
Stata 10 Tutorial 4

TOPIC: Generating and Graphing Predicted Values and Residuals After OLS Estimation

DATA: One of the following two *Stata*-format data sets you have created in previous tutorials:

auto1.dta	(the <i>Stata</i> -format data file you created in <i>Tutorial 1</i>)
auto2.dta	(the <i>Stata</i> -format data file you created in <i>Tutorial 2</i>)
auto3.dta	(the <i>Stata</i> -format data file you created in <i>Tutorial 3</i>)

TASKS: *Stata Tutorial 4* introduces you to two important *post-estimation commands* that are often used with the **regress** command. The first is the **predict** command, which is used to generate predicted values of the dependent variable and residuals following OLS estimation. The second is the **graph twoway** command, which can be used to display graphically sample data and regression results.

- The *Stata* **post-estimation and graph commands** introduced in this tutorial are:

predict	Computes estimated Y_i -values and OLS residuals.
graph twoway	Draws scatterplots of sample data points and line graphs of OLS sample regression functions.
graph save	Saves the graph currently displayed in the Graph window to a disk file in the current <i>Stata</i> working directory.
graph export	Exports the graph currently displayed in the Graph window to a file in the current <i>Stata</i> working directory.
graph use	Loads a graph on disk into memory and displays it in the Graph window.

NOTE: *Stata* commands are *case sensitive*. All *Stata* commands must be typed in the Command window in **lower case letters**.

HELP: *Stata* has an extensive on-line **Help** facility that provides fairly detailed information (including examples) on all *Stata* commands. Students should become familiar with the *Stata* on-line **Help** system. In the course of doing this tutorial, take the time to browse the **Help** information on some of the above *Stata* commands. To access the on-line **Help** for any *Stata* command:

- choose (click on) **Help** from the *Stata* main menu bar
- click on **Stata Command** in the **Help** drop down menu
- type the full name of the *Stata* command in the *Stata* command dialog box and click **OK**

Preparing for Your *Stata* Session

Before beginning your *Stata* session, use Windows Explorer to copy the *Stata*-format data set you want to use for this tutorial (**auto1.dta** or **auto2.dta**) to the *Stata working directory* on the C:-drive or D:-drive of the computer at which you are working.

- **On the computers in Dunning 350**, the default *Stata* working directory is usually **C:\data**.
- **On the computers in MC B111**, the default *Stata* working directory is usually **D:\courses**.

Start Your *Stata* Session

To start your *Stata* session, double-click on the ***Stata 10* icon** in the Windows desktop.

After you double-click the ***Stata 10* icon**, you will see the now familiar screen of four *Stata* windows.

Record Your *Stata* Session – log using

To record your *Stata* session, make a **.log** file named **351tutorial4.log**. To open (begin) the **.log** file **351tutorial4.log**, enter in the Command window:

```
log using 351tutorial4.log
```

This command opens a file called **351tutorial4.log** in the current *Stata* working directory. Once you have opened the **351tutorial4.log** file, a copy of all the

commands you enter during your *Stata* session and of all the results they produce is recorded in the **351tutorial4.log** file.

An alternative way to open the **.log** file **351tutorial4.log** is to click on the **Log** button; click on **Save as type:** and select **Log (*.log)**; click on the **File name:** box and type the file name **351tutorial4**; and click on the **Save** button.

□ Loading a Stata-Format Data Set into Stata – use

Load, or read, into memory the data set you are using. To load the *Stata*-format data file you want to use into memory, type in the Command window *any one* of the following commands:

```
use auto1  
or  
use auto2  
or  
use auto3
```

This command loads into memory the *Stata*-format data set you wish to use.

Remember that only one *Stata*-format data set can reside in memory at any one time. If you wish to use a *Stata*-format data set other than the one you initially read into memory, use the **clear** command to remove the data set currently in memory, and then read in the new data set you want to use. For example, the following sequence of commands first loads the data set **auto2.dta** into memory, then removes it from memory, and finally reads in the data set **auto3.dta**:

```
use auto2  
clear  
use auto3
```

□ Familiarize Yourself with the Current Data Set

To familiarize (or re-familiarize) yourself with the contents of the current data set, type in the Command window the following commands:

```
describe
summarize
codebook price weight mpg foreign
```

□ Computing a Simple OLS Regression Equation (again!) – regress

- The *Stata* commands covered in this tutorial are post-estimation commands – that is, they are used after a model estimation command such as a **regress** command. We therefore need an OLS sample regression equation to illustrate the use of these post-estimation commands.
- So estimate again by OLS the simple linear regression model given by the PRE

$$\text{price}_i = \beta_0 + \beta_1 \text{weight}_i + u_i \quad (1)$$

Enter in the Command window the following **regress** command:

```
regress price weight
```

□ Calculating predicted *values* and *residuals* – predict

The **predict** command is used to compute estimated Y_i values (\hat{Y}_i) and sample residuals (\hat{u}_i) for an OLS sample regression equation.

