

Stata Tutorial

Stat 100 – Spring 2008

The purpose of this tutorial is to learn how to download, install, and use Stata for data manipulation, visualization, and simple analysis.

1. Downloading and Installing onto your PC

Downloading

You can get a free site-licensed copy of Stata for your computer through Harvard's server. With this version, you will need to be able to log into the Harvard network every time you would like to use it. First, you will need to download the following items from the FAS software download website:

- go to the software download center (you will need to log in using your Harvard PIN #):

<http://www.fas.harvard.edu/cgi-bin/software/download.pl>

(If you are using a MAC, make sure you click on the correct platform for proper installation)

- Scroll down to the program: **KeyAccess**, and click on the **Download** button. Click on *I accept* and *Continue*, and your download should start automatically. Save this to a convenient location on your computer, like your desktop. (If you have a pop-up blocker installed, you may have to click on the banner that opens near the top of the screen). For Macs, the analogous program is called **KeyServer**.
- Return to the FAS software download website, and download **Stata SE** (SE = "special edition") as you did for **KeyAccess**. This is a large file (~90MB), so it may take a few minutes. For Macs, it is just called **Stata**.

Installing

Once you have downloaded the two programs as outlined above, you now need to install them. First install the program **KeyAccess**. Click on the program *k2Clientv6_1.exe*, click *OK*, click on *next* through all the prompts in the window, and then click *Install*. This should only take a few seconds. When asked, you do **not** need to re-start your computer at this time.

Next you will need to install Stata. Open up the program you just downloaded: *Stata10.exe*. Click on *OK* to being a Harvard affiliate, and the install should begin automatically. After about 1 minute, another window will pop open. Click *Next* three times, and the installation will continue. After about 30 second, a third prompt window will open up. Click *Next* through these prompts, and set-up will continue. After about 3-4 minutes, set-up should be complete. Click *OK* once the set-up finishes. You can delete the two programs you downloaded (presumably on your desktop); these were used for installation only. Now, restart your computer before booting up Stata for the first time.

*Purchasing your own Copy

There is also an option to purchase your own copy of Stata. The advantage to this is you will be able to run Stata directly from your own hard drive, and will not need to log into the Harvard server every time you want to use the software. Of course, the downside is, it is quite expensive (from \$48 to \$335). For this class, there should be no reason to purchase the software. Note that the Harvard site license can be used outside the Harvard network firewall with a VPN connection. A VPN connection can be made by downloading, installing, and running the program **VPN Client** from the FAS software download page mentioned above. If you do wish to purchase a version to use while traveling

or while outside the Harvard network, we recommend the least expensive option, Small Stata, for \$48. You only need order the product you want at the Stata website below and you will be sent an email about where to pick up the software on campus. You can find the products here:

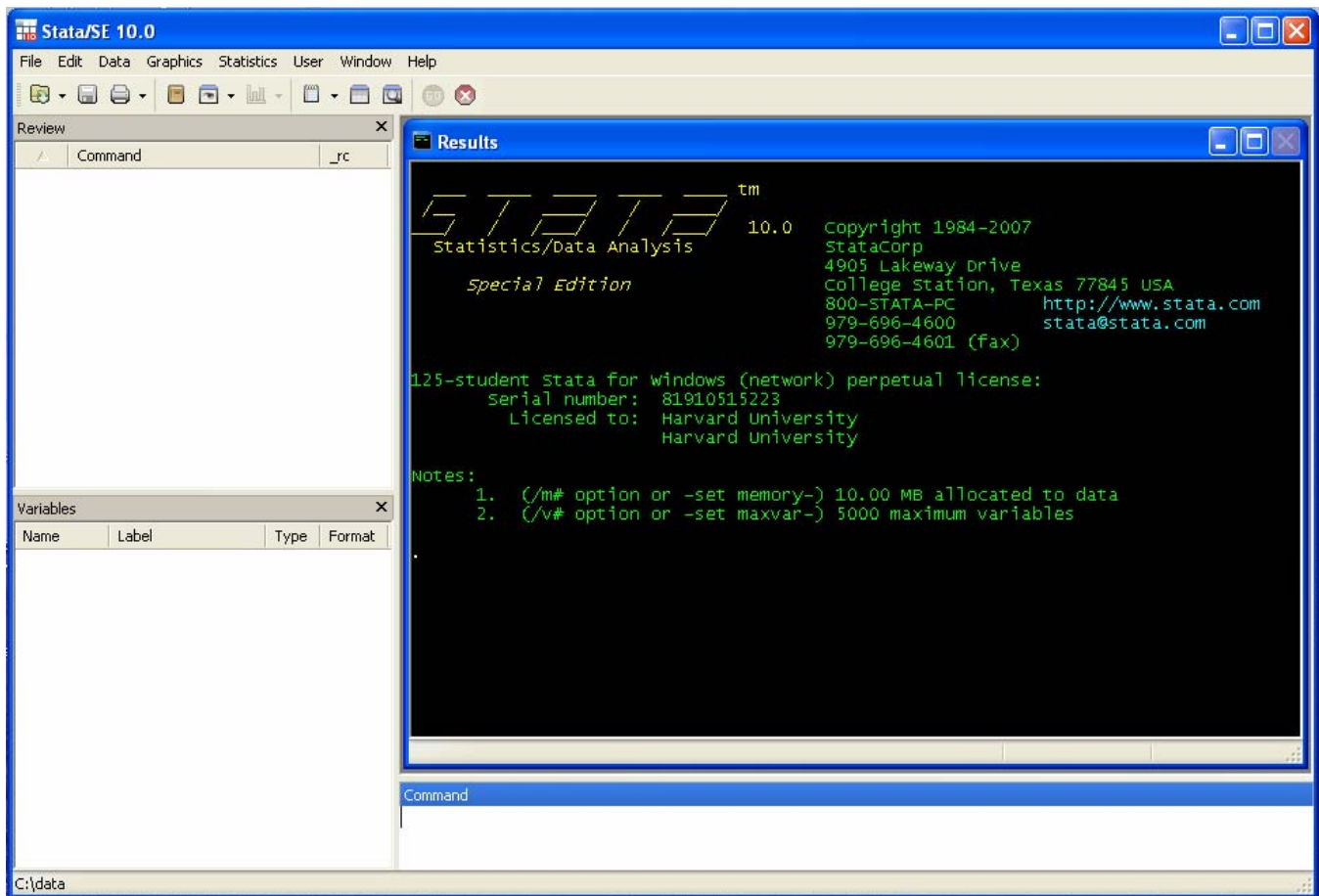
<http://www.stata.com/order/new/edu/gradplan.html>

Stata is also available in the FAS computing labs and runs as described below. You do not have to download the software on a computer in a lab.

2. Start-Up and Data Manipulation

Start-Up

To open Stata on a computer lab PC (or on your computer in which you followed the above directions), click on *Start* → *Programs* → *Stata 10* → *StataSE 10*. A screen should pop up that looks like this:



Notice, there are 4 main windows (along with the menus up top):

1. Results – Stata will print-out any analysis output or communication (like error messages)
2. Command – The user enters commands for Stata to run analysis.
3. Variables – This window will list the variables that have been entered into Stata
4. Review – This window displays the commands that Stata has processed

Data Entry

There are 3 Main Ways to bring data into Stata: by importing data created by another program or editor, by manual entry, and by reading a data set that Stata has saved in the native format of the program. We will mainly be using the importing option in this class whenever you use a data set for the first time. It is important to know about the manual entry technique, as it may be useful for when project time roles around at the end of the semester. The tutorial that is part of problem set 1 describes how to save and re-use data in Stata's native format, which is useful when you want to save your progress when doing your homework.

