Health Economics

Dr Pratap C Mohanty

Department of Humanities and Social Sciences,

Indian Institute of Technology Roorkee

Week – 11

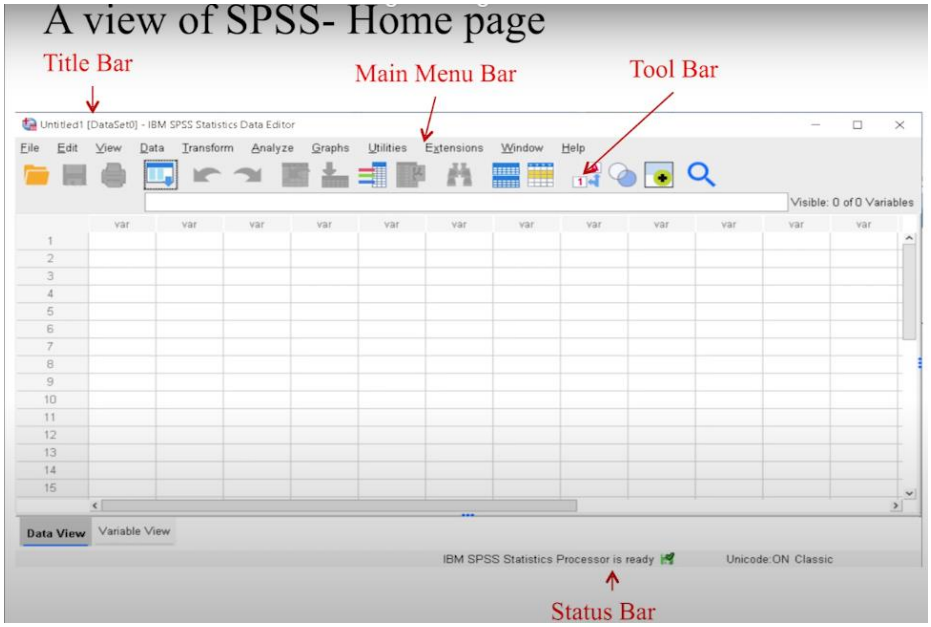
Lecture 57- Health Data Handling Packages: Licensed

Welcome friends to our NPTEL MOOC module on Health Economics. We are discussing the lecture on health data handling packages (licensed version). The licensed ones largely are here (indicated in ppt): this one (STATA), this one (SAS), this one (SPSS).

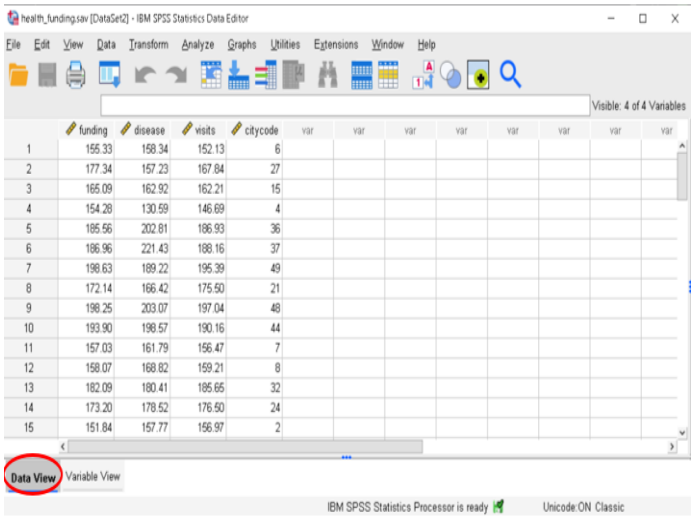
But we will be sticking to this one (STATA and SPSS) and its explanation. In the last lecture, we discussed the open-source data packages and emphasized or discussed Python and R. We also showed you the other important software available. The learning goal in this lecture is to introduce you to SPSS, as well as STATA, which scholars largely use for the data handling purposes.

I have already shown you the software available for data processing. The two most frequently used by scholars are SPSS and STATA. I will refer the hands-on materials to you, and you can also go through them and practice for your own research work. These are considered to be good, and STATA is considered to be very good, so far as graphics is concerned. Regarding the cost, SPSS is more expensive than STATA. However, even if these are paid versions, you can get the sample copy for one week by registering with the sources. Here, we will start with explaining SPSS, that is a popular programme for statistical analysis. SPSS stands for 'Statistical Package for the Social Sciences', and hence, this is called SPSS. In 2009, IBM purchased SPSS. Its proper name is now called IBM-SPSS Statistics. The data in SPSS is saved in the format called .sav.

So what does it look like? It looks like the Excel page on its data entry segment. However, entering your data using this window, then that of excel, is suggested because this also gives you a better window in terms of entering the qualitative variables.

