ECON 6300/7320/8300: ELEMENTS OF ECONOMETRICS
Dr. Dong-Hyuk Kim
Tutorial 1: Stata and Basic Statistics

At the end of this tutorial you should be able to

- use Stata to read, manipulate and save data and workfiles
- use Stata to compute descriptive statistics
- use Stata to conduct hypothesis tests concerning a population mean

1 Introduction of Stata

1.1 Starting Stata

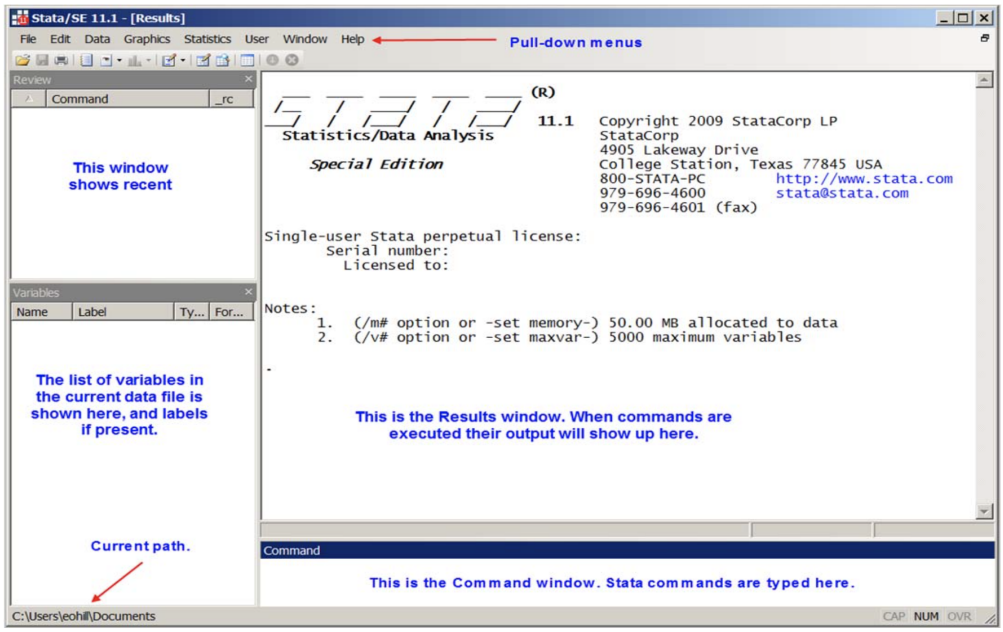
Before solving the problems, let's look around the software. Stata can be started several ways. First, there may be shortcut on the desktop that you can double- click. For the Stata/SE Release 11 it will look like

Earlier versions of Stata have a similar looking Icon, but of course with a different number. Alternatively, using the Windows menu, click the **Start** → **All Programs** → **Stata 11**. A second way is to simply locate a Stata data file, with `*.dta` extension, and double-click.

1.2 Opening Display

Once Stata is started a display will appear that contains windows titled.

Command:	this is where Stata commands are typed
Results:	output from commands, and error messages, appear here
Review:	a listing of commands recently executed
Variables:	names of variables in data and labels (if created)



Across the top are Stata pull-down menus. We will explore the use of many of these. In the lower left-hand corner is the current path to a working directory where Stata saves graphs, data files, etc. We will change this in a moment.

1.3 Exiting

To end a Stata session click on **File** → **Exit**. Alternatively, simply type

```
exit
```

in the **Command** window and press **Enter**.

1.4 Stata Data Files

Stata data files have the extension `*.dta`. These files should not be opened with any program but Stata. If you locate a `*.dta` file using double-click it will also start Stata. For the course, Econ 7310, data files and problem sets for each tutorial session will be available at the course webpage. For the exercise below we will use `consumption.dta` and `fultonfish.dta`. You should download the datasets into a convenient directory. To change the working directory use the pull-down menu **File** → **Change Working Directory**. In the resulting dialog box navigate to your preferred location and click **OK**. to this location type Stata will show the implied command

```
cd "C:\data\poe4stata"
```

This can be entered into the **Command** window and press **Enter**.

