Handling Large-Scale Unit Level Data Using STATA

Professor. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee

Lecture No. 26

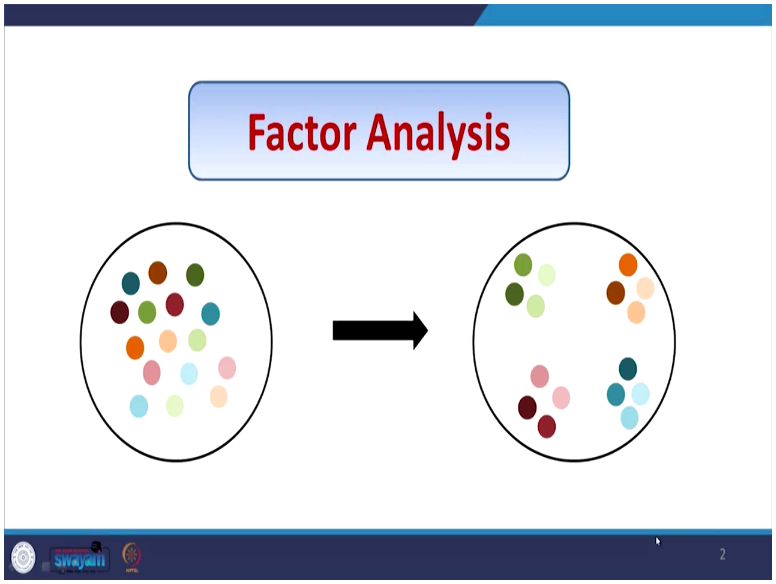
Factor Analysis with Stata-I

Welcome friends once again to the NPTEL MOOC module on Handling Large-Scale Data with Stata. We are here for explaining the analysis of unit level data. So far we have reviewed the details on unit level data from 3 different datasets along with we also referred number of other databases which are usually applied by our students. And along with that, we took couple of weeks on understanding Stata. In this particular week we will have a clarity on some of the tools which are largely used by the social science researchers, especially the researchers who go for qualitative research tools. We have included some of the modules, statistical tools, we deliberately included with the help of Stata for your better use.

In this regard, first tool we have included in our module is Factor Analysis with Stata. Let me give a backdrop to the understanding of factor analysis. If you have search in Google, you will find lots of information, but in many of the cases we have already searched, there are some short comings. And all the modules so far in our knowledge use SPSS or other, or even excel format to do factor analysis. But since our module is on Stata and Stata is very friendly if you are good at command and it derives results better.

So we are trying to present before you on the systematic steps of Stata to derive factor analysis. So, just factor analysis with Stata is not enough if you do not understand the theoretical construct behind factor analysis. So, in this regard, factor analysis is very important for researchers, especially social science researchers, health researchers.

(Refer Slide Time: 02:55)



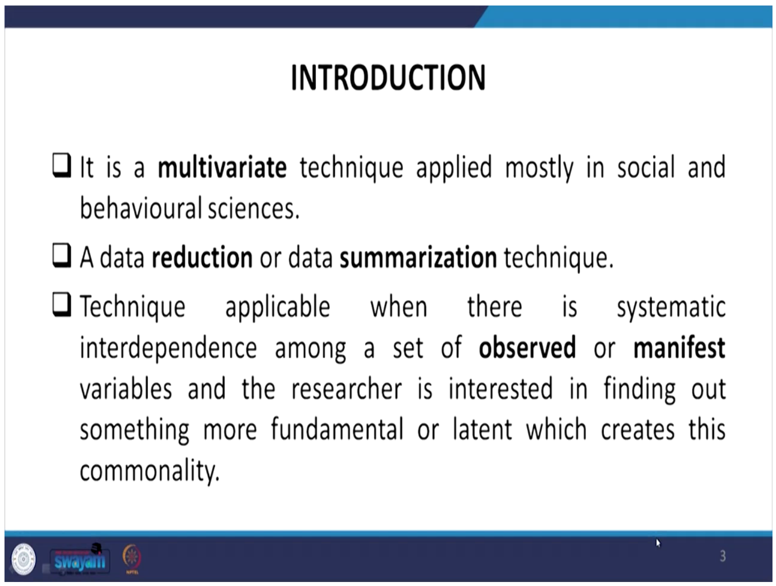
From the present picture, it is very indicative that in a set of factors if you are targeting to understand which is highlighted in the first circle, there are so many factors or at this moment for simple reason you may understand through variables, how variables are clubbed into factors we are going to guide you in a short while. But if there are so many variables and you wanted to analyze those variables for your purpose, I will give you lots of examples throughout these two lectures on factor analysis.

If there are so many variables you wanted to address some or wanted to derive some answers out of the variables, it is probably very difficult if you do not simplify those variables or in simple terms here we are saying if you do not simplify in terms of some particular factors. For me, after a simple review of those different nature of variables, I made it into 4 different groups. Those groups may be you may say 4 factors or 4 variables at this moment. Do not get confused by factors and variables.

Variables are larger set and from there we confined it to particular factors based on certain calculations, certain steps and those factors will be interpreted accordingly. So the first point of explanation for factor analysis is reduction of the variability or reduction of the differences or reduction of the factors or reduction of the variables. So, if you can reduce or make it a group or simplify, it will be easier for our analysis.

