in this video i will provide an introduction to the stata software

so i assume that you have purchased the stata license you have downloaded the software and you have executed it and you're ready to get to learn stata

so what you need to do

here is download the program introduction to stata and then double click on it to open it or alternatively

if you open stata you can go in file open and then open up this file

the easiest way to learn stata is to try to open an existing program so this program is called xero introduction to stata dot do

it has a dot dot extension which is the extension for stata programs

so here we have the program which is the dot do file and here we have what we call an output window and how we get to write in the output window is if you click on this button here which is execute or do command and basically that would execute the program

and this is what we're going to see here as the output window

so today what i will talk about in this introduction to stata video is how to set up the data sets that we have how to explore the data how to edit the data how to export regression output and we will end up with creating log files

the data set that i will use today is called wage1.dta so dot dta is the extension for data files

so dot do was the extension for the program itself

so introduction to stata dot do and the data files are dot dta extension so wage1.dta

so the first thing to do is download

the data set from the website and then store it somewhere on the computer and so for me

this data is located in the folders on my c drive econometric slash data and so you need to change this part here in the program where the data is stored on your computer in order for you to be able to run uh this program so i also have changed here um

i have used and set up a global directory so global and i'm calling this data directory because uh all of the data sets that i have are in this folder and i can call them from this folder and so later on when i want to use the data set

i will use the use command and then here this is the data directory slash wage one and this wage one would be calling up this file wage1.dta so comma clear means that if there's any other file in the memory it's going to be cleared and rewritten with this file

so now you can accomplish the same thing by just changing the directory here

if you do a cd change directory and then i use this directory here and then you could just call it simply by saying use wage one comma clear so that would accomplish the same thing

but i'm using a global statement here for the data directory

there are a couple more things to mention here

uh one of them is clear

all that also clears everything that is currently in memory so that the program could be rerun multiple times by coming here and running the program and set more off just turns off uh the more button when the program is executed so these are some of the things

so another thing that we can do here is um to create a log file to save the output so how we're doing

this is by saying log using we're giving it a name and i currently call this the log file comma replace and so comma replace meaning uh

if if this log file exists next time we run the program we're going to be replacing this file

so this is useful because as we're running the program all of our results are going to be stored in the log file so anything that you see here in the output window is going to be stored in this log file that we can later read and and basically would have all of the results stored okay

so this is basically how to set up the data so now let's go ahead and explore the data

so the first thing that we would be doing uh is to describe the data and so so here in the output windows just to backtrack a little bit

here's the log file that is starting to be created and it's saying this is where this the log file is going to be stored

this is the name of the file that's the extension of the file and um basically every time we run a file here in the output

we would see what has been done so as we're exploring the data if we execute describe

this is what we are going to see

these are all of the variables in the data set

so here we have the variable names and typically here there are the labels for these variables

however this data set in particular is not very well documented

so we have the variable names being repeated here but oftentimes this would say something like um wage measured in dollars and that would be hourly waged education measured in number of years uh so something like that

these labels are going to be like that so if we don't want to be looking at all of the data set we can actually um just describe a few variables of interest and here we would be describing wage education and experience and as you can see now we're just describing wage education and experience

so this is much much shorter if we want to list and see the first few observations of the data

what we can do here is use the command list and here we have wage and then education and experience in 110 means uh just the first 10 obs

uh means from the first to the 10th observation and here's from the first to the 10th observation

here we have wage education and experience uh

so this is for the first person for the second person and for the third person

if we want to see where the data is

we can actually click on the data editor and we can actually also look in the data just like that um

but we can also list it in the output window as well

so the next thing is to come up with summary statistics and here we would use summarize uh

if we use summarize

we will get the summary statistics for all of the variables um and then we can also summarize uh only a few a few variables so here summarize summarizes pretty much all of the variables so what we see is the number of observations mean standard deviation min and max and if we are using summarize um of these three variables that we're interested in uh here's what we're going to have as far as mean standard deviation min and max and these are the output that we can actually provide in tables if we want a little bit more information on a particular variable

we can use summarize then put the name of the variable wage common detail and here a lot more detailed statistics so we have the mean standard deviation

