Stata Introduction*

Antonio Pacifico
Federica Romei

February 29, 2012

The notes are for the first session and have been prepared in part from "Introduzione a Stata" of Professor G. Vallanti.

Contents

1 A Brief Overview

Stata is a general-purpose statistic software for data management, statistical analysis and graphic analysis.

There are a lot of resources available to help you learn and use Stata:

- Official website Resources and Support (http://www.stata.com/support/);
- The online (search) guide and the offline (help) guide;
- The official documentation (http://www.stata.com/bookstore/documentation.html);
- The non-official resources, for eg the one of the California University (http://www.ats.ucla.edu/STAT/stata);
- Stata-Journal;
- The richest non official collection of commands on the web site IDEAS (http://ideas.repec.org/s/boc/bocode.html).

Stata is full-featured statistical programming language. It has traditionally been a command-line driven package that operates in a graphical (windowed) enviroment. Stata version 11 contains graphical interface (GUI) for command entry.

1.1 Stata Layout

To start the programm follow the instruction written below
(start->program->Stata).
Gui has 4 windows:

1. **Command**: it is the window where you can write the commands (usually it is at the bottom);
2. **Results**: it is the window where you can see the results (it is usually the biggest);
3. **Variables**: it is the window where variables are displayed (it is usually at the left-bottom);
4. **Review**; it is the window in which you can see the commands that have been typed during the session (Left-Top)

If you want to close a session you have to type *exit* in the command window.

1.2 Memory

Only files with the extensions .dta can be uploaded directly in memory by Stata. These files are created by Stata and are organized in matrix form. Every row is an observation, every column a variable.

When you upload a new dataset first you have to clean the memory using the command

clear

then you have to "allocate" the right memory with the command

set memory #[b|k|m|g] [, permanently].

If, for example, the dataset is 2MB you need to write

set mem 4m

1.3 Guide(*help*)

The command *help* is the most usefull for a Stata beginner. You can write help and open a guide organized in category or you can write help followed by a command and you can look up command details.

If you type help command you will visualize a window divided in 6 part:

- **Title**: you will find command title;
- **Syntax**: you will find command syntax. Command syntax is standard and can be rapresented in this way (the square parenthesis are for the optional part):

$$\begin{array}{ccccc} command & varlist/filename & [restriction] & [,] & [options] \\ 1 & 2 & 3 & 4 & 5 \end{array}$$

1. *command* is Stata command;
2. *varlist* or *filename*: after a command usually you have to type the name of a variable or of a file;
3. *restrictions*: it is helpfull if you need to use only a part of your varlist or file;
4. *,* : comma divides the compulsory part from the optional one;
5. *options*: after the comma you have to specify the options.

- **Description**: it describes the features of the command;
- **Options**: in this part all the options you can use with the command are listed and described;
- **Examples**: you can find some examples helpfull to use the command;
- **Also see**: you can find link to other commands similar to the command you type.

If you don't know a specific command you can look for it putting some key word after the command *search*. Like

search linear regression.

Moreover it is usefull to look on the web typing on google stata and some keyword. You will find a lot of examples.

2 First practice

2.1 Basic Tools

2.1.1 Set directory

Once you open Stata, it would use the default directory. If you want to change directory you have to write

cd "name directory".

In order to check you have written the exact directory, type

pwd

and Stata will visualize the directory you are using. After you can write

ls

and Stata will visualize the files contained in the directory.

2.1.2 Log files

Log files save in a .txt format all the output and command of a Stata-working-session. You have to write the following command to create a log files

log using filename.txt.

If you are willing to save the log file in another directory (different from the one you set before) you have to specify the directory name, i.e.

log using " c:/directory/filename".

If you need to overwrite a log file previously created, add the option replace after the comma, i.e.

log using filename, replace.

To stop temporary and start again the log file you have to type *log off* and *log on* respectively.

If you want to close the log file, you can write

log close.

2.1.3 Do File

A do file is a text file in which you can write and execute stata commands. To open a .do file you need to click on the "blocknotes" in the task bar.

2.1.4 Upload the dataset

If the dataset is a .dta format, once you set the right directory, you can write

use filename, clear.

to upload the dataset.

