# Introduction to Stata

This short note is meant as a gentle introduction to Stata. Before the course, please read through this note, open Stata and try it out. Specifically, you should try to issue some commands by using the menu, by using the command window, and by creating a do-file.

User interface

When you open Stata, the interface is split into five different windows.

- The main window is the Results window. Here comes the reporting of all commands issued.
- Command: Here you may enter and execute single commands.
- Variables: This window gives an overview of the variables in the datafile that is in use.
- Properties: Here you may read and edit various properties of the variables. If you highlight one of the variables in the "Variables" window, the properties of this variable will be shown here.
- Review: This window stores the history of commands.
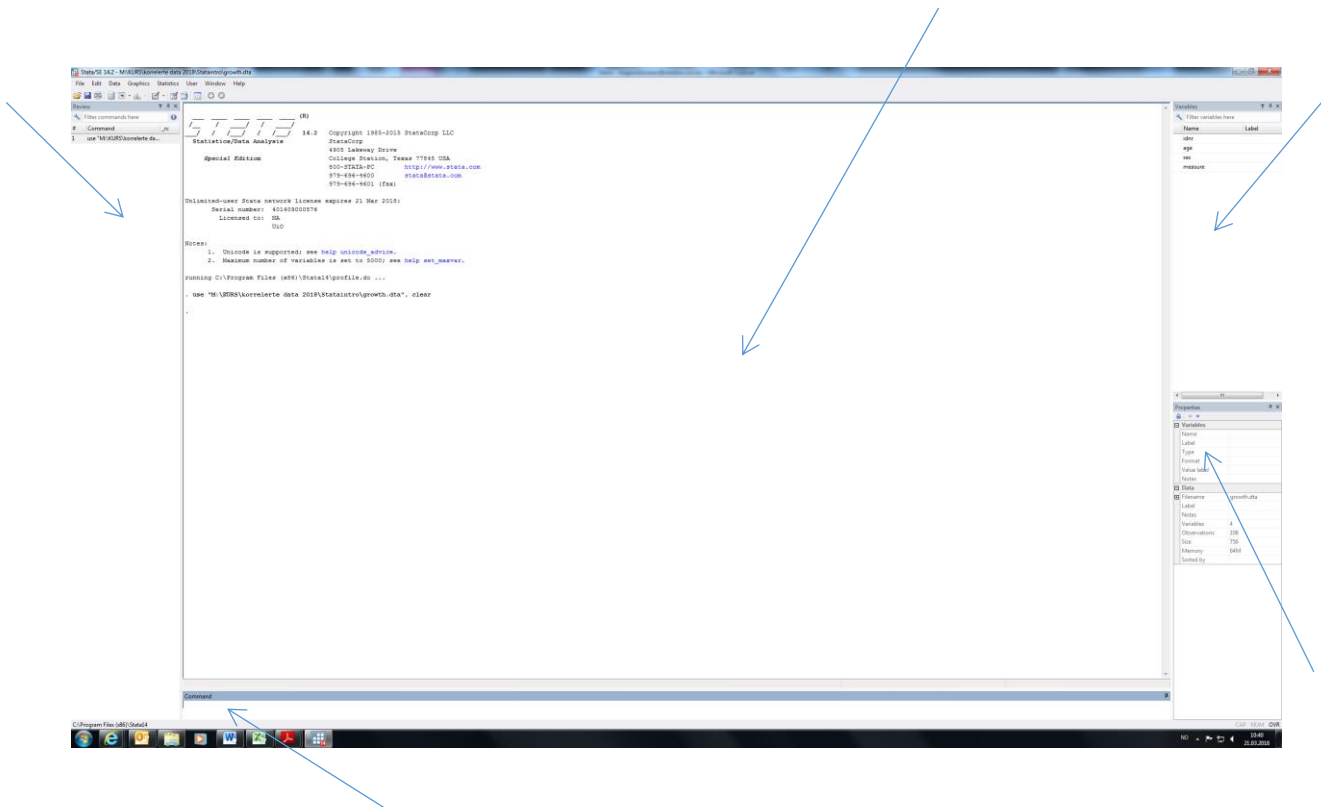
In addition to this, you can open data sheets, which looks like an Excel data sheet and you can open do-files, which we will come back to below.

Do-file   Data sheet
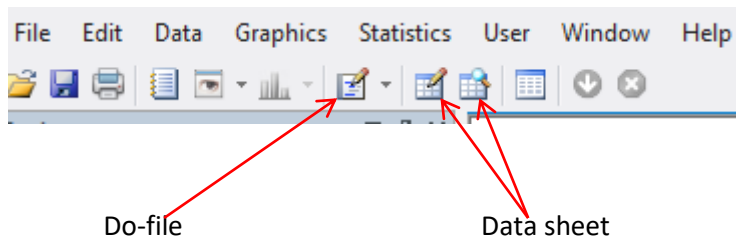
How to issue commands

When using Stata, you can issue commands in three different ways; by using the Menu, by writing commands in the "Command" window and by executing commands from a do-file. During the course, we will mainly be using do-files, but it is often useful to use the Menu as a first step, as the Menu also produces commands.

