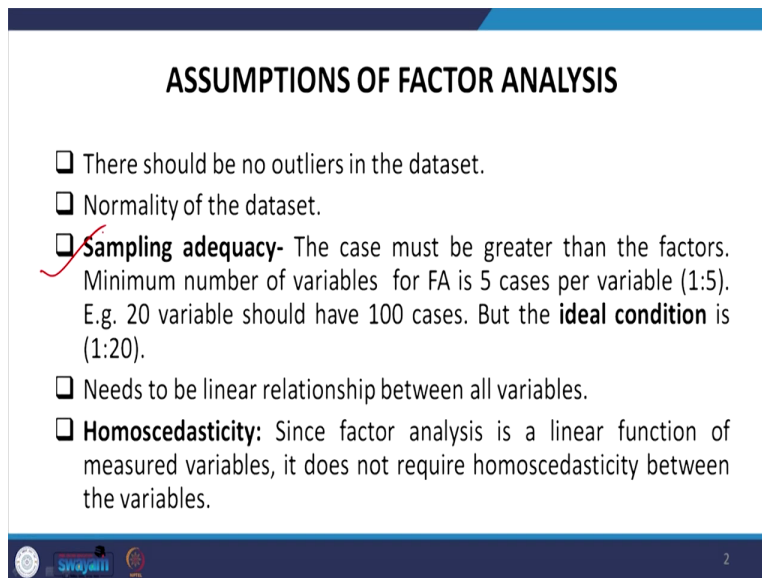


Handling Large-Scale Unit Level Data Using STATA
Professor. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee
Lecture No. 27
Factor Analysis with Stata-II

Friends once again welcome to this module on Handling Large-Scale Data Using Stata. We are at the particular week where we have been trying to analyze the unit level data after explanation of the database and the know-hows of Stata. In the lecture particularly we unfolded the discussion of factor analysis using Stata, but in the last class I particularly clarified the concepts of factor analysis, the meaning of factor analysis, especially what is called a factor, what is called a variable, what is called a latent variable and what are the thumb rules of factor loading, eigen values, commonalities we already discussed that. We are going to apply with Stata with certain assumptions to start with factor analysis in detail.

(Refer Slide Time: 01:29)



ASSUMPTIONS OF FACTOR ANALYSIS

- There should be no outliers in the dataset.
- Normality of the dataset.
- Sampling adequacy**- The case must be greater than the factors. Minimum number of variables for FA is 5 cases per variable (1:5). E.g. 20 variable should have 100 cases. But the **ideal condition** is (1:20).
- Needs to be linear relationship between all variables.
- Homoscedasticity**: Since factor analysis is a linear function of measured variables, it does not require homoscedasticity between the variables.

2

So, first assumption is that, there should be no outlier in the dataset. Dataset if it is having some outliers, the result is going to be misleading. Second assumption is that, normality of the dataset. Basically, outlier or this clarifies further that your dataset must be defining normality. There are some sampling adequacy norms we are going to discuss. The case must be greater than, so far as the adequacy is concerned and the sampling is an issue of observation sometimes for the individual researcher they go for small sampling.

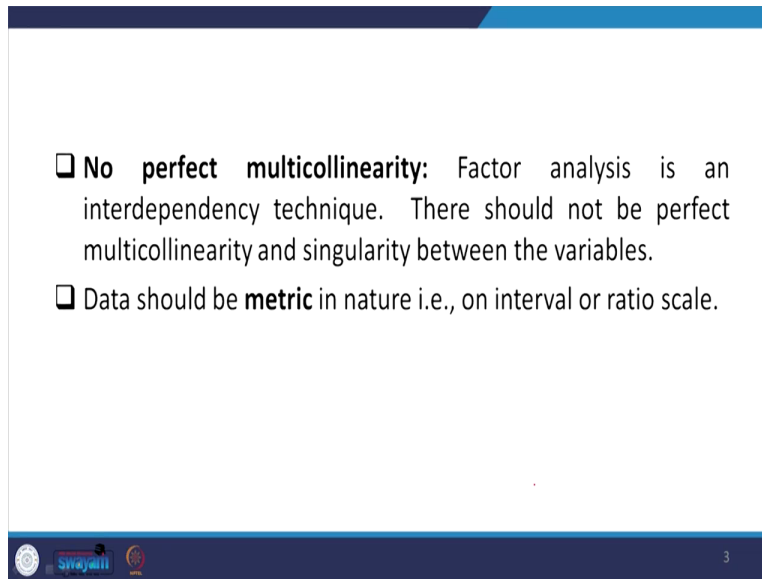
How to understand that? In this case, the sampling adequacy must be greater than the sampling must be greater than the factors, the number, that means, the minimum number of variables for the factor analysis is 5 cases per variable. That means it has to be pulling with 1:5 ratios that is if 20 variables should have 100 cases. So, the number of cases we are going to discuss or include, if you have 20 variables, so minimum of 100 cases must have been there to get certain degree of variability also commonality. But the ideal condition is 1:20.

In a large dataset usually the standard condition is followed as 1:20. For one variable 20 cases must have been there to get a sampling adequacy. This is one of the standard assumption to be followed. This needs to be having linear relationship between all the variables. The variables must have some linear relationship. Like variables must be correlated to each other linearly. If there is spurious correlation, then there is no point of taking them to the analysis through factor analysis.

Homoscedasticity is a problem. Since factor analysis is a linear combination of measured variables it does not require homoscedasticity between the variables. So, since it is already simplified the variables, I think I discussed in the last class, homoscedasticity assumptions tells us that the independent variables in the regression model should not be correlated, because in the regression analysis, each of the individual cofactor or the variable and its coefficients must be identified subject to no influence from another independent variable or explanatory variable.

If two independent variables or the explanatory variables are correlated to each other, it might be creating biasness in the model. Since in the factor analyses we already defined a factor through linear combination of variables, so homoscedasticity assumption is not required. That means in this case that homoscedasticity assumption is already valid in the model.

(Refer Slide Time: 04:59)



❑ **No perfect multicollinearity:** Factor analysis is an interdependency technique. There should not be perfect multicollinearity and singularity between the variables.

❑ Data should be **metric** in nature i.e., on interval or ratio scale.

3

So, one interesting thing to note is that there must not be perfect multicollinearity. Yes, co-relationship should be there, but if there is perfect multicollinearity or collinear to each other then there is no point of taking factor analysis. If like X plus X is equal to two y , if it is there two times of y , then that is perfect multicollinearity. Factor analysis is an interdependency technique. There should not be perfect multicollinearity and singularity between the variables. The variables must not be having perfect multicollinearity.

So, like X equal to $2y$ as I just said or X equal to 30 plus $3y$ anything, if it is perfect multicollinearity then there is no point of deriving or reducing the factors through factor analysis. Data should be metric in nature that means data should be either on interval or ratio scale. So, if it is non-metric then it might be problematic for the interpretation as well.

(Refer Slide Time: 06:20)

TYPES OF FACTOR ANALYSIS

- ❑ Broadly, two types of factor analysis:
 - ❑ Exploratory factor analysis (EFA)
 - ❑ Confirmatory factor analysis (CFA)
- ❑ **EFA-**
 - ❑ Summarizing data by grouping correlated variables.
 - ❑ Assumes that any indicator or variable may be associated with any factor.
 - ❑ This is the most common factor analysis used by researchers and it is not based on any prior theory.

Note!
The lecture is focusing on EFA

4

So, what are the types of factor analysis? Broadly there are two types, one is called exploratory factor analysis that is in short called EFA and second one is a confirmatory factor analysis. EFA stands for summarizing data by grouping correlated variables. It assumes that any indicator or variable may be associated with any factor. And the exploratory one is the most common factor analysis used by a researcher and it is not based on any prior theory. So, the lecture is focused on EFA only, not on the confirmatory factor analysis.

(Refer Slide Time: 06:57)

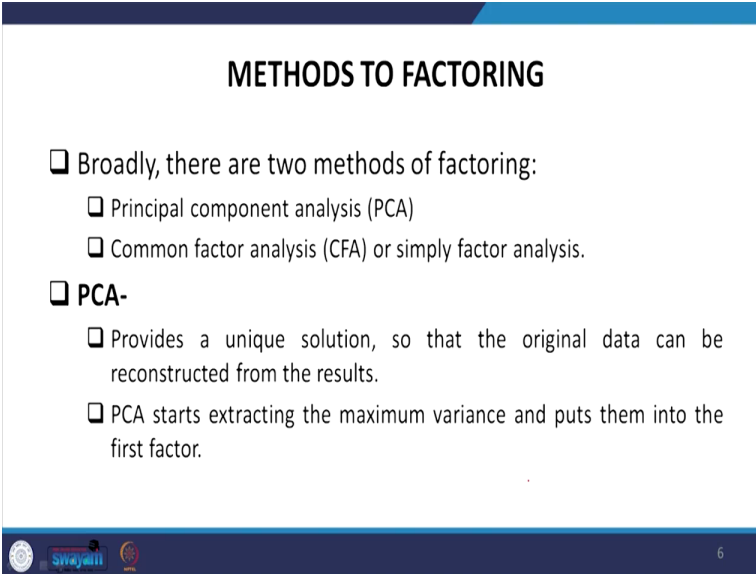
TYPES OF FACTOR ANALYSIS

- ❑ **CFA-**
 - ❑ More advanced technique.
 - ❑ Used to determine the factor and factor loading of measured variables, and to confirm what is expected on the basic or pre-established theory.
 - ❑ CFA assumes that each factor is associated with a specified subset of measured variables.
 - ❑ Needs SEM (structural equation modelling) package.