If the dataset has a format different from .dta you need to use the command

insheet using filename, (separator) clear

If the dataset is saved as a .csv file (comma separated variables) you need to write

insheet using filename, comma clear

if instead it has been saved as a .txt file you have to write

using insheet filename, tab clear.

2.2 Let's Start

We will use the dataset ceosal1.dta. The first thing to do is to set the directory and open a log file (see above).

Then you have to open the dataset typing

use ceosal1, clear.

2.2.1 Data Description and visualization

Once you upload the dataset write the command

describe

and stata will show all the data in memory. You can see description of the variables in memory.

```
Contains data from ceosal1.dta
  obs:          210
 vars:           12                          25 Feb 2010 16:06
 size:        7,140 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
pcsalary        int    %8.0g                  % change salary, 89-90
sales           float  %9.0g                  1990 firm sales, millions $
roe             float  %9.0g                  return on equity, 88-90 avg
pcroe           float  %9.0g                  % change roe, 88-90
ros             int    %8.0g                  return on firm's stock, 88-90
indus           byte   %8.0g                  =1 if industrial firm
finance         byte   %8.0g                  =1 if financial firm
consprod        byte   %8.0g                  =1 if consumer product firm
utility         byte   %8.0g                  =1 if transport. or utilties
lsalary         float  %9.0g                  natural log of salary
lsales          float  %9.0g                  natural log of sales
salary          int    %8.0g                  1990 salary, thousands $
-------------------------------------------------------------------------------
Sorted by:
```

In order to have a better understanding of a variable you can write the command

codebook varname

in our specific case we will type

codebook indus

and stata will display the range (in our case 0,1), the label (in our case type of firms) and the frequency of this variables.

```
--------------------------------------------------------------------------------
indus                                                       =1 if industrial firm
--------------------------------------------------------------------------------

                  type:  numeric (byte)

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  1/210

            tabulation:  Freq.  Value
                           142  0
                            67  1
                             1  .
```

2.2.2 Qualifiers

The qualifiers *if* and *in* are very usefull.

If you type if at the end of a command, before the comma, you are able to select a part of your data. If instead you write if, you will be able to select a subset of your dataset, specifing the position.

We can do some example with our dataset

list roe if indus==1

which means list the variables price if the variable foreign is equal to 1.

list roe in 1/10

Stata would list the first 10 observations of the variable price.

2.2.3 Summarize

If you need to have the basic statistic of your variables you can write the command

summarize

and stata will summarize the number of observation, mean, standard deviation min and max of all the variables in your dataset in a table format.

In our case you will have:

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    pcsalary |       209     13.2823    32.63392        -61        212
       sales |       209    6923.795    10633.27   175.2001    97649.9
         roe |       209    17.18422    8.518514         .5   56.30004
       pcroe |       209    10.80048    97.21943  -98.90008        977
         ros |       209    61.80383    68.17705        -58        418
-------------+--------------------------------------------------------
       indus |       209    .3205742    .4678178          0          1
     finance |       209    .2200957    .4153057          0          1
    consprod |       209    .2870813    .4534861          0          1
     utility |       209    .1722488    .3785031          0          1
     lsalary |       209    6.950386    .5663741   5.407172   9.603868
-------------+--------------------------------------------------------
      lsales |       209    8.292265     1.01316   5.165928   11.48914
      salary |       209     1281.12    1372.345        223      14822
```

If you want to see the statistics of a single variable you can type

summ varname

in our case

summ roe, det

and you would have the following output:

```
   return on equity, 88-90 avg
-------------------------------------------------------------
      Percentiles      Smallest
 1%     2.100001             .5
 5%     6.800005       1.900001
10%     8.900005       2.100001       Obs                 209
25%     12.40001       2.900002       Sum of Wgt.         209

50%         15.5                      Mean           17.18422
                        Largest       Std. Dev.      8.518514
75%           20       44.40004
90%     26.80002           44.5       Variance       72.56508
95%     35.10002       48.10004       Skewness       1.560821
99%         44.5       56.30004       Kurtosis       6.678557
```

2.2.4 Tables

You can create descriptive tables using the command: tabulate, table and tabstat. tabulate allows you to create a oneway or twoway table as shown in the syntax:

- Oneway: *tabulate varname [if] [in] [weight] [, tabulate1_options]*
- Twoway: *tabulate varname1 varname2 [if] [in] [weight] [, options].*

We can write
tabulate finance

