Stata 10 Tutorial 6

TOPICS: **Functional Form and Variable Re-scaling in Simple Linear Regression Models**

DATA: **auto1.dta** (a *Stata*-format data file)

TASKS: ***Stata Tutorial 6*** has two primary purposes: **(1)** to introduce you to some of the alternative ***functional forms*** commonly used in linear-in-coefficients regression models; and **(2)** to investigate how ***variable re-scaling*** – that is, changing the ***units of measurement*** for Y_i and/or X_i – affects OLS estimates of the slope coefficient β_1 and the intercept coefficient β_0 in a simple linear regression equation.

- The ***Stata commands*** that constitute the primary subject of this tutorial are:

regress	Used to perform OLS estimation of simple linear regression models.
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.

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

LEARNING FROM THIS TUTORIAL: Stata Tutorial 6 contains some important analytical results. You should make sure you understand them.

HELP: *Stata* has an extensive on-line **Help** facility that provides fairly detailed information (including examples) on all *Stata* commands. 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 dataset **auto1.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 B109/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, including all the *Stata* commands you enter and the results (output) produced by these commands, make a **.log** file named **351tutorial6.log**. To open (begin) the **.log** file **351tutorial6.log**, enter in the Command window:

```
log using 351tutorial6.log
```

This command opens a file called **351tutorial6.log** in the current *Stata* working directory. Remember that once you have opened the **351tutorial6.log** file, a copy of all the commands you enter during your *Stata* session and of all the results they produce is recorded in that **351tutorial6.log** file.

An alternative way to open the **.log** file **351tutorial6.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 **351tutorial6**; and click on the **Save** button.

❑ Loading a *Stata*-Format Dataset into *Stata* – use

Load, or read, into memory the dataset you are using. To load the *Stata*-format data file **auto1.dta** into memory, enter in the Command window:

```
use auto1
```

This command loads into memory the *Stata*-format dataset **auto1.dta**.

❑ Familiarize Yourself with the Current Dataset

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

```
describe
summarize
```

❑ Alternative Functional Forms for the Simple Linear Regression Model

This section demonstrates (1) how to estimate by OLS different functional forms for the simple linear regression model relating car price ($price_i$) to car weight ($weight_i$), (2) how to use the **predict** command to compute estimated or predicted values of the regressand ($\hat{Y}_i$-values) for the sample observations, and (3) how to use the **graph twoway** command to display the OLS sample regression function corresponding to the observed sample values of the regressor $weight_i$.

1. The **LIN-LIN (Linear) Model:** This model take the general form

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{1a}$$

Setting $Y_i = price_i$ and $X_i = weight_i$, PRE (1a) takes the specific form

$$price_i = \beta_0 + \beta_1 weight_i + u_i \tag{1b}$$

- To estimate this model (again!) by OLS for the full sample of observations in dataset **auto1.dta**, and to calculate the estimated (or predicted) values of $price_i$

for the sample observations, enter in the Command window the following commands:

```
regress price weight
predict yhat
```

The **yhat** variable created by the **predict** command takes the form

$$\hat{Y}_i = \hat{price}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \hat{\beta}_0 + \hat{\beta}_1 weight_i \qquad (i = 1, ..., N) \tag{1c}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the OLS coefficient estimates for the LIN-LIN model.

- To make a scatterplot of the sample data (Y_i, X_i) = ($price_i$, $weight_i$) and a line graph of the OLS sample regression function (1c), first sort the sample data by **weight** and then use the following **graph twoway** command:

```
sort weight

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

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) = ($price_i$, $weight_i$) and (2) a ***line graph*** of the *estimated* values of **price** (i.e., **yhat** = $\hat{Y}_i = \hat{price}_i$) against the sample values of **weight**, i.e, of the points $(\hat{Y}_i, X_i)$. Note that **weight** is the variable measured on the horizontal X-axis, and both **price** and **yhat** are measured on the vertical Y-axis.

2. The **<u>LOG-LOG (Double-Log) Model</u>:** This model takes the general form

$$\ln Y_i = \alpha_0 + \alpha_1 \ln X_i + u_i \tag{2a}$$

where $\ln Y_i$ is the natural logarithm of Y_i and $\ln X_i$ is the natural logarithm of X_i.

Setting $\ln Y_i = \ln(\text{price}_i)$ and $\ln X_i = \ln(\text{weight}_i)$, PRE (2a) takes the specific form

$$\ln(\text{price}_i) = \alpha_0 + \alpha_1 \ln(\text{weight}_i) + u_i, \qquad (2b)$$

where

$\ln(\text{price}_i) =$ the natural logarithm of the variable price_i;
$\ln(\text{weight}_i) =$ the natural logarithm of the variable weight_i.

Note: The natural logarithm is defined only for variables that take only positive values. This is the case for both price_i and weight_i in the dataset **auto1.dta**.

- Before estimating the LOG-LOG model (2), you must generate the natural logarithms of the variables price_i and weight_i. Use the following *Stata* **generate** commands to do this.