Importing Data

Most datasets are stored as simple text files (with extensions .csv, .txt, .dat, or even .raw) which can easily be imported into Stata. However, you will need to do a little bit of work to import an Excel file the ends with .xls.

This tutorial uses the *2004_Election.csv* data file found on the course website here:
<http://isites.harvard.edu/icb/icb.do?keyword=k16183&pageid=icb.page81929>.



Save this file on your computer's desktop to start. To import, click on the menu *File* → *Import* → *ASCII data created by a spreadsheet*. In the window that pops open, click on the *Browse* button, and select the file you want. You may have to change *Files of Type* to *Comma Separated Values (*.csv)*. Click on *OK*, and the new variables should be entered into Stata's memory.

Note: If you get an error message like "you must start with an empty dataset", then the simplest fix is to just type: *clear* in the command window and click enter. Be careful though, as this will remove any old data floating around in Stata's memory.

Creating a .csv file from a .xls Excel File

Sometimes data will come as an Excel file ending in .xls. The easiest way to deal with this is to open the file in Excel, and then save as a .csv type file. Once you have the file opened in Excel, go to the menu: *File* → *Save As*. In the window that pops open, change the option *Save as type* to the option *CSV (Comma delimited) (*.csv)*. Save the file in this format in an easy to get to location on your computer (like the desktop). In the windows that pop open, click on *OK* and *Yes*; we really do want to change this to a file that can be read into other software.

Manually Entering Data (use at your own risk)

Near the top of the Stata window, you will see what looks like a table/spreadsheet (). If you click on this button, the *Data Editor* window should open. Once this is open, you can simply click on a cell and enter data, or just copy and paste data directly from Excel. After getting the data set-up the way you want it, click on the *Preserve* button to save your changes for later. You can close this by clicking on the  like any window in MS-Windows (you must close this window to do any analysis).

3. Data Visualization

Once the dataset is read in, the main concern now is how to manipulate data. Problem set 1 discusses how to use the Stata menus to produced simple graphs and summary statistics, so that material is not reproduced here. This section of the tutorial discusses how to enter commands directly into the Stata *Command* window. Here, we will learn to get summary statistics (think measures of center, spread,

etc...) and graph/plot our data. Except for complex commands, menu choices and direct commands produce identical results.

Please note: some of the methods illustrated in this tutorial will not be used until the second week of class or later.

Summary Statistics

To get some quick statistics on a quantitative variable, use the command **summarize** followed by the list of variables you are interested, for example:

```
. summarize bush_perc gsp
```

When typing in any Stata commands, you can just copy and paste the commands from this tutorial, but do not copy the dot at the beginning of each command (that's just from the Stata output screen). Doing so will lead to an error. You will notice that when Stata prints results in the Result window, it adds the dot at the beginning of the line to indicate a command it has just executed.

Unfortunately, the above does not give the median or quartiles. To get percentiles, you have to give the option **detail**, as such:

```
. summarize bush_perc gsp, detail
```

To get frequencies of a categorical variable, use the command **table**:

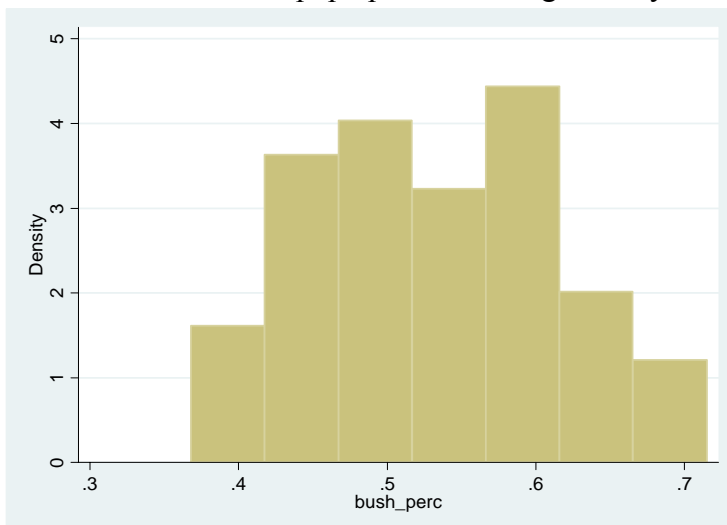
```
. table region
```

Histograms

To get a quick view of the distribution of a variable, use the command **hist**. Enter:

```
. hist bush_perc
```

