Stata 10 Tutorial 6

TOPICS: **Functional Form and Variable Re-scaling in Simple Linear Regression Models, and An Introduction to Multiple Linear Regression Models**

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

TASKS: ***Stata Tutorial 6*** has three primary purposes: **(1)** to introduce you to some of the alternative *functional forms* commonly used in linear-in-coefficients regression models; **(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; and **(3)** to introduce you to OLS estimation of *multiple linear regression models* with two or more regressors, in particular, to demonstrate how the slope coefficient estimates in a multiple linear regression model are interpreted.

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

  regress   Used to perform OLS estimation of simple and multiple 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.
  lincom   Used after estimation to compute linear combinations of coefficient estimates and associated statistics.

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.

❑ **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 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 $Y_i = \text{price}_i$ and $X_i = \text{weight}_i$, PRE (1a) takes the specific form

$$\text{price}_i = \beta_0 + \beta_1 \text{weight}_i + u_i \tag{1b}$$

• 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

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

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 $(Y_i, X_i) = (\text{price}_i, \text{weight}_i)$ 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 Line 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** $= \hat{Y}_i = \hat{\text{price}}_i$) 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 **LOG-LOG (Double-Log) Model:** This model takes the general form

$$\ln Y_i = \alpha_0 + \alpha_1 \ln X_i + u_i \qquad (2a)$$

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

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

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(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

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

where $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are the OLS coefficient estimates for the LOG-LOG model and $\ln \hat{Y}_i = \ln(\hat{\text{price}}_i)$ 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 Line and Scatterplot
  of Sample Data") legend(label(1 "Sample data points")
  label(2 "Sample regression line"))
  ```

3. The **LOG-LIN (Semi-Log) Model:** 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

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

  where $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are the OLS coefficient estimates for the LOG-LIN model and $\hat{\ln} Y_i = \ln(\hat{\text{price}}_i)$ 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 in Simple Regression Models**

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.

Analysis: (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):

$$Y_i = \beta_0 + \beta_1 X_i + u_i. \tag{4}$$

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:

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

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

$\dot{Y}_i = dY_i \quad$ (i = 1, …, N), 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:

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

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:

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

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:

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

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

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:

$$\dot{x}_i\dot{y}_i = cx_idy_i = cdx_iy_i \quad \Rightarrow \quad \sum\dot{x}_i\dot{y}_i = cd\sum x_iy_i;$$

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

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

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

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:

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

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

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:

$$\text{price}_i = \beta_0 + \beta_1 \text{weight}_i + u_i \tag{1}$$

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

- First, re-estimate by OLS the LIN-LIN model given by regression equation (1) and save as scalars the OLS coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Enter the commands:

```
regress price weight
scalar b0eq1 = _b[_cons]
scalar b1eq1 = _b[weight]
scalar list b0eq1 b1eq1
```

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. Use **scalar** and **display** commands to save the resulting OLS coefficient estimates and compare them with the coefficient estimates for the baseline model given by PRE (1). Enter the commands:

  ```
  regress newp weight
  scalar b0eq2 = _b[_cons]
  scalar b1eq2 = _b[weight]
  scalar list b0eq1 b0eq2 b1eq1 b1eq2
  display b0eq2/b0eq1
  display b1eq2/b1eq1
  ```

  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. Use **scalar** and **display** commands to save the resulting OLS coefficient estimates and compare them with the coefficient estimates for the baseline model given by PRE (1). Enter the commands:

  ```
  regress price neww
  scalar b0eq3 = _b[_cons]
  scalar b1eq3 = _b[neww]
  scalar list b0eq1 b0eq3 b1eq1 b1eq3
  display b0eq3/b0eq1
  display b1eq3/b1eq1
  ```

  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. Use **scalar** and **display** commands to save the resulting OLS coefficient estimates and compare them with the coefficient estimates for the baseline model given by PRE (1). Enter the commands:

  ```
  regress newp neww
  scalar b0eq4 = _b[_cons]
  scalar b1eq4 = _b[neww]
  scalar list b0eq1 b0eq4 b1eq1 b1eq4
  display b0eq4/b0eq1
  display b1eq4/b1eq1
  ```

  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?

❑ Interpreting Slope Coefficients in Multiple Linear Regression Models

In this section, we introduce you to the use of the *Stata* **regress** command to compute OLS estimates of *multiple* linear regression equations that contain two or more regressors. The mechanics of using the **regress** command for estimating multiple linear regression models are a straightforward extension of those for estimating simple linear regression models. This section therefore focuses more on how the *interpretation* **of slope coefficient estimates** in multiple linear regression models differs from the interpretation of slope coefficient estimates in simple linear regression models.

Consider two different linear regression models for car price, *price$_i$*.

Model 1 is the *simple* linear regression model given by population regression equation (10.1) and the corresponding population regression function (10.2):

$$\text{price}_i = \beta_0 + \beta_1 \text{weight}_i + u_i \tag{10.1}$$

$$E\left(\text{price}_i \mid \text{weight}_i\right) = \beta_0 + \beta_1 \text{weight}_i \tag{10.2}$$

Model 2 is the *multiple* linear regression model given by population regression equation (11.1) and the corresponding population regression function (11.2):

$$\text{price}_i = \beta_0 + \beta_1 \text{weight}_i + \beta_2 \text{mpg}_i + u_i \tag{11.1}$$

$$E\left(\text{price}_i \mid \text{weight}_i, \text{mpg}_i\right) = \beta_0 + \beta_1 \text{weight}_i + \beta_2 \text{mpg}_i \tag{11.2}$$

Question: How does the slope coefficient β_1 in the *simple* linear regression model 1 given by equations (10.1) and (10.2) differ from the slope coefficient β_1 in the *multiple* linear regression model 2 given by equations (11.1) and 11.2)?

Analytical Answer

♦ The slope coefficient β_1 in the *simple* linear regression model given by equations (10.1) and (10.2) is the *unadjusted* **or** *total* **marginal effect of car weight** on **mean car price**, because PRE (10.1) and PRF (10.2) do not account for, or control for, the effects on car prices of any other explanatory variables apart from *weight$_i$*.

♦ Analytically, this means that the slope coefficient β_1 in the simple linear regression model (10.1)/ (10.2) corresponds to the **total derivative of mean car price with respect to** *weight*$_i$:

$$\frac{d\,E\left(price_i\mid weight_i\right)}{d\,weight_i} = \frac{d\left(\beta_0 + \beta_1 weight_i\right)}{d\,weight_i} = \beta_1 \text{ in } \textbf{Model 1}$$

    = the *unadjusted* or *total* **marginal effect of** *weight*$_i$ on the **mean** *price* **of cars** sold in North America in 1978

    = the **change in** *mean* **car** *price*, in 1978 US dollars, associated with a **1-pound** *increase* **in car** *weight*

♦ In contrast, the slope coefficient β_1 in the **multiple** linear regression model 2 given by equations (11.1) and (11.2) is the **adjusted** or **partial** **marginal effect of car weight** on **mean** **car** *price*, because PRE (11.1) and PRF (11.2) account for, or control for, the effect on car prices of another explanatory variable apart from *weight*$_i$, namely cars' fuel efficiency as measured by miles-per-gallon *mpg*$_i$.

♦ Analytically, this means that **the slope coefficient β_1 in the *multiple* linear regression model (11.1)/ (11.2)** corresponds to the *partial* **derivative of mean car price with respect to** *weight*$_i$:

$$\frac{\partial\,E\left(price_i\mid weight_i, mpg_i\right)}{\partial\,weight_i} = \frac{\partial\left(\beta_0 + \beta_1 weight_i + \beta_2 mpg_i\right)}{\partial\,weight_i} = \beta_1 \text{ in } \textbf{Model 2}$$

    = the *adjusted* or *partial* **marginal effect of** *weight*$_i$ on the *conditional* **mean** *price* **of cars** sold in North America in 1978

    = the **change in** *conditional mean* **car price, in 1978 US dollars,** associated with a **1-pound increase in car** *weight*, **holding constant** the **fuel efficiency of cars as measured by** *mpg*

    = the **change in** *conditional mean* **car** *price*, **in 1978 US dollars,** associated with a **1-pound increase in car** *weight* **for cars of** *the same* **fuel efficiency**.