After a **regress** command you can use the *Stata* **predict** command to calculate both

- (1) the **estimated or predicted Y_i values** (i.e., the values of the OLS sample regression function) given by

$$\hat{Y}_i = \hat{\text{price}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{weight}_i \quad (i = 1, \dots, N) \quad (2)$$

and

(2) the **OLS residuals** given by

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 \text{weight}_i \quad (i = 1, \dots, N) \quad (3)$$

where in this case $Y_i = \text{price}_i$.

- **To calculate the estimated or predicted values** of the dependent variable given by equation (2), use the **predict** command without any options. For example, to calculate the estimated (or predicted) values of price_i and name them **yhat**, enter the following command:

```
predict yhat
```

The values of **yhat** generated by this **predict** command are computed according to equation (2) above using the numeric values of the coefficient estimates:

$$\hat{Y}_i = \hat{\text{price}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{weight}_i = -6.707353 + 2.044063 \text{weight}_i .$$

- **To calculate the OLS residuals** given by equation (3), use the **predict** command with the **residuals option**. For example, to calculate the residuals for the OLS sample regression equation corresponding to equation (1) above and name them **uhat**, enter the following command:

```
predict uhat, residuals
```

The values of **uhat** generated by this **predict** command are computed according to equation (3) above using the numerical values of the coefficient estimates:

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 \text{weight}_i = \text{price}_i + 6.707353 - 2.044063 \text{weight}_i .$$

- To examine the results of the above two **predict** commands, enter the following commands:

```
summarize price yhat uhat
list price yhat uhat
correlate price yhat, means
correlate weight uhat, means
correlate yhat uhat, means
```

Examine the results of the preceding **correlate** commands to verify the computational properties of the OLS sample regression equation

$$\text{price}_i = \hat{Y}_i + \hat{u}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{weight}_i + \hat{u}_i \quad (4)$$

□ **Alternative Methods of Computing *Predicted Values and Residuals***

In the previous section, you learned how to use the **predict** command to calculate *predicted values and residuals* for the OLS sample regression equation (4) corresponding to regression equation (1). This section explores alternative methods of computing predicted values and residuals.

1. **Three alternative methods of calculating predicted values**

- The **predict** command: The previous section showed you how to calculate the *estimated or predicted values* \hat{Y}_i of the dependent variable price_i using the **predict** command without any options.
- The **generate** command with **_b[*varname*]**: Another way to calculate the *predicted values* \hat{Y}_i of the dependent variable price_i is to use a **generate** command together with the coefficient estimates from the previous **regress** command. For example, to calculate the predicted values of price_i and name them **yhat1**, enter the following command:

```
generate yhat1 = _b[_cons] + _b[weight]*weight
```

- The **generate** command with **scalars**: *Predicted values* \hat{Y}_i of the dependent variable price_i can also be calculated using a **generate** command together with previously-created scalars **b0** and **b1** containing the OLS coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. To illustrate, first use scalar commands to save the OLS coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ in OLS sample regression function (2). Enter the commands:

```
scalar b0 = _b[_cons]
scalar b1 = _b[weight]
```

Now calculate the predicted values of $price_i$ and name them **yhat2** by entering the following **generate** command:

```
generate yhat2 = b0 + b1*weight
```

- To demonstrate the equivalence of these three ways of generating predicted values, enter the following commands:

```
list yhat yhat1 yhat2 in 1/10
summarize yhat yhat1 yhat2
correlate yhat yhat1 yhat2, means
```

2. Four methods of calculating OLS residuals

- The **predict** command with the **residuals option**: The previous section showed you how to calculate the *OLS residuals* \hat{u}_i given by equation (3) by using the **predict** command with the **residuals option**. These residuals were named **uhat**.
- A **generate** command with **_b[varname]**: Another way to calculate the *OLS residuals* \hat{u}_i given by equation (3) above is to use a **generate** command together with the coefficient estimates from the previous **regress** command. Enter the following **generate** command:

```
generate uhat1 = price - _b[_cons] - _b[weight]*weight
```