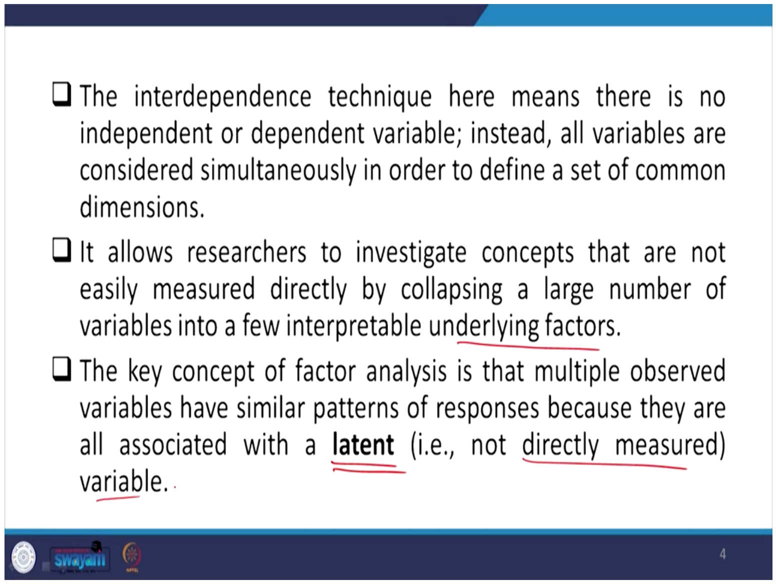
There are number of ways by which we try our best to explain the clarity of factor analysis. Let me start with the formal introduction to it. The factor analysis is a multivariate technique applied mostly in social science and behavioral sciences. It is also called a data reduction or data summarization technique which we just mentioned.

(Refer Slide Time: 05:35)



The technique applicable when there is systematic interdependence among a set of observed or manifest variables and the researcher is interested in finding out something more fundamental or the latent, we are going to clarify, latent the simple meaning is unobservable variables, which creates this commonality. So, basically set of observed or manifest variables, we are going to study in a short while.

(Refer Slide Time: 06:02)

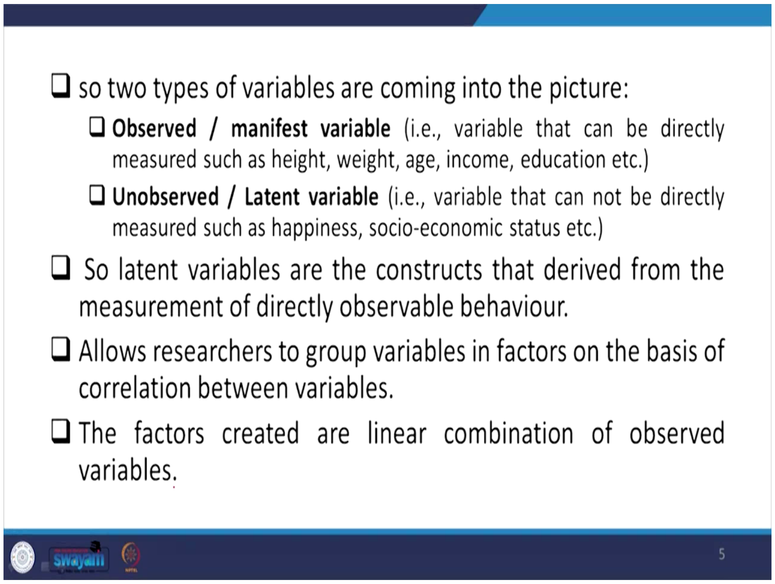


The interdependence technique here it means there is no independent of dependent variable to mark it very carefully that there is no such independent and dependent variable. We have to mention it that what sort of dependent or independent variable you want has to be very clearly mentioned, let me come to the point. This allows researchers, like as I said in a couple of minutes back that there is no clear defined independent and dependent variable, instead all variables are considered simultaneously in order to define a set of common dimensions that can explain jointly to a particular target can be understood from this analysis.

This allows researchers to investigate concepts that are not easily measured directly by collapsing a large number of variables into a few interpretable underlying factors. So, we are mentioning underlying factors for your clarity, so we are going to define very shortly. The key concept of factor analysis is that multiple observed variables have similar patterns of responses because they are all associated with a particular latent which are not directly measured or a latent variable.

So, the meaning of latent there are some observations which we are going to derive that will explain a larger goal, but since a set of variables are defining that particular variable, cannot be directly observed. So, cannot be directly measured. But that latent variable could able to explain so many things for the researchers. We are going to discuss.

(Refer Slide Time: 08:08)

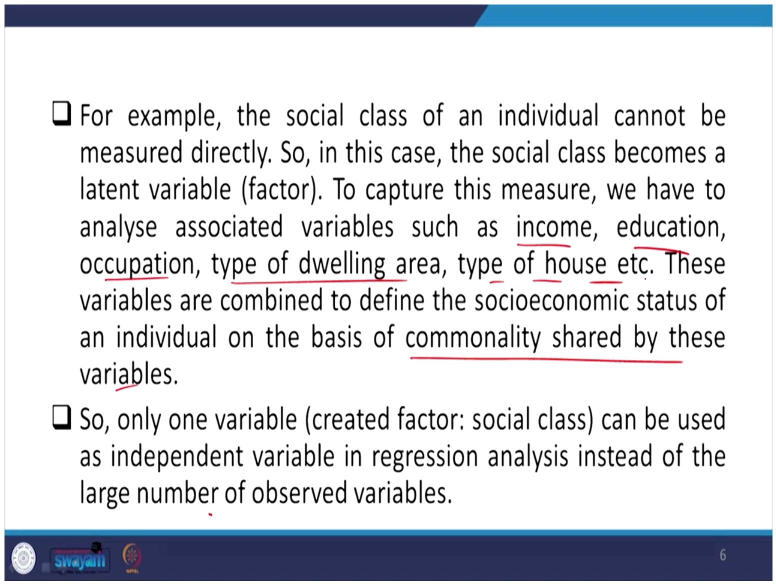


So, two types of variables are coming into the picture, that is one is observed another is unobserved. Observed we say those variables which can be manifested or can be studied, like directly we can have measurement for height of a person, weight of the person, age of the person, income, education, even the caste can be also understood, gender, there are variables, can be easily understood.