2 Answer Key

Problems:

1. The text file `consumption.dta` contains observations on the weekly family consumption expenditure (`CONS`) and income (`INC`) for a sample of 10 families.

   (a) Read the data into Stata
   (Answer)
   With Stata started, change your working directory to the where you have stored the Stata data files. In the **Command** window type

   ```
   use consumption
   ```

   and press **Enter**. If you have a data file already open, and have changed it in some way, Stata will reply with an error message.

   ```
   no; data in memory would be lost
   r(4)
   ```

   If you click on `r(4)`; you will be able to read the error message in a **Viewer box**. Sometimes this is helpful. To close the Viewer box click the **X**.

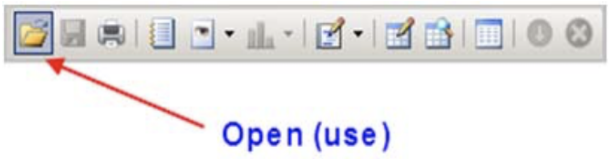
   This feature will prevent you from losing changes to a data file you may wish to save. If this happens, you can either save the previous data file [more on this below], or enter the command

   ```
   clear
   ```

   The clear command will erase what is in Stata's memory. If you want to open the data file and clear memory, enter

   ```
   use consumption, clear
   ```

You can also open a Stata data file using the tool bar click the **Open (use)** icon on the Stata toolbar.

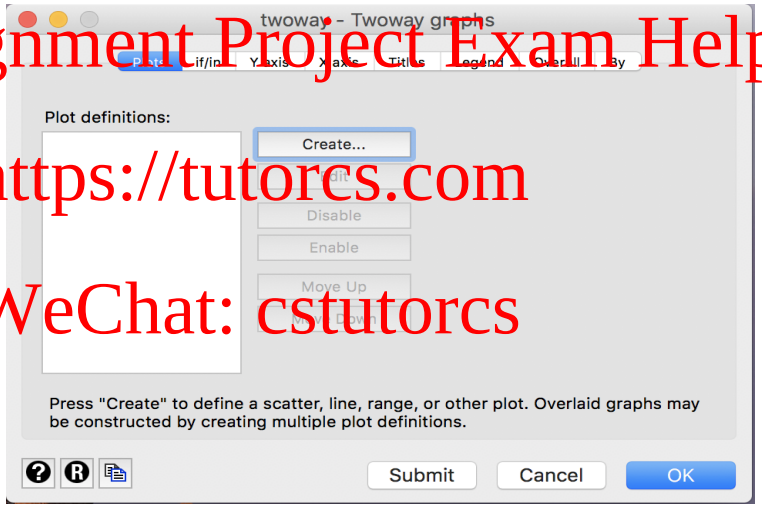


Locate the file you wish to open, select it, and click **Open**. In the Review window the implied Stata command is shown.

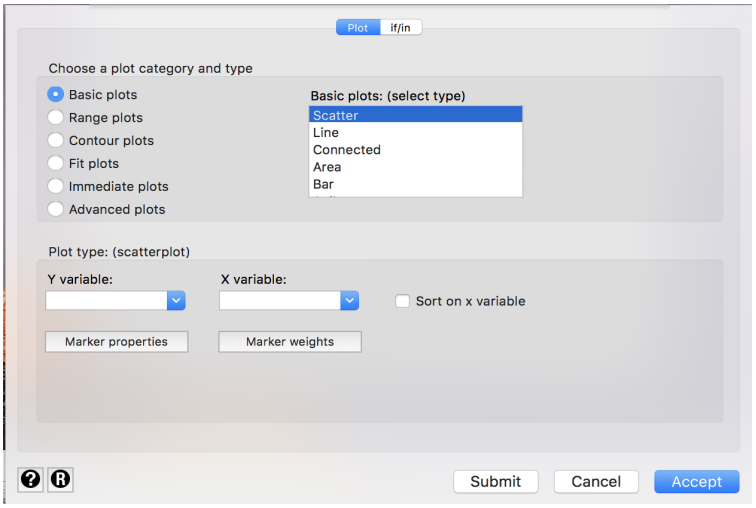
```
use "/Users/.../consumption.dta", clear
```