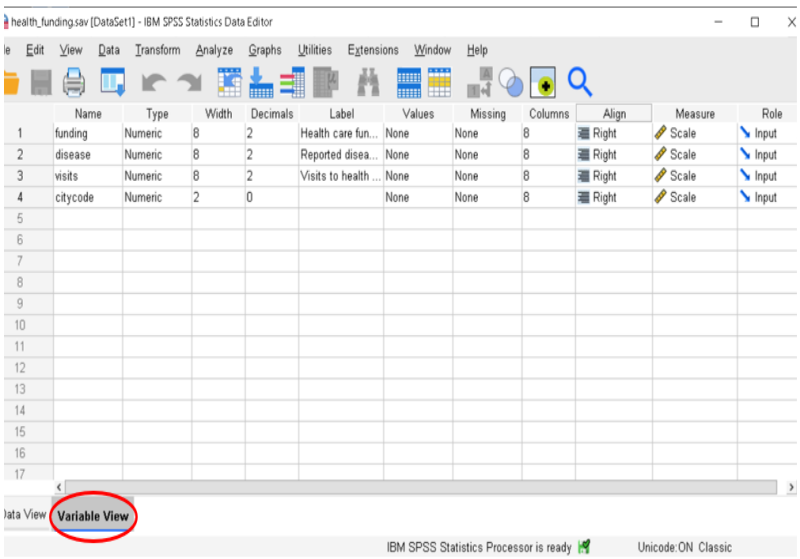


The first line is called 'Title bar' as is visible from here; the second is called 'Main menu bar', and the third is called 'Tool bar'. On the bottom, you see that 'IBM SPSS statistics processor is ready', this is called 'Status bar'. There are two view modes to SPSS. One is called data view, which I have already started discussing.



	funding	disease	visits	citycode
1	155.33	158.34	152.13	6
2	177.34	157.23	167.84	27
3	165.09	162.92	162.21	15
4	154.28	130.59	146.69	4
5	185.56	202.81	186.93	36
6	186.96	221.43	188.16	37
7	198.63	189.22	195.39	49
8	172.14	166.42	175.50	21
9	198.25	203.07	197.04	48
10	193.90	198.57	190.16	44
11	157.03	161.79	156.47	7
12	158.07	168.82	159.21	8
13	182.09	180.41	185.65	32
14	173.20	178.52	176.50	24
15	151.84	157.77	156.97	2

Data view looks like the spreadsheet. Each row one participant and each column one variable. And, the 'variable view' (the second one) is visible here.



If you open it, you will find it in SPSS, only if your institute has access. You can also request your institute to go for purchasing it, since this is useful software. The variable view gives subcategories about the variables, divided into other subcategories, like its name, type, width, decimal, label, values, missing, columns, align, measure, and role, etc. These are all given here. (Highlighting in ppt)- Name, then type, width, decimal, label, values, missing, columns, align, measure and role.

Most importantly, we type the name of the variable here. Then, you specify→ what is the width of that variable? Which character type? Is it string or numeric? Then, does it have decimals? Then, the details about that variable, you can just write it here. If we say funding, we are referring to 'healthcare funding' from this period. This is simply called a label. Others are not that important, but whenever you start using it, you will find the differences. In the variable view window, each of the subcategories has features listed below.

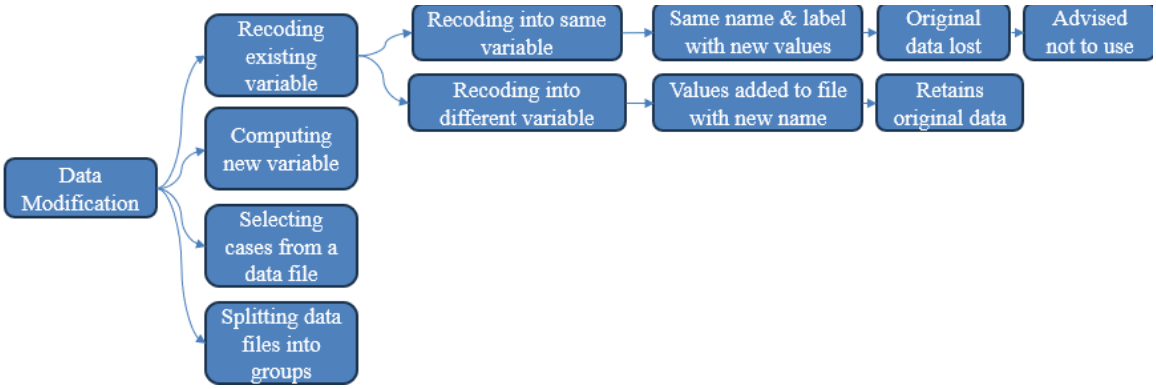
- **Variable name: No space**, **No special character**/ symbol, **First character→ can't** be a **number**, must be **unique** within dataset
- **Variable type**: Numeric, comma, dot, scientific-notation, data, dollar, string, custom currency, & restricted numeric (integer with leading zeros)
- **Width**: Maximum characters allowed for each response
- **Decimals**: Number of decimal places (for numerical variable)
- **Label**: Display name (unlike variable name→ it can contain spaces)
- **Values**: List of valid options for the variable. Ex: Area→(1- Rural, 2-Urban)
- **Missing**: Value being used→ when response is not applicable or not answered

Variable name: No space should be given. No special character or symbol, first character cannot be a number, must be unique within the dataset. Hence, in the variable view, you should not use the space, and it has to be consecutive. You can just see, it has to be consecutive, and special characters etc. should not be written. If you want to clarify, you can just write down in this box (all details).