Whereas, some variables like if you just try to understand happiness of the person, empowerment of women, it composed of so many variables, socio-economic status of the persons or employer for example you are walking in any organization and employer has taken feedback from different perspectives, different directions, just feedback with one direction can be observed if it is there. But if there are so many directions and so many dimensions of the feedback and they have collated with different dimension to a single variable those are called latent variables.

So, latent variables are the constructs that derived from measurement of directly observable behaviors of different variables. So, it allows researchers to group variables in factors on the basis of correlation between the variables. So, any variables cannot just be taken together to define the latent variables, there must be similarities among the combination of variables which are going to define a latent variable. So, there must be some similarities. We are going to define what kind of similarities we are going to study. The factors created are linear combination of the observed variables that the linear combination jointly defining another variable or another factor is nothing but called latent variable.

(Refer Slide Time: 10:24)



So, for example, the social class of an individual cannot be measured directly. So, in this case, the social class becomes a latent variable if it is not understood directly. To capture this measure, we have to analyze associated variables such as income. Here if you are getting little confused with social class, with social sector or their caste, if you are little confused, this refers to social status. So, class refers to status here. So, that may be proxied by income, may be education, occupation, but jointly proxied, it not individually proxied, income, education, occupation, then type of dwelling, locality, housing type also. If you are combining all those dimensions together that may explain the social caste of a particular person.

So, individual with the index value, with a new factor value and defined as a new variable is called latent variable in simple term. These variables are combined to define the socioeconomic status of an individual on the basis of commonalities shared by these variables. So, interesting part is commonality must be shared by these variables. These variables we mean, income, education, occupation, type of dwelling areas etcetera.

So, only one variable created factor that is social caste can be used as an independent variable in regression analysis. Interesting part besides this factor analysis is that, factor analysis is not going to give you a detailed understanding of inferences. It only simplifies the variables. You cannot derive many inferences out of factor analysis. At best, you can simplify the variables, reduce the variables since there is no such dependent and independent variables.

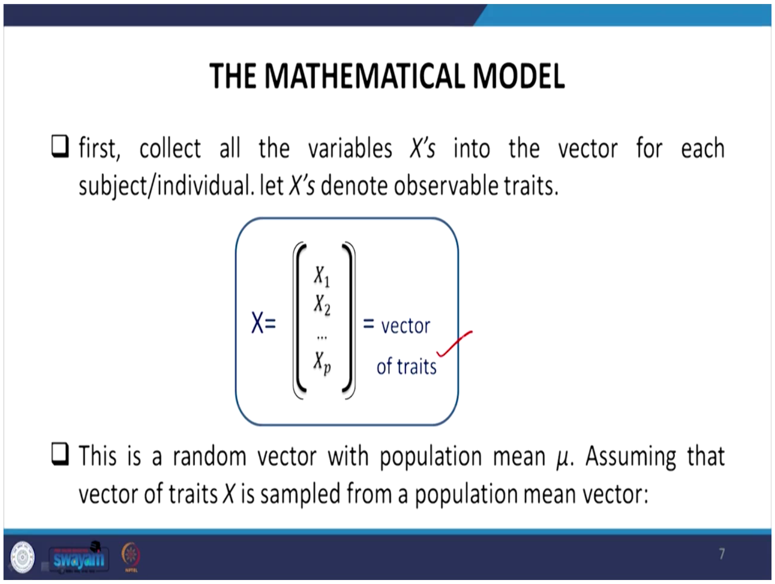
In social science research, even in management research, we largely go for a regression analysis. Regression straight away requires dependent and independent variables. So in that case, independent variable we generally take the variables which are observed and which are also not observed. even there are lots of interpretations for unobserved variables. We will also explain some of them in the regression analysis.

At this moment what I wanted to say, if you can collate so many variables together to a fantastic different variable that variable will narrate important interpretations and can explain the dependent variable in a very holistic manner. So, it attracts the researcher, the readers to understand the ideas very differently that also helps in policy making as well.

So, the latent variable or the single variable which we have defined by collating so many common variables based on their commonality and that will be going to give a very good robust results. We will also discuss why it is going to define some forms of robust result, because it avoids the possibility of multicollinearity among the independent variables. Because since you have collated so many variables to another single variable so the linear relationship with that latent variable with other independent variable hardly exist.

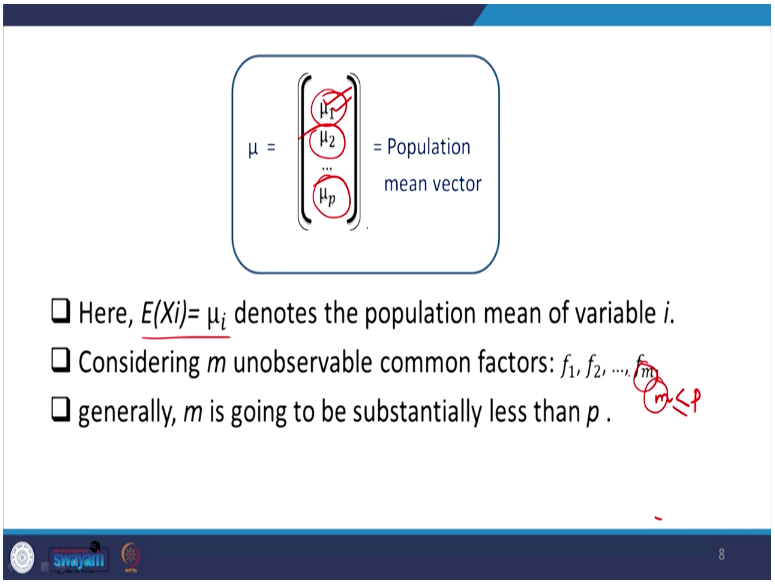
So, the extent of multicollinearity is expected to be very very less with this approach. So, this factor analysis is also considered to be one of the strongest approach of regression for pretesting of its variables.

