

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

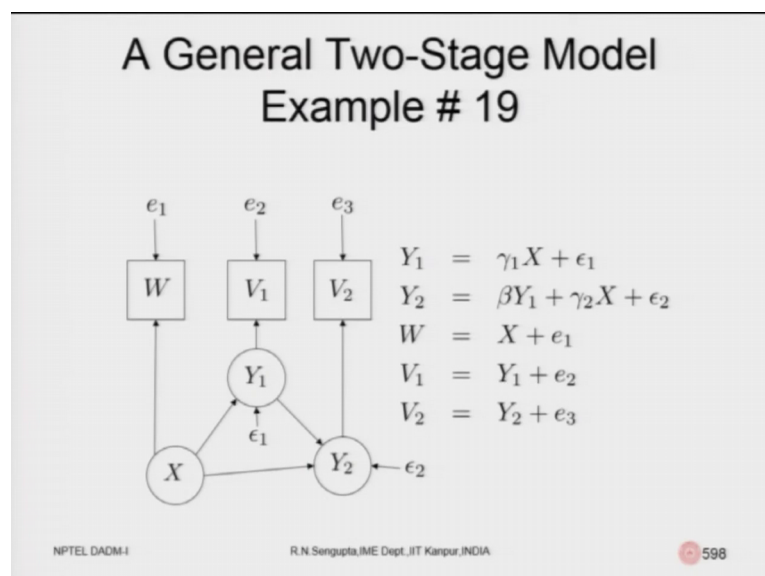
**Lecture - 60**  
**SEM**

Welcome back my dear friends and dear students, a very good morning, good afternoon, good evening to all of you. And this is the DADM one course on the NPTEL series. And DADM means Data Analysis and Decision Making. And as you know this course is for 12 weeks of total duration of 30 hours, 60 number of lectures, each lecture being for half an hour. And we are in the last lecture.

So, as you know that each week we have 5 lecture each for half an hour, and I am Raghu Nanndan Sengupta from the IME department, IIT Kanpur. So, we are discussing structural equation modeling; where unlike multiple linear regression, there is a set of latent variables, and the relationship between the variables are such they are cyclic. Or they would be some structural relationship through a diagram, which can basically give or explain what is the level of dependency or what is the level of relationship between the variables.

So, if you consider actually our regression model, you will basically have Y is dependent on X 1 to X k depending of k being the number of independent variables.

(Refer Slide Time: 01:22)



And error has a particular distribution normal with 0 mean and some standard deviation. X's are all independent, X's are also independent of the errors, errors are independent of each other and based on that we find out what is the variance covariance matrix; the mean value of the Y which is the dependent variable and also the variance of Y which is the dependent variable. Now when you come to the structural equation modeling; obviously, there would be a set of errors which would basically have an effect on the latent variable, set of errors which will have an effect on the measurement variables, and it will also have a white noise.

So, the errors will technically be divided into 2 sets. One is the white noise which cannot be control to and one is the set of errors which can be controlled, but it is basically the measurement error.

(Refer Slide Time: 02:15)

**Example: Wheaton, Muthén, Alwin, Summers (1977) Study of Stability of Alienation**

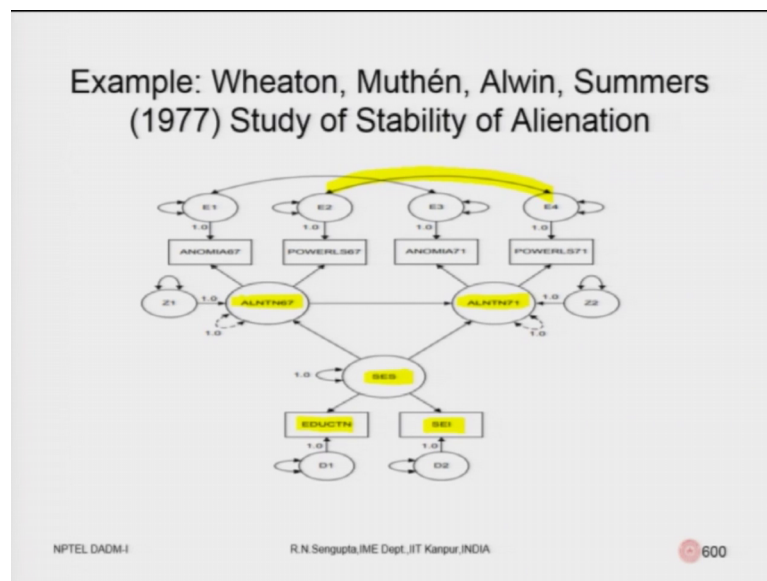
- There are three LVs:
  - Socioeconomic Status (SES)
  - Alienation 1967 (Alntn67)
  - Alienation 1971 (Alntn71)
    - Each LV has two indicators.
    - For SES: Education and SEI
    - For Alntn: Anomia and Powerlessness (attitude scales) measured at each occasion.
- On the following page is a fully specified path diagram representing a model of the stability of the attitude of alienation over two points in time.