Then, the variable type includes numeric, commas, dots, scientific notations, data, dollars, strings, custom currency, restricted numerics, etc. Then, the width can be specified. The maximum characters allowed for each response is called the width (width about the variable). Decimals here refers to the number of decimal places (for numeric variables). Then, label, basically you can write it down the name of that variable. Here, you can use anything, maybe space, maybe special characters, whichever is suitable to you, you can write down as the label. As I told you, the label clarifies about the variables. Then, values are list of valid options for the variable. If it is like this, I will just show you. Values: At this moment we have entered as none, but actually like if you have a categorical variable, you can enter its values. If it is rural-urban, then it is 1 or 2. Then you have to define that, within these, there will be options. You can just click, and it gives directions for entering the codes and its clarifications.

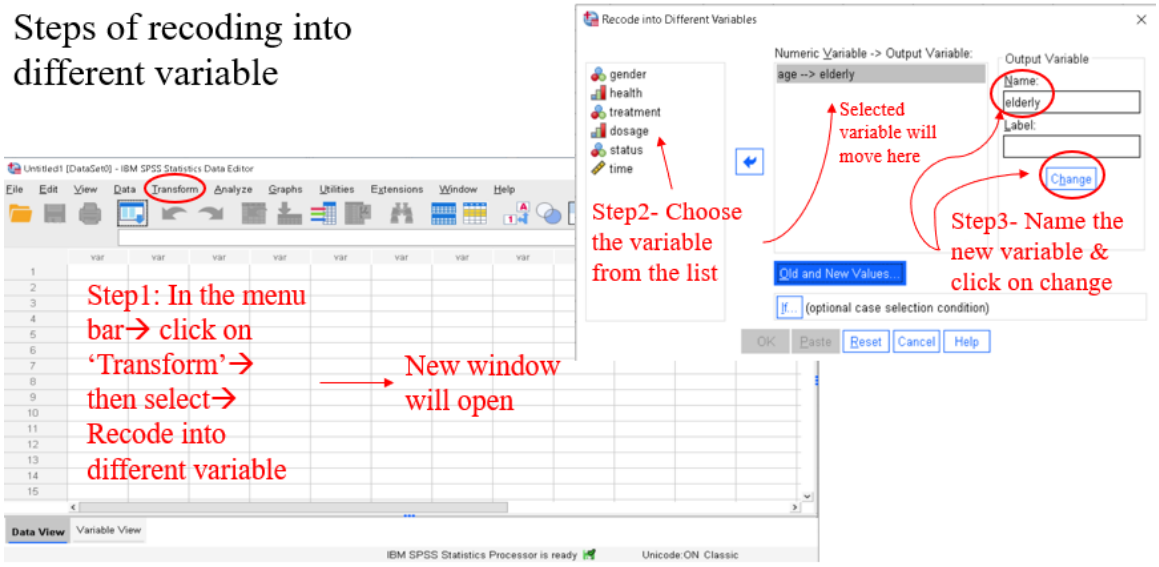
Then, width values I have already mentioned. Missing value, value being used when response is not applicable or not answered. Align the data, whether the data should be left or right aligned. Measurement in scale (nominal-ordinal etc.). Then, these specifications are clear. I am sure, you can able to clarify, we are not explaining each of them. And if you want, you can also follow my own module. These are all available in YouTube. You will find out further details.

Data modification in SPSS, how to do it? Data modification includes recoding the existing variable, computing new variable, selecting cases from data file, or splitting data files into groups.



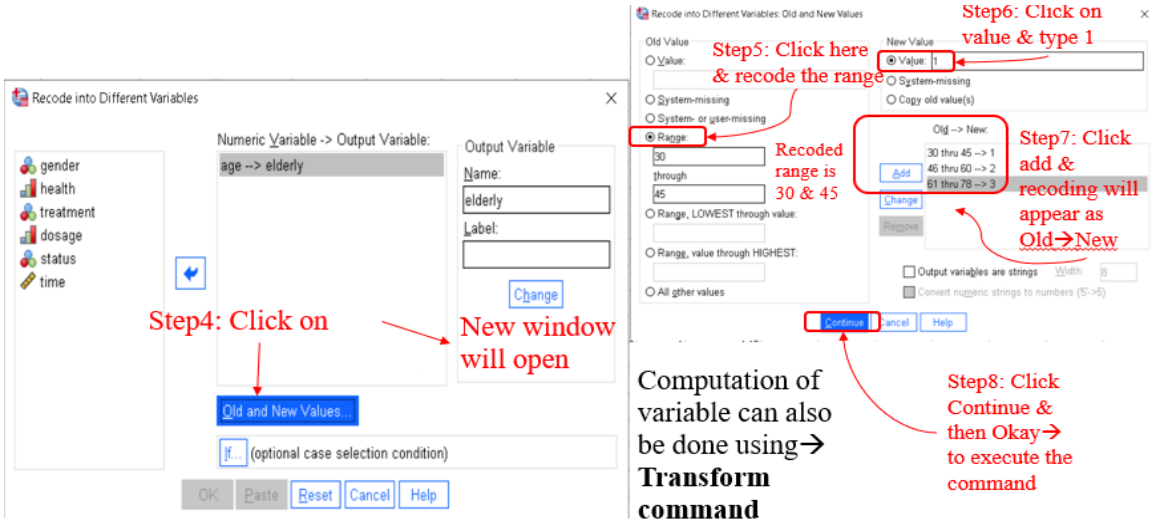
Recoding existing variable, in that case, it might be recoding into the same variable with the same name and label with new values, then original data is lost, so advised not to use if it is of the same name. Then, if you are recoding into a different variable, then values are added to the file with a new name, and a new variable is created. Then, it retains the original, as well as new.