- The **generate** command with **scalars**: The *OLS residuals* \hat{u}_i given by equation (3) can also be computed using a **generate** command together with the previously-created scalars **b0** and **b1** containing the OLS coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Enter the command:

```
generate uhat2 = price - b0 - b1*weight
```

- The **generate** command with **predicted values**: A last way to calculate the *OLS residuals* \hat{u}_i given by equation (3) is to use a **generate** command together

with the previously-created predicted values $\hat{Y}_i = \hat{\text{price}}_i$ of the dependent variable $Y_i = \text{price}_i$. For example, enter the command:

```
generate uhat3 = price - yhat1
```

- To demonstrate the equivalence of these four commands for calculating the OLS residuals, enter the following commands:

```
list uhat uhat1 uhat2 uhat3 in 1/10
summarize uhat uhat1 uhat2 uhat3
correlate uhat uhat1 uhat2 uhat3, means
```

□ Graphing sample data – graph twoway scatter

Simple forms of the **graph** command can be used to obtain a graphical representation of the sample data. Such a graphical representation of sample data points is called a *scatterplot*.

- To generate a *scatterplot* of the sample data $(Y_i, X_i) = (\text{price}_i, \text{weight}_i)$, enter the following **graph** command:

```
graph twoway scatter price weight
```

This **graph twoway scatter** command graphs price_i against weight_i . That is, the sample values of the dependent variable **price** are measured on the vertical Y-axis, and the sample values of the independent variable **weight** are measured on the horizontal X-axis.

Note:

When a **graph** command is executed, the results are displayed in the Graph window, which usually covers up the Results window. To put the Results window back on top, click on the **Results** button. To see the graph again, click on the **Graph** button. In this way, you can use the **Results** and **Graph** buttons to toggle between the Results and Graph windows.

- Use the **ytitle()** and **xtitle()** options to add your own titles for the Y-axis and X-axis to the scatterplot of the sample data generated by the preceding **graph twoway scatter** command. Enter the command:

```
graph twoway scatter price weight, ytitle("car price (U.S.
dollars)") xtitle("car weight (pounds)")
```

- Generate separate graphs (or scatterplots) of the sample data on $price_i$ and $weight_i$ for foreign and domestic cars by using the **by(varname)** option on the **graph twoway scatter** command, where *varname* is the name of the categorical variable that identifies the subsamples for which separate data scatterplots are desired.

Before issuing the **graph** command, enter a **sort** command to sort the data set by the values of the variable to be used in the **by()** option. Enter the commands:

```
sort foreign
graph twoway scatter price weight, by(foreign)
```

This **graph** command draws separate scatterplots of the data for foreign and domestic cars in a single screen, or frame.

□ Graphing the sample regression function – graph twoway line

The **graph** command can also be used to obtain a graphical representation of the **OLS sample regression function** given by equation (2) above:

$$\hat{Y}_i = \text{price}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{weight}_i \quad (i = 1, \dots, N) \quad (2)$$

- Use the following **graph twoway scatter** command to generate a *scatterplot* of the estimated (predicted) values of $price_i$ against the observed values of $weight_i$ – i.e., a scatterplot of the points $(\hat{Y}_i, X_i) = (\text{price}_i, \text{weight}_i)$ for $i = 1, \dots, N$.

Enter the command:

```
graph twoway scatter yhat weight
```

- To draw a *line graph* representing the OLS sample regression function (2), use the **graph twoway line** command. This command instructs *Stata* to draw a smooth line through the points $(\hat{Y}_i, X_i) = (\text{price}_i, \text{weight}_i)$. Enter any of the following commands:

```
graph twoway line yhat weight
twoway line yhat weight
line yhat weight
```

Note that the command name **graph twoway line** can be abbreviated as **twoway line** or simply as **line**.