Let's try to illustrate:

I have opened a file called "growth.dta", which you can find in Timeplan (download the file, e.g. to your desktop and open it from Stata, e.g. by File – Open). The file contains four different variables; idnr, age, sex and measure.

If we click on one of the data sheet buttons (see above), we will have a look at the data:

Assume we are interested in the mean value of the variable "measure". We can compute this by going to the Menu (in the main interface) and click:

Statistics – Summaries, tables, and tests – Summary and descriptive statistics – Means

Specify the variable of interest, in this case "measure", and click "OK". Stata gives the following output:

```
. mean measure

Mean estimation                       Number of obs   =        108

                     Mean    Std. Err.     [95% Conf. Interval]

      measure     24.02315    .2818024      23.46451     24.58179
```

First, we see the relevant command ("mean measure") and next the results, showing the mean value (24.0) and the standard error (0.28).

We could have produced the same result by typing this command into the "Command" window and pressed "Enter" (try that).

If we are interested in the mean measure by sex, we can again go by the menu. When you come to the panel where you specify the variable of interest, click the if/in/over label and specify sex as group variable.
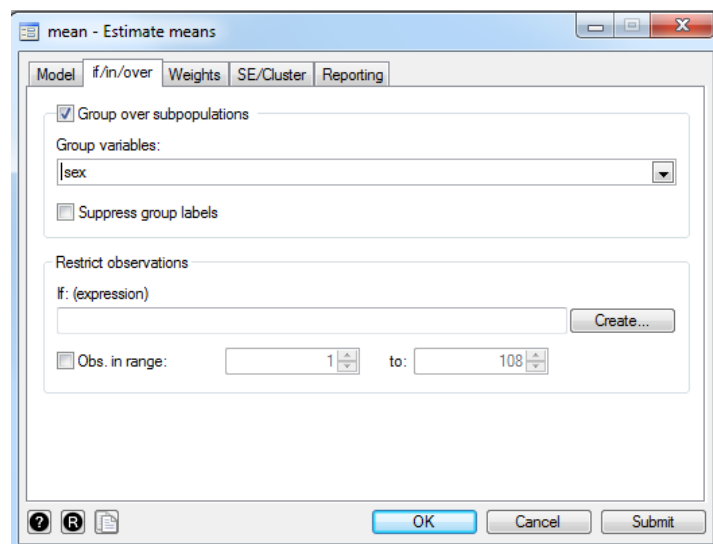
This will produce the following command: "mean measure, over(sex)".

Now, we will obviously be interested in saving this set of commands, so that the results can be reproduced afterwards. This can be done in a do-file.

You may open a new do-file by clicking the do-file icon, as shown below. When you have the new file open, you may enter the relevant commands:
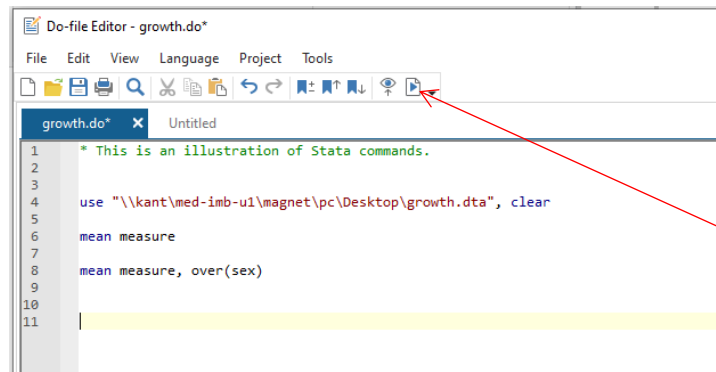
Lines starting with * are used for comments.
The first command her (use "\\kant ....) is loading the relevant data file.
You run the commands by marking them by the mouse and click the icon indicated by the red arrow above.

The idea is now to fill in all commands that are relevant for your analysis in this file and save it for later use. You will find the file "growth.do" in Timeplan.

Let's have a look at some other useful commands.