Let us see what the steps are for recoding different variables. You can just see it-



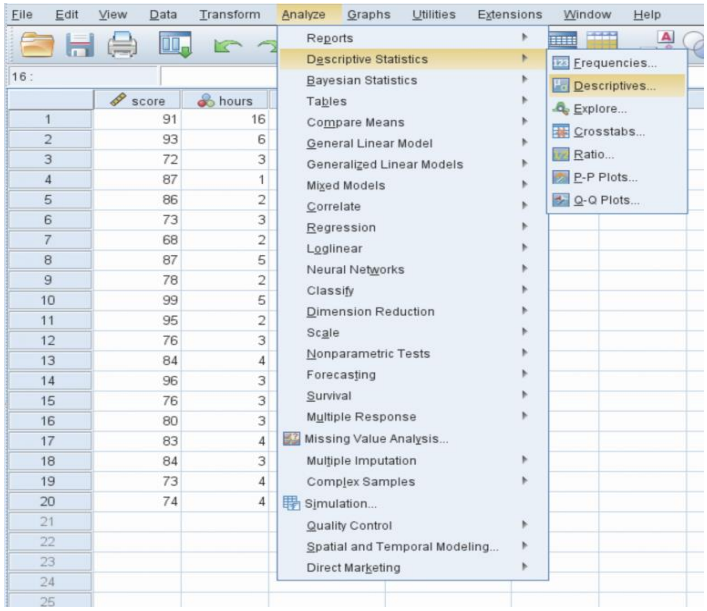
You have to go to transform on the front page itself. So, you simply click transform. We have written it clearly. Click here on transform, and then it will display recode into a different variable, and it will open a new window, and the new window gives the window like this. It gives like the small window and displayed with a small window like a numeric variable to output variable, if any you are giving. Then in that case, the selected variable, like choose the variable from this list. Once you see the list here, you just select which one you want to recode or transform. Then, the selected variables will move here. Once you click here, it will move here. Earlier if it was there, if you click here, it will move to this side, otherwise there will be this kind of arrow if nothing is available in this window.

So, once it is there, then you can name that new variable, and the label can be specified. Label means the meaning of that variable. In short, you have to write down the variable name. Then, in long words, you can also write down the label of that variable.



Then similarly, if you want to change this old code to new code. Say, you want to combine some code and make an aggregate variable. Like in India, states are there, there are different states, they are there, out of that you want to divide into northern states and southern states. So, from that variable, you can find out or change the values of the northern or southern entries. You can just see it here. So, new and old variable and optional are given. If you click on it, it now gives range, etc., and values.

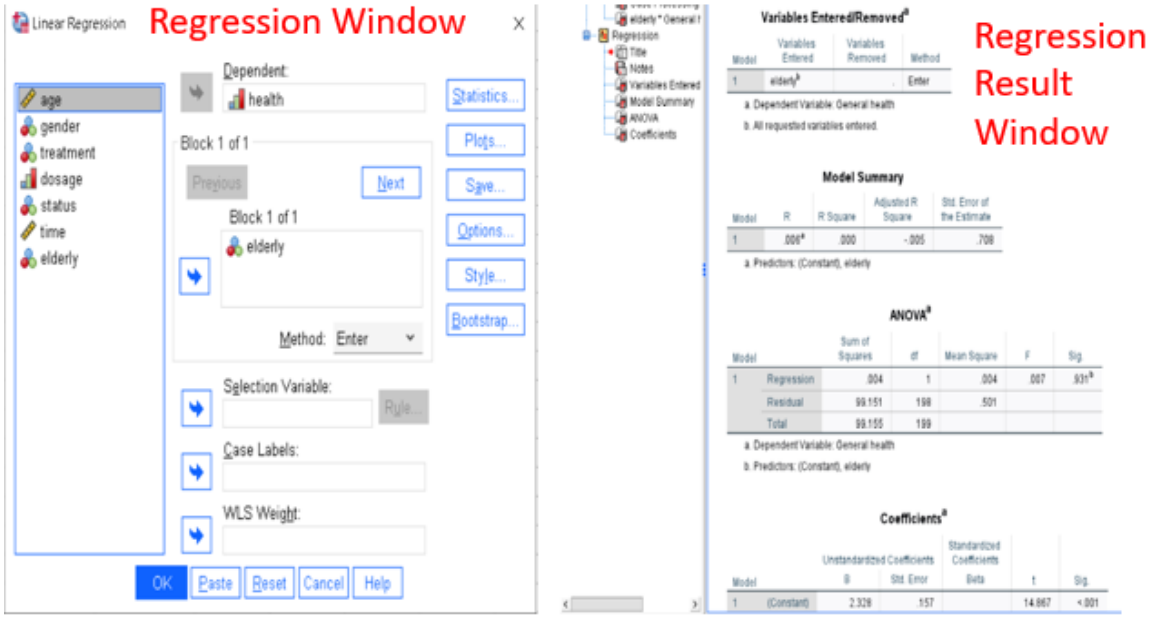
Since it is categorical, you either want to just enter all the states into one or two, based on north and south, and then you have to define accordingly. So, range you can specify, and based on that, like here, in an example, given '30 to 45' one code is given, '46 to 60' two code is given. Similarly, if you want to say 1 to 5, then even through lowest value through any number, if you just add 1 till 10, it should be 1, then 10 to 15 should be 2 and so on. If you wanted to make it correctly and you know the state codes correctly, then you can assign only into binary numbers. This is what is given, you can use it. Then you are simply supposed to add it. Then, finally, if you continue, click on continue, that variable with the new code generated. Compilation of variable can also be done using transform command. To transform command, just through the command on the window, we can also transform it, instead of putting clicks on that button.