(Refer Slide Time: 15:01)



One simplified mathematical model we are presenting for you to highlight the fact that if you have so many variables in the model and if that could be presented in a vector and let X be, X denoting the vector of traits or the features of the variables. This is random vector with population mean mu. Assuming that vector of traits that is X is sampled from a population mean vector.

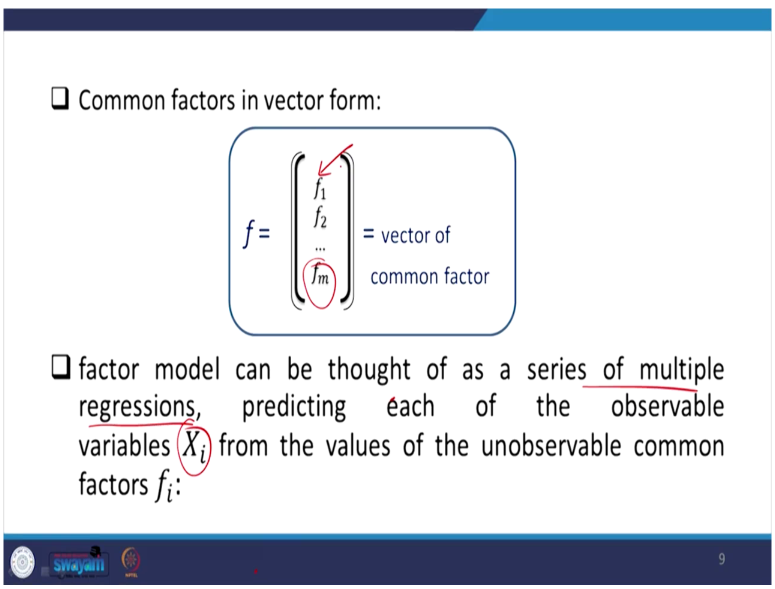
(Refer Slide Time: 15:16)



Let us understand a bit on the population mean vector. It distributed over a vector of information for the first explanatory variable for mu one, then second set of explanatory variable for X_2 is mu_2 that is the average mean of the population for the X_2 type of variables and that is for mu_p for the pth information of independent variables that is X_i. So, the expected value of X_i is denoted by mu_i. So, mu_i we are referring to either mu_i stands for mu one or mu two or mu p, population mean.

So, considering m unobservable common factors if it stands to explain there are set of common factors like starting from one to m, so not necessarily the number of variables as explained with the vector till p will be similar till m. This m might be lesser than that of p, or may be similar also, less than equal to. But in general since we are trying to reduce the number of factors, so we always expect that m is lesser than that of the p. So, generally, m is going to be substantially less than that of p. But that not necessarily might be equal, but if it is equal then what is the point of doing factor analysis. So, it should be substantially lesser than that of p.

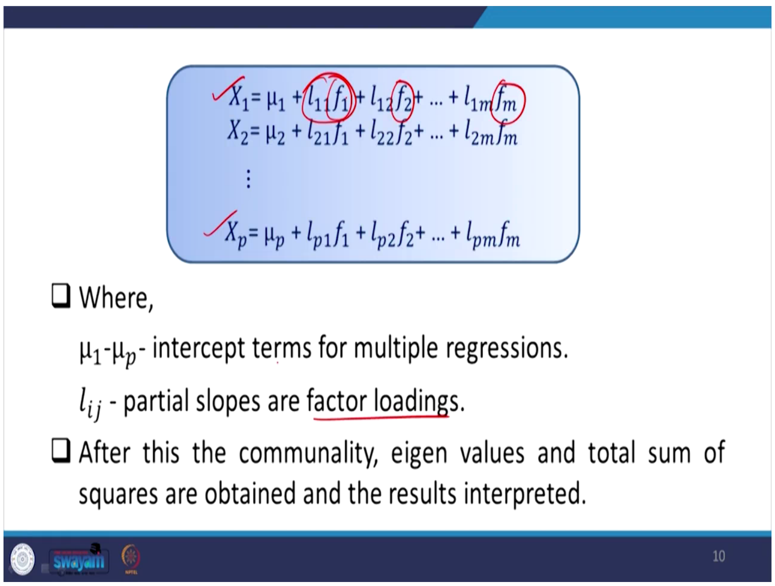
(Refer Slide Time: 16:54)



So, the common factors can be also presented in a vector, in a column vector we are mentioning here and that represents till m^{th} common factors. But m^{th} common factors are derived using the p^{th} independent variable vector X. So, factor model can be thought of as a series of multiple regressions. So, we are mentioning multiple regressions because we are including so many variables or so many factors together. So, predicting each of the observable variables that is X_i is the observable variables here not necessarily observable variables, these are unobservable factors, but from using the observable variables that is X_i.

And these are the observable variables from the values of the unobservable common factor f_i. So, there are some confusion X_i if you have gone through the understanding as observable variables then from the X_i you may limit it to the factors called f_i and f may be your unobservable common factors.