NPTEL DADM-I R.N.Sengupta, IIT Kanpur, INDIA 599

So, let us consider the example of Wheaton, Muthen, Alwin and Summers. This is study of stability of alienation. So, there are 3 latent variables, and we consider socioeconomic status, the alienations of 1967 and 1960 71. So, these are the 3 variables we are take we take. And each latent variable have 2 indicators. So, say for example, for education you will have SEs which is the socioeconomic status. You will basically have the education system and the social economic positions which they are for alternative alienations concept they would be anomia and powerlessness, which is attitudinal scales.

And these scales would basically depend on what is your attitude towards different responses which you get. And what we will do is that? We will basically fully the specified a path diagram and try to analyze the problems accordingly. So, representing a model of the stability of the attitude of alienation over 2 points in time, one in the 1967 on one in the 1971 we will try to basic and analyze; the struck using structural equation modeling, how this alienation how the relationship could be analyzed.

(Refer Slide Time: 03:28)



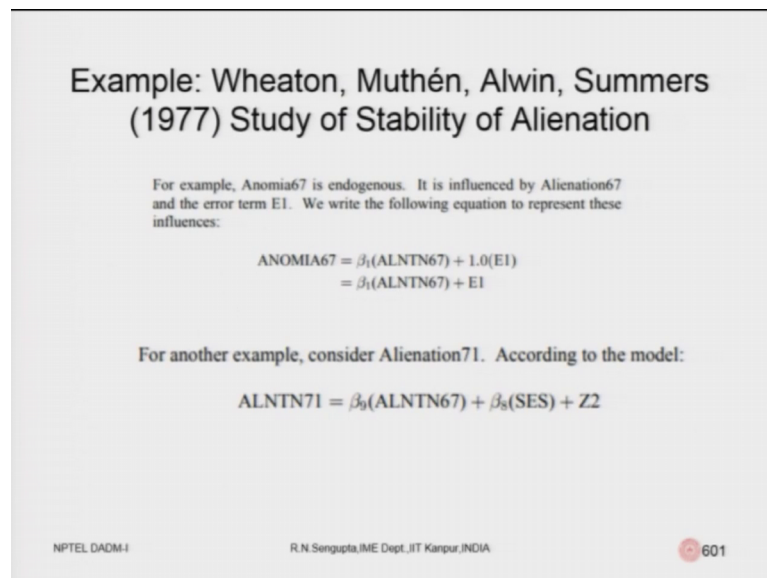
So, we will have basically the models specified using the structural equation modeling like this. So, this is the SEs which you are considered the education a levels. We have the alienation on 1967 and 1971. And we will consider 1967 is going to affect 1971, that is why the arrow goes from 1967 to 1971, point 1. Point number 2 we will consider the level of relationship which is existing between the alienations of 67 and 71 are the type of variables which is anomia of 67 and the power structures of 67, along with the concept of power structure of 71 and the anomia of 71.

But once you understand that the relationship are also inter in intra, in the sense e 2 which basically gives you the power structure of a 67 affects and is indirectly also affected by the powers structure of 71. So obviously, they would be a relationship or a level of dependence depending on the correlation coefficients between e 2 and e 4. Similarly, you will have educations SEs being a dictated by education and SE 1 I. So,

they would basically be denoted by D 1 and D 2 with levels depending on what is the level of dependence structure.

So, here you have socioeconomic status SEs, and that SEs would have basically education and socioeconomic indicators. And alienations as I said would be from 67 and 71 also.