Steps of some other SPSS computations: From Main menu, we will go to analyze. Then, we can choose the analysis such as frequency, descriptive analysis, and cross-tabulation. Then we can analyze the inferential statistics such as regression, regression analysis and their, executions etc. We can do it. In the main window, after transform, you will see analyze. There are so many analysis possible, but if you are interested in regression analysis or inferential statistics, you have to click here. Otherwise, for basic frequency distribution, descriptive statistics are more than enough.

And within descriptive statistics, you just see it has frequencies, descriptives, explore, cross tab etc.

The regression window and the result obtained in the result window is illustrated below:



If you are going through a regression analysis, you have to define the dependent variable on the window and the set of independent variables. Here, we have derived the dependent variable (we have mentioned the variable), the dependent variable is health. And then, other variables like elderly, etc., we are just taking them as I will tell you just now as per the window. Similarly, if you click on statistics, then plot for diagram, graph, etc., you will get many other directions for analysis. Here, the predictors are elderly, and I think it is possible to derive, and if you have any queries, I will certainly address it in our query class.

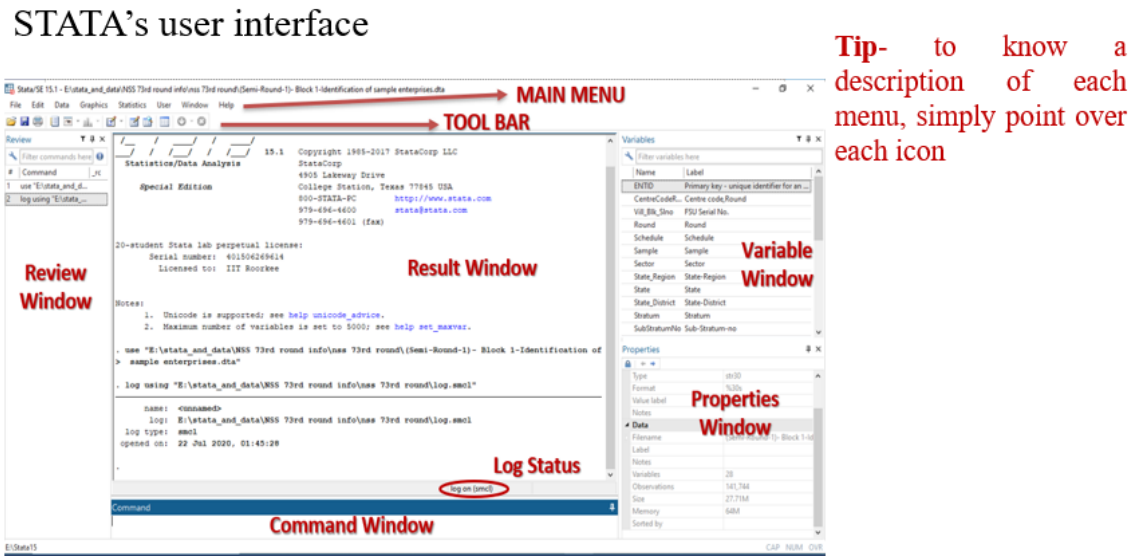
Let us move to another statistical software called STATA. I have used a number of times in my earlier modules, that is another NPTEL module of mine is also floated during the same time. It is available for registration at this moment, that is called 'Survey data in healthcare' or, 'Using analysis of survey data in healthcare', that is another module if you see, we have discussed about the different available survey data. We also used statistical packages like STATA.

"STATA is a complete integrated statistical software package that provides everything you need for data analysis, data management and graphics", as per the definition of STATA itself in 2016. The name STATA is a syllabic abbreviation of the word statistics and data. Official site is available here for you to download. The student or instructor for short term license request can also be given in this link.

Different versions of STATA are available for work. I will just clarify which versions are available. One is called the standard version of STATA. Usually, this handles very less

variables, only 2048 variables, out of which 798 should be at maximum independent variables at a time, and 2 billion observations. Whereas, in the case of special edition, STATA special edition version can handle 32,767 variables, 10,000 independent variables, and 2 billion observations. This allows longer string variables and larger matrices. STATA MP version, that is also called multiprocessor or multicore version. I will also simultaneously show you how it looks if I open a STATA version before you. The best one is called the MP version, the multiprocessor, or multicore version. It handles up to 1,20,000 variables, 65,532 independent variables, and covers 10 to 20 billion observations.