5

Whereas the CFA is more advanced in nature. It includes advanced techniques. This used to determine the factor and factor loading of measured variables and to confirm what is expected on the basic and pre-established theory, because without that theory this confirmatory factor analysis is not possible. CFA assumes that each factor is associated with a specified subset of measured variables. This needs structural equation modeling package, SEM is useful for understanding CFA. Since we are not explaining CFA at this moment and structural equation modeling at this moment, so CFA is not being explained. We are only explaining EFA here.

(Refer Slide Time: 07:56)



METHODS TO FACTORING

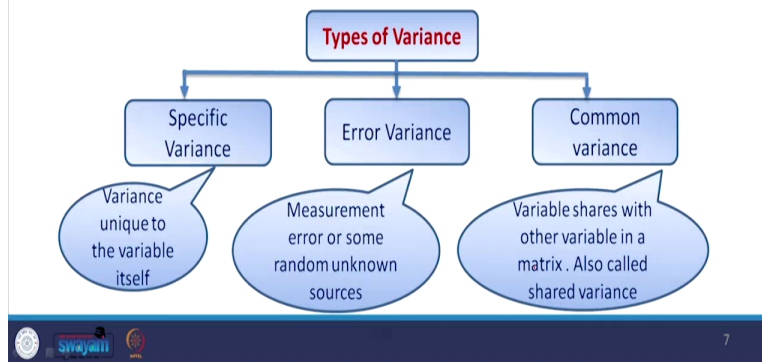
- Broadly, there are two methods of factoring:
 - Principal component analysis (PCA)
 - Common factor analysis (CFA) or simply factor analysis.
- PCA-**
 - Provides a unique solution, so that the original data can be reconstructed from the results.
 - PCA starts extracting the maximum variance and puts them into the first factor.

So, broadly there are two methods of factoring. What are the approaches? One is called principal component analysis. I think you might have heard in different journal article they use principal component analysis to develop an index. From the index they interpret the results in a very qualitative manner. Another is called CFA, common factor analysis or simply factor analysis.

So, PCA, if you are discussing this, it provides a unique solution so that the original data can be restructured from the result, reconstructed from the results. PCA starts extracting the maximum variance and puts them into the first factor that is important.

(Refer Slide Time: 08:41)

- ❑ After that, it removes that variance explained by the first factors and then starts extracting maximum variance for the second factor. This process goes to the last factor.
- ❑ This considers the total variance in the data.



After that, it removes that variance explained by the first factor and then starts extracting the maximum variance of the second factor. That is the approach by which first factor then second factor, we will explain unless you do not understand first factor, second factor. It might be little confusing to you. From your data we will clarify. This process goes to the last factor, accordingly the way first factor and second factor is extracted. This considered the total variance in the data. That is the type of variance of rather specific variance or what kind of variance are there.

Variance if you are discussing it may be a specific variance, error variance or common variance. There are three ways of understanding the variance. So, the specific variance is generally unique to the particular variable itself, whereas error variance is a stochastic variance and this is also called random unknown variance derived from unknown sources. Common variance, the variable shares with other variable in a matrix and also called shared variance.

(Refer Slide Time: 10:05)

Total Variance = Common Variance + Specific Variance + Special Variance

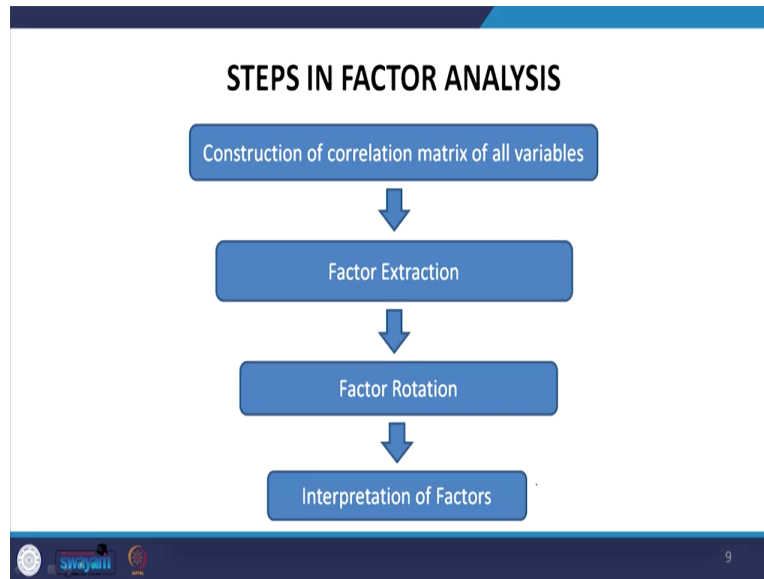
- the most common method used by researchers.
- CFA-**
 - The second most preferred method by researchers.
 - extracts the **common variance** and puts them into factors.
 - This method does not include the unique variance of all variables.
 - This method is used in SEM.

8

So, the total variance we are going to explain is equal to common variance plus specific variance plus special variance if it is there. The most common method used by the researcher that is the PCA. In the CFA, that is basically we are saying the common factor analysis or simply the factor analysis we are discussing. CFA, the second most preferred method by a researcher after PCA. It extracts the common variance and puts them into factors or the factor analysis as we mentioned.

This method does not include the unique variance of all the variances, whereas PCA, it includes first then second accordingly, PCA considers weight as per the extent of variance as their weight of that particular variable. So, PCA includes entire variables and their loadings, whereas in case of CFA, we need not include entire variance, it is the approach of reducing the number of variables. This method does not include the unique variance of all the variables. This method is used in SEM. The common factor analysis largely used in structural equation modeling.

(Refer Slide Time: 11:34)



So, what are the steps involved in factor analysis. The first and foremost important step is used is correlation matrix of all the variables follows with factor extraction, then factor rotation, to validate your factor extraction factor rotation is used, and last but not the least one is interpretation of the factors.

(Refer Slide Time: 11:50)

FACTOR ANALYSIS IN STATA

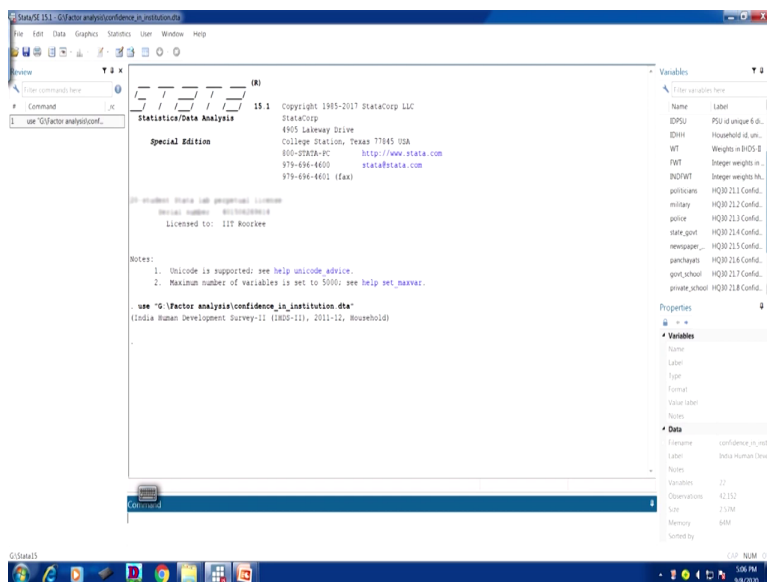
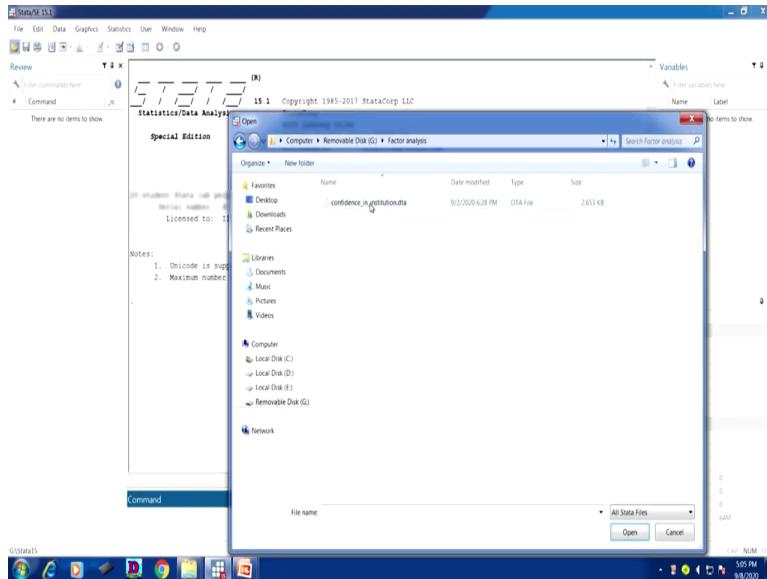
- Launch STATA->
- Open a dataset you want to perform factor analysis on->

Note!

We are using IHDS 2 household dataset. The dataset contains variables for confidence in institution. 12 Different questions were asked to the respondents. We are trying to minimize the number of variables by converting into the factors. For the simplicity we kept only those variable which are relevant to this analysis.

Use "confidence_in_institution", clear

The slide has a white background with a blue header. The title 'FACTOR ANALYSIS IN STATA' is centered. Below the title are two bullet points with checkboxes. A blue rounded rectangular box contains a 'Note!' section with text and a code snippet. At the bottom, there are logos for 'Sri Jayanti' and a page number '10'.