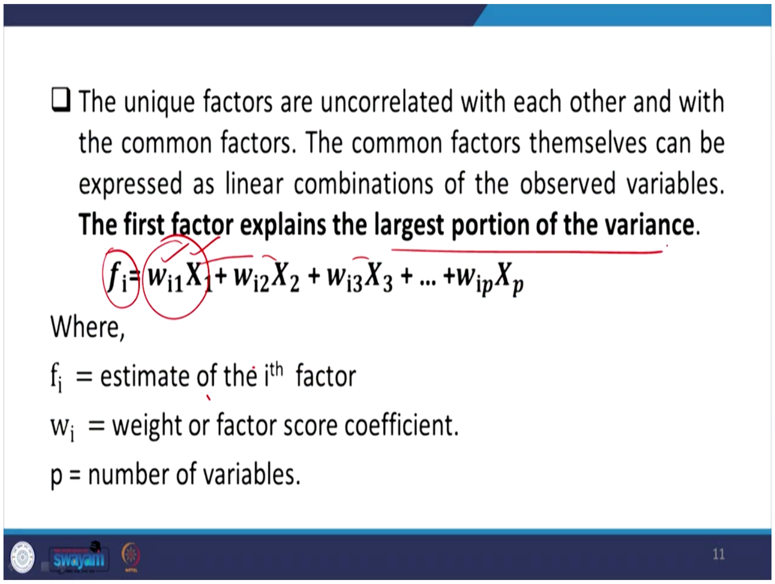
(Refer Slide Time: 18:15)



That can be represented in linear equation, as I already mentioned that it explains a multiple regression with multiple variables together, so with its factors, factor 1, factor 2 and factor m for the first kind of variable if you have so, and for the first model and for the second model and similarly other set of factors are included and accordingly we can explain till p^{th} of the variable we are trying to explain.

So, or mu_1 till mu of p represents intercept terms, the constant terms of the multiple regressions, whereas the l_{ij} stands for the partial slopes in the regression coefficient, we will also explain those in detail, are also called factor loadings, each of the factor loadings of the individual factor loadings to the result of X_1 in that model for. In this model, we are taking the outcome variable as the X_1. After this the commonality Eigen values and total sum of squares are obtained and the results interpreted accordingly.

(Refer Slide Time: 19:40)

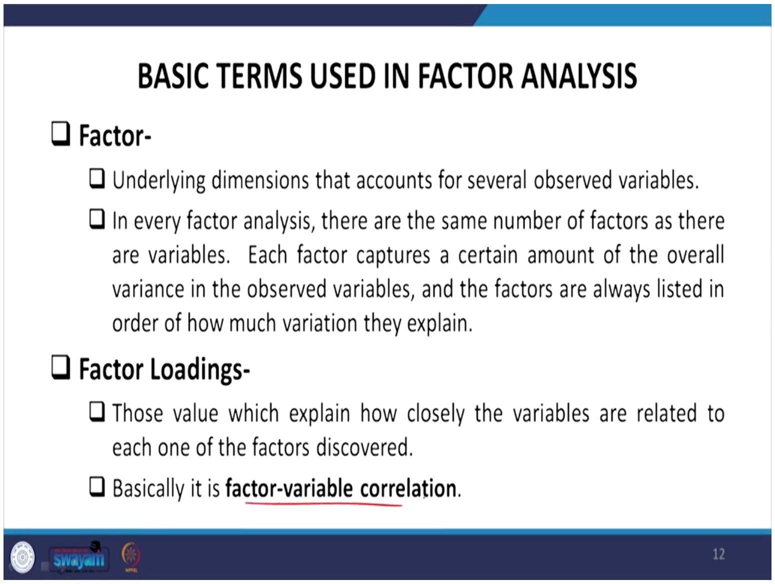


We are going to discuss what is the necessity of sum of squares and how sum of squares are useful in understanding the Eigen values or so, we will explain in a short while. So, the unique factors are uncorrelated with each other and with the common factors. The common factors themselves can be expressed as linear combination, which we have mentioned, combinations of the observed variables.

The first factor explains, after getting those factor variables through a set of other factors, after getting that and its individual weight if you can able to define, we can define an individual factor. So, those common factors themselves can be expressed as linear combinations of the observed variables. Here we mean these are the observed variables and the linear combination based on their respective weightage.

So, but the first factor if you can define, I will explain with the help of example through Stata, the first factor if we explain that represents the largest proportion of the variance. The largest disturbances in the model should be explained by the first factor f_i, if f_i, maybe f_1, f_2, f_3 depending upon the number of factors it has defined. But the first factor generally defines the larger proportion of variance. So, f_i is the estimate of i^{th} factor and w_i is the weight or the factor coefficient, we will also discuss that, then p is the number of variables in the model.

(Refer Slide Time: 21:37)

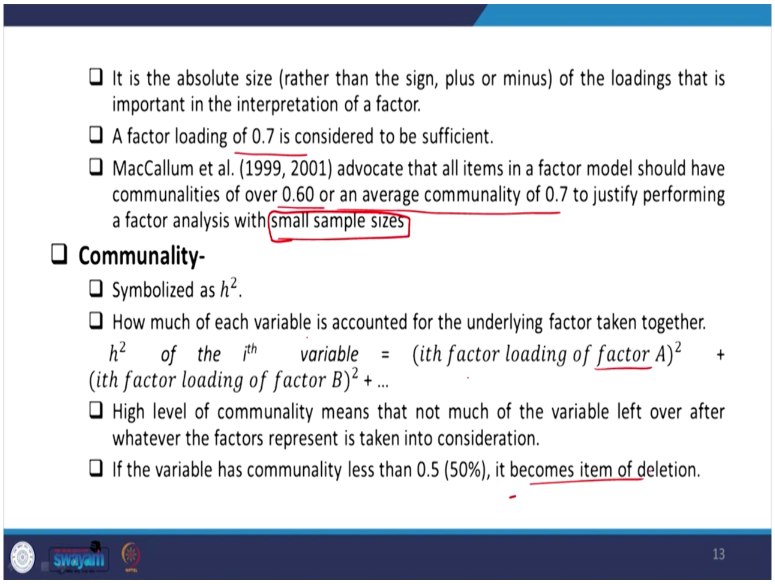


so far we have used some of the terminologies and we are also going to discuss important terminologies which are very, very necessary for the factor analysis are the following. One is factor, factor loadings and then commonality, we will also discuss eigen values, we will also discuss factor, rotated scree plot, then rotated matrix, then we will also discuss factor loading, taking it here, factor score as well, so and also rotation, there are some thumb rules of rotation but we will not be emphasizing much on it, but at this moment let me start how we should proceed for it.