- Use the **ytitle()** and **xtitle() options** to add your own titles for the Y-axis and X-axis to the line graph of the sample regression function (SRF) generated by the preceding **graph twoway line** command. Enter the command:

```
graph twoway line yhat weight, ytitle("predicted mean car
price" "(U.S. dollars)") xtitle("car weight (pounds)")
```

- Finally, use the **title()** and **subtitle() options** to add your own title and subtitle to the **line graph** of the sample regression function (SRF) generated by the preceding **graph twoway line** command. Enter *on one line* the command:

```
graph twoway line yhat weight, ytitle("predicted mean car
price" "(U.S. dollars)") xtitle("car weight (pounds)")
title("OLS Regression of Car Price on Car Weight")
subtitle("North America, 1978")
```

□ Graphing the sample data and sample regression function in one diagram

You can combine in one diagram the scatterplot of the sample data points $(Y_i, X_i) = (\text{price}_i, \text{weight}_i)$ and the straight line representing the OLS sample regression function (2), which is $\hat{Y}_i = \hat{\text{price}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{weight}_i$.

- To do this, enter the following **graph twoway** command:

```
graph twoway scatter price weight || line yhat weight
```

This command instructs *Stata* to draw on the same set of coordinate axes both (1) a **scatterplot** of the sample data points $(Y_i, X_i) = (\text{price}_i, \text{weight}_i)$ and (2) a **line graph** of the *estimated* values of **price** (i.e., **yhat** = $\hat{Y}_i = \hat{\text{price}}_i$) against the sample values of **weight**. Note that **weight** is the variable measured on the horizontal X-axis, and both **price** and **yhat** are measured on the vertical Y-axis.

- Augment the figure produced by the preceding **graph twoway** command by using the **ytitle()** and **xtitle()** *options* to add your own titles to the vertical Y-axis and the horizontal X-axis and by using the **title()** and **subtitle()** *options* to add an appropriate title to the figure. Enter *on one line* the command:

```
graph twoway scatter price weight || line yhat weight,
ytitle("car price (U.S. dollars)," "observed and estimated")
xtitle("car weight (pounds)") title("OLS Regression of Car
Price on Car Weight") subtitle("North America, 1978")
```

- The **legend()** *option* can be added to the preceding **graph twoway** command to override the default legend labels and replace them with your own labels. Enter *on one line* the command:

```
graph twoway scatter price weight || line yhat weight,
ytitle("car price (U.S. dollars)," "observed and estimated")
xtitle("car weight (pounds)") title("OLS Regression of Car
Price on Car Weight") subtitle("North America, 1978")
legend(label(1 "Sample data points") label(2 "Sample
regression line"))
```

Carefully inspect the graph produced by this **graph twoway** command. Note in particular how the legend labels are different from those in the immediately preceding graph.

- The title of a *Stata* graph can be more than one line in length. To illustrate, make the title of the graph produced with the above **graph** command more accurate by making it a two-line title that better describes the contents of the figure you are creating. In particular, modify the **title()** *option* by adding the second line “and Scatterplot of Sample Data” in a second set of double quotation marks. Enter *on one line* the command:

```
graph twoway scatter price weight || line yhat weight,
ytitle("car price (U.S. dollars)," "observed and estimated")
xtitle("car weight (pounds)") title("OLS Regression of Car
Price on Car Weight" "and Scatterplot of Sample Data")
subtitle("North America, 1978") legend(label(1 "Sample data
points") label(2 "Sample regression line"))
```

Carefully inspect the graph produced by this **graph twoway** command. Note the new two-line title of your figure.

□ Saving a graph to disk – graph save

You may want to save a graph to disk in a format such that you can use it, and perhaps modify it, either later in your current *Stata* session or during some future *Stata* session. The **graph save** command can be used to do this.

- First, enter the following **graph** command to create a scatterplot of the sample data on car price and car weight. Enter the **graph twoway scatter** command:

```
graph twoway scatter price weight, title("Sample Data on Car  
Price and Car Weight" "North America, 1978")
```

- To save the graph created by the above **graph** command to a disk file in the current *Stata* working directory and give that disk file the name **graph1_tutorial4.gph**, enter the following **graph save** command:

```
graph save graph1_tutorial4.gph
```

Note that specifying the file extension **.gph** in the name given to the disk file is optional; all **graph save** commands will create disk files with the file extension **.gph**.

□ Exporting a graph to disk – graph export

You may also want to export a graph created by *Stata* to a file on disk so that you can insert that graph into a document created by a word processor such as MS Word. The **graph export** command can be used to do this.