We are going to give you the handholding of the factor analysis with the help of Stata. And we are here to suggest you that you should have a dataset, the dataset must contain metric variables or metric values in that variables. And in order to get this factor analysis through our existing data, we are going to take the help of IHDS, India Human Development Survey Phase 2 household dataset which contains an indicator variables called confidence in the institutions. Different institutions are there in India.

Not the physical institutions like institutions related to a person's defining confidence, like confidence on police, confidence on doctors, confidence on our local leaders, maybe there are

different indicators, maybe confidence on the public distribution system. There are various ways. We have 12 different questions in the IHDS dataset. So, first step for the factor analysis to do is to start or launch the Stata platform and open the dataset we want to operate. So here we are going to operate the dataset like this. So, we have opened the Stata.

We are going to open the Stata, the dataset, which is there. So in IHDS, household data. I think we have already filtered. We will find out. It is G file. It is in our pen drive. Factor analysis, it is within the factor analysis. So, this is the dataset we are opening. Why we are opening this instead of the original dataset, because the dataset contains so many variables those are not in metric.

The dataset which we have just opened for your clarity is that this has metric information and we are going to show you in the PPT of the dataset. The variables that is of interest for us is starting from here confidence on the politicians, on the military, police, state government, similarly till courts and banks. So, let me proceed further.

For simplicity, we kept those variables which are relevant to the analysis as I pointed out. And why we opened only 12 questions because then only we can able to operate without consuming more time and those questions are relevant for factor analysis. All variables just without understanding the features of it, we cannot run factor analysis. It is not advisable to run factor analysis.

(Refer Slide Time: 15:26)

Confidence in Institutions

I am going to name some institutions in the country. As far as the people running these institutions are concerned, would you say you have

A great deal of confidence=1
Only some confidence=2
Hardly any confidence at all=3

21.1 Politicians – to fulfil promises	□	21
21.2 Military – to defend the country	□	22
21.3 Police – to enforce the law	□	23
21.4 State government – to look after the people (e.g. UP, AP, etc.)	□	24
21.5 Newspapers/News media – to print/broadcast the truth	□	25
21.6 Village Panchayats / Nagarpalika / Nagar Panchayat – to implement public projects	□	26
21.7 Govt. Schools – to provide good education	□	27
21.8 Private Schools – to provide good education	□	28
21.9 Govt. Hospitals and doctors – to provide good treatment	□	29
21.10 Private Hospitals and doctors – to provide good treatment	□	30
21.11 Courts – to deliver justice	□	31
21.12 Banks – to keep money safe	□	32

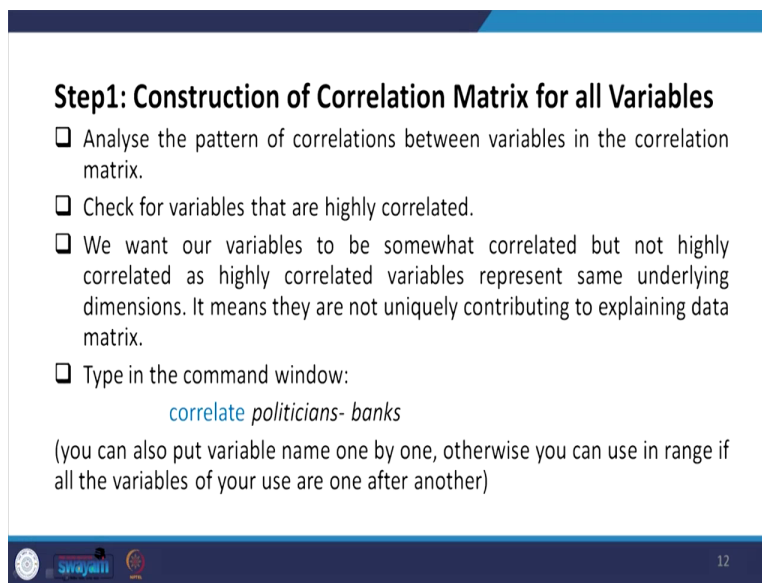
12 questions were asked related to the confidence in institution category.

Swayamii IIT Bombay 11

What are those variables? In the questionnaire, in the schedule, if you take from the original schedule which we have already shown to you, there are 12 questions asked related to the confidence in the institutions and under the head of institutions category. So, politicians they asked a questions to the respondent that what is your take on politician regarding fulfilling promises.

So, the respondent with the help of a great deal of confidence with 1 as the code then with some degree of confidence or some confidence is 2 and hardly any confidence is 3. Similarly, military, police, for all categories also they ask questions to media as well village panchayat, hospitals, private hospitals, courts and banks. So, all the records are noted in IHDS data for all those 12 indicators.



(Refer Slide Time: 16:30)

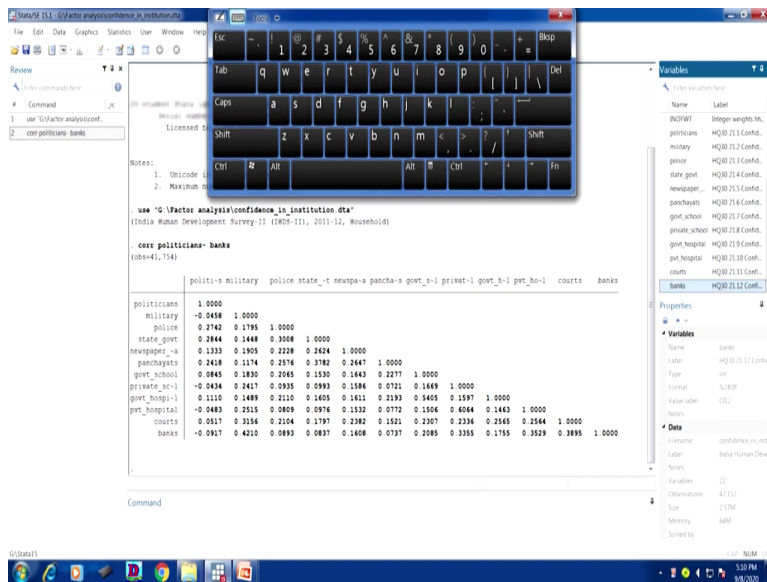


Step1: Construction of Correlation Matrix for all Variables

- Analyse the pattern of correlations between variables in the correlation matrix.
- Check for variables that are highly correlated.
- We want our variables to be somewhat correlated but not highly correlated as highly correlated variables represent same underlying dimensions. It means they are not uniquely contributing to explaining data matrix.
- Type in the command window:
`correlate politicians- banks`

(you can also put variable name one by one, otherwise you can use in range if all the variables of your use are one after another)

 Sriyanti  12



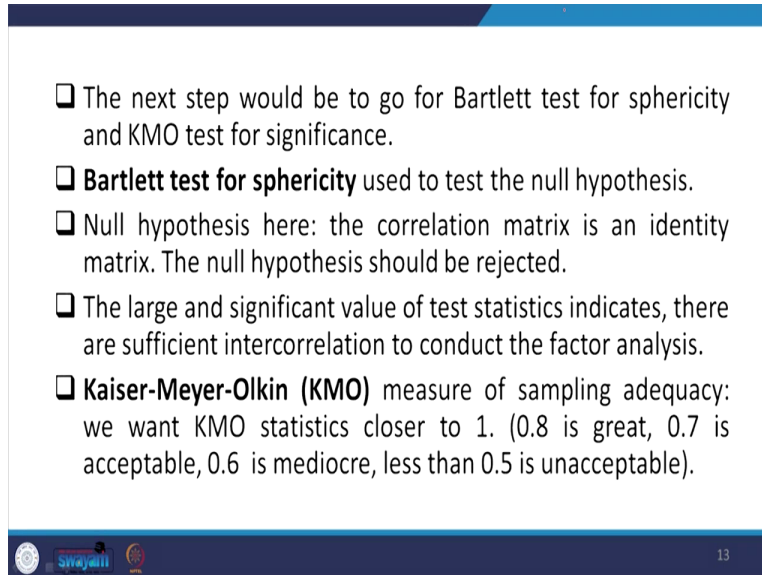
So, the construction of correlation matrix first, then the first step after opening the Stata and data, the step here is to have a correlation matrix, first step is correlation as we already suggested. So, analyze the pattern of correlation between the variables that is among the 12 variables we are suggesting and develop a correlation matrix. And check for variables that are highly correlated. We want our variables to be somewhat correlated, but not highly correlated as highly correlated variables represents same underlying dimensions. It means that they are not uniquely contributing to explaining the data matrix.

So, as I told you, if it is perfectly correlated or perfect correlation is there then that is also not relevant for running any factor analysis. Highly correlated is also problematic. If you have some doubts, if doubt on some of the variables and you are confused which variable to be taken then factor analysis is most useful. We have already opened the dataset, in the command simply you do a correlation, you please derive a correlation matrix. Simply type and then starting from here politician, you simply put a hyphen not underscore it should be this, then till the last variable if you take enter you will get the results.

Here in short we have given entire variable starting from politician till banks instead. If you want only specific variables you need not give this command, you have to enter each variable individually. Since we know that entire, all those variables are in metric scale and we are going to use these so you have given the hyphen, politician hyphen banks. Accordingly we derived the correlation matrix and we are going to tell you how we can interpret and how we can proceed

further. I think you can also put variable name one by one as I just said instead of continuous series the way we did.