(Refer Slide Time: 05:17)



**Example: Wheaton, Muthén, Alwin, Summers (1977) Study of Stability of Alienation**

For example, Anomia67 is endogenous. It is influenced by Alienation67 and the error term E1. We write the following equation to represent these influences:

$$\begin{aligned} \text{ANOMIA67} &= \beta_1(\text{ALNTN67}) + 1.0(\text{E1}) \\ &= \beta_1(\text{ALNTN67}) + \text{E1} \end{aligned}$$

For another example, consider Alienation71. According to the model:

$$\text{ALNTN71} = \beta_9(\text{ALNTN67}) + \beta_8(\text{SES}) + \text{Z2}$$

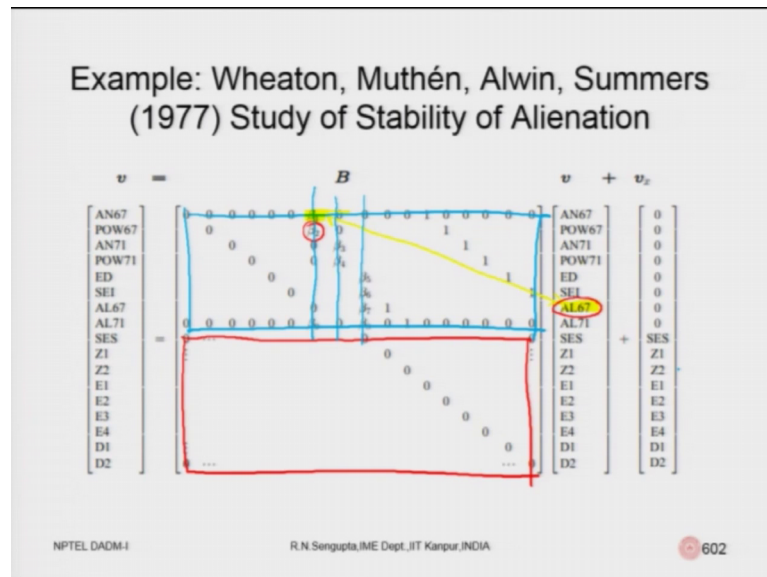
NPTEL DADM-I R.N.Sengupta, I.M.E Dept., IIT Kanpur, INDIA 601

Now, let us consider the example further. So, you will consider anomia of the 67 basically being a linear structural equation of the following form. What are the following forms would be? Alienations of 67 on 1967 would be would be considered. And that would have the error term is e 1 and the error term factor is one, while the factor related to alienation of 67 would basically be beta 1 which we need to find out and that will give us the level of a dependent structure which is there.

So, let me continue reading it. For example, anomia 67 it is an endogenous variable. It is influenced technically by alienation 67 and the error term which would basically have a particular distribution with a certain mean and a certain standard deviation. So, we write the concept of anomia of 67 as dependent on alienation of 67 and the error term. And for the other example we will consider. That alienation of 71 which is also being effected by 67 would be alienation of 71 is dependent on alienation of 67 and the education SEs levels while the factors to which they will be multiplied is given by beta 9 and beta 8.

So, we are taking the suffixes depending on the number of variables which are there and obviously there would be error term which will considering as  $z^2$  it can be standardized also.

(Refer Slide Time: 06:48)



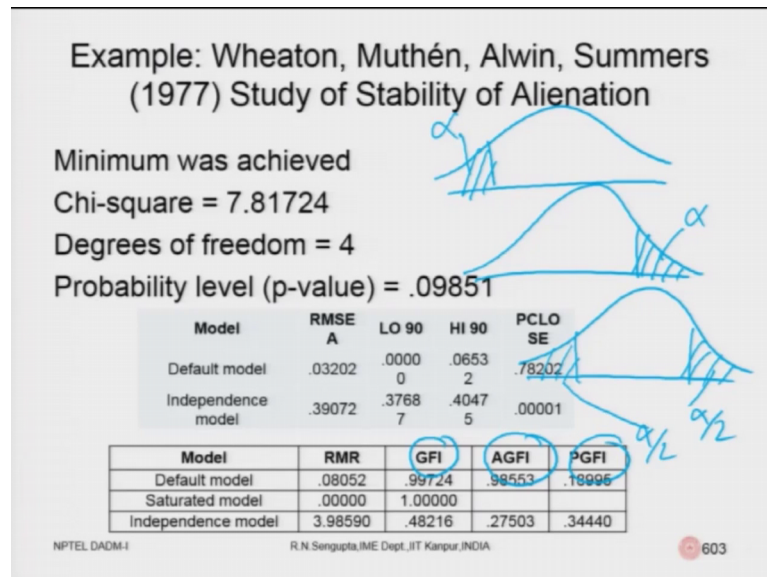
Now when you find out the factors corresponding to the structural equation modeling, what you have actually is this. So, beta 1 would basically denote which is which one of the elements in the first row would basically denote the affects which is going to happen 1, 2, 3, 4, 5, 6; so 1, 2, 3, 4, 5, 6, so this one. So, beta 1 would be multiplied by let me count it out; 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, 6, 7.

