Stata 12 Tutorial 6

TOPIC: **Representing Multi-Category Categorical Variables with Dummy Variable Regressors**

DATA: **wage1_econ452.dta** (a *Stata*-format dataset)

TASKS: ***Stata 12 Tutorial 6*** deals with issues concerning the interpretation, testing and graphical representation of the conditional/marginal effects of multi-category categorical variables and with differences in the conditional/marginal effects of multi-category categorical variables between two mutually exclusive population subgroups (e.g., males and females). It illustrates these matters in terms of a simple ln-wage regression model for male and female employees in which the multi-categorical variable of interest is industry of employment, as represented by a seven-category explanatory variable.

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

  regress Used to perform OLS estimation of multiple linear regression models.
  lincom Used after estimation to compute linear combinations of coefficient estimates and associated test statistics.
  test Used to compute Wald F-tests of linear coefficient restrictions, with **notest** and **accumulate** options.
  return list Used to display all temporarily saved results from the most recent **test** or **lincom** command.
  graph bar Used to create bar graphs of the conditional/marginal effects of a multi-category categorical variable in linear regression models.
  graph export Exports the graph currently displayed in the Graph window to a file in the current *Stata* working directory.
  margins Used after OLS estimation to compute estimates of the ***conditional* and *marginal* effects of *categorical*, or *indicator*, explanatory variables**.

- No ***Stata* statistical functions** are used in this tutorial.

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

❑ Preparing for Your *Stata* Session

Before beginning your *Stata* session, use Windows Explorer to copy the *Stata*-format dataset **wage1_econ452.dta** to the *Stata* ***working directory*** on the C:-drive or D:-drive of the computer at which you are working.

- **<u>On the computers in Dunning 350</u>**, the default *Stata* working directory is usually **C:\data**.

- **<u>On the computers in MC B111</u>**, the default *Stata* working directory is usually **D:\courses**.

❑ Start Your *Stata* Session

<u>To start your *Stata* session</u>, double-click on the ***Stata* icon** on the Windows desktop or in the **Start** menu under **<u>P</u>rograms**.

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

❑ Record Your *Stata* Session -- log using

<u>To record your *Stata* session</u>, including all the *Stata* commands you enter and the results (output) produced by these commands, make a text-format **.log** file named **452tutorial6.log**. To open (begin) the log file **452tutorial6.log**, enter in the Command window:

```
log using 452tutorial6.log
```

This command opens a plain text-format (ASCII) file called **452tutorial6.log** in the current *Stata* working directory.

Note: It is important to include the **.log** file extension when opening a log file; if you do not, your log file will be in smcl format, a format that only *Stata* can read. Once you have opened the **452tutorial6.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 **452tutorial6.log** file.

❑ Record Only Your *Stata* Commands -- cmdlog using

To record only the *Stata* commands you type during your *Stata* session, you can use the *Stata* **cmdlog using** command. To start (open) the command log file **452tutorial6.txt**, enter in the Command window:

```
cmdlog using 452tutorial6
```

This command opens a plain text-format (ASCII) file called **452tutorial6.txt** in the current *Stata* working directory. All commands you enter during your *Stata* session are recorded in this file.

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

Be certain that you have downloaded the *Stata*-format dataset **wage1_econ452.dta** from the ECON 452 course web site, and have placed it in the *Stata* working directory.

To check that the *Stata*-format dataset 'auto1.dta' is in the current *Stata* working directory of the computer at which you are working, type in the Command window:

```
dir wage1_econ452.*
```

You should see in the *Stata* Results window the filename 'wage1_econ452.dta'.

To load, or read, into memory the *Stata*-format dataset auto1.dta, type in the Command window:

```
use wage1_econ452
```

This command loads into memory the *Stata*-format dataset **wage1_econ452.dta**.

To summarize the contents of the current dataset, use the **describe** and **summarize** commands. Type in the Command window the following commands:

```
describe
summarize
```

❑ Model 1 – Different Intercepts for Male and Female Employees

The dataset **wage1_econ452.dta** contains a **binary indicator (dummy) variable** ***female*** $\textit{female}_i$ that distinguishes between female and male workers. The ***female indicator*, or *dummy*, variable** is defined as follows:

$$\text{female}_i \;= 1 \text{ if the i-th worker is female}$$
$$\quad\quad\;\; = 0 \text{ if the i-th worker is male}$$

- To see for yourself how the dummy variable $\textit{female}_i$ is coded, as well as how many workers in the sample are male and how many are female, enter the following commands:

```
codebook female
tab1 female
summarize female, detail
```