(Refer Slide Time: 19:12)



- ❑ The next step would be to go for Bartlett test for sphericity and KMO test for significance.
- ❑ **Bartlett test for sphericity** used to test the null hypothesis.
- ❑ Null hypothesis here: the correlation matrix is an identity matrix. The null hypothesis should be rejected.
- ❑ The large and significant value of test statistics indicates, there are sufficient intercorrelation to conduct the factor analysis.
- ❑ **Kaiser-Meyer-Olkin (KMO)** measure of sampling adequacy: we want KMO statistics closer to 1. (0.8 is great, 0.7 is acceptable, 0.6 is mediocre, less than 0.5 is unacceptable).

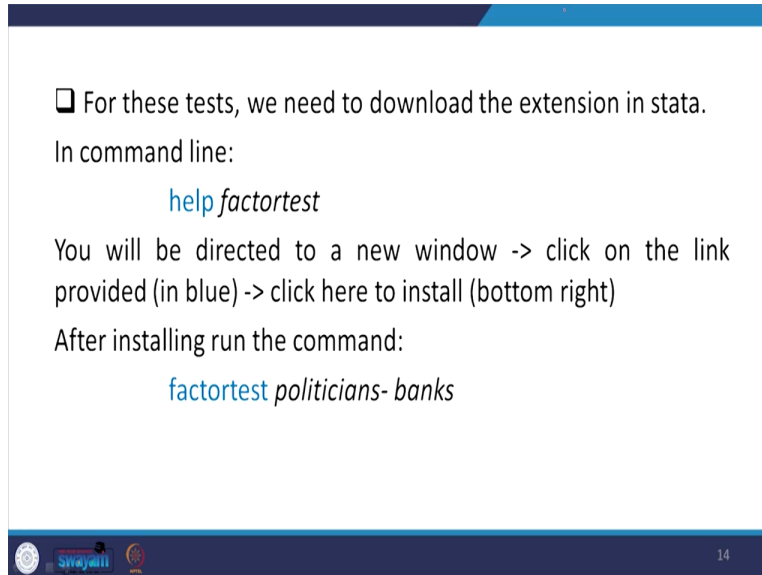
The next step would be to go for a Bartlett test for sphericity and KMO test for significance. The extent of variance and what is the extent of variance is acceptable and how this is useful and whether the significance justify our sample's size correctly or not, sampling adequacy or not is going to be very useful. Bartlett test for sphericity used to test the null hypothesis. The null hypothesis here that is the correlation matrix should be an identity matrix. The correlation matrix which we defined should be in identity matrix. But as per our requirement, in this case, the null hypothesis should be rejected.

That means our result Bartlett test for sphericity should be significant then only we can reject our null hypothesis that the correlation matrix is going to be identity one. The large and significant value of test statistics indicates that there is sufficient inter-correlation to conduct the factor analysis. This will guarantee that factor analysis is justified and you can go for conducting factor analysis.

Similarly, KMO matrix, Kaiser Meyer Olkin measure guarantees regarding sampling adequacy, where there is sample size you have considered 1: 20 or even 1:5, if you are going to take it, we have already discussed that, but there is a particular test if that boils down closer to one that is

really great, if it is 0.8 and above it is very good or it is great. If it is 0.7 and above till 0.8 it is acceptable limit, 0.6 is at the media core, less than 0.5 is certainly unacceptable. This is also an important test before running or operating the factor analysis.

(Refer Slide Time: 21:33)



For these tests, we need to download the extension in stata.

In command line:

`help factortest`

You will be directed to a new window -> click on the link provided (in blue) -> click here to install (bottom right)

After installing run the command:

`factortest politicians- banks`

14

For these tests we need to download the extension in Stata. Why? This is important aspect to be noted for all of you. You might be confused that all the Stata window may not take these commands. You need to download some of the commands, some patching must have been done with the original Stata format then only it will read. It may not read these commands. Factor test we are going to do; it does not come by default with the software. You have to download first.

(Refer Slide Time: 22:15)

Viewer: search factortest

search factortest

help for factortest not found

search for factortest (manual: [R] search)

Search of official help files, FAQs, Examples, Sts, and STBs

Web resources from Stata and other users

(contacting <http://www.stata.com>)

1 package found (Stata Journal and STB listed first)

factortest from <http://fmwww.bc.edu/RePEc/boocode/f>

"FACTORTEST": module to perform tests for appropriateness of factor analysis // factortest performs Bartlett's test for sphericity and calculates // the Kaiser-Meyer-Olkin Measure of Sampling Adequacy. Both tests // should be used prior to a factor or a principal component //

(click here to return to the previous screen)

end of search

```
set_maxvar,
di
rcha-s govt_3-1 privat-1 govt_3-1 prt_3o-1 courts banks

0.0000
0.2217 1.0000
0.0721 0.1669 1.0000
0.2193 0.5405 0.1597 1.0000
0.0772 0.1504 0.4044 0.1443 1.0000
0.1521 0.2307 0.2334 0.2545 0.2544 1.0000
0.0737 0.2085 0.3355 0.1755 0.3529 0.3895 1.0000
```

Variables

Name	Label
INDWRT	Integer weights, M...
politcons	HQ0 21.1 Confid...
military	HQ0 21.2 Confid...
police	HQ0 21.3 Confid...
state govt	HQ0 21.4 Confid...
newspaper...	HQ0 21.5 Confid...
parachyats	HQ0 21.6 Confid...
govt_school	HQ0 21.7 Confid...
private_school	HQ0 21.8 Confid...
govt_hospital	HQ0 21.9 Confid...
courts	HQ0 21.10 Confid...
banks	HQ0 21.12 Confid...

Properties

Name: banks

Label: HQ0 21.12 Confid...

Type: int

Format: %12.0P

Value label: C02

Notes:

Data

Filename: confidence_in_instu...

Label: India Human Develo...

Notes:

Variables: 22

Observations: 42,812

Size: 257M

Memory: 64M

Sorted by:

Viewer: net describe factortest, from(<http://fmwww.bc.edu/RePEc/boocode/f>)

net describe factortest, from(<http://fmwww.bc.edu/RePEc/boocode/f>)

package factortest from <http://fmwww.bc.edu/RePEc/boocode/f>

TITLE

"FACTORTEST": module to perform tests for appropriateness of factor analysis

DESCRIPTION/AUTHOR(S)

factortest performs Bartlett's test for sphericity and calculates the Kaiser-Meyer-Olkin Measure of Sampling Adequacy. Both tests should be used prior to a factor or a principal component analysis.

KEYWORDS

factortest: factor analysis
factortest: principal components
factortest: sphericity
factortest: sampling adequacy

Requires: Stata version 7.0

Author: Joao Pedro Azevedo, University of Newcastle-upon-Tyne, UK
Support: email: j.p.azevedo@ncl.ac.uk

Distribution-Date: 20060427

INSTALLATION FILES (click here to install)

factortest ado
factortest help

(click here to return to the previous screen)

```
set_maxvar,
di
rcha-s govt_3-1 privat-1 govt_3-1 prt_3o-1 courts banks

0.0000
0.2217 1.0000
0.0721 0.1669 1.0000
0.2193 0.5405 0.1597 1.0000
0.0772 0.1504 0.4044 0.1443 1.0000
0.1521 0.2307 0.2334 0.2545 0.2544 1.0000
0.0737 0.2085 0.3355 0.1755 0.3529 0.3895 1.0000
```

Variables

Name	Label
INDWRT	Integer weights, M...
politcons	HQ0 21.1 Confid...
military	HQ0 21.2 Confid...
police	HQ0 21.3 Confid...
state govt	HQ0 21.4 Confid...
newspaper...	HQ0 21.5 Confid...
parachyats	HQ0 21.6 Confid...
govt_school	HQ0 21.7 Confid...
private_school	HQ0 21.8 Confid...
govt_hospital	HQ0 21.9 Confid...
courts	HQ0 21.10 Confid...
banks	HQ0 21.12 Confid...

Properties

Name: banks

Label: HQ0 21.12 Confid...

Type: int

Format: %12.0P

Value label: C02

Notes:

Data

Filename: confidence_in_instu...

Label: India Human Develo...