So, what do you mean by factor? So, the underlining dimensions that accounts for several observed variables are called factors. In every factor analysis, there are the same number of factors as there are variables. Each factor captures a certain amount of the overall variance in the observed variables, so we have already discussed that, and the factors are always listed in order of how much variation they accounted for.

Come to factor loadings. Factor loadings are those value which explain how closely the variables are related to each one of the factors discovered. So, how closely these variables are related to each other. So, based on their loading, we can find out how much they explain. And basically, it is a factor variable correlation. So, that correlation aspect can be captured from the factor loadings estimation.

(Refer Slide Time: 23:43)



And it is the absolute size rather than the sign, plus or minus, is not going to divert the discussion, so the absolute size of the loadings that is important in the interpretation of a factor. to mention it very clearly as a standard thumb rule for interpretation or for explanation, factor loading of 0.7 is considered to be sufficient as per different estimation.

By MacCallum et al. 1999, 2001 advocate that all items in a factor model should have commonalities of over at least 0.6, minimum of 0.6 is required or an average commonality of 0.7, if individuality is not of 0.6, but on the average the factor loadings are of 0.7 that is considered to be sufficient and that 0.7 justifies performing a factor analysis with even small sample size.

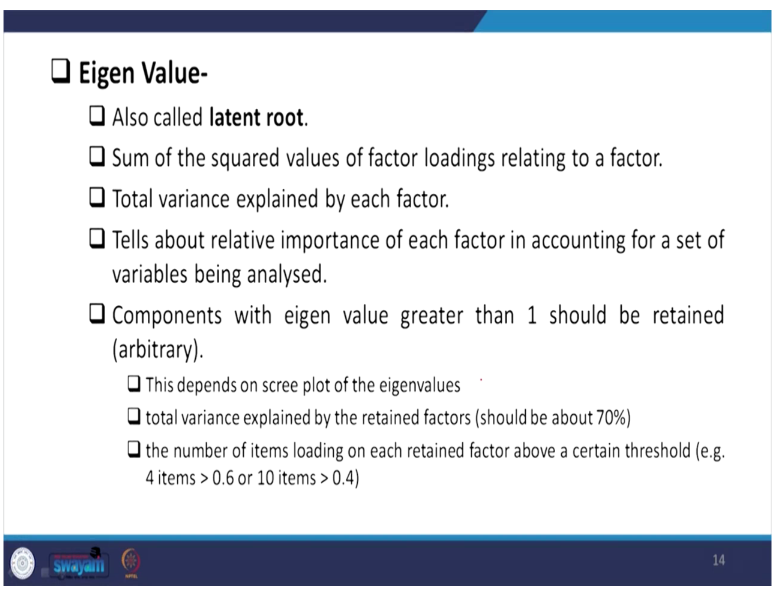
Usually it is suggested to go for a large sample size estimation, but for individual researchers who does for their own study usually go for small sample size. So, in that case the standard thumb rule for your approach of going for factor analysis to test whether is it fine or not, you must stick to the standard rule that is 0.7. 0.7 of factor loading must have been maintained then only your factor analysis is expected to be correct.

Another aspect is called commonality. So, commonality the word has to be very clearly identified through the factor analysis. It is symbolized usually by h square. How much of each variable is accounted for the underlying factor taken together? So, the h square of the i^{th} variable, if you are discussing about any particular variable, h square of that is basically i^{th} factor loading of the factor A square that is if the factor A is taken, if that square plus the i^{th} factor loading of factor B square etc. If you take the square of each of the factors of that particular order ^{ith} or first, second, third of the factor, if you take the total square of the variables then that is nothing but the commonality.

High level of commonality means that not much of the variable left over after whatever the factors represent is taken into consideration. So, high commonality if the square boils down to be much higher, then you need not worry, you have already counted the specific variables that represents the highest loading and the square will be much higher and any other if you are adding is not going to divert much from the model or from the factor analysis.

We are going to claritfy with the help of example. If the variable has commonality less than that, it is interesting to note if it is less than 50 percent or 0.5, it becomes item of deletion. The variable if that has a commonality less than 0.5, so that should not be considered for the analysis, should be deleted.

(Refer Slide Time: 27:21)



Another most important aspect for explanation in the factor analysis is Eigen value. Eigen value explains the latent root. It is also called the latent root. The sum of the square values of the factor loadings relating to a factor is also sum of the squares, once you have taken the sum of the squares of the factor loadings that defines the Eigen value. And it also explained the total variance by each of the factor in the model in the analysis.

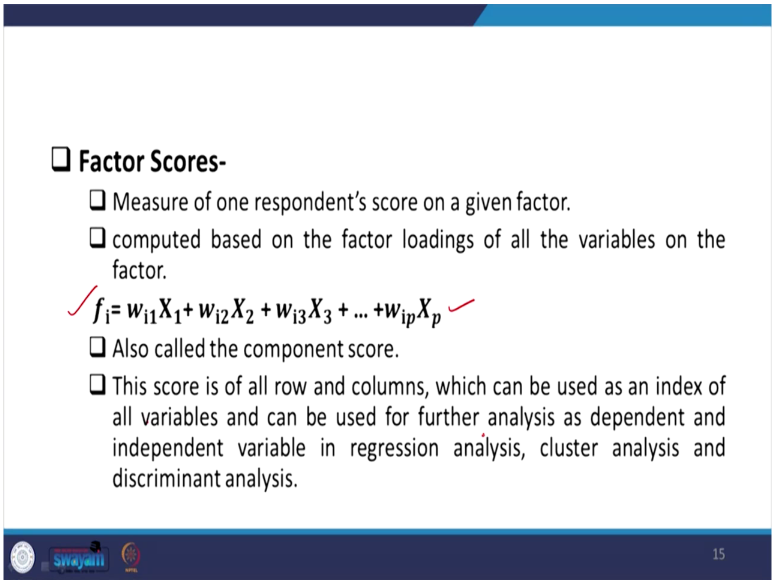
So, we are explaining Eigen value. In Eigen values we are going to count down the square of the factor loading. Sum of the square of the factor loadings are explaining that is nothing but called the Eigen values and this is also called latent root. And when you have taken sum of the square of the factor loadings, obviously factor loading is basically the extent of the variance loaded in a particular factor.