A new window should pop up with a histogram of your chosen variable like this one:



Boxplots

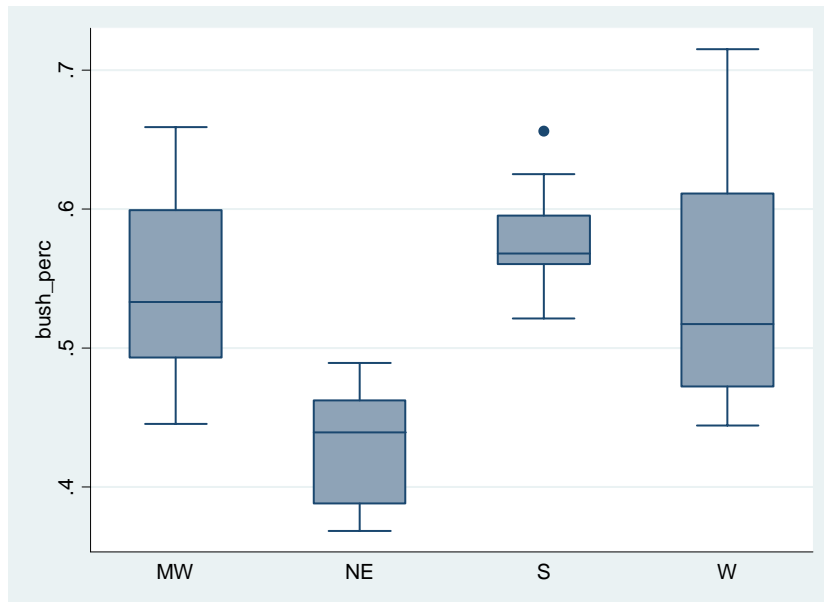
To produce a boxplot of a variable, use the command `graph box`. Enter:

```
. graph box bush_perc
```

You can also split a boxplot into different categories. For example, we can do:

```
. graph box bush_perc, over(region)
```

And you should get the following graph:



Scatterplots

To get a quick visual of how two variables are related, use the command `plot` to get a basic plot in the results window (Note, the first variable is the y-variable, and the second is the x-variable):

```
. plot bush_perc gsp
```