Basic Syntax

```
graph export newfilename.suffix [ , options ]
```

This command exports to a disk file in the current *Stata* working directory the graph currently displayed in the Graph window. The output format of this file is determined by the suffix of *newfilename.suffix*. An alternative way to specify the output format of the exported file is to use the **as(fileformat)** option, where *fileformat* specifies the desired format of the exported file.

The available *output formats* that can be specified with the **graph export** command are listed in the following table.

<i>suffix</i>	implied option	output format
.ps	as(ps)	PostScript
.eps	as(eps)	EPS (Encapsulated PostScript)
.wmf	as(wmf)	Windows Metafile
.emf	as(emf)	Windows Enhanced Metafile
.pict	as(pict)	Mackintosh PICT format
.pdf	as(pdf)	PDF
.png	as(png)	PNG (Portable Network Graphics)
.tif	as(tif)	TIFF

Notes: **ps** and **eps** are available with all versions of *Stata*; **png** and **tif** are available for all versions of *Stata* except *Stata(console)* for Unix; **wmf** and **emf** are available only with *Stata for Windows*; and **pict** and **pdf** are available only with *Stata for Mackintosh*.

Source: StataCorp 2007. *Stata Statistical Software Release 10: Graphics*. College Station, TX: StataCorp LP, p. 124.

- To illustrate how to use the **graph export** command, first re-enter the **graph** command you used previously to create a line graph of the OLS sample regression function of car weight on car price. Enter *on one line* the command:

```
graph twoway line yhat weight, ytitle("predicted mean car
price" "(U.S. dollars)") xtitle("car weight (pounds)")
title("OLS Regression of Car Price on Car Weight")
subtitle("North America, 1978")
```

- To save this line graph to a file named **graph2_tutorial14.gph** in the current *Stata* working directory, enter the **graph save** command:

```
graph save graph2_tutorial14
```

- To export this line graph in EPS (Encapsulated PostScript) format to a file named **graph2_tutorial14.eps** in the current *Stata* working directory, enter the **graph export** command:

```
graph export graph2_tutorial14.eps
```

- To export the same line graph in Windows Metafile format to a file named `graph2_tutorial4.wmf` in the current *Stata* working directory, enter the **graph export** command:

```
graph export graph2_tutorial4.wmf
```

- Suppose you want to export the graph that you previously saved on disk in the file `graph1_tutorial4.gph`. To do this, you first have to retrieve the graph and place it in the current Graph window by entering the following **graph use** command:

```
graph use graph1_tutorial4.gph
```

- You can now save this graph in TIFF format to a file in the current *Stata* working directory named `graph1_tutorial4.tif` by entering the **graph export** command:

```
graph export graph1_tutorial4.tif
```

- To illustrate how the different output formats for a *Stata* graph work, re-enter the previously issued **graph** command that creates both a scatterplot of the sample data and a line graph of the estimated OLS regression of car price on car weight. Enter *on one line* the command:

```
graph twoway scatter price weight || line yhat weight,  
ytitle("car price (U.S. dollars)," "observed and estimated")  
xtitle("car weight (pounds)") title("OLS Regression of Car  
Price on Car Weight" "and Scatterplot of Sample Data")  
subtitle("North America, 1978") legend(label(1 "Sample data  
points") label(2 "Sample regression line"))
```

- To save the graph created by the above **graph** command to a disk file in the current *Stata* working directory and give it the name `graph3_tutorial4.gph`, enter the following **graph save** command:

```
graph save graph3_tutorial4
```

Note that it was not necessary to specify the file extension `.gph` in the above **graph save** command; the saved file will nonetheless have the *filename.suffix* `graph3_tutorial4.gph`.

- To save this graph in Windows Metafile format to a file in the current *Stata* working directory named `graph3_tutorial14.wmf`, enter the **graph export** command:

```
graph export graph3_tutorial14.wmf
```

- To save the same graph in Windows Enhanced Metafile format to a file in the current *Stata* working directory named `graph3_tutorial14.emf`, enter the **graph export** command:

```
graph export graph3_tutorial14.emf
```

- To save the same graph in PNG (Portable Network Graphics) format to a file in the current *Stata* working directory named `graph3_tutorial14.png`, enter the **graph export** command:

```
graph export graph3_tutorial14.png
```

- Finally, to save the same graph in EPS (Encapsulated PostScript) format to a file in the current *Stata* working directory named `graph3_tutorial14.eps`, enter the **graph export** command:

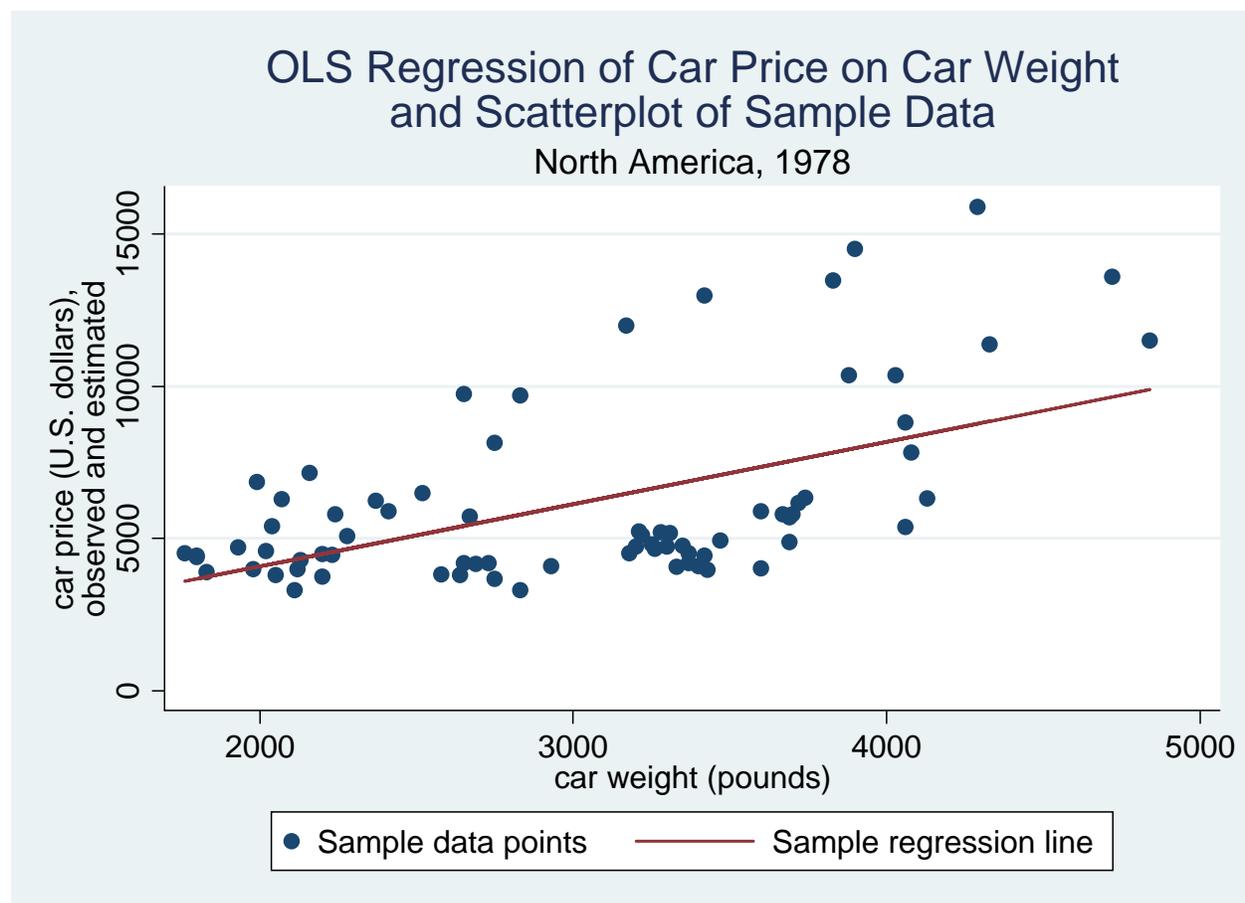
```
graph export graph3_tutorial14.eps
```