In Stata opening a data file is achieved with the use command. The path of the data file is shown in quotes. The quotes are necessary if the path name has spaces included. The option `clear` indicates that any existing data is cleared from memory. □

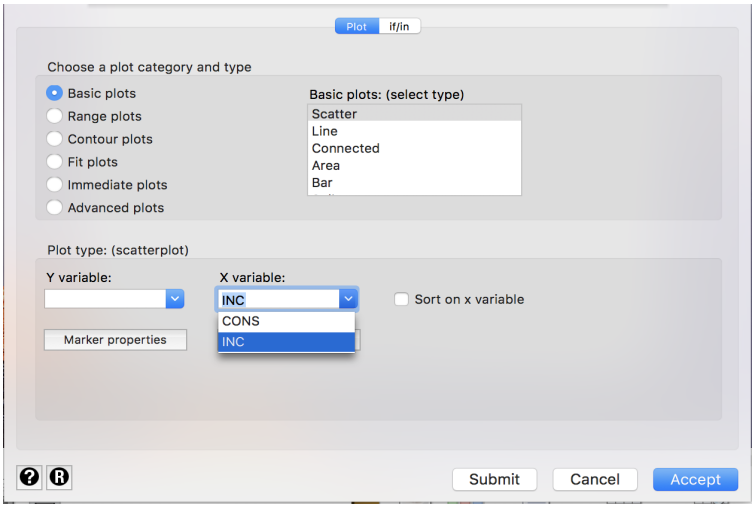
(b) Draw a scatter diagram of `CONS` against `INC`.
(Answer)
From the pull-down menu, click on **Graphics → two way graphs (scatter, line, etc)**



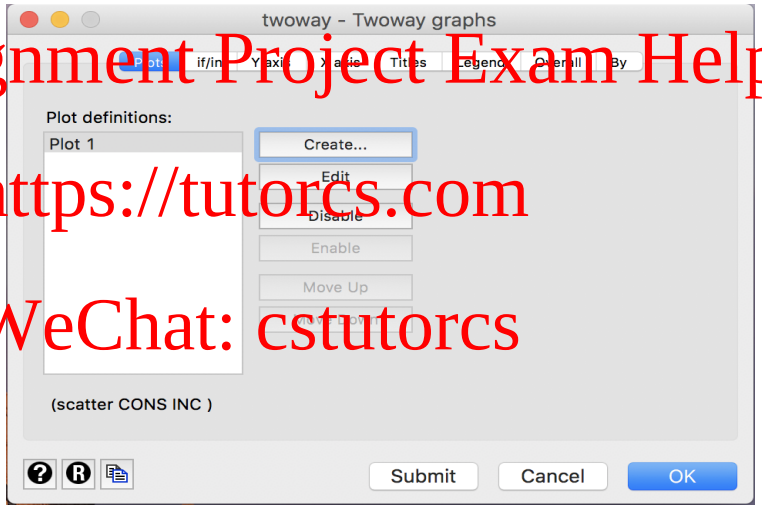
Then, click on **Create** and choose **Basic plots → Scatter**



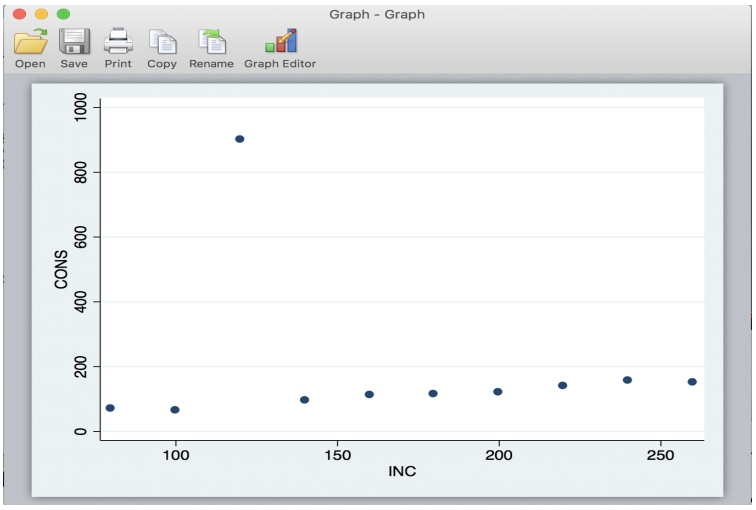
Then, under choose `INC` under **X variable** and choose `CONS` under **Y variable**; Then, click on **Accept**.



