# Introduction to STATA

**Introduction to Econometrics W3412**

## Begin

Create a folder in the My Documents folder called "stata_recitation". Download auto.dta and example1.do from Courseworks into this folder. We will save all of our output into this folder.

## Introduction to STATA

STATA is a statistical package that performs data analysis, data management, and graphics. This introduction should allow you to complete basic tasks. More detailed tutorials can be found at the following websites:

- http://data.princeton.edu/stata/
- http://www.ats.ucla.edu/stat/stata/
- http://staskolenikov.net/stata/Duke/
- http://www.cpc.unc.edu/research/tools/data_analysis/statatutorial/

This introduction borrows from these various tutorials.

## 1 Basics

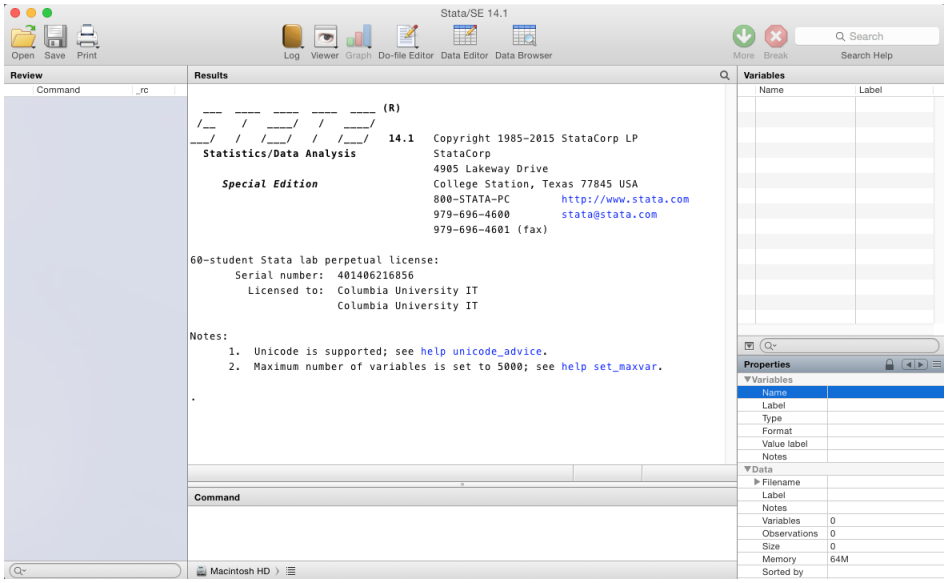
### 1.1 Opening STATA

Mac Users: In the CU computer labs, open STATA using Spotlight Search > StataSE 14.

Windows Users: open STATA from its program directory. In the CU computer labs, go to Start > All Programs > StataMP 13.

When STATA starts up, you will see a menu bar and usually 4 or 5 windows. You can select commands from the drop down menus (e.g. go to File > Open). You can also type commands in the *Command* window. STATA shows the results of your commands in the *Results* window. The

commands are added to the *Review* window so you can keep track of the commands you have used. The *Variables* window lists the variables in your data. *Properties* window shows the properties of a chosen variable, and the properties of the data in use.




Alternatively, you can type your commands in a separate file called a do-file. More on this later.

## 1.2 Typing Commands

In the *Command* window, type

```
display 2+2
```

The `display` command causes STATA to show the results in the *Results* window.

Now try

```
Display 2+2
```

You will get a red error message since STATA does not recognize this command; commands are case-sensitive.

## 1.3 Help

To search for a concept, go to Help > Search. Or type something like:

```
search linear regression
```

in the command window (`findit <keywords>` searches both STATA's help and the internet; Substitute the words you want for `<keywords>`).

If you know the command you need help with, type: `help <commandname>`.

To interrupt any STATA command, press Q or hit the Break button  .

### 1.4 Command Syntax

Most STATA commands have the following syntax:

`command [variable(s)] [if expression] [in obs. range] [[weights]] [using filename], [options]`

The `command` is performed on the *variable(s)*. If applicable, a particular subset of observations may be selected by the `if` and `in` modifiers, and the specific ways the command should behave are controlled by *options*.

