Getting started with the Stata

1. Begin by going to a computer lab.

Stata is installed on all campus computers. You can use either a PC or a Mac to do your assignments, Stata should work the same regardless.

For a list of campus computer labs, see:

http://www.columbia.edu/acis/facilities/labs/locations/

2. Getting started – Your first Stata session.

Begin by starting Stata on your computer.

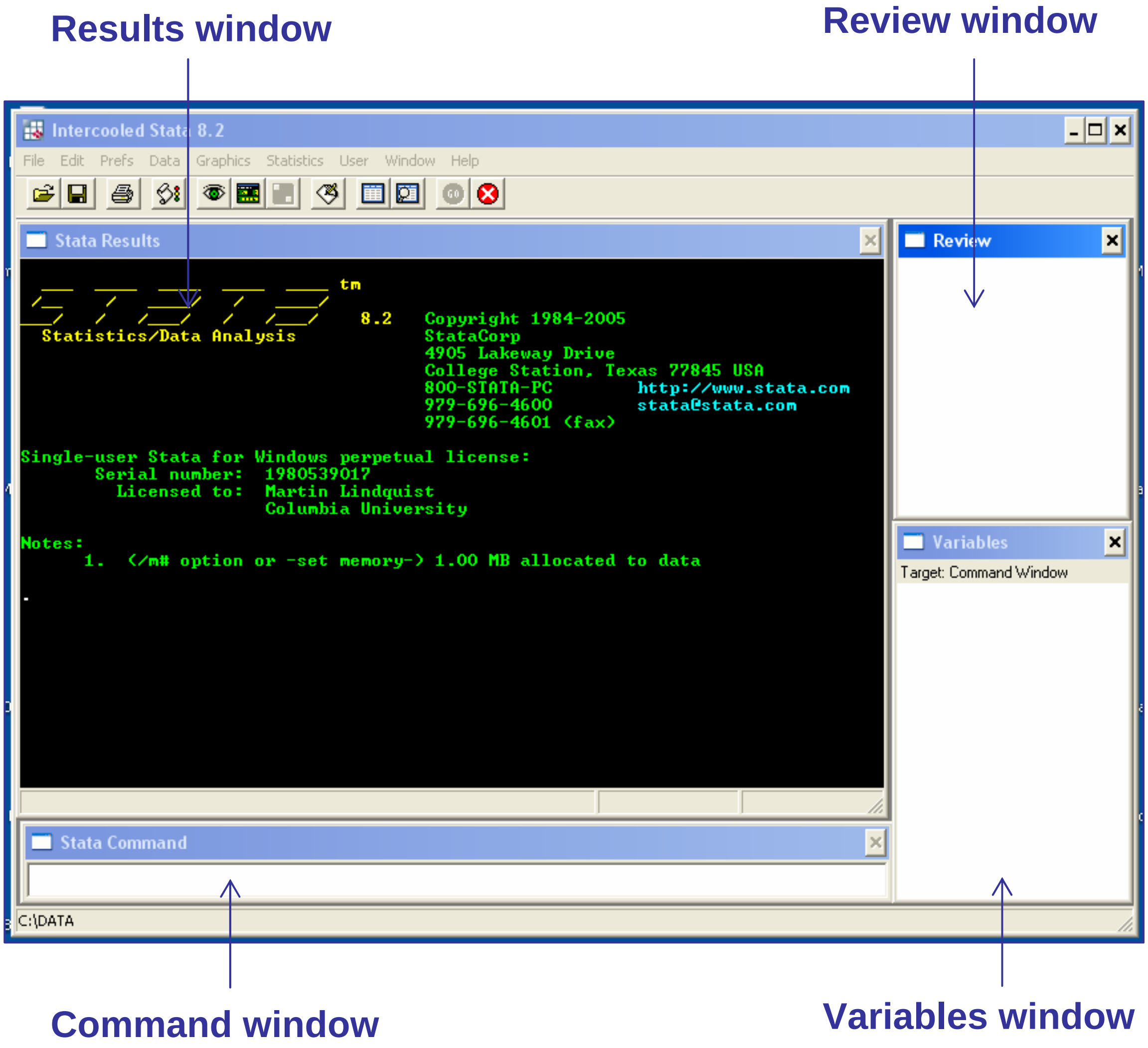
Using a PC:

1. Click on the Start menu
2. Click on Programs
3. Click on Statistical Applications
4. Click on "Intercooled Stata 10". Alternatively click on “Special Edition STATA 10”, if this is an option on your computer. (Note that the version of STATA that is available may differ between computers. You can use any version that is available without a problem.)