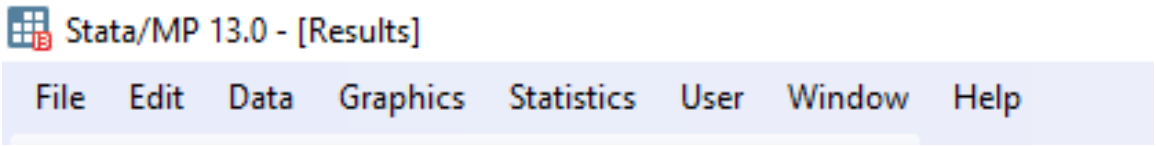
This is how STATA looks like:



I am just going to show you how it looks like. So, I have opened the STATA. I hope it is partially visible to you. It's very difficult to make the screen big. However, I am just showing you the actual window. I have opened STATA of my own version. So, you can see how you can work. I will just guide you what those different components are. This is the main window, and this side, here is the result window. This side is the result window. I will show you. The right-hand side are the variables and its properties window, left hand-side is screen view window, and the bottom line is command window. Otherwise, it has also files, edit, data, graphics etc. I will come to it one by one. So, STATA user interface is here. I have just written all those details, what is called result window, then what is called review window, and what is called command window, and what is variable window and its properties.

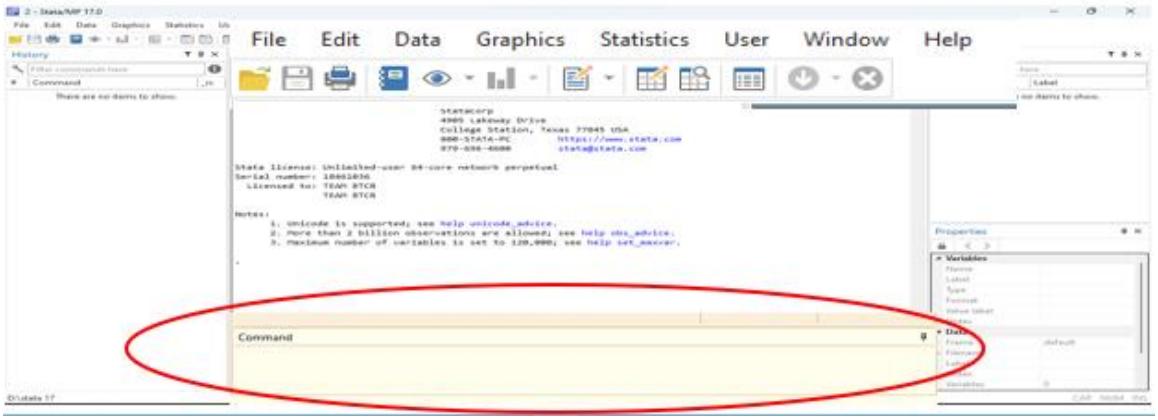
This sides are their main menu; these are their toolbars. One of the important aspects, since this NPTEL health economics course does not give you enough space for explaining the very basics of STATA, etc., you can follow our module, you can search handling large-scale data, this is my course where I have given one week, I have step by step guided you about STATA. So, at this moment it is very difficult to explain, given the time constraint. If you still miss here, then you can also follow my another module called survey data in healthcare (which I just mentioned survey data in healthcare).

One tip is there: To learn the description of each menu, simply pointing over each icon can also give you further details. I will tell you how these are in the STATA main menu bar.



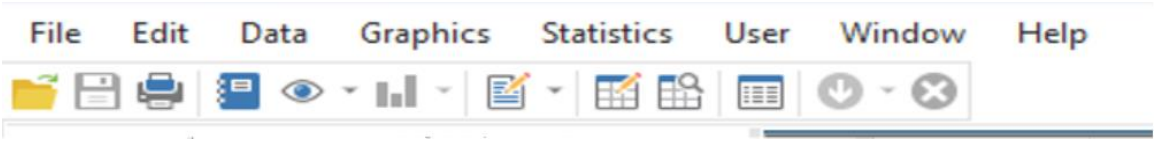
File, Edit, Data, Graphics, etc. I think I have already shown to you just now. So, not explaining much of it, I think we should go through. And I am just showing you for your interest only. If you open our respective module, you will benefit.

Here, this is the command side, where we can type the command, since STATA is a command-driven package.



It has also drop down menus like from this side. If you click here, click there, you will get drop down menus, just to click and find the result based on the variables and data.

This is what is called the command box, and this is the menu bar and the toolbar-