Then, you will see that "Plot 1" is ready under **Plot definitions:**.



By clicking on **OK**, you will see the scatter diagram;

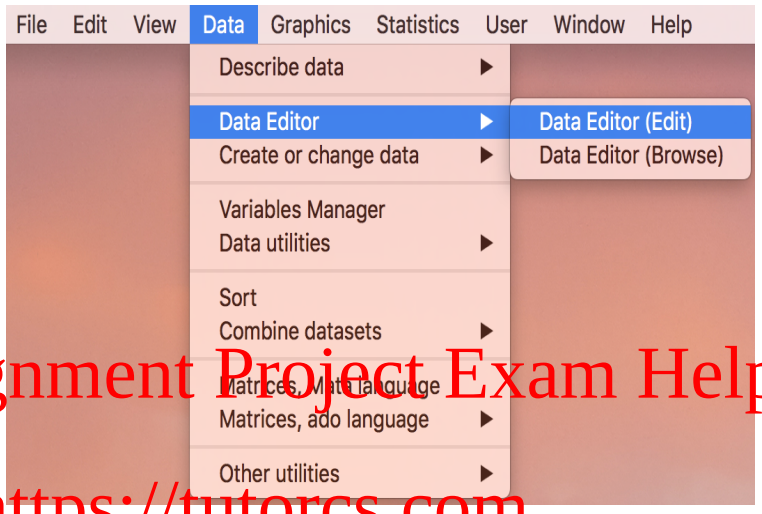


In the **Results** window, you can find

```
twoway (scatter CONS INC)
```

is echoed, which means that by typing this expression in the **Command** window, you can generate the same diagram. Try it! □

(c) On checking the data, you find that your assistant has recorded the weekly consumption expenditure for Family 8 as \$900 instead of \$90. Correct this error and redraw the scatter diagram (Answer)
From the pull-down menu, click on **Data** → **Data Editor** → **Data Editor (Edit)**

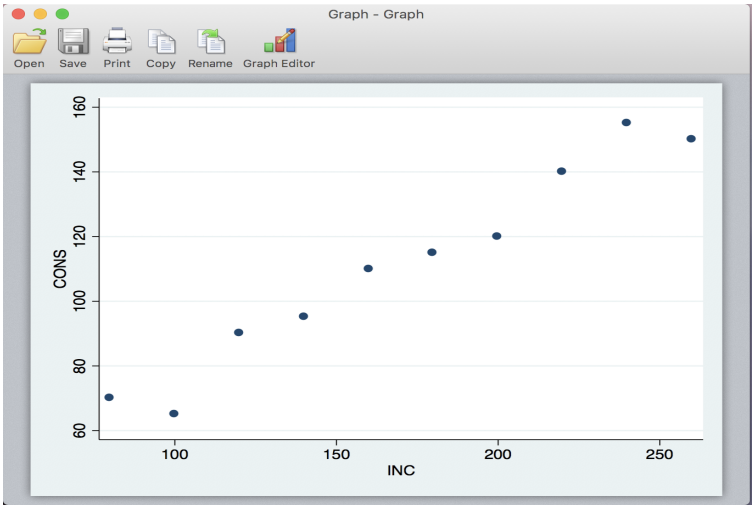


Then, you will see a spreadsheet with data;

Go to the 8$^{\text{th}}$ row of the variable `CONS`, change the entry `900` to `90`, and hit **Enter**. Finally, type

```
twoway (scatter CONS INC)
```

in the **Command** window, which will generate the modified scatter diagram;