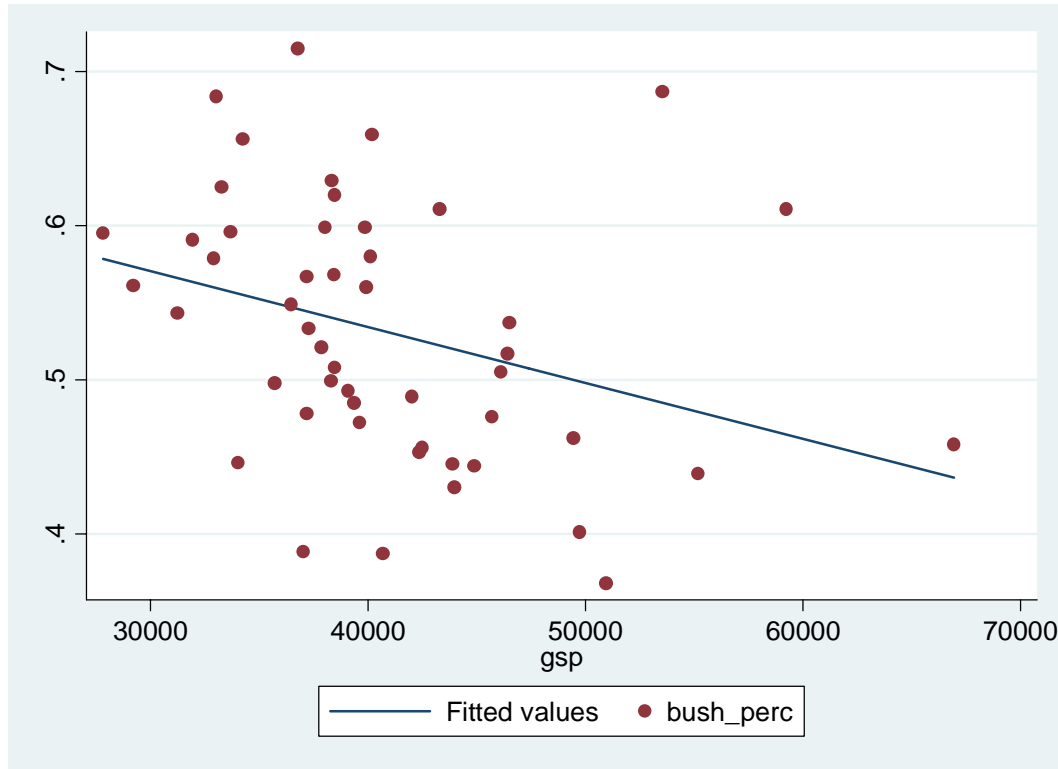
To get a much more detailed view of the scatterplot, use the command `graph twoway` instead:

```
. graph twoway scatter bush_perc gsp
```

To add the regression line to the scatterplot, add the command `lfit` like this:

```
. graph twoway (lfit bush_perc gsp) (scatter bush_perc gsp)
```

And the result should look like this:



Saving and Printing Graphs

The easiest way to print a histogram, scatterplot, etc... is to right-click on the graph window itself (in the middle), and then copy and paste into a word processor. From there you can add comments, adjust the size, etc... Graphs can also be saved by Stata with the extension `.gph` and re-opened during a session or at a later session.

4. Data Analysis

Next week, we will learn to measure and analyze the association between two variables (correlation and regression). Later in the course, we will see many more ways to do analysis (confidence intervals, hypothesis testing, ANOVA, etc...). Let's do some work on what we know for now:

Correlation

To find the correlation coefficient between two (or more) variables, use the command `corr`:

```
. corr bush_perc gsp
```

Regression

To get the printout of a regression (to find the estimates for the slope and intercept of a line), use the command `regress`:

```
. regress bush_perc gsp
```