## 2 In Class Example

Let's do an example to practice using STATA.

### 2.1 Loading Data

#### 2.1.a Directory

Type `pwd` to find STATA's current directory. This is the directory from which STATA opens or saves files if you do not specify an alternate directory. You can see the contents of this directory by typing `ls`.

You can change directories by typing `cd <newdirectory>`.

Change the directory to the stata_recitation folder:

`cd "D:\My Documents\stata_recitation\"`

From now on, we will use and save files in the stata_recitation folder.

#### 2.1.b Log File

Let's open a log file. The log file will keep a permanent record of your results. Do this by typing:

`log using example1.log, replace`

The replace option means that the log file will be overwritten if it already exists. We can later close the log file by typing:

`log close`

#### 2.1.c Load Data

Clear the memory by typing:

`clear`

(If you already have data in memory, you will get an error message if you do not clear the memory.)

Load the data by going to File > Open > My Documents > stata_recitation > auto.dta.

Stata displays the command

```
use "D:\My Documents\stata_recitation\auto.dta", clear
```

Note that you could have typed this command into the *Command* window to open the file. The `use` command opens files in STATA format (*.dta).

Also note that you can use files that are not in your current directory if you specify where the file is located, as above.

If you need to read in a different type of file (e.g. *.csv, *.txt), see

http://www.ats.ucla.edu/stat/stata/modules/input.htm for help.

## 2.2 Data Management and Analysis

### 2.2.a Describe the Data

To describe the data, type:

```
desc
```

STATA will display the number of observations, variable names and labels.

### 2.2.b Look at the Data

To look at the *make* of the first 5 observations, sorted by *make*, type:

```
sort make
```

```
list make in 1/5
```

The `in` option allows you to specify the observations to list. You can also view the data by using Data > Data Editor > Data Editor (Browse), or hitting the data browser button 

.

You can both view and manually edit your data in the Data Editor (Data > Data Editor > Data Editor (Edit), or 

). However, manually editing data is not recommended.

### 2.2.c Summarize the Data

To obtain summary statistics of the data, type:

```
sum
```

STATA will give the means, standard deviations, minimum and maximum of all the variables in the sample. For example, the mean *mpg* is 21.3.

Note: *make* says it has 0 observations since it is a string variable (letters). The variable *rep78* has 69 observations because the values of *rep78* are missing for some observations. Stata denotes missing with a "." for numeric variables and " " for string variables.

In order to obtain a more detailed summary of the variable *mpg*, type:

```
sum mpg, detail
```

The `detail` option tells STATA to also give the percentiles, skewness, and kurtosis of mpg. Since the skewness is positive, we know that *mpg* has a long right tail.

To summarize *mpg* for only foreign cars, type:

```
sum mpg if foreign==1
```

Note that `if` is followed by a **logical statement** `(foreign==1)`:

The symbols for logical statements include:

| Symbol | Meaning |
|---|---|
| `>` | Strictly greater |
| `<` | Strictly less |
| `>=` | Greater or equal |
| `<=` | Less or equal |
| `==` | Equal |
| `!=` or `~=` | Not equal |

To summarize *mpg* for domestic cars, type:

```
sum mpg if foreign!=1
```

Alternatively, you could have generated summary statistics by group:

```
sort foreign
by foreign:  sum mpg
```

To obtain more detailed summary statistics for foreign cars, type:

```
sum mpg if foreign==1, detail
```

Note how the stata command structure allows you to use an `if` statement followed by the option detail.

### 2.2.d Graphs

To see the empirical distribution of *mpg*, type:

```
histogram mpg
```

The graph shows us that *mpg* has a long right tail.

You can save this graph by typing:

`graph save histogram_mpg.gph, replace`

Make a scatterplot of *mpg* vs. *weight* by typing:

`scatter mpg weight`

The scatterplot shows heavier cars have lower *mpg*.