To illustrate the difference between the slope coefficient of *weight* in the simple linear regression model, Model 1, and the slope coefficient of *weight* in the

multiple linear regression model, Model 2, we will use the *Stata* **regress** command to compute OLS estimates of the two models.

Interpreting the slope coefficient estimate of *weight* in Model 1

♦ Write the **OLS sample regression function (SRF) for Model 1** as

$$\tilde{\text{price}}_i = \tilde{\beta}_0 + \tilde{\beta}_1 \text{weight}_i \tag{10.3}$$

where $\tilde{\beta}_j$ denotes the OLS estimate of β_j in Model 1, and $\tilde{\text{price}}_i$ is the OLS estimate of mean car price given by the population regression function (PRF) $E(\text{price}_i \,|\, \text{weight}_i) = \beta_0 + \beta_1 \text{weight}_i$ for Model 1, the simple linear regression model given by PRE (10.1) and PRF (10.2).

♦ The OLS SRF (10.3) for Model 1 implies that the *estimated* **change in** *mean* **car** *price* associated with **a change in car** *weight* of Δweight is

$$\Delta\tilde{\text{price}} = \tilde{\beta}_1 \Delta\text{weight} \tag{10.4}$$

♦ In Model 1, the estimated effect on mean car price of a 1-pound increase in car weight can be obtained by setting $\Delta\text{weight} = 1$ in equation (10.4):

$$\Delta\tilde{\text{price}} = \tilde{\beta}_1 \quad \text{when } \Delta\text{weight} = 1 \tag{10.5}$$

The slope coefficient estimate $\tilde{\beta}_1$ in Model 1 is therefore **an *estimate* of the change in** *mean* **car** *price* associated with **a 1-pound increase in** *weight*, holding constant no other explanatory variables that may be related to car price.

Interpreting the slope coefficient estimate of *weight* in Model 2

♦ Write the **OLS sample regression function (SRF) for Model 2**, obtained by OLS estimation of the multiple linear regression equation (11.1), as

$$\hat{\text{price}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{weight}_i + \hat{\beta}_2 \text{mpg}_i \tag{11.3}$$

where $\hat{\beta}_j$ denotes the OLS estimate of β_j in Model 2, and $\hat{price}_i$ is the OLS estimate of mean car price given by the population regression function (PRF) $E\left(price_i \mid weight_i\right) = \beta_0 + \beta_1 weight_i + \beta_2 mpg_i$ for Model 2, the multiple linear regression model given by PRE (11.1) and PRF (11.2).

♦ The OLS SRF (11.3) implies that the *estimated* **change in *mean* car *price*** associated with **a change in car *weight*** of $\Delta weight$ *and* **a simultaneous change in *fuel efficiency*** of Δmpg is

$$\Delta \hat{price} = \hat{\beta}_1 \Delta weight + \hat{\beta}_2 \Delta mpg \qquad (11.4)$$

♦ We can hold constant the fuel efficiency of cars by setting $\Delta mpg = 0$ in equation (11.4); the resulting change in estimated mean car price is then

$$\Delta \hat{price} = \hat{\beta}_1 \Delta weight \quad \text{when } \Delta mpg = 0 \qquad (11.5)$$

♦ In Model 2, the estimated effect on mean car price of a 1-pound increase in car weight when fuel efficiency is held constant can be obtained by setting $\Delta weight = 1$ in equation (11.5), or equivalently by setting $\Delta weight = 1$ and $\Delta mpg = 0$ in equation (11.4):

$$\Delta \hat{price} = \hat{\beta}_1 \quad \text{when } \Delta weight = 1 \text{ } and \text{ } \Delta mpg = 0 \qquad (11.6)$$

The slope coefficient estimate $\hat{\beta}_1$ in Model 2 is therefore **an *estimate* of the change in *mean* car *price*** associated with a **1-pound increase in *weight*, <u>holding constant</u> fuel efficiency** as measured by *mpg*.

The following *Stata* exercises are designed to illustrate the preceding analysis of how the slope coefficient estimate $\tilde{\beta}_1$ in the simple linear regression model, Model 1, differs from the slope coefficient estimate $\hat{\beta}_1$ in the multiple linear regression model, Model 2. It also illustrates the meaning of "holding constant other variables" in multiple linear regression models. These exercises also introduce you to an important post-estimation *Stata* command, the **lincom** command.

Stata **Exercise 1: Model 1**

- First, estimate (again!) by OLS the simple linear regression model, Model 1, on the full sample of observations in dataset **auto1.dta**. Enter in the Command window the following **regress** command:

  ```
  regress price weight
  ```

- Use a *Stata* **lincom** command to compute the estimated mean price of a car that weighs 3,000 pounds (for which *weight* = 3000), i.e., to compute $\tilde{E}(\text{price}_i \mid \text{weight}_i = 3000) = \tilde{\beta}_0 + \tilde{\beta}_1 3000$. Enter the **lincom** command:

  ```
  lincom _b[_cons] + _b[weight]*3000
  ```

- Next, increase *weight* by 1 pound, from 3,000 to 3,001 pounds. Use another *Stata* **lincom** command to compute the estimated mean price of a car that weighs 3,001 pounds (for which *weight* = 3001), i.e., to compute $\tilde{E}(\text{price}_i \mid \text{weight}_i = 3001) = \tilde{\beta}_0 + \tilde{\beta}_1 3001$. Enter the **lincom** command:

  ```
  lincom _b[_cons] + _b[weight]*3001
  ```

- Finally, use a *Stata* **lincom** command to compute **the *difference*** between the estimated mean price of a car that weighs 3,001 pounds and the estimated mean price of a car that weighs 3,000 pounds, i.e., to compute

$$\tilde{E}(\text{price}_i \mid \text{weight}_i = 3001) - \tilde{E}(\text{price}_i \mid \text{weight}_i = 3000).$$

$$= \tilde{\beta}_0 + \tilde{\beta}_1 3001 - (\tilde{\beta}_0 + \tilde{\beta}_1 3000).$$

$$= \tilde{\beta}_0 - \tilde{\beta}_0 + \tilde{\beta}_1 (3001 - 3000).$$

$$= \tilde{\beta}_1$$

Enter on one line the **lincom** command:

  ```
  lincom _b[_cons] + _b[weight]*3001 - (_b[_cons] +
  _b[weight]*3000)
  ```

Compare the output of this last **lincom** command with the coefficient estimate $\widetilde{\beta}_1$ for the regressor ***weight**_i* produced by the **regress** command used to estimate Model 1 by OLS. You will see that they are identical.

Result of Exercise 1: We have demonstrated in this exercise that the slope coefficient estimate $\widetilde{\beta}_1$ of the regressor ***weight**_i* in Model 1 is an estimate of the ***total* effect** on ***mean* car *price*** of **a 1-pound increase in car *weight* when no other explanatory variables that might affect car *price* are held constant.**

Stata **Exercise 2: Model 2**

- Estimate by OLS the multiple linear regression model, Model 2, on the full sample of 74 observations in dataset **auto1.dta**. Enter in the Command window the following **regress** command:

  ```
  regress price weight mpg
  ```

- Use a *Stata* **lincom** command to compute for Model 2 the estimated mean price of a car that weighs 3,000 pounds (for which *weight* = 3000) and has fuel efficiency equal to 21 miles per gallon (for which *mpg* = 21), i.e., to compute $\hat{E}\left(\text{price}_i \mid \text{weight}_i = 3000, \text{mpg}_i = 21\right) = \hat{\beta}_0 + \hat{\beta}_1 3000 + \hat{\beta}_2 21$. Enter the **lincom** command:

  ```
  lincom _b[_cons] + _b[weight]*3000 + _b[mpg]*21
  ```

- Next, increase *weight* by 1 pound, from 3,000 to 3,001 pounds, while holding constant fuel efficiency at a value of 21 miles-per-gallon. Use another *Stata* **lincom** command to compute for Model 2 the estimated mean price of a car that weighs 3,001 pounds (for which *weight* = 3001) and has fuel efficiency equal to 21 miles per gallon (for which *mpg* = 21), i.e., to compute $\hat{E}\left(\text{price}_i \mid \text{weight}_i = 3001, \text{mpg}_i = 21\right) = \hat{\beta}_0 + \hat{\beta}_1 3001 + \hat{\beta}_2 21$. Note that we are holding fuel efficiency constant at *mpg* = 21. Enter the **lincom** command:

  ```
  lincom _b[_cons] + _b[weight]*3001 + _b[mpg]*21
  ```

- Finally, use a *Stata* **lincom** command to compute **the *difference*** between (1) the estimated mean price of a car that weighs **3,001 pounds** and has fuel efficiency of 21 miles-per-gallon and (2) the estimated mean price of a car that

weighs **3,000 pounds** and has fuel efficiency of 21 miles-per-gallon, i.e., to compute

$$\hat{E}\left(price_i \mid weight_i = 3001, mpg_i = 21\right) - \hat{E}\left(price_i \mid weight_i = 3000, mpg_i = 21\right)$$

$$= \hat{\beta}_0 + \hat{\beta}_1 3001 + \hat{\beta}_2 21 - (\hat{\beta}_0 + \hat{\beta}_1 3000 + \hat{\beta}_2 21)$$

$$= \hat{\beta}_0 + \hat{\beta}_1 3001 + \hat{\beta}_2 21 - \hat{\beta}_0 - \hat{\beta}_1 3000 - \hat{\beta}_2 21$$

$$= \hat{\beta}_1 (3001 - 3000)$$

$$= \hat{\beta}_1$$

Enter on one line the **lincom** command:

```
lincom _b[_cons] + _b[weight]*3001 + _b[mpg]*21 - (_b[_cons]
+ _b[weight]*3000 + _b[mpg]*21)
```

Compare the output of this last **lincom** command with the slope coefficient estimate $\hat{\beta}_1$ for the regressor *weight_i* produced by the **regress** command used to estimate Model 2 by OLS. You will see that they are identical.

Result of Exercise 2: We have demonstrated in this exercise that the slope coefficient estimate $\hat{\beta}_1$ of the regressor *weight_i* in Model 2 is an estimate of the *partial* **effect** on *mean* **car** *price* of **a 1-pound increase in car *weight* when fuel efficiency as measured by *mpg* is held constant** at some fixed value.

❑ **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.

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.