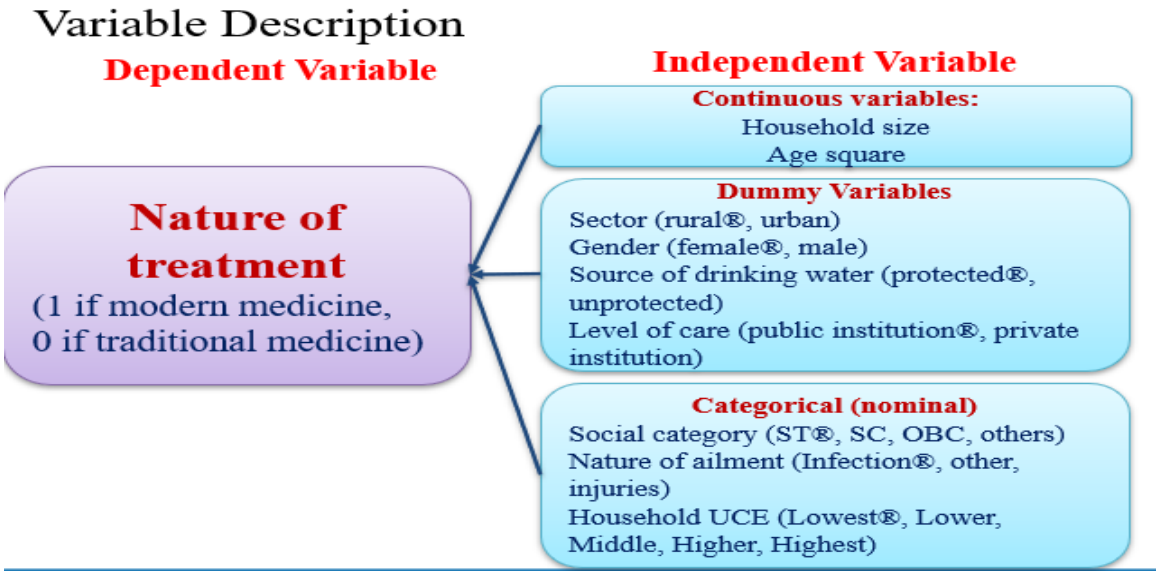
Estimation and modeling in STATA. In this unit's third lecture, we discussed linear and non-linear regression modelling. I am just giving you a glimpse of how STATA works, how beautiful the software is. I think it might still preserve your interest in this lecture. Here I am just giving you the answer, one module I will just simply run, but in the third lecture particularly, we will clarify the concept, what is this econometrics modelling.

So, at this moment, I am giving you just one answer based on my own database, which I have been working. So, it is based on the nature of the treatment. The data is taken from the National Sample Survey, where the latest round on healthcare is considered. The variable of interest for us to explain is the 'nature of treatment', which is in binary form, traditional

versus modern treatment. If somebody is suffering with some ailment, then the person takes the help of different methods of treatment, those may be traditional or modern. Hence, we have categorized them into binary forms.

For this purpose, we are using the sample, as I just said the latest round of NSS, it is called the NSS 75th round. You can also cross-check from the previous round, i.e., the 70th round on healthcare, if you want to check for some comparisons for your work. Here, I am just giving you the result. The NSS 75th round, conducted in 2017-18, is for the pan India level.

As I just said, the dependent variable is a dummy one. We have mentioned traditional versus modern and categorized it into traditional and modern treatment.



A mixture of qualitative and quantitative variables are used as regressors for the analysis. So, here is the complete model. Regarding formation of this model, independent variable and its concept etc., we will make it in the third lecture. So, here the code for their dummy variable dependent model is: 1 if modern medicine is used, and 0 if traditional medicine is used. And the independent or the control variables we have taken are these all, some of are continuous variables, some of are dummy variables, and some of are categorical variables.

These all variables are taken in that model and here is that model specification.

Model Specification

$$\text{trad\_modern(Y)} = \beta_0 \text{HouseholdSize} + \beta_1 \text{square\_age} + \beta_2 \text{i.Sector} + \beta_3 \text{i.male\_female} + \beta_4 \text{i.public\_privatehospital} + \beta_5 \text{i.drinkingwater} + \beta_6 \text{i.SocialGroup} + \beta_7 \text{i.nature\_ailment} + \beta_8 \text{i.UMPCE}$$

So, Y is your 'traditional medicine versus modern medicine'. Then, household size, age square, sector (its conditional variable is subject to the reference category), gender, public

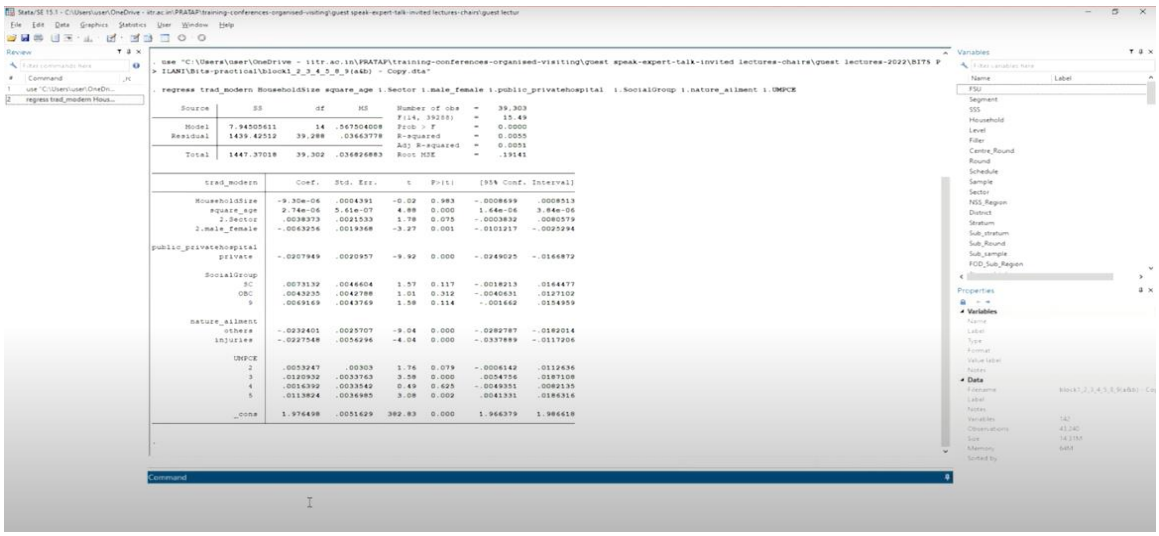
versus private hospital access, etc., are taken into account. Let us see how it works in STATA. I will just tell you. I am going to just show you how it actually works.