□

(d) Compute the mean, median, maximum and minimum values of `INC` and `CONS`.
(Answer)
There are a few things you should do each time a data file is opened. First, enter the **Command**

```
describe
```


This produces a summary of the dataset in memory, including a listing of the variables, information about them, and their labels. A portion of the results is


```
Contains data from consumption.dta
  obs:            10
 vars:             2                          22 Feb 2017 16:50
 size:            80
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
CONS            float   %9.0g
INC             float   %9.0g
-------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```


Next, enter the **Command**

```
summarize
```

In the **Results** window we find the summary statistics. A portion is

```
. summarize

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        CONS |         10         111    31.42893         65        155
         INC |         10         170    60.55301         80        260

.
```

where you find the mean, maximum, and minimum values of the variables. But, this result does not give you the median values. Alternatively, enter the **Command**

```
summarize CONS, detail
```

```
. summarize CONS, detail

                            CONS
-------------------------------------------------------------
      Percentiles      Smallest
 1%           65             65
 5%           65             70
10%         67.5             90       Obs                  10
25%           90             95       Sum of Wgt.          10

50%         112.5                      Mean                111
                        Largest       Std. Dev.       31.42893
75%           140            120
90%         152.5            140       Variance        987.7778
95%           155            150       Skewness       -.0374625
99%           155            155       Kurtosis        1.81431
```

which gives a detailed distributional information on the variable CONS, including the median (50^{th} percentile).

Activities:

Use the pull-down menus to obtain summary statistics. Also, explore the **Commands**

```
su
sum INC
summ INC, detail
summar, detail
```


What's the difference between `summarize` and other commands here?

See if you can shorten the **Command** `describe` similarly, e.g., `de`?

Draw a histogram of CONS (**Graphics** → **Histogram**) and play around the main options (e.g., use 20 bins). See if you can draw the histogram using the **Command** window.

(e) Compute the correlation coefficient between CONS and INC. Comment on the result.
(Answer)
Type in the **Command** window the following


```
correlate CONS INC
```

In the **Results** window we find the correlation matrix. A portion is

```
. correlate CONS INC
(obs=10)

             |     CONS      INC
-------------+------------------
        CONS |   1.0000
         INC |   0.9808   1.0000
```

Activities:

Use the pull-down menus to have the same result. **Statistics** → **Summaries, tables, tests** → **Summary and descriptive statistics** → **Correlations and covariances**.

Discuss how to calculate covariance.

(f) Create the following new variables

$$\texttt{DCONS} = 0.5\texttt{CONS}$$
$$\texttt{LCONS} = \log(\texttt{CONS})$$
$$\texttt{INC2} = \texttt{INC}^2$$
$$\texttt{SQRTINC} = \sqrt{\texttt{INC}}$$

(Answer)
Type in the **Command** window

```
gen DCONS = 0.5 * CONS
generate LCONS = log(CONS)
gen INC2 = INC^2
gen SQRTINC = sqrt(INC)
```

(g) Delete the variable `DCONS` and `SQRTINC` from the workfile
(Answer)
Type in the **Command** window

```
drop DCONS SQRTINC
```

(h) Delete this workfile. Type in the **Command** window
(Answer)