For more on graphs, see `help graph.`

### 2.2.e Generating Variables

The variable *weight* is in pounds. Let's generate a variable for weight in 1000s of pounds.

`gen weightdiv1000 = weight/1000`

Summarize the variables *weight* and *weightdiv1000* to see that the variable was created correctly.

Note that some mathematical expressions in STATA include:

| Symbol | Expression | Symbol | Expression |
|---|---|---|---|
| `+` | Add | `^` | power |
| `-` | Substract | `sqrt()` | square root |
| `*` | Multiply | `exp()` | exponential |
| `/` | Divide | `ln()` | natural log |

There is a command `egen` which is an extension to the generate command (See `help egen`).

Generate a variable equal to the sample mean of *mpg*:

`egen mpg_bar = mean(mpg)`

Other useful data manipulation commands include:

| Command | Description |
|---|---|
| `drop x y` | drop variables named x and y |
| `keep x y` | drop every variable except x and y |
| `replace x=y` | replace variable x with expression y |
| `rename x y` | rename variable x "y" |

### 2.2.f Test of the Null Hypothesis

Suppose you wanted to test whether foreign and domestic cars have the same mean mpg.

The null hypothesis is: $H_0 : \mu_{domestic} - \mu_{foreign} = 0$

The alternative hypothesis is: $H_1 : \mu_{domestic} - \mu_{foreign} \neq 0$

Test this hypothesis using Student's t-test in STATA:

`ttest mpg, by(foreign)`

Due to the very small p-value, we reject the null hypothesis that foreign and domestic cars have the same mean mpg.

### 2.2.g Linear Regression

(copied/modified from http://www.ats.ucla.edu/stat/stata/output/reg_output.htm)

Do a linear regression of *mpg* on *weight*:

```
reg mpg weight
```

You should see the following output:

```
. reg mpg weight

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  1,    72) =  134.62
       Model |   1591.9902     1   1591.9902           Prob > F      =  0.0000
    Residual |  851.469256    72  11.8259619           R-squared     =  0.6515
-------------+------------------------------           Adj R-squared =  0.6467
       Total |  2443.45946    73  33.4720474           Root MSE      =  3.4389

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0060087   .0005179   -11.60   0.000    -.0070411   -.0049763
       _cons |   39.44028   1.614003    24.44   0.000     36.22283    42.65774
------------------------------------------------------------------------------
```

Note that you can now obtain the fitted values for *mpg*:

```
predict mpghat
```

and the residuals of the regression:

```
predict ehat, resid
```

which will be stored in the new variables *mpghat* and *ehat*.

Here's what your regression output means: In the top left hand corner, there is the Analysis of Variance (ANOVA) Table

```
      [a] Source |       [b] SS       [c] df     [d] MS
-----------------+------------------------------
           Model |   1591.9902     1   1591.9902
        Residual |  851.469256    72  11.8259619
-----------------+------------------------------
           Total |  2443.45946    73  33.4720474
```

[a] **Source** - Looking at the breakdown of variance in the outcome variable, these are the categories we will examine: Model, Residual, and Total. The Total variance is partitioned into the variance that can be explained by the independent variables (Model) and the variance that is not explained by the independent variables (the Residual, sometimes called Error).

[b] **SS** - These are the Sum of Squares associated with the three sources of variance, Total (TSS), Model (ESS), and Residual (SSR).

[c] **df** - These are the degrees of freedom associated with the sources of variance. The total variance has $N - 1$ degrees of freedom. The Model degrees of freedom corresponds to the number of coefficients estimated minus 1. Including the intercept, there are 2 coefficients, so the model has $2 - 1 = 1$ degree of freedom. The Residual degrees of freedom is the DF Total minus the DF Model, $73 - 1 = 72$.

[d] **MS** - These are the Mean Squares, the Sum of Squares divided by their respective df.