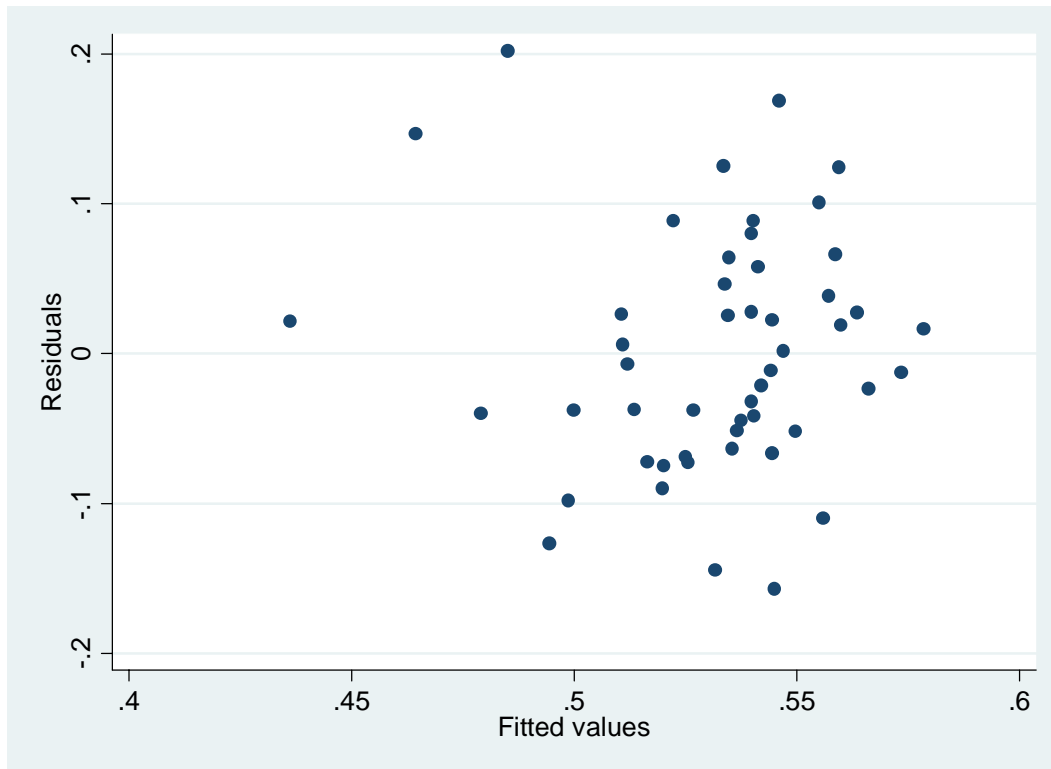
The results should look like this:

Source	SS	df	MS			
Model	.036676619	1	.036676619	Number of obs =	50	
Residual	.312433678	48	.006509035	F(1, 48) =	5.63	
				Prob > F =	0.0217	
				R-squared =	0.1051	
				Adj R-squared =	0.0864	
Total	.349110297	49	.0071247	Root MSE =	.08068	

bush_perc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
gsp	-3.63e-06	1.53e-06	-2.37	0.022	-6.71e-06	-5.56e-07
_cons	.6795575	.0634325	10.71	0.000	.5520179	.807097

We know we often would like to look at the residual plot for a regression to see if the assumptions are met (to see if there is a pattern in the residuals, like a U-shape). To get the residual vs. fitted plot (the fitted variable being your \hat{y}_i), use the command `rvfplot`. Note, this command should be entered directly following the `regress` command, since it refers back to it.

`. rvfplot`



Practice Problem

S&P 500 Stock Index. This dataset can be downloaded here:

ichart.finance.yahoo.com/table.csv?s=%5EGSPC&a=00&b=3&c=1950&d=03&e=12&f=2007&g=m&ignore=.csv

For this problem we will be using the **monthly** S&P 500 Index Prices Since 1950. We are going to see if we can predict the S&P 500 price by the volume traded that day.

Download the above chart onto your desktop (I called it SP500.csv). Open up Stata. Within Stata, read in the file using the menu: *File* → *Import* → *ASCII data created by a spreadsheet* like above.

Alternatively, you can copy and paste the file directly from Excel. Open the file in Excel. It should look like this:

	A	B	C	D	E	F	G	H
1	Date	Open	High	Low	Close	Volume	Adj Close	
2	4/2/2007	1420.83	1448.73	1416.37	1447.8	3.02E+09	1447.8	
3	3/1/2007	1406.8	1438.89	1363.98	1420.86	3.21E+09	1420.86	

Copy the columns of interest, paste into Stata’s Data Editor, and then hit *Preserve*; it should look like this:

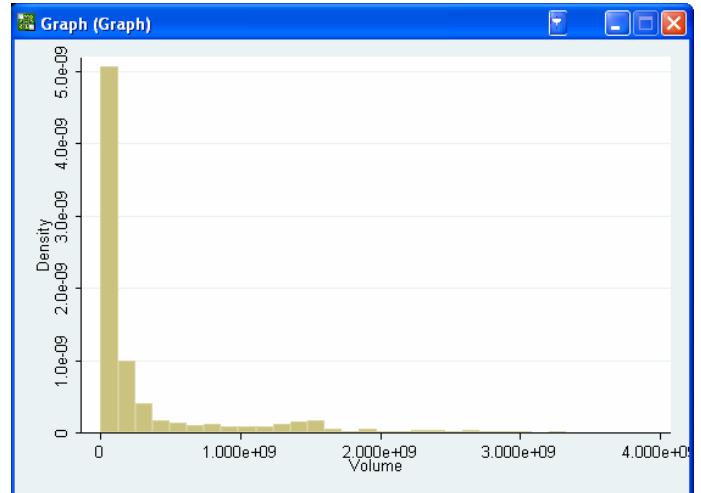
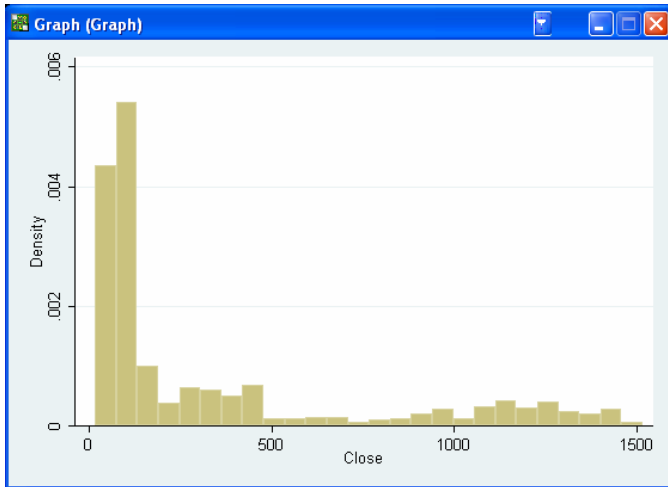
	date	open	high	low	close
1	4/2/2007	1420.83	1448.73	1416.37	1447.8
2	3/1/2007	1406.8	1438.89	1363.98	1420.86

Once you have the file read into Stata correctly, we can start analyzing the data. Now type in the commands into the command window (one at a time & without the dot):

```
. summarize close volume , detail
. summarize close volume
```

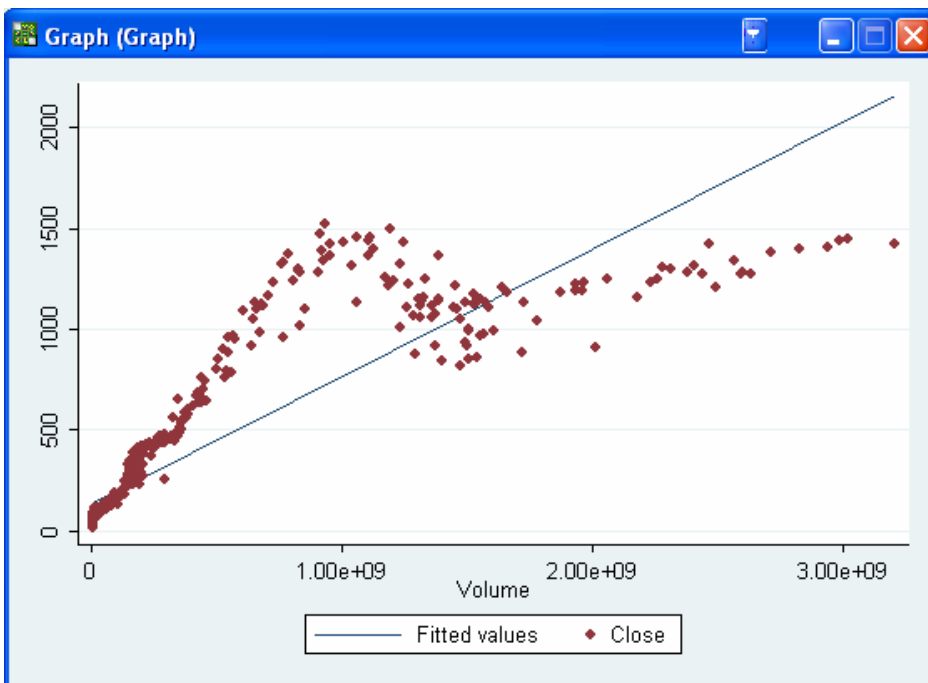
Variable	Obs	Mean	Std. Dev.	Min	Max
close	688	337.951	416.9518	17.05	1517.68
volume	688	3.12e+08	5.89e+08	1024300	3.21e+09

```
. hist close  
. hist volume
```