```
exit, clear
```

Activities: why do we need the option `clear`?

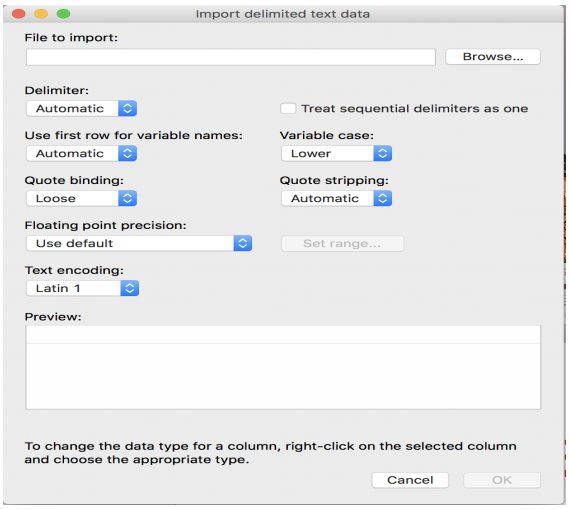
2. At the Fulton Fish Market in New York city, sales of whiting (a type of fish) vary from day to day. Over a period of several months, daily quantities sold (in pounds) were observed. These data are in the file `fultonfish.dat`. Description of the data is in the file `flutonfish.def`. Describe the first four columns.

(a) Use Stata to open the data file and name the series in the first four columns as `date`, `lprice`, `quan` and `lquan`
(Answer)
`fultonfish.dat` is not in `*.dta` format. So, we can't load it by typing

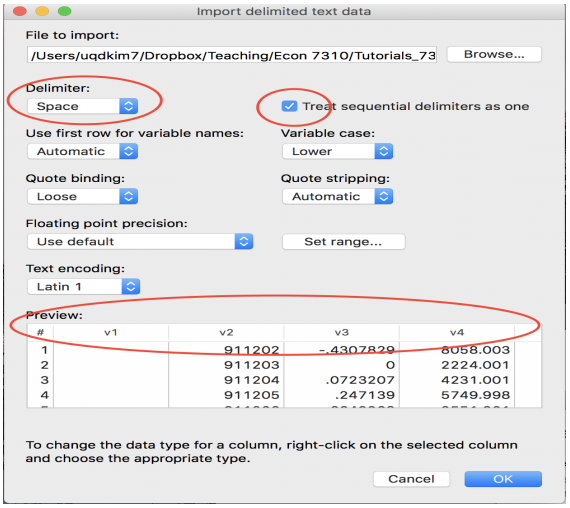

```
use fultonfish
```

When the data is not in Stata format, the way to import the file into the software depends on the structure/format of the data file. Fortunately, Stata offers a number of options to do this job. For this particular data file, from the pull-down menus, click on

File → Import → Text data (delimited, *.cvs, ...)



Click on **Browse** and choose `fultonfish.dat` after setting **File Format** as **All Files**.

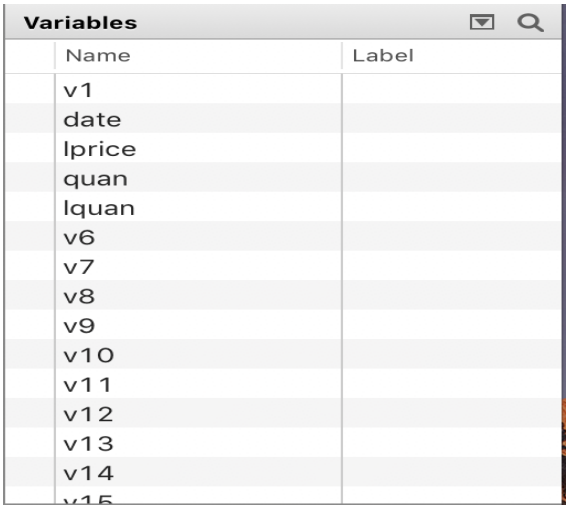


Then, set **Delimiter:** to **Space** and choose the option to treat sequential delimiters as one by checking the box next to **Delimiter:**. Then, in the preview you will see the data are nicely aligned with the default names such as v1, v2, v3, ... Notice that v1 is an empty column. So, the first variable appears in v2, the second in v3, and so on. Now, import the data by clicking on **OK**. Then, the **Variables** window shows something like

Now, we change the variable names by typing in the **Command** window

```
rename v2 date
ren v3 lprice
ren v4 quan
ren v5 lquan
```

Whenever you change, check the **Variables** window. At the end, you must have



Since we will use only the four variables, type in the **Command** window

```
keep date lprice quan lquan
```

(b) Compute the sample mean and standard deviation of the quantity sold (`quan`).
(Answer)
In the **Command** window,

```
su quan
```

to have

```
. su quan

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       quan |        111    6334.667     4040.12    490.0003   21619.99

