1 capital A 2 and capital A 3 are the deterministic solution considering the reliabilities changing.

So, first solve find out capital A 1 or capital A 2 capital A 3 depending on simple optimization problem. Then what you do is that, basically you try to use the reliability model to find out the best possible point for utilising this A 1 and A 2 such that the distance moved from A 1 or A 2 or 3 as the case may be would basically be the smallest distance in order to basically reach the tangential point of the circle, circle of the ellipse whatever it is, so you keep repeating it. First find out A 1, then try to find out the optimum point using the tangential concept for a circle or ellipse.

Then finding out the tangential point you basically go in to the next step again try to optimize and find out the best possible solution of A 2, and then again find out the tangential point. Keep repeating it till you are basically certain that there is no such further improvement in solution within a certain bound. So, in the simulation case, and then you basically put that value, at the best possible level of solution, depending on the level of alpha 1 and alpha 2; which has been basically pre decided by the person who is doing the optimization problem and when he is trying to basically solve it. These are blank slide I will just skip and I will come to the saturation modelling concept.

(Refer Slide Time: 25:10)



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

- Given a path diagram, write the model equations and say which exogenous variables are correlated with each other.
- Given the model equations and information about which exogenous variables are correlated with each other, draw the path diagram.
- Given either piece of information, write the model in matrix form and say what all the matrices are.
- Calculate model covariance matrices
- Check identifiability

NPTEL DADM-I R.N.Sengupta,IME Dept.,IIT Kanpur,INDIA 587

So, given a path diagram; path diagram basically means the relationship which you have between the explanative variables and the independent and dependent variables. And remember that explanative variables can be the dependent variables in the next set up of equations in the (Refer Time: 25:25) modelling case. So, given a path diagram write the model equations, and say which are the exogenous variables and how they are correlated with each other.

And if the correlation coefficient can be found out using the simple correlation coefficient matrix; so, given the model equations and information about which exogenous variables are correlated with each other, you basically again draw the path diagram. So, first you basically define the relationship, draw the path diagram and proceed accordingly. So, given either piece of information write the models in matrix form, and see what are the matrixes based on which you are trying to basically optimize.

So, these matrices would give you the type of equations which you are which you are trying to find out, and also try to find out what is the maximum amount of variability to a maximum to relationship we can find out using this path diagrams, and finding out the equations which are there. You calculate the models covariance, because covariance would definitely give you the level of significance which is applicable.

And you will check the identifiability of the problem, try to find out whether it is feasible and then basically commit accordingly. You will go step by step, as the variability changes you will basically report the answers and find out the relationship at each and every step, what are the exogenous variables, what are the independent variables, what are the dependent variables and basically say that how the overall path diagram of the interrelationship is there. For each different set of equations, the as I mentioned and I did mentioned it twice, the level of significance will change point 1. Point number 2 it will also matter that how does the explanative variables and the levels of the independent and dependent variable structure changes.

So because that will have an effect on the path diagrams the relationship and in which the in which way the arrows point. And it will also give you the level to which the dependent structure is basically being model using the covariance variance matrix. So, with this I will end the lecture; with 2 lectures still left to end this course I will try to basically wrap up the structural equation modelling. And if possible in the last lecture give you an overview what things we have covered.

And there are lot of things we may have not of covered, but I will basically still try to highlight that considering the overall coverage, where you can basically concentrate on and basically look into the and other topics and if you find any interesting one, you are most welcome to write us on the forum and we will reply.

And obviously, it will matter that you go these are notes are the slides which I am teaching are not the overall concept which would basically make one person master in statistical method which are required. But, obviously a little bit practice, little bit reading would make things much better on your parts such that if you solve the problems, what we have done and lot of the cases, we have solved or the assignments we solved it will definitely clear many of the doubts which you are facing. And obviously, practicing would make things much more better and much more perfect for all of you. With this I will close this class and have a nice day.

And thank you very much.