In this section, we estimate by OLS a regression model for the natural logarithm of employees' wage rates that constrains all the *slope* coefficients to be the same for male and female workers. Write the **population regression equation for Model 1** as:

$$\begin{aligned}\ln(\text{wage}_i) = \beta_0 + \beta_1\text{ed}_i + \beta_2\exp_i + \beta_3\exp_i^2 + \beta_4\text{ind}2_i + \beta_5\text{ind}3_i + \beta_6\text{ind}4_i \\ + \beta_7\text{ind}5_i + \beta_8\text{ind}6_i + \beta_9\text{ind}7_i + \delta_0\text{female}_i + u_i\end{aligned} \quad \textbf{(1)}$$

Regression equation (1) allows only the intercept coefficient to differ between male and female workers; it restricts all the slope coefficients β_j (j = 1, …, 9) to be equal or identical for male and female employees. The male intercept coefficient is β_0, and the female intercept coefficient is $\beta_0 + \delta_0$; the slope coefficient of the female indicator variable female_i is therefore the female-male difference in intercept coefficients, or equivalently the female-male difference in conditional mean ln-wages for given values of years of completed schooling, potential work experience, and industry of employment.

- First, generate the regressand (or dependent variable) $\ln(\text{wage}_i)$ in Model 1. Enter the commands:

```
generate lnwage = ln(wage)
summarize wage lnwage
```

- Estimate by OLS regression equation (1) on the full sample of 526 observations for both female and male employees. Enter *on one line* the **regress** command:

  ```
  regress lnwage ed exp expsq ind2 ind3 ind4 ind5 ind6 ind7 female
  ```

  You will want to refer back to the OLS estimates of Model 1 produced by this **regress** command, as Model 1 represents the benchmark for all subsequent models in this tutorial.

❑ Model 2 – Different Industry Effects for Male and Female Employees

In this section, we consider a regression model that allows not only the intercept coefficient, but also the set of six industry slope coefficients, to differ between male and female employees. In other words, Model 2 allows for ***different* industry effects for *male* and *female* workers**, but restricts the marginal effects of the continuous explanatory variables $\boldsymbol{ed_i}$ and $\boldsymbol{exp_i}$ to be equal for male and female workers. Model 2 adds to the regressors of Model 1 a set of interactions of the female indicator $\boldsymbol{female_i}$ with each of the six included industry dummy variables; these **female-industry interaction variables** are defined as follows:

$$
\begin{aligned}
f_i ind2_i &= female_i ind2_i \\
f_i ind3_i &= female_i ind3_i \\
f_i ind4_i &= female_i ind4_i \\
f_i ind5_i &= female_i ind5_i \\
f_i ind6_i &= female_i ind6_i \\
f_i ind7_i &= female_i ind7_i
\end{aligned}
$$

The **population regression equation for Model 2** can be written as:

$$
\begin{aligned}
\ln(wage_i) = \beta_0 + \beta_1 ed_i + \beta_2 \exp_i + \beta_3 \exp_i^2 + \beta_4 ind2_i + \beta_5 ind3_i + \beta_6 ind4_i \\
+ \beta_7 ind5_i + \beta_8 ind6_i + \beta_9 ind7_i + \delta_0 female_i + \delta_4 f_i ind2_i + \delta_5 f_i ind3_i \qquad \textbf{(2)} \\
+ \delta_6 f_i ind4_i + \delta_7 f_i ind5_i + \delta_8 f_i ind6_i + \delta_9 f_i ind7_i + u_i
\end{aligned}
$$

- Before estimating Model 2 by OLS, you will need to create the female-industry interaction variables. Enter the following **generate** commands:

```
generate find2 = female*ind2
generate find3 = female*ind3
generate find4 = female*ind4
generate find5 = female*ind5
generate find6 = female*ind6
generate find7 = female*ind7
```

♦ ***Interpretation of Model 2: Industry base group is industry 1***

Before proceeding to estimation of Model 2, we should make sure that we understand how to interpret the industry coefficients in Model 2.

The **population regression equation for Model 2** can be written as:

$$\begin{aligned}\ln(\text{wage}_i) = {} & \beta_0 + \beta_1 \text{ed}_i + \beta_2 \exp_i + \beta_3 \exp_i^2 + \beta_4 \text{ind2}_i + \beta_5 \text{ind3}_i + \beta_6 \text{ind4}_i \\ & + \beta_7 \text{ind5}_i + \beta_8 \text{ind6}_i + \beta_9 \text{ind7}_i + \delta_0 \text{female}_i + \delta_4 f_i \text{ind2}_i + \delta_5 f_i \text{ind3}_i \qquad \textbf{(2)} \\ & + \delta_6 f_i \text{ind4}_i + \delta_7 f_i \text{ind5}_i + \delta_8 f_i \text{ind6}_i + \delta_9 f_i \text{ind7}_i + u_i\end{aligned}$$