So, beta 1 would be related to alienation of 67. So, the factor is beta 1; when I go to beta 2 again this is beta 2, which is in second row. So, that would basically when you multiply. Again it will come into alienation, but with the factor of beta 2, provided we are considering that for 71 and 67 separately. So, once you do that, you will basically have a set of factors which are like this. The second set would all be 0's because the affects are not considered. And the first set would basically have the factors depicted normalized; where beta 1, beta 2, beta 3, beta 4, till beta 9, beta 8 on beta 9 would give you the effects which are specific to the factors considered on a on a standalone basis.

So, beta 1, beta 2 and beta 9 would be based on the fact they are on the 6th level. 6th level means 6th entry onto the right. Beta 3, beta 4 on the 7th level of the 7th onto the right; and beta 5 beta 6 beta 8 would be on the 8h level onto the right. So, they would

basically be multiplied accordingly as you consider the variables which are given. And obviously, the errors are given by  $z_1, z_2, e_1, e_2, e_3, e_4$ , and so on and so forth with has a certain mean and a certain standard deviation.

(Refer Slide Time: 09:01)



Now what we need to find out is basically, we need to minimize the errors and minimize the deviations and find out the maximum variation which is going to come out. So, from here we find out the chi square.

So, the chi square is basically if you remember when we considering the chi square values it was related to the distribution which was from considering the normal distribution case. And our main concern was to understand that how the standard deviations of the variance would basically depend on a particular distribution. So, they are not comparing distribution we are only considering one certain value of the standard deviation and trying to find out that based on the sample how we can consider. The degrees of freedom would give you to what level or what are the different effects which we have.

So, if what technically what is the number of such so called variables on which level they can be considered. So, if there is dependence structure obviously, the degrees of freedom would decrease accordingly. The probability level obviously, that would depend on the level of confidence which you want.

So, if you remember in the case of hypothesis testing, in the case of interval estimation we considered, the level of significance, depending on whether is the left hand test or a right hand test or a or a boattail test, we considered that the values could be 5 percent, 10 percent, 15 percent and so on and so forth. So, we will basically divide it accordingly. So, if it is a left hand test, we will consider the whole level of significance being on the left hand side, right hand test will consider the whole level of significance of the right hand side, and if does a 2 tailed test will basically divide it equally.

So, if you remember for the left hand test on the left hand side, on the right hand test on the right hand side. And if we consider there both levels it will be on the both sides, but the overall area would be added one value with the same. So, this is alpha, this is alpha and the sum of them would be alpha such that it they would be equally divided by alpha by 2 alpha by 2 on the left hand the right hand side.

So, once we consider the root mean square errors corresponding to the values. So, for the default model it will be about 32 percent. Not 32 percent about 3.2 percent and for the independent modulate it will be about 39 percent depending on the level of significance which we have.

So, o the values of GFI, AGFI and PGFI I will basically come to this details in few minutes. So, the default model saturation model and the independence in independence model will be considered accordingly.

(Refer Slide Time: 11:45)

**Example: Wheaton, Muthén, Alwin, Summers (1977)  
Study of Stability of Alienation: Regression Weights**

			Estimate	S.E.	C.R.	P	Label
Alienation1967	<---	SES	-.64495	.05350	-12.05418	***	
Alienation1971	<---	SES	-.22497	.05509	-4.08390	***	
Alienation1971	<---	Alienation1967	.58916	.05580	10.55811	***	
Educ	<---	SES	1.00000				
SEI	<---	SES	.58409	.04264	13.69760	***	
Powles67	<---	Alienation1967	1.00000				
Anomia67	<---	Alienation1967	1.12575	.06772	16.62422	***	
Powles71	<---	Alienation1971	1.00000				
Anomia71	<---	Alienation1971	1.13332	.07111	15.93816	***	

NPTEL DADM-I R N Sengupta IIT Kharagpur, INDIA 604

So, first let me go to the general measures or matrix which we considered. So, let me go into it detailed first. So, these 2 would be important for means. So, I will basically take it here, so let me go one by one.