we also have the percentiles the first from the first to the 99th percentile

we also have the smallest and the largest values so sometimes we need to look into more details for a particular variable

we can also provide summary statistics by group if we use tabulate a female

this provides a frequency table for females

so here we have a female equals one

we have 252 observations and if female equals zero we have 274 observations

these are the percent frequency and the total number of observations we we have here so

now

we can also summarize the data uh if female is equal to one uh so summarize the variable wage if female is equal to one notice here that we use the double equals sign if it's equal to one uh and we need to say which variable here

so we're summarizing uh wage for females and this is the average wage for females or if we want to uh summarize for both males and females how we're going to do this uh using by sort females

so here we're sorting the data by the variable female and then we will be summarizing wage

so what comes with this command

as far as output is we will summarize the wage if female is equal to zero

this is basically the wage for males and then we will summarize the wage if female is equal to one

so this is the average wage for females here

the next thing that we can do now is uh edit the data so we can keep or drop variables and um so here we can use the command keep and these are all of the variables that we would keep here and so the rest of the variables are going to be dropped from the data or if it's going to be easier

perhaps we can just drop a particular variable such as tenure and here we kept it but now we're dropping it

so we we can also drop a particular variable if we no longer need it

so another thing that we can do is drop if wage is less than 2.

so if we have certain values for a variable that we want to drop we can use this command so notice that while keep and drop here refer to the variable itself or to the columns in the data drop of wage is less than 2 would actually delete certain observations in in data that have wage less than two

so it would delete certain rows in the data

so the next thing we can do is we can actually label the variables because again this data set in particular was not very well labeled

so we can put label variable and then we can say the name of the variable and then how we want it to be labeled

so here we could put hourly wage in dollars and you see now

when we use describe wage

we can see wage here and then this is the label of of the wage so the nice thing about using dta files with stata is that not only that the variable names and all of the values of data are stored like an excel file

but it also has this other useful information such as variable labels

the next thing we can do here is we can generate new variables so that is done with the command generate or gen for shorter

so here we can generate log wage which would be equal to log of wage

we can also generate education squared which would be equal to education to the power of two or we can generate a region called southwest by summing up south and the west uh the observations that are in the south and in the west

so this is a way to just generate new variables

so the last thing we are going to do here is export regression output because sometimes it's useful to do that if we need to create tables so um these um lines

here uh

i have put a star in front of them

so anything that has a star in front of them is commented out so stata is not reading it

so here we would need to do it only one time and this would mean to install the outreg2 package and so if we remove the star in front of these commands here we we would be installing this outreg 2 package

so actually we need to remove the star here and execute this line only once

so this ssc

install

uh this would be installing the outreg 2 package and this package actually later on

we could be calling it so that we can generate a nice table so

i'm going to put the star back on here because i have it installed on my computer so again you need to do this only once

so how are we going to do that the first thing is let's estimate a simple regression model so the way we're doing

this is by using reg or regress and here wage is the dependent

variable and education is the independent variable

so reg of wage and education would estimate the regression model and the regression model is given here and so here we have wage dependent variable education

the independent variable this is the coefficient that we have here so it means for one additional year of education

the wage would increase by 53 cents so there are a lot of also additional uh statistics here from this table that we will be discussing later on but for now we're just interested in exporting results in tables

so here's how we're going to do this

we will call this outrigger2 package

then we will say using and i'm giving this file name

i'm going to call it a regressiontable.doc and here comma replace if we have any file

we would actually be replacing this file and so then let's say that now we want to estimate the multiple regression model so how we are going to do

this is uh estimating a multiple regression model so now

in addition to education

we have an experience so here we're running a regression of wage which is the dependent variable education and experience which are independent variables and these would be the coefficients that we get from this model so let's export

also these results in table and so we are going to be using outreg2 using the same file and if we put comma append here this is going to append to the um the previous file that we have so again

these are nice tables that get generated that could be used later on um later on

so the final thing that we we could be doing

here is uh log close and if you remember we opened the log file before and now we're closing it so everything that was actually done in this output

uh window here was saved in this log file so if you go on your computer where the data files are where where currently the programs are you could actually see this log file in there and you can basically save all of this regression output here

so this was a very quick introduction to stata thank you very much for watching and please see the other videos