```
   =1 if |
financial |
     firm |      Freq.     Percent        Cum.
------------+-----------------------------------
        0 |        163       77.99       77.99
        1 |         46       22.01      100.00
------------+-----------------------------------
    Total |        209      100.00
```

or
tabulate finance indus

```
     =1 if |
 financial | =1 if industrial firm
      firm |         0          1 |     Total
-----------+----------------------+----------
         0 |        96         67 |       163
         1 |        46          0 |        46
-----------+----------------------+----------
     Total |       142         67 |       209
```

The command table allows you to choose the content of the table. Table syntax is
table rowvar [colvar [supercolvar]] [if] [in] [weight] [, options].
You can type
table utility, content (mean roe sd roe)

```
----------------------------------
=1 if     |
transport |
. or      |
utilties  |  mean(roe)     sd(roe)
----------+-----------------------
        0 |   18.38671    8.762107
        1 |   11.40556    3.529546
----------------------------------
```

The command tabstat joins the principal characteristic of summ and tabulate allowing for greater flexibility.

tabstat varlist [if] [in] [weight] [, options]

In our example we can write

tabstat roe ros, stat(mean) by (utility).

```
 utility |       roe       ros
---------+--------------------
       0 |  18.38672  63.02312
       1 |  11.40556  55.94444
---------+--------------------
   Total |  17.18422  61.80383
------------------------------
```

2.2.5 Test Hypothesis

We can now test some hypothesis using the L.L. Central Limit Theorem. Suppose we think that transport firms have on average same roe as non transport firms against the hypothesis that transport firms have on average less roe than non-transport ones. This means test

$$H_0 \ \Delta roe = roe_{Transport} - roe_{non\ transport} = 0$$

$$H_1 \Delta roe = roe_{Transport} - roe_{non\ transport} < 0.$$

We can use the sample counterpart to test this hypothesis. We take the sample mean of roe for the transport firms and for the non transport ones.

Under regolarity condition, (observations are i.i.d., $E(roe_{i,transport}) < \infty$, $E(roe_{i,non\ transport}) < \infty$, $Var(roe_{i,transport}) < \infty$ and $Var(roe_{i,non\ transport}) < \infty$) we can apply the L.L. Central Limit Theorem, which means

$$\Delta \hat{roe} = \frac{roe_{Tr\bar{a}nsport} - roe_{Non\ \bar{T}ransport}}{\sqrt{Se(roe_{Transport})^2 + Se(roe_{Non\ Transport})^2}} \xrightarrow{D} N(0,1)$$

where $roe_{Tr\bar{a}nsport}$ and $roe_{Non\ \bar{T}ransport}$ are sample mean of roe of transport and non transport firms respectively.

Typing in Stata the command

$$mean\ roe\ if\ utility == 1$$

and

$$mean\ roe\ if\ utility == 0$$

we can visualize the following output and have all the possible information we need to compute $\Delta\hat{roe}$, i.e.

```
 mean roe if utility==1

Mean estimation                     Number of obs    =      36

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
         roe |   11.40556   .5882577      10.21134    12.59979
--------------------------------------------------------------