If we have taken a square of it, we can find out the deviation from that particular factor and how much deviation is explained with the help of variance is explained in the Eigen value, because the total sum of the squares can be derived. It tells us the relative importance of each factor in accounting for a set of variables being analyzed since we are taking the individual square as well. And the components with Eigen value greater than 1 that is the standard thumb rule here, greater than 1 eigen value should be retained. This is again arbitrary. This is also not a standard rule everywhere. But largely if it is greater than 1, usually those factors are considered.

So the value of Eigen values to be considered depends on another aspect called scree plot of the eigen values. We are going to show, also explain it. The total variance explained by the retained factors that should be about 70 percent, 0.7 we have already shown you. That is sufficient. And the number of items loading on each retained factor above a certain threshold level that is, if it is 0.6, 4 items greater than 0.6 or 10 items there are some standard rule given. If there are 4 items if it is 0.6 that is perfectly fine. And if there are 10 items if it is 0.4 still that is also fine.

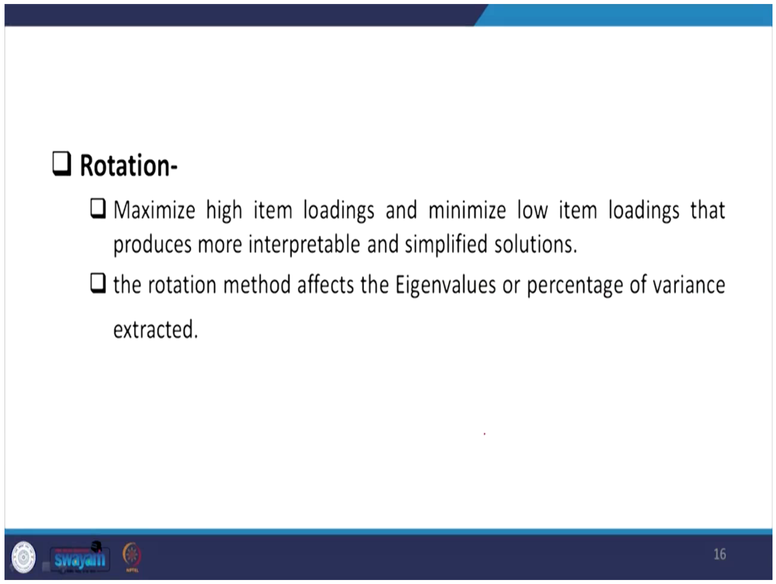
So, we mention that it is arbitrary. There are no defined rules of taking only one. It depends on the researchers and how they are going to be using their variables or factors for further analysis. So, factor scores are going to be very important at the end of the factor analysis. Factor scores is the measure of one respondent score on a given factor.

(Refer Slide Time: 30:55)



This is computed based on the factor loadings of all the variables on the factor that we have already mentioned with the help of individual factor with their individual weight of the observable variables that is X_i and that we have already explained. And also these factor scores are also called component score. And this score is of all row and columns which can be used as an index of all the variables and can be used for further analysis as dependent and independent variable in regression analysis, cluster analysis or discriminant analysis. Once the scores are known we can use it anywhere for our better purpose.

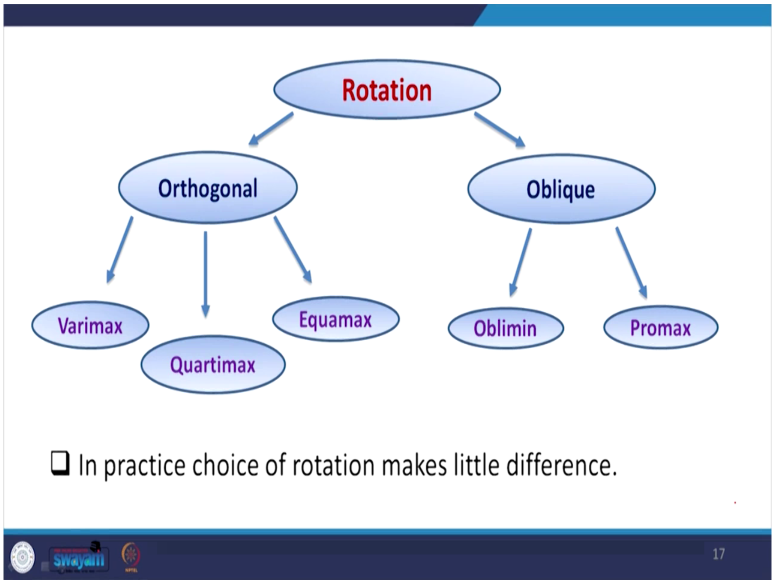
(Refer Slide Time: 31:41)



Now, once factor scores are determined and we understood the commonalities, we understood the factor loadings, we understood the factor scores, once that is done how assured we are to carry forward the factor loadings for our result, how to assure ourselves, what are the standard rules of our assurance. Can it be generalizing with the rules for all kind of dataset will be taking?

So, in order to know that factor rotation is important, factor rotation is cross test with different set ups. And the rotation maximizes high item loadings and minimizes low item loading that produces more interpretable and simplified solutions. The rotation method affects the Eigen values or percentage of variance extract. It rotates first and based on the rotation we can find out by sorting their variance and accordingly you can interpret. There are some standard rules of rotation as well we will explore through our explanation.