(Refer Slide Time: 12:06)

### SEM: Model Identification

Model-Fit Criteria and Acceptable Fit Interpretation

Model-Fit Criterion	Acceptable Level	Interpretation
Chi-square	Tabled $\chi^2$ value	Compares obtained $\chi^2$ value with tabled value for given $df$
Goodness-of-fit index (GFI)	0 (no fit) to 1 (perfect fit)	Value close to .90 or .95 reflect a good fit
Adjusted GFI (AGFI)	0 (no fit) to 1 (perfect fit)	Value adjusted for $df$ , with .90 or .95 a good model fit
Root-mean square residual (RMR)	Researcher defines level	Indicates the closeness of $\Sigma$ to $S$ matrices
Standardized RMR (SRMR)	< .05	Value less than .05 indicates a good model fit
Root-mean-square error of approximation (RMSEA)	.05 to .08	Value of .05 to .08 indicate close fit
Tucker-Lewis Index (TLI)	0 (no fit) to 1 (perfect fit)	Value close to .90 or .95 reflects a good model fit
Normed fit index (NFI)	0 (no fit) to 1 (perfect fit)	Value close to .90 or .95 reflects a good model fit
Parsimony fit index (PNFI)	0 (no fit) to 1 (perfect fit)	Compares values in alternative models
Akaike information criterion (AIC)	0 (perfect fit) to positive value (poor fit)	Compares values in alternative models

NPTEL DADM-I                      R.N. Sengupta, IIME Dept., IIT Kanpur, INDIA                      604

For the model fit criteria or the acceptance fit interpretation for the structural equation modeling would be like this. For the chi square would be considered for the chi square tabled on the case that what is the degrees of freedom which you consider.

So, obviously, will consider the central limit theorem not to be true, and will case the take the chi square depending on the degrees of freedom which you have. So, if it is  $n$  or  $n$  minus 1 it will basically define that for the univariate case  $um$ , if  $n$  minus 1 being a case that you are utilizing the sample to find out the first moment. If it is  $n$  only; that means, you have all the information for the first moment.

So, chi square which would be acceptable level would be at table in the chi square values. And the interpretations would be compares, it will compare the obtained chi square values with the table values for a given degrees of freedom. So, given their degrees of freedom you are changing; obviously, the chi square values with change, and the level of acceptance would also change.

Now, the goodness of fit index would be the depending on the levels it can be either binary one. With a 0 for no fit and one for a perfect fit, depending on their values of



goodness of fit which we give. So, close value is to 0.9 and 0.95 or 0.975 would give you that the values reflect a very good fit. So, obviously, one may not be the possible case, because one would be the best fit.

So, if the values are almost closer to one. About 90 percent, 97 percent, 95 percent, 99 percent it will give you what is the level of degrees of your fit of the model with respect to the existing data which you have. The adjusted they have goodness of fit would basically be somewhat value depending on the degrees of freedom.

So, if you remember in regression model technically we have  $r$  square and adjusted.  $R$  square,  $r$  square is for the case where we do not consider the degrees of freedom, and for the adjusted  $r$  square is basically the case when we consider degrees of freedom. So, if that is the value. So we will consider the degrees of freedom, and consider the adjusted goodness of fit which will give us the level of significance to which our model, actually fits the existing data or a existing criteria or existing analysis we are doing.

Now, this goodness of fit should not be analyzed or should not be confused with the goodness of fit which we have considered in the balance loss function. When we consider the precision of estimation and the goodness of fit consider corresponding to the fact that the precision of estimation was to do with how we are able to find out. The beta values with respect to their estimated values being  $\hat{\beta}$ .

So, goodness precision of estimation would give me the errors pertaining to  $\hat{\beta}$  and  $\beta$ . So, that can be as I discuss it can be linear loss or squared error loss. And the goodness of fit was basically based on the fact that what we want to predict what you want to forecast that how good or bad it is with respect to the actual value. So, that was basically the difference between  $\hat{Y}$  and  $Y$ ; where  $Y$  with actual value and  $\hat{Y}$  was the predicted value. Now adjusted goodness of fit would basically be how good or bad your model is.

So, here there we consider the error concept, here we will consider the level of significance based on the fact that one is the best fit and 0 is the worst fit. So, any close any values closer to one would give you that how good your actual prediction is. The root means square residual values would basically depend indicate the closeness of the overall covariance variance matrix of the population with respect to the sample.