Notes:

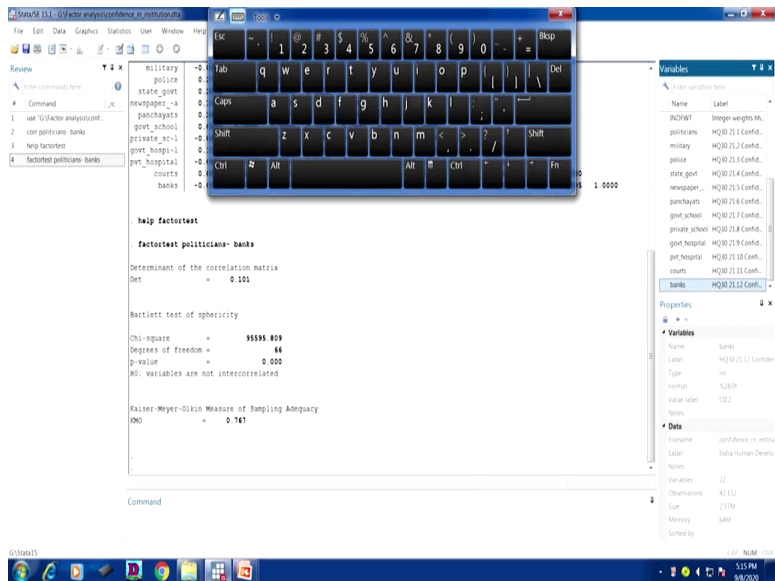
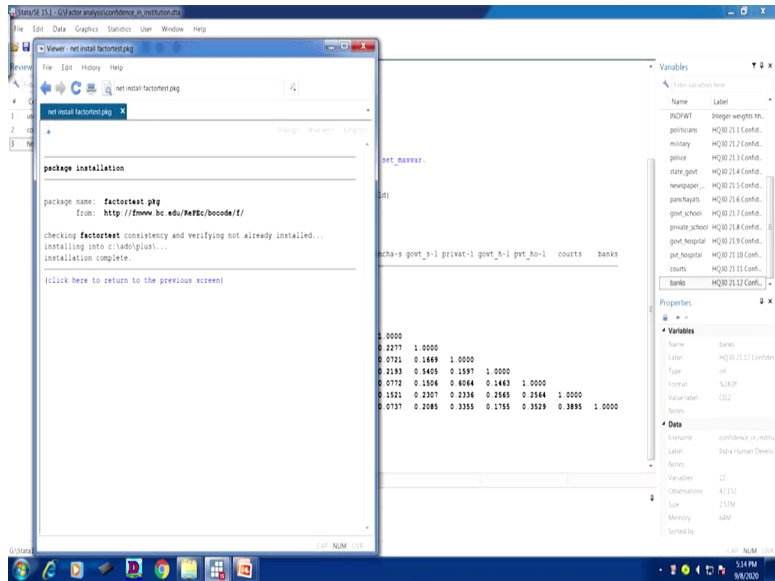
Variables: 22

Observations: 42,812

Size: 257M

Memory: 64M

Sorted by:



So, help factor test if you just do that then it will redirect you to a window with some commands, if you just click the suggestions it will suggest you to install. Here we are supposed to install it. Click here for installation. You will get the information like installation is complete. So, that means you are ready with this command. Factor test command is going to be operational.

Once that is done, you can run the factor test with the same variable since we have already defined a correlation matrix. So, you can go for factor test. So here it is. From the politician to banks, politicians and banks if you just enter, you will get the result. What results you have derived. Very interesting to note that you have derived the significance level which we want

related to sphericity. Bartlett test, this is the command for Bartlett test as well as KMO measure of sampling adequacy.

These are two standard tools before factor analysis. Bartlett test guarantees with the fact that your null hypothesis is rejected. Since it is significant at even 0.01 level, even less than that also, it is 0.000 it is perfectly significant that does mean you are rejecting the null hypothesis. Your variables in the matrix form are not in identity format. That means one is not overlapping with each other. There are some correlations, some degree of correlation. They are not perfectly correlated.

Second one is KMO matrix, it defines with the sampling adequacy, the standard range we have defined. If it is less than 0.5 straightaway, you drop the factor analysis because your sample is not adequate. Here we know that we are taking from a large sample size data it is 42,152, but for you it might be very less and the number we derived here it is 0.7 and above, so that means it is perfectly sufficient to run factor analysis. And for your case, you might be dealing with an individual survey data, please be careful about 0.5 here.

(Refer Slide Time: 25:08)

The image shows a screenshot of SPSS output for a factor analysis. The output is as follows:

```
. factorstest politicians- banks  
  
Determinant of the correlation matrix  
Det          =    0.101  
  
Bartlett test of sphericity  
  
Chi-square    =    95595.809  
Degrees of freedom =    66  
p-value       =    0.000  
H0: variables are not intercorrelated  
  
Kaiser-Meyer-Olkin Measure of Sampling Adequacy  
KMO          =    0.767
```

Two callout boxes provide additional context:

- The top callout box states: "The p-value is highly statistically significant. It means there are sufficient correlation to conduct the factor analysis".
- The bottom callout box states: "KMO measure is 0.767, which is in the acceptable category".

The slide footer includes the Swayamii logo and the number 15.

So, let me proceed. So, that we have done. And this is what we have explained. the p-value, I have already mentioned. This is also highlighted in the PPT. Variables are not inter-correlated. It is clear in the HO, suggests that they are not inter-correlated, but it is rejecting that means there

are some degree of correlations. The p-value is highly statistically significant. It means that there is sufficient correlation to conduct the factor analysis. So, the minimum level of correlation is guaranteed by the Bartlett test. KMO measure is of 0.767, which is the acceptable category, as we already mentioned.

(Refer Slide Time: 25:50)

Step 2: Factor Extraction

- ❑ Here, we have to decide which method to be used for factoring.
- ❑ The principal component analysis is the most commonly used extraction method. Principal component factors can also be used for this purpose.
- ❑ Then the decision can be made about the number of factors underlying a set of measured variables based on eigen value and scree plot.

factor politicians- banks, pcf

16

IR test: independent vs. saturated: chi2(166) = 9.6e+04 ProbChi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
politicians	0.2124	0.6166	0.1956	0.5365
military	0.3450	-0.2732	0.3927	0.6189
police	0.4750	0.4036	0.2265	0.5942
state_govt	0.4759	0.4592	0.3208	0.4598
newspaper_a	0.4947	0.1912	0.2506	0.6539
parthayats	0.4488	0.4704	0.1285	0.5403
govt_school	0.5881	0.2326	-0.4443	0.2584
private_sc-1	0.5416	-0.4704	0.2041	0.4429
govt_hosp-1	0.5499	0.1545	-0.4445	0.2584
pvt_hospital	0.5442	-0.4883	0.2272	0.4148
counts	0.4039	-0.1339	-0.0253	0.6294
banks	0.5692	-0.4504	0.0148	0.4727

The step two followed in the factor analysis is that we have to decide which method we use for factoring. The principal component analysis is the most commonly used extraction method. Directly it boils down to some indexes with the help of the principal component. The component

carries with the highest amount of weight and the weight is used for defining the principal component, defining the index score.

The principal component factor can be used for this purpose that is one method, then the decision can be made about the number of factors underlining a set of measured variables based on Eigen values or scree plot. If you are going by the factor analysis, we have to stick to the Eigen values and scree plot, because in factor analysis, we are supposed to drop some of the variables. They are not relevant for the analysis. If the Eigen value is less than one, that is a standard thumb rule we are going to discuss now.

So, for the factor analysis with the same variables we have already defined that our pretesting of the factor analysis through KMO and Bartlett test confirms to the fact that factor analysis is doable and can be run. So, for us we are going to explain you with the help of Stata. With the same command we are going to change the factor, here it will be only factor and PCF, there is two additional things to be added, factor and here it is comma PCF. This will give you the result of factor analysis.

This has given a random allocation of variables within different factors. There is no systematic ordering of the factor loadings. We are going to discuss the factor loadings. So, there are Eigen values, there are proportion or the cumulative values. The Eigen values is going to be very useful. We are going to discuss with the help of PPT.

(Refer Slide Time: 28:06)

❑ Else you can also use, *pca politicians- banks*

❑ In case of pcf

```

factor politicians- banks, pcf
(observe=41,754)

Factor Analysis/correlation          Number of obs = 41,754
Method: principal-component factors  Retained factors = 3
Rotation: (unrotated)              Number of params = 33

```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	3.16158	1.35272	0.2635	0.2635
Factor2	1.80886	0.64466	0.1507	0.4142
Factor3	1.16420	0.17588	0.0970	0.5112
Factor4	0.98832	0.17082	0.0824	0.5936
Factor5	0.81750	0.07272	0.0681	0.6617
Factor6	0.74478	0.04443	0.0621	0.7238
Factor7	0.70034	0.04193	0.0584	0.7821
Factor8	0.63841	0.03632	0.0532	0.8353
Factor9	0.60209	0.07280	0.0502	0.8855
Factor10	0.52928	0.07661	0.0441	0.9296
Factor11	0.45267	0.06071	0.0377	0.9673
Factor12	0.39197	.	0.0327	1.0000

LR test: independent vs. saturated: $\chi^2(66) = 9.6e+04$ Prob $\chi^2 = 0.0000$

We are using pcf just to identify the underlying dimensions, so here common variance is in picture.

Proportion tells us how much variation is explained by each factor

We are looking for an eigen value above 1.0

So, this is what we derived in the Stata result. So, you can go through PCF for factor analysis or you can apply the PCA with the same variables. PCA is going to define the values of individual respondents with a certain index values after getting the factor loadings and each of the factor loadings are considered to be certain extent of weight in the PCA.

In case of our PCF, that is factor analysis result, which we have shown you here. This is important to note that Eigen values, then proportions, how much proportion explained. I am going to discuss you in a short while. So, we are using PCF just to identify the underlining dimensions. So, here the common variance is in picture. The proportion tells us how much variation is explained by each factor is given by the proportion. Each factor, like usually the starting factors from the top at least with the hierarchical order. Highest the proportion of their loading that gives the highest factor value.

Basically, the factor one has highest proportion of explanation. The factor two that represents 26 percent of the total explanation, then 15 percent, then 9 percent. Others are our miniscule percentage of proportion is explained. We are looking for an Eigen values with the value over 1. So, in this case it is highlighted these 3. These 3 are having more than 1 Eigen values, already explained.