(Refer Slide Time: 33:01)



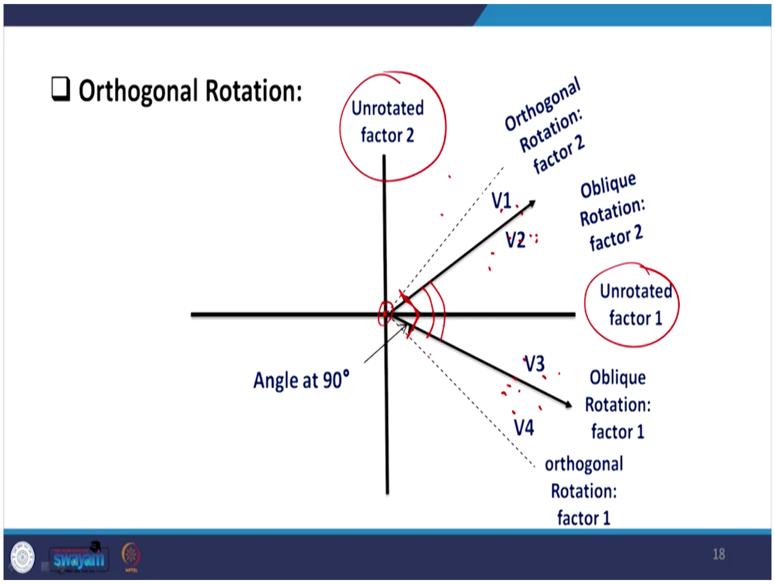
One is that rotation is observed through two approaches, one is called orthogonal, another is called oblique rotation. And orthogonal, they are usually of 3 types and one is called varimax, second one is called quartimax and third one is called equamax. So, another one we called oblique method of rotation. Broadly there are oblimin and promax.

Let me give you a choice of rotation which is practiced. There has been an evolution on it. When there was no computer, people used to do their calculation with the help of manual methods, with the help of paper pen, they usually go for varimax approach of the orthogonal type, where the rotation is 90 degree type, 90 degree rotation then they check whether the diagram is close to another variable or not by 90 degree change. Every 90 degree change which variable are close to the average factor value which we have defined.

Once that is there accordingly we can assure yes these kind of variables are very important for our analysis. With the use of computers and randomizations there are many likelihood estimations by high-end computing techniques. Authors, there are some authors who are very good in research for numbers, they suggest that oblique method is very good.

So, I am not going to the depth of it, because we are interested to find out the factor scores and its interpretation largely with the help of Stata and there are many geometrical methods of understating orthogonal to the oblique method, but I have only given you the backdrop of which is used and how they are used, you can refer other articles for it. So, in practice the choice of rotation makes very little difference. Rotation even not required much.

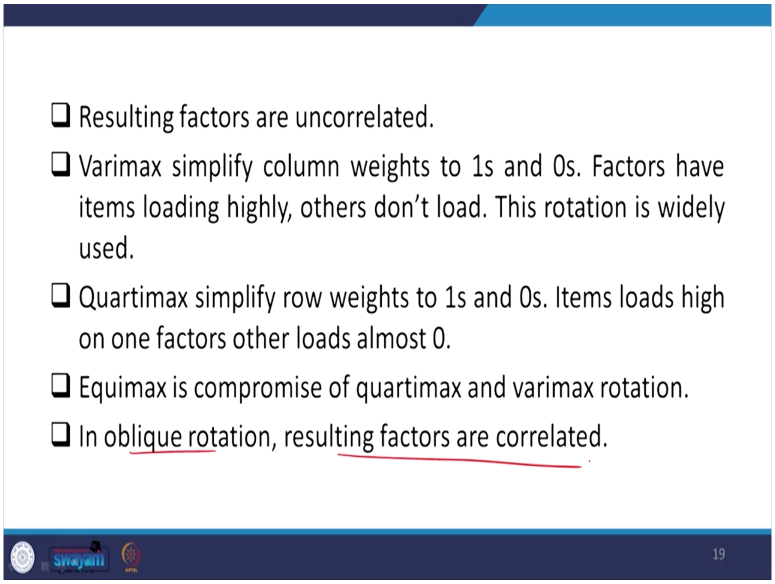
(Refer Slide Time: 35:17)



From the figure you can find out like we have, before we defined a factor one and factor two and that once we have plotted at a starting point at zero with a 90-degree angle initially, we are simply going to rotate how many variables are in which level, where the variables are mostly concentrated. if you take a 90-degree rotation for our understanding which earlier researcher used to do when they apply for the manual techniques, so they define the 90-degree rotation, like here the 90 degree rotation is being explained.

But in reality when you have high-end computing any degree, this may not be 90 degree. So, oblique rotation is also applied these days by researchers those who are very good at computing. Otherwise in standard technique or in practice it is the orthogonal rotation.

(Refer Slide Time: 36:30)



The rotation resulting factor, the resulting factors are uncorrelated. Varimax simply column weights to one or zeros. Factors have items loading highly and others do not load. This rotation is widely used. Quartimax simplify row weights to one or zeros which we have mentioned. Items load high on one factors, other loads almost zero depending upon where you are rotating and how the computer is rotating. Equimax is compromise of quartimax and varimax rotation. In oblique rotation, resulting factors are correlated. There are some rules of rotation. So, especially in oblique rotation method, factors that correlated much. So, any rotation you can do and find out the better answer.

So far I discussed about very important aspects of factor analysis. From the next lecture, I think it will be better to discuss about the assumptions and from the assumption we will clarify where to stick and what kind of assumptions are required, like some standard rules of factor loading and commonality and Eigen values we have discussed. There are some other assumptions also for factor analysis we will carry forward to the next class and also we will discuss the data using Stata for the analysis. Let me stop here. We will discuss in the next class. Thank you.