 mean roe if utility==0

Mean estimation                     Number of obs    =     173

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
         roe |   18.38672   .6661706      17.07179    19.70164
--------------------------------------------------------------
```

Now we are able to compute $\Delta\hat{roe}$, that is

$$\frac{11.40 - 18.38}{\sqrt{0.588^2 + 0.666^2}} = -7.85.$$

We know that p-value of $\Delta\hat{roe}$ is very closed to zero and that -7.85<-1.64, hence we can easily reject the Null Hypothesis.

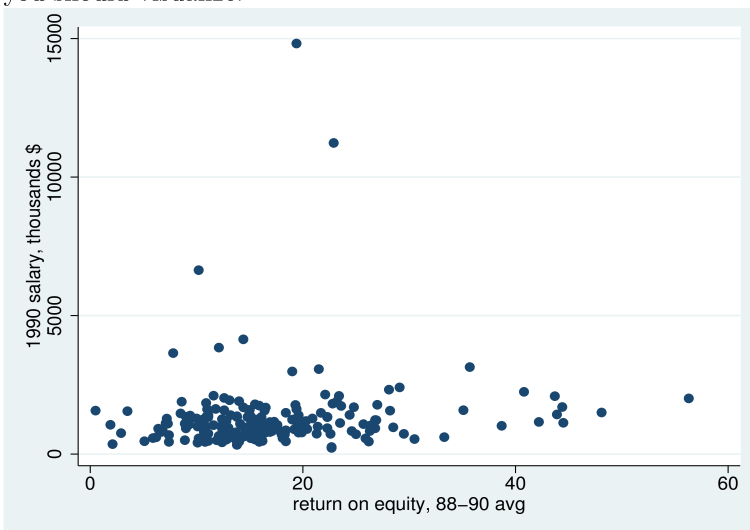
2.2.6 Graph

We can use the graph in order to establish if there is a relation between roe and salary of Ceo. We argue, indeed, that the Ceo wage is higher when the roe is higher.

If you write

$$scatter\ salary\ roe$$

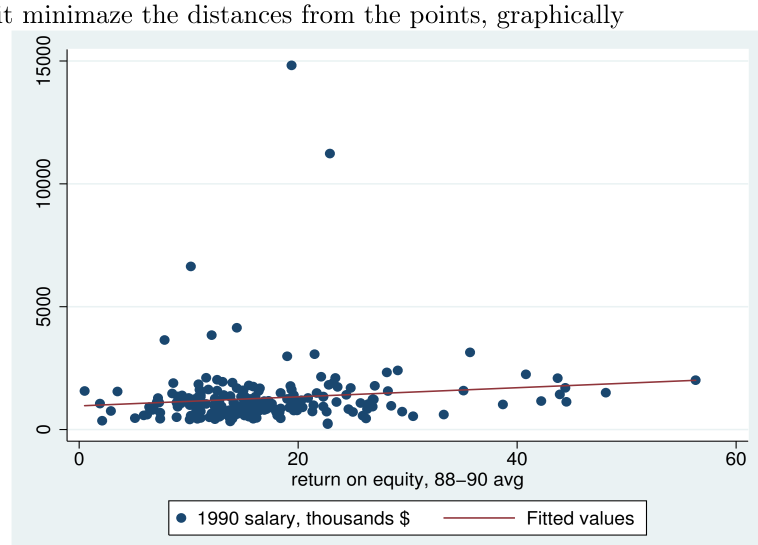
you should visualize:



It seems to be a slightly positive relation. To be more sure we can use the command

$$twoway\ (scatter\ salary\ roe)\ (lfit\ salary\ roe).$$

Lfit minimaze the distances from the points, graphically



The relation between salary and roe seems to be positive.

2.2.7 Regression

To be sure about the relation between salary and roe we can run a regression which means to build a model as

$$salary = \beta_0 + \beta_1 roe + u$$

where u is the error term.

Using the command

reg salary roe

you can visualize this output

```
Linear regression                                      Number of obs =     209
                                                       F(  1,   207) =    7.34
                                                       Prob > F      =  0.0073
                                                       R-squared     =  0.0132
                                                       Root MSE      =  1366.6

------------------------------------------------------------------------------
             |               Robust
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         roe |   18.50118   6.829445     2.71   0.007     5.036991    31.96536
       _cons |   963.1913   121.1062     7.95   0.000     724.4315    1201.951
------------------------------------------------------------------------------
```

The first column shows the coefficient estimates. The last coefficient in Stata is always the constant (the intercept). In this case roe seems to have a positive correlation with the salary of Ceo. Stricty speaking a unitary increse in roe seems to increase the wage of ceo of 18 $.

Second column rapresent the Standard Error of β- coefficient.[1]

The contents of the other columns will be explained in the next practice.

[1] Standard Error is the "Sample Standard Deviation" divided by square root of n, where n rapresents the number of observations in the sample.