a) What do you notice about the two variables we are interested in?

```
. graph twoway (lfit close volume) (scatter close volume)
```

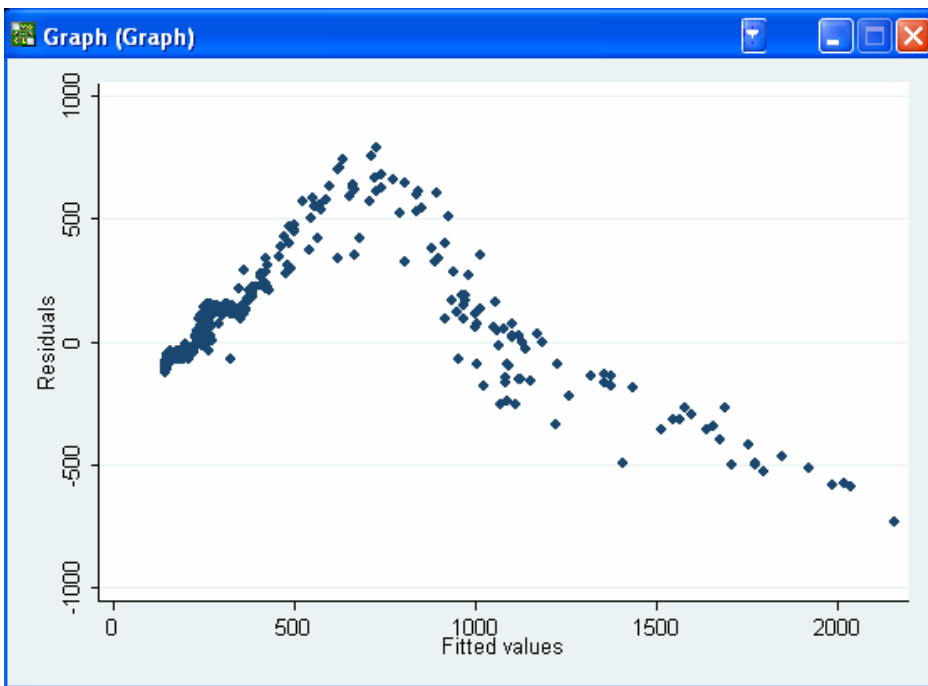


. regress close volume

Source	SS	df	MS			
Model	93906698.9	1	93906698.9	Number of obs =	688	
Residual	25527430.4	686	37211.9977	F(1, 686) =	2523.56	
				Prob > F =	0.0000	
				R-squared =	0.7863	
				Adj R-squared =	0.7860	
Total	119434129	687	173848.805	Root MSE =	192.9	

	close	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
volume		6.28e-07	1.25e-08	50.24	0.000	6.04e-07	6.53e-07
_cons		142.1741	8.323191	17.08	0.000	125.8321	158.5161

. rvfplot



b) What is the equation for the least squares regression line? What does this mean?

c) What is the predicted closing price for a day that had a billion (10^9) shares traded? What about for 3.12 billion ($3.12 \cdot 10^9$)?

d) What are our results? Does this seem to be a good fit? How do you know?