Using a MAC:

1. Click on the hard drive icon on the desktop (located top right).
2. Click on the "courseware" icon
3. Click on the Stata Icon.

Once you have started Stata, you will see a large black window that is surrounded by a number of smaller white windows; see Figure 1 on the next page for an example. The large window is called the **Results window**. Here all the results from your Stata session, except graphs which are shown in a separate window, will appear. New commands are entered in the **Command window**. The **Review window** records all previous Stata commands you enter. You can repeat a previous command by simply double-clicking the relevant command in the Review window. Finally, the **Variables window** shows a record of all the variables in the data set that you are currently using.



```
  ___  ____  ____  ____  ____ tm
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   8.2   Copyright 1984-2005
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
                                      College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user Stata for Windows perpetual license:
       Serial number:  1980539017
         Licensed to:  Martin Lindquist
                       Columbia University

Notes:
      1.  (/m# option or -set memory-) 1.00 MB allocated to data

.
```

Figure 1. When you start Stata the following windows should appear. The large window is called the Results window. Here all the results of your Stata session, except graphs which are shown in a separate window, will appear. New commands are entered in the Command window. The Review window records all previous Stata commands you enter. You can repeat a previous command by simply double-clicking the relevant command in the Review window. Finally, the Variables window shows a record of all the variables in the data set that you are currently using.

Always begin your STATA session by placing a USB key or equivalent into the computer (This is often the E or F drive). Alternatively, you can save directly to the computers hard drive (the C drive) if the computer permits it. In this tutorial we assume we are saving onto a key placed in the E drive. Be sure to make the appropriate edits in the code if you are saving to another drive.

Next, type the following command in the command window:

```
log using "e:/filename.log"
```

where “filename” can be any name you chose to give your file. This command opens a log file called “filename.log” where all your work is saved. This command essentially stores all your commands and results (including errors) in a log file. This is the file that at the end of your session you are going to print and hand in for homework. **If you don't start a log file you will lose your work.** This file can be found on the USB key (the E drive) as we started the command using “e:/”. If you are saving to another drive make sure you use the appropriate drive name (e.g. “a:”, “c:” or “f:”). The file will often contain a fair share of unwanted output and should not be submitted without further editing. You can edit the file using any word processing program, such as Notepad, WordPad or Microsoft Word.

In naming your log file there are a few rules that you need to follow:

- Your filename should be a single word (no spaces)
- Your filename should be made up of letters and numbers - no symbols
- The file name is case-sensitive so remember the case you use: i.e. a file named “Log1” is different from one named “log1”.

Once you've finished your Stata session you can type the following command:

```
log close
```

This stops the log from recording.

You can then exit Stata. To exit Stata from the command line you have two choices. Either exit without saving the data by typing

```
exit, clear
```

or save the data and then exit by typing, for example,

```
save “e:/yourdata”
exit
```

Afterwards, open your USB key and double click on the log file. This will open the file in Notepad or some other similar program. Here you can edit your mistakes and print your work from this program.