□ Inserting exported graph files into an MS Word document

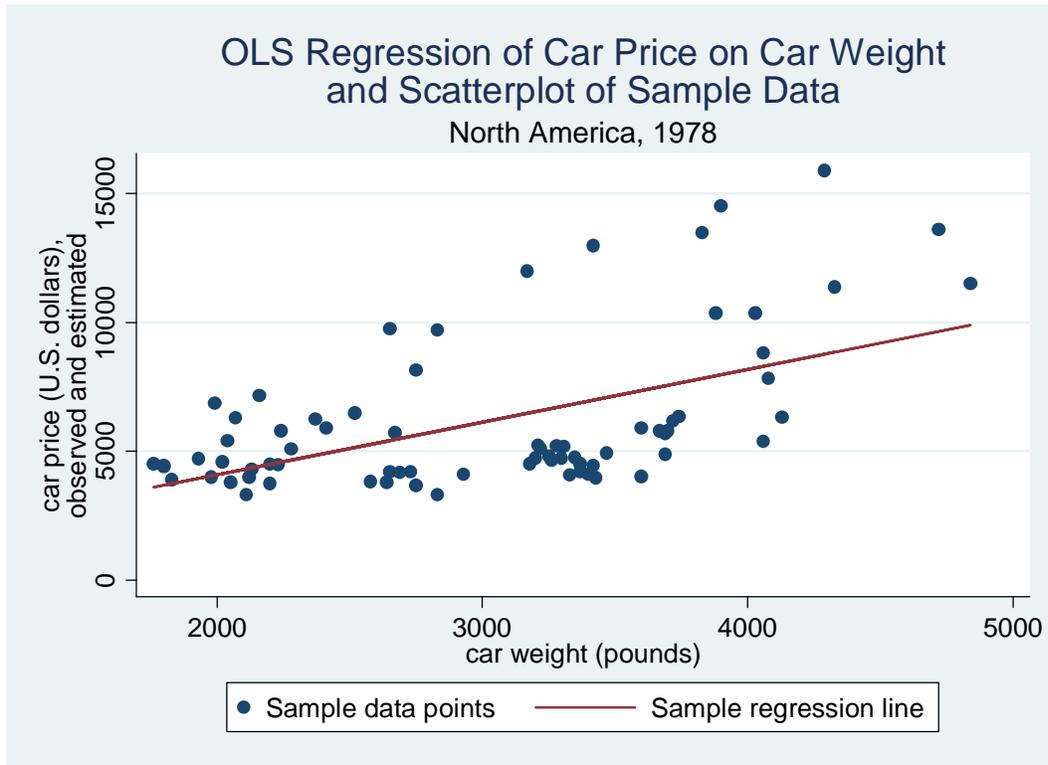
In this section, we illustrate how exported *Stata* graph files in various output formats appear when they are inserted into the current document, which is a Microsoft Word document. The exported files we insert are the following: the file `graph3_tutorial14.wmf`, which is in Windows Metafile format; the file `graph3_tutorial14.emf`, which is in Windows Enhanced Metafile format; the file `graph3_tutorial14.png`, which is in PNG (Portable Network Graphics) format; and the file `graph3_tutorial14.eps`, which is in EPS (Encapsulated PostScript) format.

In MS Word, each of these *Stata* graph files can be inserted into a document by selecting **I**nsert > **P**icture > **F**rom File... and then choosing the filename of the graph you want.

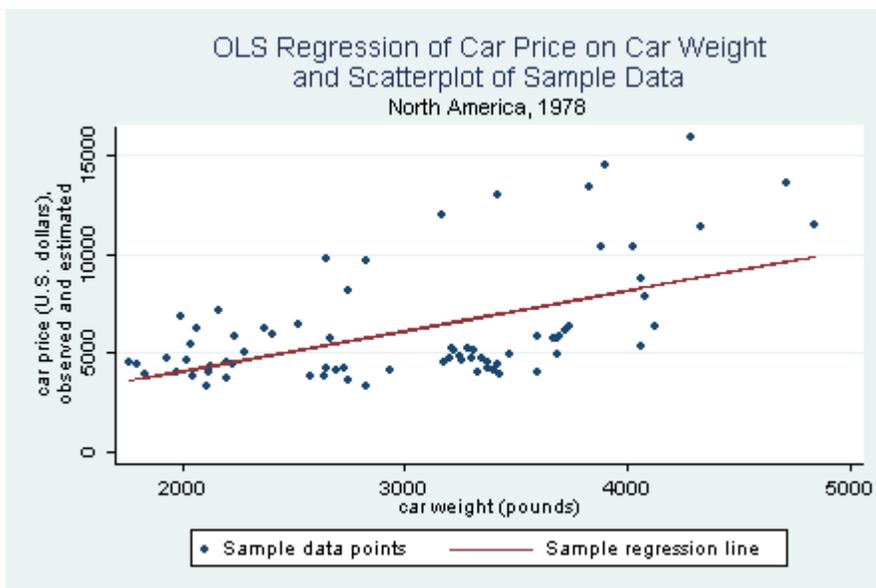
- ◆ The file `graph3_tutorial14.wmf` in Windows Metafile format:



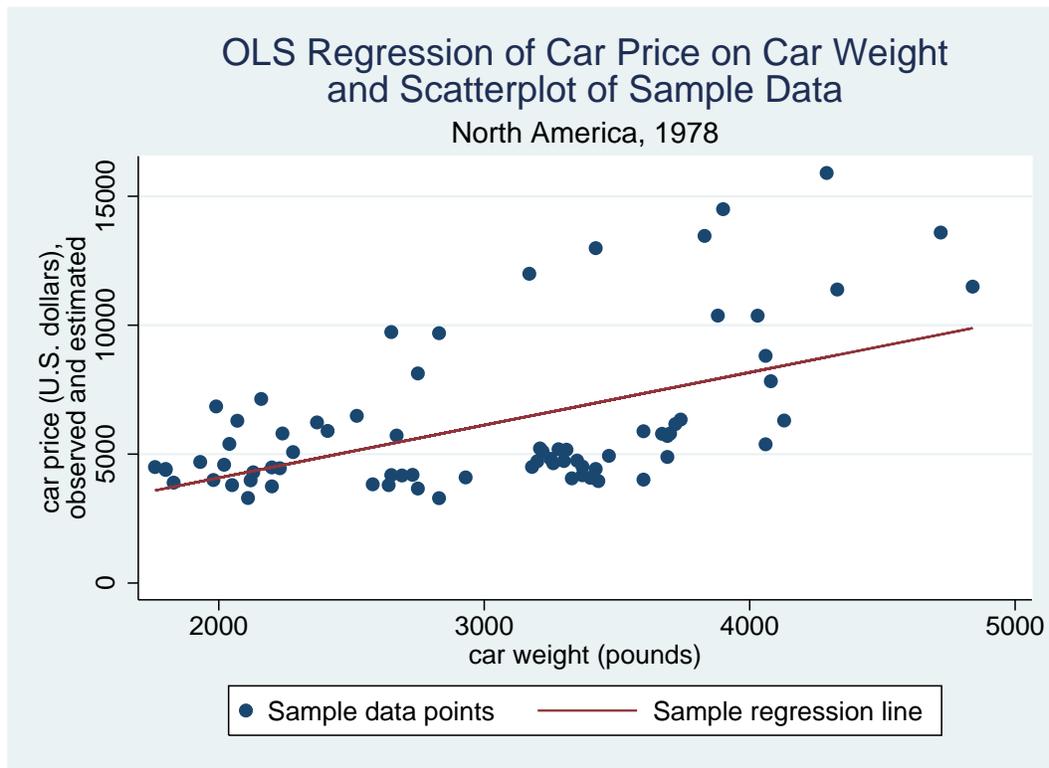
- ◆ The file `graph3_tutorial14.emf` in Windows Enhanced Metafile format:



- ◆ The file `graph3_tutorial14.png` in PNG (Portable Network Graphics) format:



- ◆ The file `graph3_tutorial14.eps` in EPS (Encapsulated PostScript) format:



□ Preparing to End Your Stata Session

Before you end your Stata session, you should do two things.

- First, you should **save the current data set**. Enter the following **save** command to save the current data set as *Stata*-format data set **auto4.dta**:

```
save auto4
```

- Second, **close the .log file** you have been recording. Enter the command:

```
log close
```

□ End Your Stata Session – exit

- **To end your Stata session**, use the **exit** command. Enter the command:

```
exit    or    exit, clear
```

□ Cleaning Up and Clearing Out

After returning to Windows, you should copy all the files you have used and created during your *Stata* session to your own diskette. These files will be found in the *Stata working directory*, which is usually **C:\data** on the computers in Dunning 350, and **D:\courses** on the computers in MC B111. There is one file you will want to be sure you have: the *Stata* log file **351tutorial4.log**. If you saved the *Stata*-format data set **auto4.dta**, you will want to take it with you as well. Finally, you will probably want to take with you the several *Stata* graph files you saved to disk during this tutorial. Use the Windows **copy** command to copy any files you want to keep to your own portable electronic storage device (e.g., flash memory stick) in the E:-drive (or to a diskette in the A:-drive).

Finally, **as a courtesy to other users** of the computing classroom, please delete all the files you have used or created from the *Stata* working directory.