First, our measure variable is actually height, measured in inches. We would like to create a new variable 'height' that gives height in cm. We do that by the generate (or gen for short) command:

        gen height=2.54*measure

Assume we are interested in relating our measure (of growth) to age. We would like to start out by producing a scatter plot. The following command will produce the relevant plot, which is shown below:
   twoway (scatter height age) [or simply 'scatter height age']

You can of course again produce this command by use of the Meny (Graphics – Twoway graph. Click 'Create' and choose the correct plot and variables). Note that the Stata Menu is not always easy to navigate in, so once you learn how to use the commands (and do-files) this is much quicker and easier.
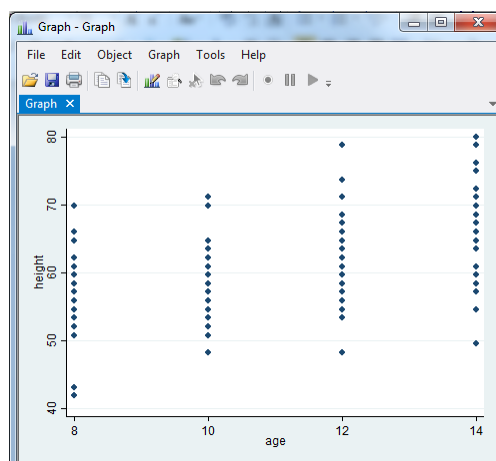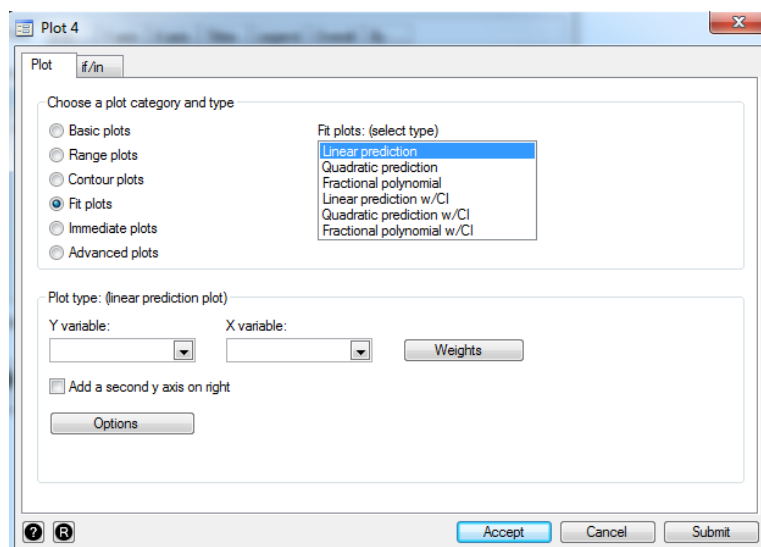
We can also impose a regression line (which we will come to later in the course) here by:

twoway (scatter height age)(lfit height age)
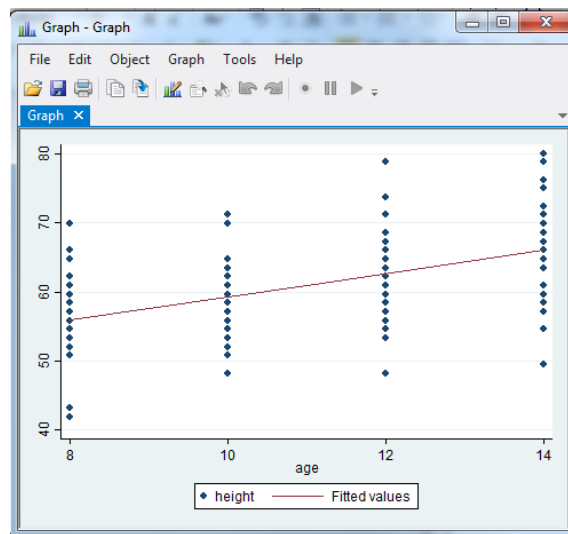
If you want to produce this plot by using the Menu, you will have to choose a fit plot instead of the default "basic plot":

Furthermore, we can be interested in making the same plot, but with separate regression lines for each gender. All of this can be done within the twoway plots, which are pretty flexible. However, the commands become longer ... (see the file growth.do for this specific command).

Stata has a help-function that is quite good. If you type e.g. "help twoway" in the Command window, you will have a lot of information about how to produce different plots.

Formal command structure

Although we have seen examples of the opposite above, most Stata commands follow the same general form:

[by *varlist*:] *command* [*varlist*] [if *exp*] [in *range*] [*,options*]

To illustrate this, let us have a look at a very basic command; summarize. This command gives summary statistics of the variables listed (assuming continuous variables).
Instead of the mean command we used previously, we could use the summarize command to get descriptive measures of height:

summarize height

If we want it by gender we do the following:

by sex: summarize height

However, for this command to work, the file has to be sorted by sex. This can be done simply by

sort sex

Or if we want to see males only (sex=1):

summarize height if sex==1

Notice the logical operator "==" that is being used for comparisons. Remember that we used "=" in the generate command, for assigning values to a variable.

Next, say we are, of some reason only interested in males above 50cm:

summarize height if sex==1 & height>50

Finally, we can look at a subset of the full data file by specifying which lines we are interested in. Say we are interested in the first 50 records:

summarize height in 1/50

The following logical operators are allowed:

== equal to (note that we use 2 equality signs)
~= not equal to
>= larger than or equal to
> larger than
< less than
& and
| or