(Refer Slide Time: 30:19)

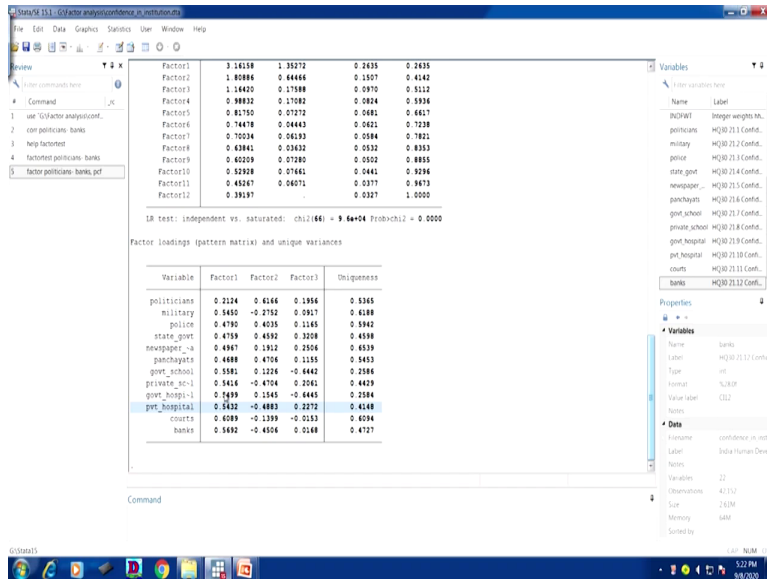
□ Our concern here is how many factors to be extracted:

□ Based on **eigen value** exceed 1.0, 3 factors are to be extracted.

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
politicians	0.2124	0.6166	0.1956	0.5365
military	0.5450	-0.2752	0.0917	0.6188
police	0.4790	0.4035	0.1165	0.5942
state_govt	0.4759	0.4592	0.3208	0.4598
newspaper_~a	0.4967	0.1912	0.2506	0.6539
panchayats	0.4688	0.4706	0.1155	0.5453
govt_school	0.5581	0.1226	-0.6442	0.2586
private_sc-1	0.5416	-0.4704	0.2061	0.4429
govt_hospi-1	0.5499	0.1545	-0.6445	0.2584
pvt_hospital	0.5432	-0.4883	0.2272	0.4148
courts	0.6089	-0.1399	-0.0153	0.6094
banks	0.5692	-0.4506	0.0168	0.4727

Here, we can see that stata has already retained 3 factors based on eigen values.



So, our concern here is how many factors to be extracted? So, the three factors, based on the Eigen values, 3 factors are extracted. I think in the factor analysis result, with the help of Stata already identify that till factor 3 we have the loadings identified by the factor analysis. These three are being explained and each factor with their individual loading is mentioned.

So, which particular variable has highest loading can be defined. Instead of if there are so many factors, you might be confused. We should put them in an order. So, let us proceed further to put them in order. So, here we can see that Stata has already retained three factors based on the Eigen values.

(Refer Slide Time: 31:23)

- Uniqueness is percentage of variance for the variable that is not explained by the common factors.
- Value of uniqueness more than 0.6 are generally considered high, it means variable is not well explained by the factors.
- **1- uniqueness = communality**
- Another method for looking at the factors to be extracted is screeplot. Scree plot is a line plot eigen values of the factors and principal components in an analysis.

screeplot

screeplot, yline(1)

(option yline(1), puts a line at one horizontal to x-axis)



19

- Our concern here is how many factors to be extracted:
 - Based on **eigen value** exceed 1.0, 3 factors are to be extracted.

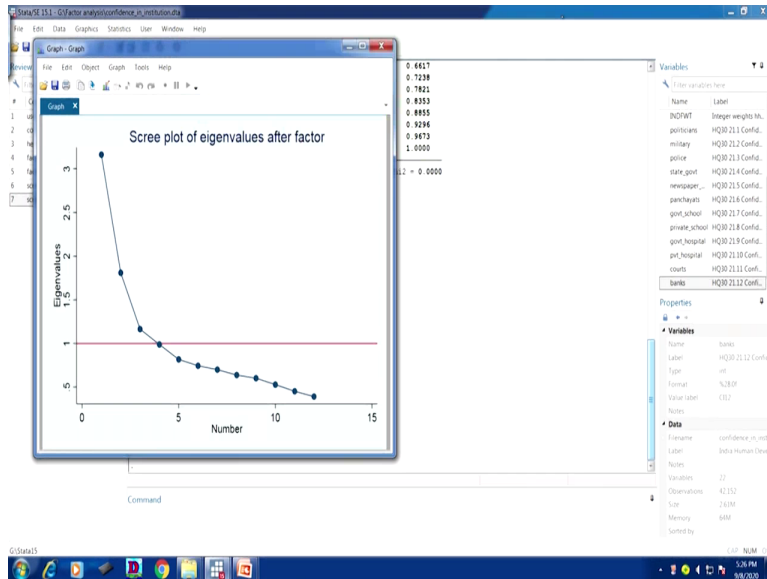
Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
politicians	0.2124	0.6166	0.1956	0.5365
military	0.5450	-0.2752	0.0917	0.6188
police	0.4790	0.4035	0.1165	0.5942
state_govt	0.4759	0.4592	0.3208	0.4598
newspaper_a	0.4967	0.1912	0.2506	0.6539
panchayats	0.4688	0.4706	0.1155	0.5453
govt_school	0.5581	0.1226	-0.6442	0.2586
private_sc-l	0.5416	-0.4704	0.2061	0.4429
govt_hospi-l	0.5499	0.1545	-0.6445	0.2584
pvt_hospital	0.5432	-0.4883	0.2272	0.4148
courts	0.6089	-0.1399	-0.0153	0.6094
banks	0.5692	-0.4506	0.0168	0.4727

Here, we can see that stata has already retained 3 factors based on eigen values.



18



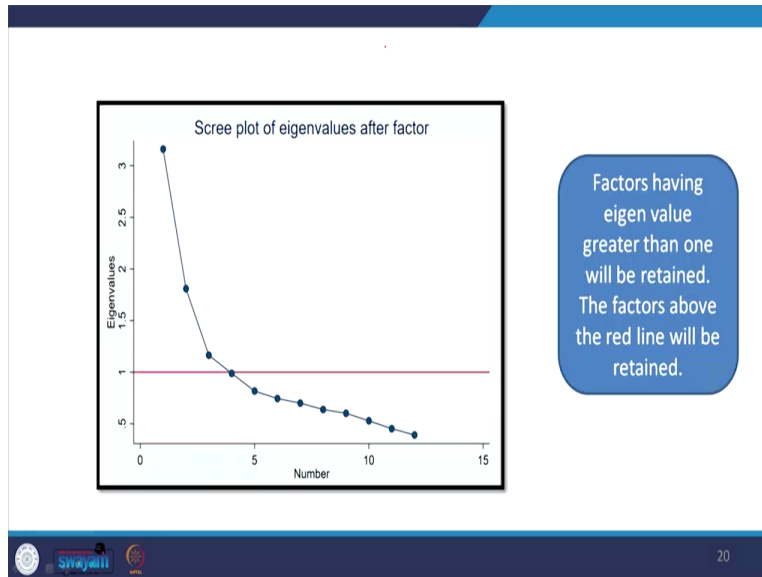
So, what do you mean by uniqueness in the result. The uniqueness is percentage of variance of that particular variable that is not explained by the common factors. The common factors which we have taken here, the three factors we have taken here. These are 53 percent of that particular variables are not explained by the three common factors. That means one minus the common factor explains how much of the percentage they have explained. So, that is nothing but called the commonality, 1 minus uniqueness. 53 percent is not explained that means rest is explained by the 3 factors. So, that is basically called commonality.

Another standard rule we must be very careful about it, the value of uniqueness more than 0.6 are generally considered high. It means variable is not well explained by the factors. Anything that is more than 0.6 that is not explained by the factors. you can see this is one, this is another, this is here third. They are not clearly explained by these three factors which have been identified by the Eigen values or even through the scree plot also we are going to discuss it.

Another method for looking at the factors to be extracted is scree plot. The scree plot is line plot of Eigen values of the factors and the principal component of that analysis. So, after this result in the Stata, you simply write down scree plot. Scree plot so if you do it, the result in another window with the scree plot will be derived. Usually it, here is the scree plot. question here how can I identify the Eigen level which according to that we define the important factors, three important factors that can be benchmarked with the help of our command. Here is the command. You have to specify.

In the y-axis we have taken Eigen values, isn't it? So, your scree plot with y-line that is at one level, one is your Eigen value level, if you can highlight that will give you the result. So, this is y-line basically, so scree y-line then within bracket one and this gives you a benchmark level of your understanding regarding the important factors. This is derived. I think you guys can able to do this correctly without any problem.

(Refer Slide Time: 34:59)



That we have already discussed this is how it looks like and we also given you in the PPT. We operated through the Stata. And factors having Eigen value greater than one will be retained. The factors above the red line with more than one Eigen value is here, more than one is retained for the factor analysis.

(Refer Slide Time: 35:20)

Step 3: Factor Rotation

- Once you obtain minimal number of factors, you have to interpret them.
- Factor rotation makes interpretation more meaningful and easier.
- If factors shows cross-loading (more or less similar factor loading in more than one variable), in this case orthogonal rotation will be helpful.
- If nothing changes with orthogonal rotation then there might be correlation between factors, in this scenario oblique rotation works.



The third step that is important also for further validation of your result that is three factors though you have defined how could you able to validate in different contexts this is one form of validation called factor rotation. Once you obtain minimum number of factors, you have to interpret them. Factor rotation makes interpretation more meaningful and easier. If factors show cross-loading, more or less similar factor loading in more than one variable, in this case orthogonal rotation will be very helpful.