3. Data Management

A. Entering Data Directly into Stata

Suppose we want to enter the following set of data into STATA

Name	**Age**	**Number of siblings**
Bob	27	2
Sue	33	1
BobBill	21	0
John	56	4

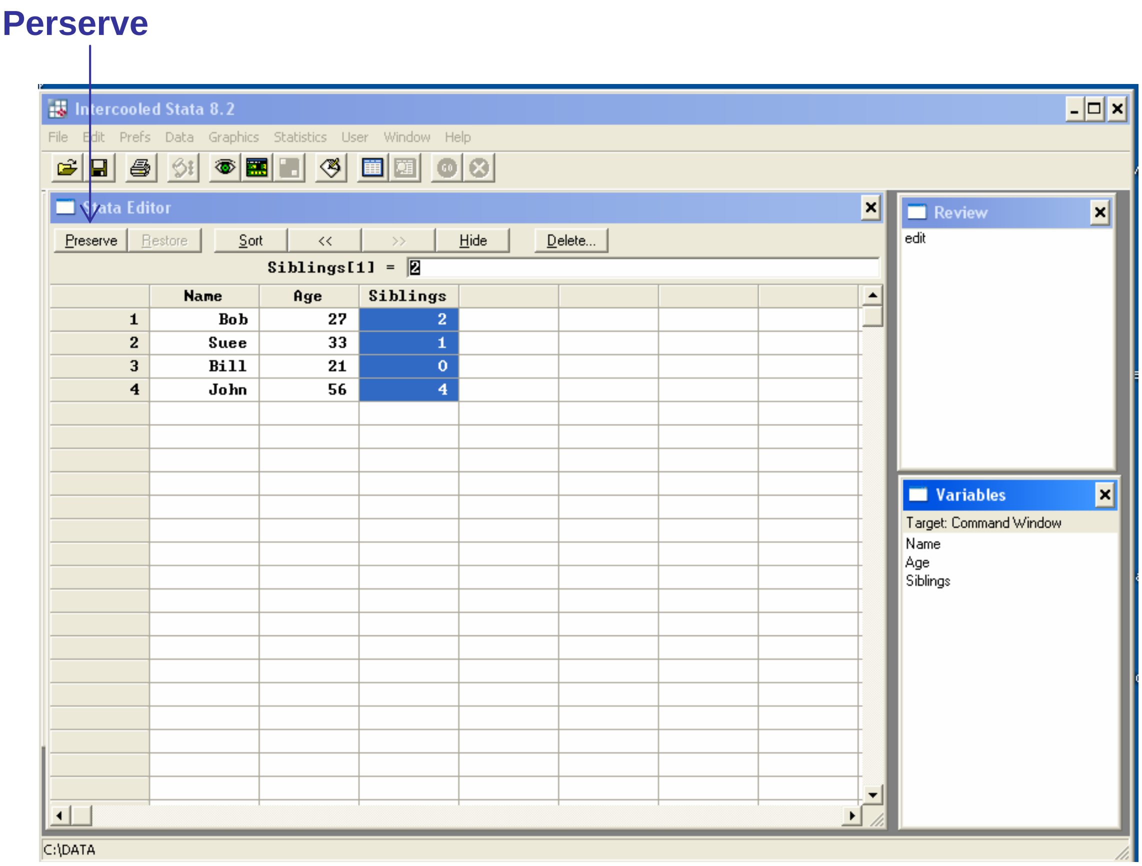
To enter the data set directly by hand, type edit from the command window. This opens up a spreadsheet where you can type in your data; see Figure 2 for an example. Use the mouse or arrows to move around the spreadsheet and press enter after each entry. On the first row, first column of the white area of the spreadsheet write Bob and press enter. In the second column write 27 and in the third column write 2.

Continue to the second row, in the first column write Sue. In the second column write 33, etc…..

Once you are finished typing in the values, the data in your spreadsheet will consist of 4 rows and 3 columns. The first column will be named var1, the second column var2 and the third var3. You can rename the variables by double-clicking on their columns. A box will appear where you can write the new name. For example, if you double click on the second column a box appears and you can change the name from var2 to Age.

You can correct any mistakes you make by simply clicking on the entry in the table that you want to change. When you are done entering your data, click on the **Preserve** button on the upper left hand side of the editor window, then close the Editor window by pressing the **X** in the upper right side of the window. If you want to go back and make more changes, simply type edit again from the command window.

Exit Stata or type clear before continuing with the next section



Perserve

edit

Siblings[1] = 2

	Name	Age	Siblings
1	Bob	27	2
2	Suee	33	1
3	Bill	21	0
4	John	56	4

Name
Age
Siblings

Figure 2. To enter the data set directly by hand, type edit from the command window. This opens up a spreadsheet where you can type in your data.

B. Creating and Reading Datasets from an External File

You should use the command infile if you want to read data from a file. The data in the file must be separated by spaces and all the string variables (i.e., non-numeric characters) must either consist of one word or be surrounded by quotation marks. A period (.) is understood to mean a numeric missing value; double quotes ("") to mean a missing string variable.

Suppose we want to read an ASCII file saved on your USB key named “numbers.txt” which contains data on three numerical variables we want to name x, y and z. The command:

```
infile x y z using “e:/numbers.txt”
```

will read the file “numbers.txt” from the USB key and name the three numerical variables x, y and z.

You can use Notepad or some other word processing program to create this type of external file. When entering the data make sure to only enter the data and not the variable names. If we wanted to make a file containing the example discussed in the previous section, the file should look like this:

```
Bob 27 2
Sue 33 1
Bill 21 0
John 56 4
```

Make sure to save the file as an ASCII or text file. Suppose we have saved the data above onto our USB key in a file which we called “data.txt”. In this case it is important to note that the first column of data consists of a string variable. We have to forewarn Stata about this otherwise it will try to read the names as numbers. Prior to the variable Name we need to write str4. This tells Stata that the next variable, Name, is a string and the longest word is 4 characters long.