```
generate lnprice = ln(price)
generate lnweight = ln(weight)
summarize lnprice lnweight
```

- To estimate the LOG-LOG model by OLS for the full sample of observations and to calculate the estimated (or predicted) values of $\ln(\text{price}_i)$ for the sample observations, enter in the Command window:

```
regress lnprice lnweight
predict lnyhatdl
```

The **lnyhatdl** variable created by the **predict** command takes the form

$$\hat{\ln Y_i} = \ln(\hat{\text{price}}_i) = \hat{\alpha}_0 + \hat{\alpha}_1 \ln X_i = \hat{\alpha}_0 + \hat{\alpha}_1 \ln(\text{weight}_i) \quad (i = 1, ..., N) \qquad (2c)$$

where $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are the OLS coefficient estimates for the LOG-LOG model and $\hat{\ln Y_i} = \ln(\hat{\text{price}}_i)$ denotes the predicted values of $\ln Y_i$.

- To make a scatterplot of the sample data ($\ln Y_i$, $\ln X_i$) and a line graph of the OLS sample regression function (2c), use the following **graph twoway** command:

```
graph twoway scatter lnprice lnweight || line lnyhatdl
lnweight, ytitle("ln(price)," "observed and estimated")
xtitle("ln(weight)") title("LOG-LOG Model of Car Price on
Car Weight") subtitle("OLS Regression and Scatterplot of
Sample Data") legend(label(1 "Sample data points") label(2
"Sample regression line"))
```

3. The **<u>LOG-LIN (Semi-Log) Model</u>:** This model takes the general form

$$\ln Y_i = \gamma_0 + \gamma_1 X_i + u_i \tag{3a}$$

Setting $\ln Y_i = \ln(\text{price}_i)$ and $X_i = \text{weight}_i$, PRE (3a) takes the specific form

$$\ln(\text{price}_i) = \gamma_0 + \gamma_1 \text{weight}_i + u_i \tag{3b}$$

- To estimate the LOG-LIN model by OLS for the full sample of observations and to calculate the estimated (or predicted) values of $\ln(\text{price}_i)$ for the sample observations, type in the Command window:

```
regress lnprice weight
predict lnyhatsl
```

The **lnyhatsl** variable created by the **predict** command takes the form

$$\hat{\ln} Y_i = \widehat{\ln(\text{price}_i)} = \hat{\gamma}_0 + \hat{\gamma}_1 X_i = \hat{\gamma}_0 + \hat{\gamma}_1 \text{weight}_i \quad (i = 1, ..., N) \tag{3c}$$

where $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are the OLS coefficient estimates for the LOG-LIN model and $\hat{\ln} Y_i = \widehat{\ln(\text{price}_i)}$ denotes the predicted values of $\ln Y_i$.

- To make a scatterplot of the sample data ($\ln Y_i$, X_i) and a graph of the OLS sample regression function (3c), use the following **graph twoway** command:

```
graph twoway scatter lnprice weight || line lnyhatsl weight,
ytitle("ln(car price)," "observed and estimated")
xtitle("car weight (pounds)") title("LOG-LIN Model of Car
Price on Car Weight") subtitle("OLS Regression and
Scatterplot of Sample Data") legend(label(1 "Sample data
points") label(2 "Sample regression line"))
```

❑ Units of Measurement and Re-scaling of Variables

The coefficient estimates in linear (LIN-LIN) regression models depend on the units of measurement for the dependent variable Y_i and the independent variable X_i. This section presents some analytical results on how changing units of measurement for Y_i and/or X_i affects the OLS estimates of the slope coefficient β_1 and the intercept coefficient β_0 in a simple linear regression equation. It then illustrates these results with a simple linear regression model.

<u>Analysis</u>: (There are no *Stata* commands in this section.)

The term "re-scaling a variable" means multiplying that variable by a constant; this is what happens when we change the units in which a variable is measured.

Write the original regression equation, expressed in terms of the original variables Y_i and X_i, as equation (4):

$$Y_i = \beta_0 + \beta_1 X_i + u_i \,. \qquad (4)$$

Re-scale the original variables Y_i and X_i by multiplying each by some arbitrarily-selected constant. Create the re-scaled variable $\dot{X}_i$ by multiplying X_i by the constant c:

$$\dot{X}_i = cX_i \qquad (i = 1, \ldots, N), \text{ where c is a specified constant.}$$

Similarly, create the re-scaled variable $\dot{Y}_i$ by multiplying Y_i by the constant d:

$$\dot{Y}_i = dY_i \qquad (i = 1, \ldots, N), \text{ where d is a specified constant.}$$

The new regression equation written in terms of the re-scaled variables $\dot{X}_i$ and $\dot{Y}_i$ can be written as:

$$\dot{Y}_i = \beta_{0\bullet} + \beta_{1\bullet}\dot{X}_i + \dot{u}_i . \tag{5}$$

Questions:

How is the OLS estimate of the slope coefficient $\beta_{1\bullet}$ in equation (5) related to the OLS estimate of β_1 in equation (4)?

How is the OLS estimate of the intercept coefficient $\beta_{0\bullet}$ in equation (5) related to the OLS estimate of β_0 in equation (4)?

Answers:

The formula for the OLS estimator of β_1 in the original equation (4) is:

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \quad \text{where} \quad x_i = X_i - \overline{X} \text{ and } y_i = Y_i - \overline{Y} \quad (i = 1, \ldots, N).$$

The formula for the OLS estimator of β_0 in the original equation (4) is:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X} \quad \text{where} \quad \overline{Y} = \sum Y_i / N \text{ and } \overline{X} = \sum X_i / N .$$

The formula for the OLS estimator of $\beta_{1\bullet}$ in the re-scaled equation (5) is:

$$\hat{\beta}_{1\bullet} = \frac{\sum \dot{x}_i \dot{y}_i}{\sum \dot{x}_i^2} \quad \text{where} \quad \dot{x}_i = \dot{X}_i - \overline{\dot{X}} \text{ and } \dot{y}_i = \dot{Y}_i - \overline{\dot{Y}} \quad (i = 1, \ldots, N). \tag{6}$$

To see how the new slope coefficient estimator $\hat{\beta}_{1\bullet}$ is related to the original slope coefficient estimator $\hat{\beta}_1$, we need to determine how the re-scaled deviations-from-means variables $\dot{x}_i = \dot{X}_i - \overline{\dot{X}}$ and $\dot{y}_i = \dot{Y}_i - \overline{\dot{Y}}$ are related to the original deviations-from-means variables $x_i = X_i - \overline{X}$ and $y_i = Y_i - \overline{Y}$. Here is the algebra:

$$\dot{X}_i = cX_i \;\Rightarrow\; \overline{\dot{X}} = c\overline{X} \;\Rightarrow\; \dot{x}_i = \dot{X}_i - \overline{\dot{X}} = cX_i - c\overline{X} = c(X_i - \overline{X}) = cx_i;$$

$$\dot{Y}_i = dY_i \;\Rightarrow\; \overline{\dot{Y}} = d\overline{Y} \;\Rightarrow\; \dot{y}_i = \dot{Y}_i - \overline{\dot{Y}} = dY_i - d\overline{Y} = d(Y_i - \overline{Y}) = dy_i.$$

Thus, we see that $\dot{x}_i = cx_i$ and $\dot{y}_i = dy_i$ (i = 1, …, N). These two equalities in turn imply the following results:

$$\dot{x}_i\dot{y}_i = cx_i dy_i = cdx_i y_i \;\Rightarrow\; \sum \dot{x}_i\dot{y}_i = cd\sum x_i y_i;$$

$$\dot{x}_i^2 = (cx_i)^2 = c^2 x_i^2 \;\Rightarrow\; \sum \dot{x}_i^2 = c^2 \sum x_i^2.$$

Now substitute these results into expression (6) for $\hat{\beta}_{1\bullet}$:

$$\hat{\beta}_{1\bullet} = \frac{\sum \dot{x}_i\dot{y}_i}{\sum \dot{x}_i^2} = \frac{cd\sum x_i y_i}{c^2\sum x_i^2} = \frac{d}{c}\frac{\sum x_i y_i}{\sum x_i^2} = \frac{d}{c}\hat{\beta}_1.$$

The formula for the OLS estimator of $\beta_{0\bullet}$ in the re-scaled equation (5) is:

$$\hat{\beta}_{0\bullet} = \overline{\dot{Y}} - \hat{\beta}_{1\bullet}\overline{\dot{X}}. \tag{7}$$

To see how the new intercept coefficient estimator $\hat{\beta}_{0\bullet}$ is related to the original intercept coefficient estimator $\hat{\beta}_0$, substitute into expression (7) for $\hat{\beta}_{0\bullet}$ the previous results showing that $\hat{\beta}_{1\bullet} = \frac{d}{c}\hat{\beta}_1$, $\overline{\dot{Y}} = d\overline{Y}$ and $\overline{\dot{X}} = c\overline{X}$:

$$\hat{\beta}_{0\bullet} = \overline{\dot{Y}} - \hat{\beta}_{1\bullet}\overline{\dot{X}} = d\overline{Y} - \frac{d}{c}\hat{\beta}_1 c\overline{X} = d\overline{Y} - d\hat{\beta}_1\overline{X} = d(\overline{Y} - \hat{\beta}_1\overline{X}) = d\hat{\beta}_0.$$

Results:

$$\hat{\beta}_{1\bullet} = \frac{d}{c}\hat{\beta}_1 \quad \Rightarrow \quad \hat{\beta}_{1\bullet} \text{ is affected by the re-scaling of both } Y_i \text{ and } X_i. \qquad (8)$$

$$\hat{\beta}_{0\bullet} = d\hat{\beta}_0 \quad \Rightarrow \quad \hat{\beta}_{0\bullet} \text{ is affected only by the re-scaling of } Y_i. \qquad (9)$$

Some Examples

To illustrate the effects of variable re-scaling – i.e., of changing the units of measurement for Y_i and/or X_i – we investigate how changing the units of measurement for the variables in regression equation (1) affect the OLS coefficient estimates. For convenience, the original equation (1) is rewritten here as:

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

where price_i = car price measured in US dollars and weight_i = car weight measured in pounds.

1. Re-scale *only* the *dependent* variable.

Re-scale the dependent variable price_i so that it is measured in ***hundreds* of US dollars** instead of US dollars.

- Generate the re-scaled price_i variable newp_i = car price measured in hundreds of US dollars, where $\text{newp}_i = \text{price}_i/100$. Enter the command:

```
generate newp = price/100
```

- Compare the sample values of the original price_i variable with those of the re-scaled price variable newp_i. Enter the commands:

```
summarize price newp
regress newp price
```

- Estimate by OLS the regression equation with $newp_i$ as dependent variable and $weight_i$ as the independent variable. Enter the command:

```
regress newp weight
```

Carefully compare the results of this command with those from OLS estimation of the original regression equation (1). Which results have changed as a result of re-scaling only the dependent variable?

2. Re-scale *only* the *independent* variable.

Re-scale the independent variable $weight_i$ so that it is measured in ***kilograms*** instead of pounds, where 1 kilogram = 2.2 pounds.

- Generate the re-scaled $weight_i$ variable $neww_i$ = car weight measured in kilograms, where $neww_i = weight_i/2.2$. Enter the command:

```
generate neww = weight/2.2
```

- Compare the sample values of the original $weight_i$ variable with those of the re-scaled weight variable $neww_i$. Enter the commands:

```
summarize weight neww
regress neww weight
regress weight neww
```

- Estimate by OLS the regression equation with $price_i$ as dependent variable and $neww_i$ as the independent variable. Enter the command:

```
regress price neww
```

Carefully compare the results of this command with those from OLS estimation of the original regression equation (1). Which results have changed as a result of re-scaling only the independent variable?

3. Re-scale *both* the *dependent* variable and the *independent* variable.

Re-scale both the dependent variable $price_i$ and the independent variable $weight_i$ as above. The re-scaled dependent variable is $newp_i$ = car price measured in *hundreds* of US dollars, where $newp_i = price_i/100$. The re-scaled independent variable is $neww_i$ = car weight measured in *kilograms*, where $neww_i = weight_i/2.2$.

- Estimate by OLS the regression equation with $newp_i$ as dependent variable and $neww_i$ as the independent variable. Enter the command:

```
regress newp neww
```

Carefully compare the results of this command with those from OLS estimation of the original regression equation (1). Which results have changed as a result of re-scaling both the dependent and independent variables?

❑ Preparing to End Your *Stata* Session

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

- First, you may want to **save the current dataset** (although you will not need it for future tutorials). Enter the following **save** command to save the current dataset as *Stata*-format dataset **auto6.dta**:

```
save auto6
```

- 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 **351tutorial6.log**. If you saved the *Stata*-format data set **auto6.dta**, you will probably want to take it with you as well. 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.