. .
```

So, the sample size is $n = 111$ and the sample mean $\overline{X} = 6,334.667$ and the sample standard deviation is $\hat{\sigma} = 4,040.12$.

(c) Test the null hypothesis that the mean quantity sold is equal to 7,200 pounds a day at the 5% level of significance.
(Answer)
The null is $H_0 : \mu = 7,200$ and the alternative is $H_1 : \mu \neq 7,200$.
Since

$$\left|\frac{\overline{X}_n - \mu_0}{\hat{\sigma}/\sqrt{n}}\right| = \left|\frac{6,334.67 - 7,200}{4,040.12/\sqrt{111}}\right| \approx 2.26 > 1.96,$$

we reject H_0 □

(d) Construct the 95% confidence interval for part (c)
(Answer)

$$6,334.67 \pm 1.96 \times 4040.12/\sqrt{111} = 6,334.67 \pm 751.58$$

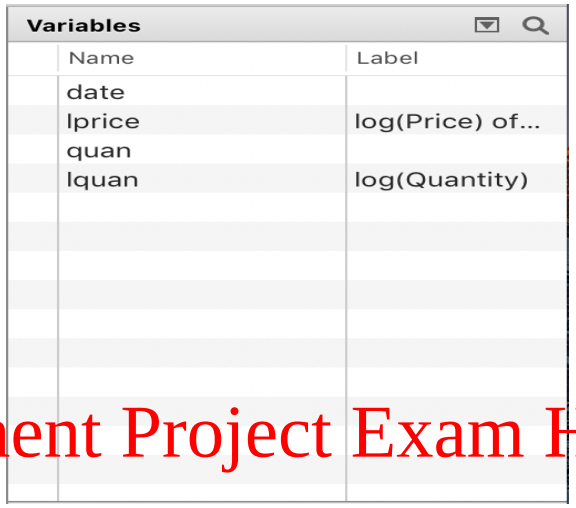
□

(e) Label the variable `lprice` as "log(Price) of whiting per pound" and `lquan` as "log(Quantity)". Then, plot `lprice` against `lquan`. Comment on the nature of the relationship between these two variables.

(Answer)
In the **Command** window, type

```
label variable lprice "log(Price) of whiting per pound"
label variable lquan "log(Quantity)"
twoway(scatter lprice lquan)
```

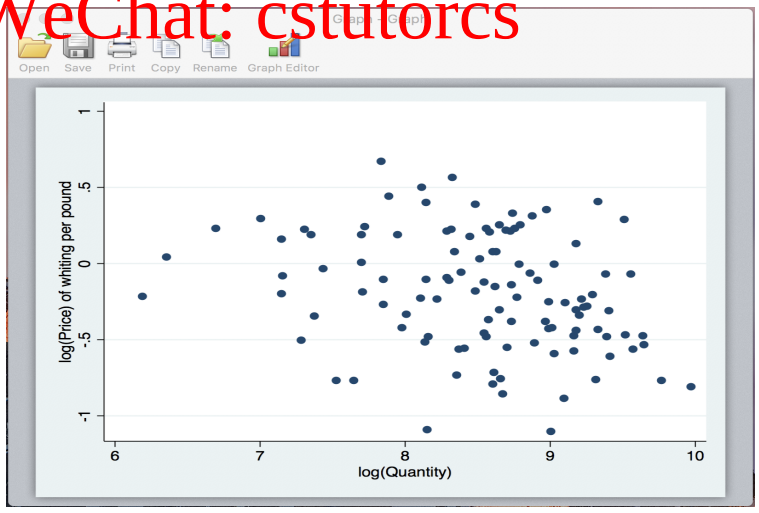
Then, the **Variables** window shows the labels for the variables;



Now, in the **Command** window, type


```
twoway(scatter lprice lquan)
```

and hit **Enter**. Then the diagram below will pop up;



Activities
Draw a histogram of `lprice`: **Graphics** → **Histogram**

(f) Save this workfile to any folder on any drive.
(Answer)
File → **Saveas**