In the Stata command window type:

```
infile str4 Name Age Siblings using "e:/data.txt"
```

Stata will now recognize three variables Name (containing Bob, Sue, Bill and John), Age (containing 27, 33, 21 and 56) and Siblings (containing 2, 1, 0 and 4).

C. Saving and Reusing Datasets for your next session

Once you have entered your data set it is possible to save it in Stata for a later session. The command:

```
save “e:/newfile”
```

saves the data currently in memory on your key as a file named “newfile.dta”.

If you want to retrieve the data set at a later time, simply type:

```
use “e:/newfile”
```

4. Structure of Stata Commands

In general, you tell Stata what you want to do by typing commands in the Command Window. The general form of all Stata commands is

```
command variables, options
```

- **command** tells STATA which command you want to execute.
- **variables** (each variable name separated by a space) are the list of variables used to perform the command.
- **options** tells Stata how you want to execute the command.

A. Some Basic Stata Commands

Suppose we use the data set of ages and number of sibling that we introduced in the previous section. To list all the values in the data set simply type:

```
list
```

This will result in a table containing the data which can be seen in the Results window. If we are only interested in the names of the people in the data set we can write:

```
list Name
```

This command will result in a list of the names contained in the data set. The variable, Name, is case-sensitive so be sure to enter it correctly.

To obtain summary statistics about the age of the people contained in the data, such as the mean and standard deviation, simply type:

```
summarize Age
```

This command produces summary statistics only for the variable Age. If we had simply written the command summarize, without specifying Age, we would have gotten summary statistics for all the variables.

If also want to calculate the median we need to add the option detail. The following command

```
summarize Age, detail
```

will calculate both the mean and median of the variable Age, as well as some other interesting information.

B. More Stata Commands

Here are some commands that you may find useful in this course. We will discuss them in more detail at a later time.

by: repeat operation for categories of a variable
clear: clears previous dataset (if this doesn't work use drop _all)
correlate: correlation between variables
describe: briefly describes the data (#of obs, var names, etc.)
drop: eliminate variables from memory
edit: enter data, or alter an existing data set
exit: leave stata
generate: creates new variables (e.g., generate years = close - start)
graph: general graphing command (this command has many options!)
help: online help
if: lets you select a subset of observations (e.g., list if radius >= 3000)
infile: read non-Stata-format dataset (ASCII or text file)
input: type raw data
list: lists the whole dataset in memory
log:save or print stata output (except graphs)
lookup: keyword search of commands, often precursor to *help*
plot: text-mode (crude) scatterplots
predict: calculated predicted values (y-hat)
regress: regression
sort: sorts observations from smallest to largest
summarize: produces summary statistics, has detail option
use: retrieve previously saved Stata datasets

5. A guided exercise

Make sure you have read the material in this tutorial. You will need it to perform this exercise.

(a) Start Stata. If you are continuing a previous session write clear in the command window, to erase all old data.

(b) Create a log file called assignment1.log. Use the following command:

```
log using "e:/assignment1.log"
```

Note if you are not using the E drive you will need to make the appropriate changes to the command above.

(c) Enter the data below using the edit command.

After you type the edit command enter the data below into the spreadsheet.

Brief background on data: Climatologists interested in flooding gather statistics on the daily rainfall in various cities. The following data set gives the maximum daily rainfall (in inches) for the years 1941 to 1964 in South Bend, Indiana.

Data:

1.88 2.23 2.58 2.07 2.94 2.29 3.14 2.14 1.95 2.51 2.86 1.48 1.12 2.76
1.48 1.12 2.76 1.50 2.99 3.48 2.12 4.69 2.29 2.12

Be sure to enter the data one observation at a time (vertically in one column):

To name the column, double click on the grey cell at the top titled **var 1** and type the appropriate variable name. In this case we will name the variable rain. When you have entered all your data and named the variable, click on the **Preserve** button and close the window by clicking on the **X** in the upper right hand corner.

(d) Save the data set on your USB key as **rainfall**. Use the command:

```
save “e:/rainfall”
```

(e) Compute the values of the mean and median. The command should be:

```
summarize rain, detail
```

(f) Close the log file by using the command

```
log close
```

(g) Open the log file on your USB key (in this case “assignment1.log”) by double clicking on its icon. Edit and print your work.

(h) Hand in your edited log file.