That means if factor cross-loading and more or less similar factor loading in more than one variable, similar factor loading as I already said there are different factor loadings, if similar is there, then orthogonal rotation will clarify which factors are going to be better. If nothing changes with orthogonal rotation, then there might be correlation between factors. If still the factor loadings remain same, then factor rotation is important. But if it is not changing after factor rotation in this scenario oblique rotation works. So, oblique rotation and the orthogonal rotation varimax which we mentioned some clarification has already been given in this slide.

(Refer Slide Time: 36:54)

- Simply commanding `rotate` to stata, will gives result of orthogonal rotation.
- But one of the problem with rotate command is that it does not show factor loading in order.
- To sort the factor loading from high to low, install a command called `sortl`.

`help sortl`

Another help window will open with three links, click on the first link and install the command.

Simply command stata `sortl`

Command window:

```
1 use "G:\factor analysis\confidence_in_institutions"
2 factor politicians banks
3 help factorlist
4 factorlist politicians banks
5 factor politicians banks pdf
6 sortl
7 sortl pdf
8 rotate
```

Rotated factor loadings (pattern matrix) and unique variances:

Variable	Factor1	Factor2	Factor3	Uniqueness
politicians	-0.2075	0.6470	0.0429	0.5365
military	0.5823	0.1202	0.1664	0.6188
police	0.0921	0.5922	0.2159	0.5942
state_govt	0.1247	0.7342	0.0518	0.4508
newspaper_m	0.2864	0.5075	0.0801	0.4539
panchayats	0.0401	0.6355	0.2218	0.5453
govt_school	0.1359	0.0819	0.8510	0.2586
private_hosp-1	0.7449	0.0382	0.0386	0.4429
govt_hospit-1	0.0759	0.1059	0.8519	0.2584
priv_hospital	0.7643	0.0255	0.0190	0.4149
courts	0.5031	0.2027	0.3104	0.4094
banks	0.4931	-0.0294	0.2147	0.4727

Factor rotation matrix:

	Factor1	Factor2	Factor3
Factor1	0.6806	0.5118	0.5243
Factor2	-0.6671	0.7287	0.1546
Factor3	0.3029	0.4549	-0.8374

Variables window:

Name	Label
INDFWT	Integer weights M...
politicians	HQ09 21.1 Confid...
military	HQ09 21.2 Confid...
police	HQ09 21.3 Confid...
state_govt	HQ09 21.4 Confid...
newspaper_m	HQ09 21.5 Confid...
panchayats	HQ09 21.6 Confid...
govt_school	HQ09 21.7 Confid...
private_school	HQ09 21.8 Confid...
govt_hospital	HQ09 21.9 Confid...
priv_hospital	HQ09 21.10 Confid...
courts	HQ09 21.11 Confid...
banks	HQ09 21.12 Confid...

Stata/11.1 - C:\factor_analysis\conference_p_methadone

Viewer: search sort

```

search sort
-----
help for sort1 not found
search for sort1
-----
Search of official help files, FAQs, Examples, Sts, and STBs
Web resources from Stata and other users
(contacting http://www.stata.com)
3 packages found (Stata Journal and STB listed first)
-----
sort1 from http://fmwww.bc.edu/R/fEz/boccode/s
"sort1": module to sort factor loadings or rotated matrix from PCA or
factor / To make an interpretation of a factor solution easier, sort1 /
sorts the rotated loadings (pattern matrix) or rotated components / stored
by rotate into the matrix e(r_1). It also sorts the matrix / e(r_1) of the
sort1lib by http://fmwww.bc.edu/R/fEz/boccode/s
"sort1lib": module to sort by random or by ancillary numlist /
sort1lib will sort the items in list at random or by the / values of an
ancillary numlist. sort1lib does the same / however, it is a little
bit faster if the number of items in list / is small (and much slower if
listlib from http://fmwww.bc.edu/R/fEz/boccode/s
"listlib": modules to manipulate lists of words / These functions
manipulate lists of words. For details, see the / help file. / Author:
Nicholas J. Cox, University of Durham / Support: email
N.J.Cox@durham.ac.uk / Distribution-Date: 20090523
(click here to return to the previous screen)
end of search!

```

Variables

Name	Label
INDWAT	Integer weights M...
politcans	HQ0 21.1 Conf...
military	HQ0 21.2 Conf...
police	HQ0 21.3 Conf...
state_govt	HQ0 21.4 Conf...
newspaper...	HQ0 21.5 Conf...
parachyats	HQ0 21.6 Conf...
govt_school	HQ0 21.7 Conf...
private_school	HQ0 21.8 Conf...
govt_hospital	HQ0 21.9 Conf...
priv_hospital	HQ0 21.10 Conf...
courts	HQ0 21.11 Conf...
banks	HQ0 21.12 Conf...

Properties

Variables

Name: banks

Label: HQ0 21.12 Conf...

Type: int

Format: %3.0P

Value label: C02

Notes:

Data

Filename: conference_p_metha...

Label: India Human Develo...

Notes:

Variables: 22

Observations: 43,312

Size: 283M

Memory: 64M

Sorted by:

Stata/11.1 - C:\factor_analysis\conference_p_methadone

Viewer: net describe sort, from http://fmwww.bc.edu/R/fEz/boccode/s

```

net describe sort, from http://fmwww.bc.edu/R/fEz/boccode/s
-----
package sort1 from http://fmwww.bc.edu/R/fEz/boccode/s
-----
TITLE
"sort1": module to sort factor loadings or rotated matrix from PCA or
factor
-----
DESCRIPTION/AUTHOR(S)
To make an interpretation of a factor solution easier, sort1
sorts the rotated loadings (pattern matrix) or rotated components
stored by rotate into the matrix e(r_1). It also sorts the matrix
e(r_1) of the unique or unexplained variances created by factor
or by pca into the same order.
KW: factor loadings
KW: factor
KW: pca
KW: rotate
Requires: Stata version 9
Distribution-Date: 20091120
Author: Dirk Enzmann, University of Hamburg
Support: email dirk.enzmann@uni-hamburg.de
-----
INSTALLATION FILES
sort1.ado
sort1.hlp
(click here to install)
(click here to return to the previous screen)

```

Variables

Name	Label
INDWAT	Integer weights M...
politcans	HQ0 21.1 Conf...
military	HQ0 21.2 Conf...
police	HQ0 21.3 Conf...
state_govt	HQ0 21.4 Conf...
newspaper...	HQ0 21.5 Conf...
parachyats	HQ0 21.6 Conf...
govt_school	HQ0 21.7 Conf...
private_school	HQ0 21.8 Conf...
govt_hospital	HQ0 21.9 Conf...
priv_hospital	HQ0 21.10 Conf...
courts	HQ0 21.11 Conf...
banks	HQ0 21.12 Conf...

Properties

Variables

Name: banks

Label: HQ0 21.12 Conf...

Type: int

Format: %3.0P

Value label: C02

Notes:

Data

Filename: conference_p_metha...

Label: India Human Develo...

Notes:

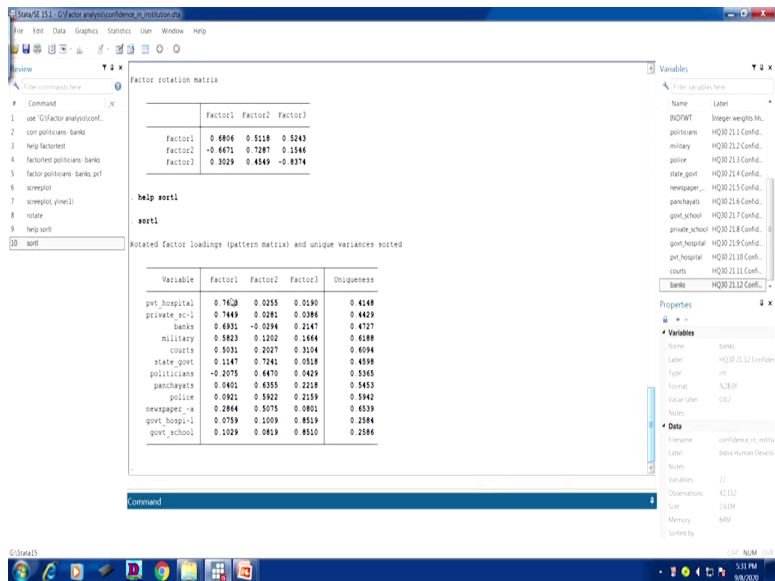
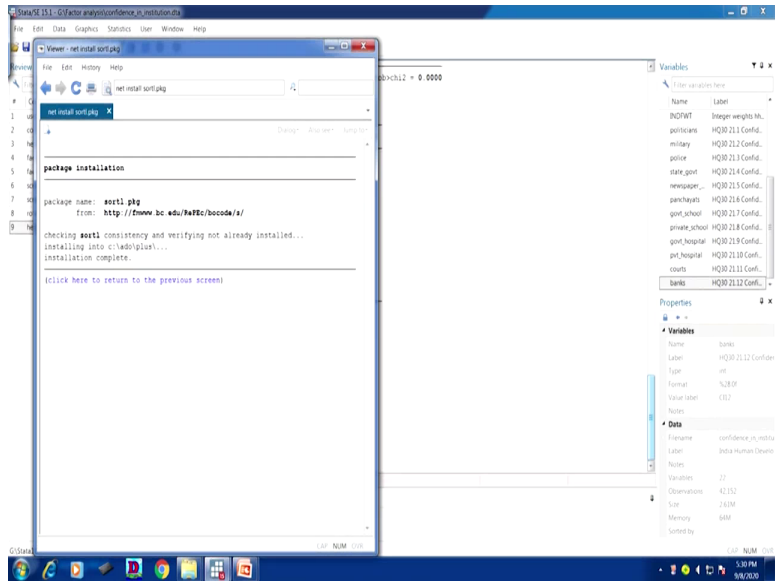
Variables: 22

Observations: 43,312

Size: 283M

Memory: 64M

Sorted by:



Simply commanding rotate to Stata will give results of orthogonal rotation, like here simply after that you write down rotate you will get the result. Rotate, it gives you the rotated matrix of the factor loading. So, it clearly retains with rotated factor loadings. But one of the problems of the rotate command is that it does not show factor loading in order. So, the factor loadings are not in order. After rotation we need to sort those factor loadings for better interpretation and for sorting up the variables for the factors definition.