In the top right hand corner, there is the table for overall model fit.

```
[e] Number of obs =      74
[f] F(  1,    72) =  134.62
[g] Prob > F      =  0.0000
[h] R-squared     =  0.6515
[i] Adj R-squared =  0.6467
[j] Root MSE      =  3.4389
```

[e] **Number of obs** - This is the number of observations used in the regression analysis.

[f] **F( 1, 72)** - F-statistic is the Mean Square Model (1591.9902) divided by the Mean Square Residual (11.8259619), yielding F=134.62. The numbers in parentheses are the Model and Residual degrees of freedom are from the ANOVA table above.

[g] **Prob > F** - This is the p-value associated with the above F-statistic. It is used in testing the null hypothesis that all of the model coefficients are 0.

[h] **R-squared** - R-Squared is the proportion of variance in the dependent variable (*mpg*) which can be explained by the independent variable (*weight*). This is an overall measure of the strength of association.

[i] **Adj R-squared** - This is an adjustment of the R-squared that penalizes the addition of extraneous predictors to the model. Adjusted R-squared is computed using the formula $1 - ((1 - Rsq)((N - 1)/(N - k - 1))$ where $k$ is the number of predictors.

[j] **Root MSE** - Root MSE is the standard deviation of the error term, and is the square root of the Mean Squares of the Residual ($\sqrt{(11.83)} = 3.4389$).

In the bottom, there is the table for the regression coefficients.

```
------------------------------------------------------------------------------
      [k] mpg |    [l] Coef. [m]Std. Err.    [n] t  [o] P>|t|    [p] [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0060087   .0005179   -11.60   0.000    -.0070411   -.0049763
       _cons |   39.44028   1.614003    24.44   0.000     36.22283    42.65774
------------------------------------------------------------------------------
```

[k] **mpg** - This column shows the dependent variable at the top (*mpg*) with the predictor variables below it (*weight* and *_cons*). The last variable (*_cons*) represents the constant or intercept.

[l] **Coef.** - These are the values for the regression equation for predicting the dependent variable from the independent variable(s). The regression equation is presented in many different ways, for example: $Y_{predicted} = b_0 + b_1 * x_1$

The column of estimates provides the values for $b_0$ and $b_1$ for this equation.

The coefficient on *weight* is -.0060087. So for every one pound **increase** in weight, a - .0060087 unit **decrease** in *mpg* is predicted, holding all other variables constant.

[m] **Std. Err.** - These are the standard errors associated with the coefficients.

[n] **t** - These are the t-statistics used in testing whether a given coefficient is significantly different from zero.

[o] **P > | t |** - This column shows the 2-tailed p-values used in testing the null hypothesis that the coefficient (parameter) is 0. Using an alpha of 0.05: the coefficient for *weight* is significantly different from 0 because its p-value is 0.000, which is smaller than 0.05. The constant (*_cons*) is significantly different from 0 at the 0.05 alpha level.

[p] **[95% Conf. Interval]** - These are the 95% confidence intervals for the coefficients. The confidence intervals are related to the p-values such that the coefficient will not be statistically significant at the 0.05 level if the confidence interval includes 0.

### 2.2.h Save

Save the data to a new dataset named auto2. Alternatively you can overwrite auto by typing auto in place of auto2.

```
save auto2, replace
```

### 2.2.i Close the Log File and Exit STATA

Now close your log file by typing:

```
log close
```

The log file provides a record of the commands you typed. Open the log from the stata_recitation folder.

Exit STATA by typing:

```
clear
exit
```

### 2.2.j Do Files

Instead of typing your commands one by one, you can write a do file and run all those commands at once.

A sample do-file that corresponds to the In-Class Example is provided on courseworks (example1.do). You have already downloaded this file. This file was created and can be edited in a text editor. When you save this file, make sure you change the extension to *.do. It will not work if it has a different extension.

Open a do file editor by going to Window > Do-file editor > Top do-file (for Windows users: New do-file Editor), or hitting 

. From the editor, open (example1.do) and review it. You can run the do file by typing:

```
do example1.do
```