So, obviously, and the greater the difference is; that means, the variability is very high in the sample with respect to the variability in the population which is low. So, as you compare the mean values of the population with the sample. You also consider the difference of the closeness of the standard deviation of the covariances of the population with respect to the standard error square, and the covariances of the sample. That means, you are trying to compare the difference of the dispersion difference of the variability. And closer the values are, better is the fit and we will basically consider the values of root mean square which will give me that how good or bad the fit of the overall sample covariance is with respect to the population covariance matrix.

So, there will be matrix depending on the number of variables which we have. Standardized mean square error would basically give you that you are trying to basically standardize depending on the degrees of freedom of value if it is less than 0.05. So, it will indicate a good model fit with respect to the root mean squared or they are standardized to the case, that you are able to find out per unit what is the overall error which you are trying to phase. The root mean squared error or approximation which is basically between 0.05 to 0.08 would basically mean when that what is the overall approximation you are trying to utilize for the root mean squared error.

So, if you remember that it is some something like this. In general, the overall mean value of the sample should be exactly equal to the population mean, or the some standard error or the variance of the sample should be exactly equal to the population standard deviation of population covariance. Now as the difference occurs so, obviously, you will try to find out what is the probability.

If you remember the concept of consistency, consistency was as the sample size increases, what is the probability of the difference by between the population parameter and the sample parameter. And it should be less than some epsilon where epsilon depends on the sample size. So, if it is basically the error is less than that epsilon, where epsilon start decreasing and then as the  $n$  sample size increases and this probability tends to one; obviously, will say that the variance is on a actually decreases and becomes 0.

So, this would be true for the case as the sample size increases and approaches the population value. The other fits are tuckers Dewey's index, normalize phase index on the parsimony fit index which are used to in order to find out what is the value of the of the

fit with respect to the normalized in the standard normalized structural equation model mean method or the values.

Now here I would like to mention that even though in the overall holistic sense it may not be in have been possible to cover different methods in it is a absolute details in order to give the results. But we have tried to basically cover many of the things which are there in multivariate statistics; considering multiple linear regression congenital this canonical correlation method, factor analysis, then PCA method, principal component analysis method; then different type of distributions which you have considered for the multivariate statistics and also we consider the structural equation modeling.

And if you go back little bit more into the multivariate univariate cases, we considered the concept or different type of hypothesis testing, considering we use the chi square distribution f distribution, t distribution and the z distribution. And before that we did also consider the different above of univariate discrete and univariate continuous distribution.

So, considering the overall coverage of DADM 1, it has been a huge amount of concept which you was considered. In many of the cases we just went and trying to cover the concepts. In other cases we basically went to the concept and solve different problems.

But a we think that considering the overall coverage consider and the time frame which we had for the course, we have been able to do definitely some justice in the overall coverage of the course. And obviously, we will be open to feedback from the students in the forums if they write and ask questions will be most happy to answer them. And a definitely as things are fine tuned in the sense that the students learned this topics try to solve the problems.

I am sure that many of the concepts which you have covered fleetingly definitely would be much clearer as they continue in using this in the different type of problem setting; maybe it is in production, maybe it is in quality control, maybe it is say for example operations, maybe it is in supply chain, maybe it is in the area of social sciences. Such that they will be able to utilize the concept of a DM data analysis and decision making trying to utilize different univariate and multivariate statistics to the best possible extend.

With this I will close this class. And I hope that we will and I am sure that will be there to help the students. And hope that all the queries which are there coming up in the forum will try our level best to answer them to the best possible extents such that they benefits the students in the long run. With this I will close the course and I will definitely like to thank all of you for your attention, I would like to thank my TAS and the whole NPTEL team form the institute where I am teaching an IIT Kanpur and definite IIT madras. And I am sure that they with the able help with all the staff all the members will be able to do a much better job in order to propagate a this concepts of DADM 1 to all the students who definitely want to do such course.

And to give a small brief background will also be going to the DADM 2 course; which will consider different-different topics in operation research in reliability optimization, robust optimization, different type of parametric and non-parametric tools, but obviously, the coverage would be huge, but considering the time a constraints which will have considering we are 30 hours we will try to do our level best. In order to give the both the theory as well as the problem solving techniques, and also solve few problems who do the best interest of the students. Have a nice day.

Thank you very much for your attention.