To sort the factors loading from high to low, we have to install another command that is called sortl. Generally, it is not carrying with the standard database of Stata, it has to be passed again.

So, how do you do it? You have to take the help of Stata with the same help command then sortl. So, it will redirect to the window. You go by the first one and try to install it. Click here to install. It gives you the link for installing. installation is complete. That means now you can able to operate this link.

So, another help window, this is what we have already said. You can go through these lines. In the Stata command simply if you type the same command sortl, it will give you the result. Since it is there, simply delete this. The result is in sort. This is very clearly understood that your rotated factor loadings are in hierarchical order. So, this gives in chronological order with highest to the lowest factor loadings in each of the cases. Accordingly, we can define and take some decisions for our interpretation.

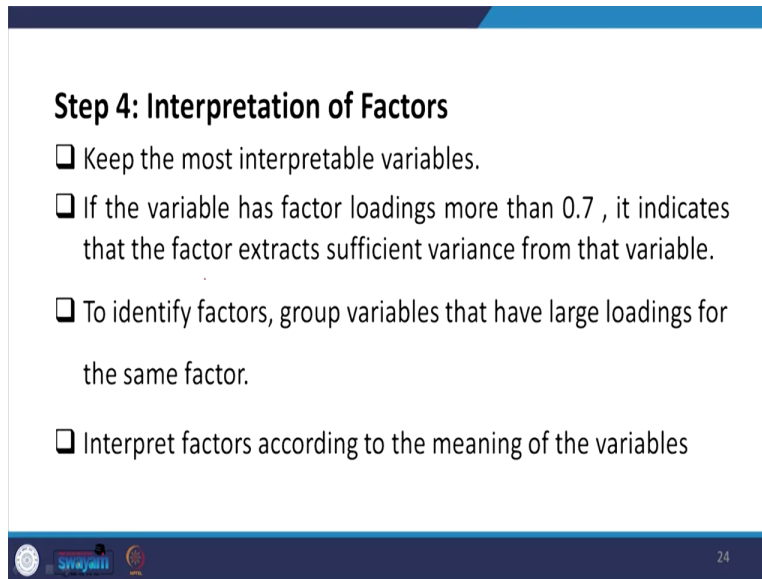
(Refer Slide Time: 39:36)

```
. sortl
Rotated factor loadings (pattern matrix) and unique variances sorted
```

Variable	Factor1	Factor2	Factor3	Uniqueness
pvt_hospital	0.7643	0.0255	0.0190	0.4148
private_sc-1	0.7449	0.0281	0.0386	0.4429
banks	0.6931	-0.0294	0.2147	0.4727
military	0.5823	0.1202	0.1664	0.6188
courts	0.5031	0.2027	0.3104	0.6094
state_govt	0.1147	0.7241	0.0518	0.4598
politicians	-0.2075	0.6470	0.0429	0.5365
panchayats	0.0401	0.6355	0.2218	0.5453
police	0.0921	0.5922	0.2159	0.5942
newspaper_-a	0.2864	0.5075	0.0801	0.6539
govt_hospi-1	0.0759	0.1009	0.8519	0.2584
govt_school	0.1029	0.0819	0.8510	0.2586

So, this is what is defined. the standard rule as I already mentioned for 0.5 and above factor loadings should be considered. And 0.5 and above has been sorted out and highlighted in the data and in the result, we can also do it from the Stata window as well. So, from the first factor which variables are important, first that is private hospital, then this is called private school, banks, military and courts. Similarly, others variables, I think if you check the original data of IHDS, those 12 variables we have also given you and that will guide you correctly. And now we have defined the important factors and their variables also.

(Refer Slide Time: 40:34)



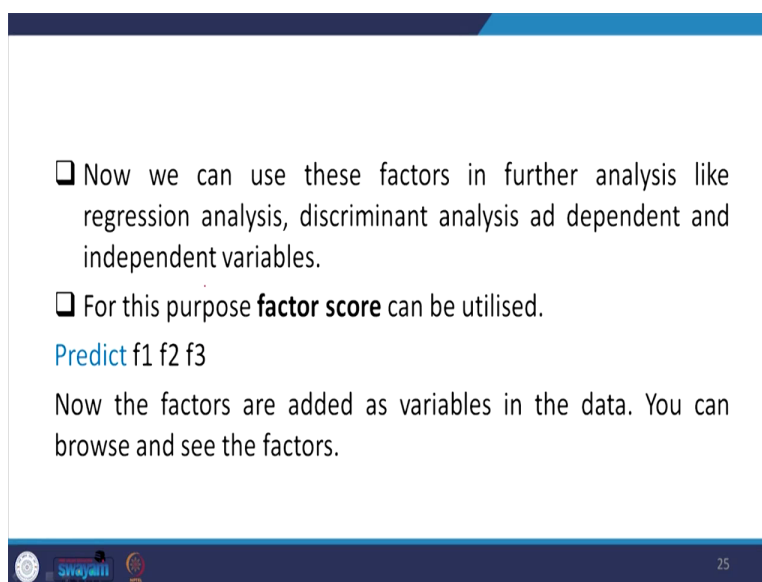
Step 4: Interpretation of Factors

- Keep the most interpretable variables.
- If the variable has factor loadings more than 0.7 , it indicates that the factor extracts sufficient variance from that variable.
- To identify factors, group variables that have large loadings for the same factor.
- Interpret factors according to the meaning of the variables

24

The fourth step is important to note that the interpretation of the factors at the end is very important. Keep the most interpretable variables as per the standard rule, it may be 0.7. If the variables have factor loadings more than 0.7, it is very good. It indicates that the factor extracts sufficient variance from that variable. Otherwise, 0.5 is the standard benchmark. Below that I think factor loading is not suggested. Interpret factors according to the meaning of that variable otherwise it will be completely redundant.

(Refer Slide Time: 41:09)

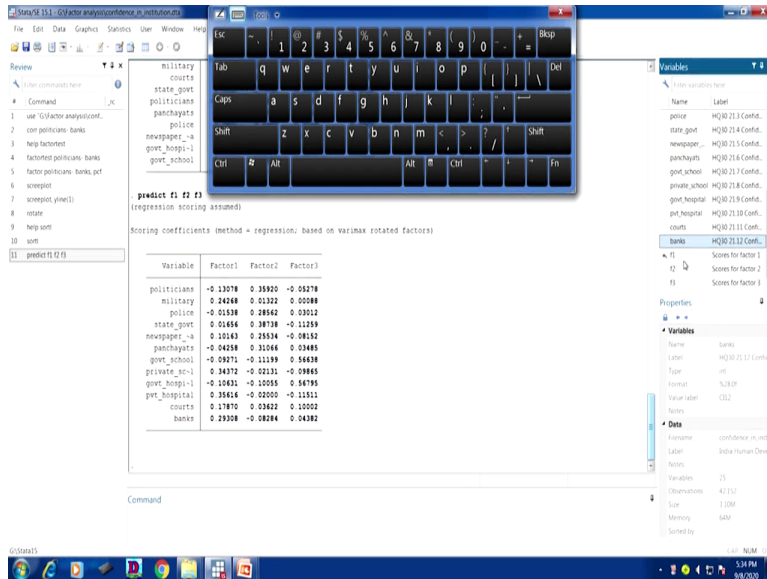


- Now we can use these factors in further analysis like regression analysis, discriminant analysis ad dependent and independent variables.
- For this purpose **factor score** can be utilised.

Predict f1 f2 f3

Now the factors are added as variables in the data. You can browse and see the factors.

25



We can use these factors in further analysis like regression once you have sorted the variables and for discriminant analysis maybe as a dependent or independent variables as per your choice. For for this purpose factor score can be also utilized. Factor scores can be defined with predict.

If you simply predict, since three factors are there, predict f1, f2 and f3 with this command you will get the result. Predict, isn't it, then f1, f2 and f3, isn't it? So this will give you the predicted values and those are also highlighted the scores of each of the predicted scores are highlighted in our variable listing, that you can use it for further analysis. f1, f2 and f3 are there. The factors are added as variables in the data as we have already shown to you. You can browse and see these variables.

A detailed analysis we have made with the help of Stata and those factor scores are usually very helpful for regression analysis and you must take the use of it. One important serious aspect of factor analysis is that please take care of your sample adequacy, please take care of your sphericity and the assumption behind the sphericity, then also try to look at the benchmark level of each of the indicator we mentioned. Then with that there will be no problem with the factor analysis. With this, I think we have explained factor analysis with the help of Stata very clearly and we should close the lecture here. Thank you.