So, here is my window. I opened the refined dataset of National Sample Survey 75th round, and here I have loaded it in my computer.

You will learn from my other module what is called do-file, and this is how the do-file I have.

I will just run this regression model based on my data. I will copy this. I will copy the first command. This is the command. I will just tell you how it works.

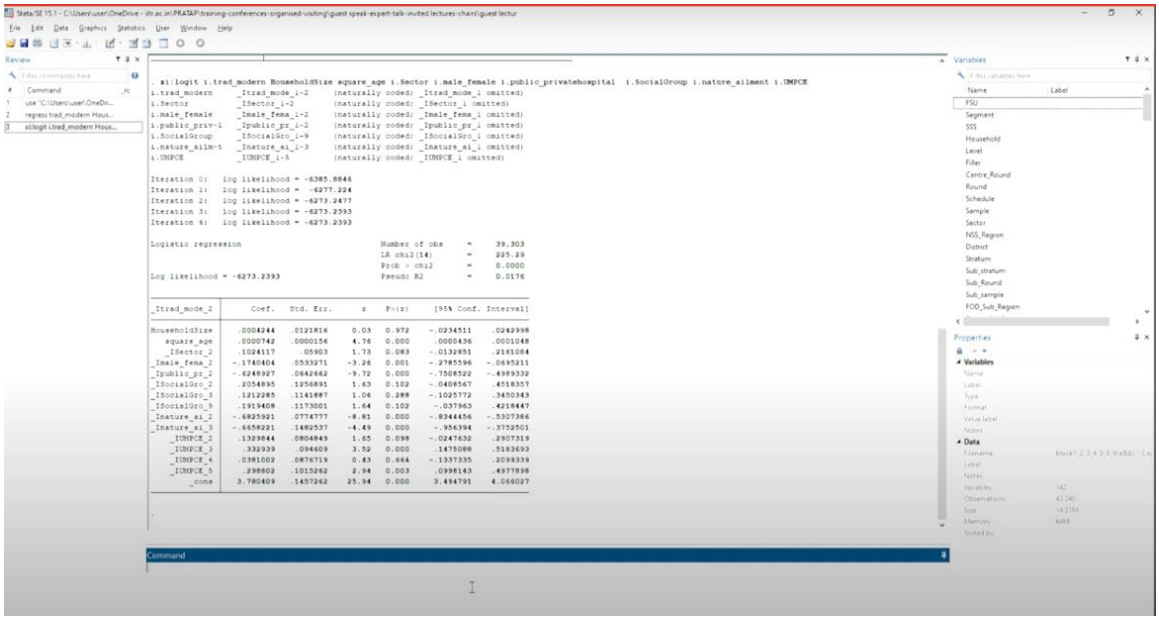
Suppose I apply the ordinary least square method to the model, even if the dependent variable is dummy categorical. I am using the OLS method. It will also display results, but it may not be robust. You can enter. We will simply enter, and now the results are actually displayed.



So, you can see the number of observations, and its F-statistics, its R square. The model is derived. I cannot interpret much because of the time constraint. As I told you, these are available already step by step, and even its interpretation in my NPTEL module on 'Survey data in Healthcare'. Please go through it.

The second command, I will just run. Since the first command we have treated as ordinary least square regression, even if our dependent variable is dummy, which should not be done, that is an error. Now, we are running one of the appropriate models. So, we will take another regression which is most fitted (considered to be most fitted), since it is a limited dependent variable model. I have just started running it. This is more important than that of the first model, because this is a better fit, since that dependent variable is limited categorical.

And this is how it looks. I have tried my best to show you what our results look like.



Since we do not have much time, I think it is better to proceed from the regression coefficients, etc. We can easily interpret its significance, etc. The interpretation is very difficult when the dependent variable is categorical. But in the logit model, we actually derived the logit coefficients in the second one. Coefficients with the p values, we can interpret very correctly. So, it is better not to spend too much time. I am referring to my lecture and going to explain the other things.

So, this is how it works. I am sure I am not justifying much at this moment. Therefore, I am referring to my previous NPTEL modules by (Prof. Pratap Mohanty), which we have explained in detail. I am sure you will get it.

What to conclude then? We are saying that there are powerful data packages that are available, and although they are paid, used to be very work-friendly. Both SPSS and STATA can be helpful in various health and healthcare analyses. Moreover, econometric software is

important for analysing the model's inferential statistics. However, understanding of the data is crucial.

For example, when we run an OLS regression, OLS follows an assumption called linearity. But the results are not that good because we have only gotten average figures. However, when we apply an appropriate model (say- binary response model), i.e., logit at this moment, we get better results. I wanted to apply other important models, but because of time constraints, I was not able to do so. The results show that the type of data available has direct implications for model specification and estimation. We have also clarified this in our lecture number three of this week.

So, what will we do it in the next lecture? That is on the public health system and policies. So, these are the readings. I hope even our own NPTEL module link is given for your reference. I hope you will learn the best. If you are still having some trouble, we are there to support you through our query lectures. Thank you. Let me stop here. Thank you.