- The **population regression function for Model 2.1** is obtained by taking the conditional expectation of regression equation (2.1) for any given values of the four explanatory variables ed_i, exp_i, industry_i and female_i:

$$\begin{aligned}& E(\ln(\text{wage}_i) \mid \text{ed}_i, \exp_i, \text{ind2}, \ldots, \text{ind7}, \text{female}_i) \\ & \quad = \beta_0 + \beta_1 \text{ed}_i + \beta_2 \exp_i + \beta_3 \exp_i^2 + \beta_4 \text{ind2}_i + \beta_5 \text{ind3}_i + \beta_6 \text{ind4}_i \\ & \qquad + \beta_7 \text{ind5}_i + \beta_8 \text{ind6}_i + \beta_9 \text{ind7}_i + \delta_0 \text{female}_i + \delta_4 f_i \text{ind2}_i + \delta_5 f_i \text{ind3}_i \qquad (2^*) \\ & \qquad + \delta_6 f_i \text{ind4}_i + \delta_7 f_i \text{ind5}_i + \delta_8 f_i \text{ind6}_i + \delta_9 f_i \text{ind7}_i\end{aligned}$$

- The ***male*** **population regression function** implied by Model 2 is obtained by setting the *female* indicator variable $\text{female}_i = 0$ in (2*):

$$E(\ln(\text{wage}_i) \mid \text{ed}_i, \exp_i, \text{ind2}, \ldots, \text{ind7}, \text{female}_i = 0)$$

$$= \beta_0 + \beta_1 ed_i + \beta_2 \exp_i + \beta_3 \exp_i^2 + \beta_4 ind2_i + \beta_5 ind3_i + \beta_6 ind4_i + \beta_7 ind5_i + \beta_8 ind6_i + \beta_9 ind7_i \qquad (2m)$$

The ***male*** **industry intercept coefficients** are:

Industry 1 intercept coefficient for males = β_0
Industry 2 intercept coefficient for males = $\beta_0 + \beta_4$
Industry 3 intercept coefficient for males = $\beta_0 + \beta_5$
Industry 4 intercept coefficient for males = $\beta_0 + \beta_6$
Industry 5 intercept coefficient for males = $\beta_0 + \beta_7$
Industry 6 intercept coefficient for males = $\beta_0 + \beta_8$
Industry 7 intercept coefficient for males = $\beta_0 + \beta_9$

The set of **industry effects for male workers in Model 2** – i.e., the inter-industry differences in conditional mean ln-wages for male employees with given values of *ed* and *exp* – are given by the ***male*** **slope coefficients of the industry dummy variables** $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$ that are included as regressors in Model 2. From the industry intercept coefficients for males given above, it follows that these male industry effects are:

β_4 = the industry 2 – industry 1 difference in mean ln-wages for males;
β_5 = the industry 3 – industry 1 difference in mean ln-wages for males;
β_6 = the industry 4 – industry 1 difference in mean ln-wages for males;
β_7 = the industry 5 – industry 1 difference in mean ln-wages for males;
β_8 = the industry 6 – industry 1 difference in mean ln-wages for males;
β_9 = the industry 7 – industry 1 difference in mean ln-wages for males.

- The ***female*** **population regression function** implied by Model 2 is obtained by setting the *female* indicator variable $\text{female}_i = 1$ in (2*):

$$\mathrm{E}(\ln(\text{wage}_i) \mid \text{ed}_i, \exp_i, \text{ind}2, \ldots, \text{ind}7, \text{female}_i = 1)$$

$$\begin{aligned} &= \beta_0 + \beta_1 \text{ed}_i + \beta_2 \exp_i + \beta_3 \exp_i^2 + \beta_4 \text{ind}2_i + \beta_5 \text{ind}3_i + \beta_6 \text{ind}4_i \\ &\qquad + \beta_7 \text{ind}5_i + \beta_8 \text{ind}6_i + \beta_9 \text{ind}7_i + \delta_0 + \delta_4 \text{ind}2_i + \delta_5 \text{ind}3_i \\ &\qquad + \delta_6 \text{ind}4_i + \delta_7 \text{ind}5_i + \delta_8 \text{ind}6_i + \delta_9 \text{ind}7_i \end{aligned}$$

$$\begin{aligned} &= (\beta_0 + \delta_0) + \beta_1 \text{ed}_i + \beta_2 \exp_i + \beta_3 \exp_i^2 + (\beta_4 + \delta_4)\text{ind}2_i + (\beta_5 + \delta_5)\text{ind}3_i \\ &\quad + (\beta_6 + \delta_6)\text{ind}4_i + (\beta_7 + \delta_7)\text{ind}5_i + (\beta_8 + \delta_8)\text{ind}6_i + (\beta_9 + \delta_9)\text{ind}7_i \end{aligned} \qquad (2f)$$

The ***female*** **industry intercept coefficients** are:

Industry 1 intercept coefficient for females = $\beta_0 + \delta_0$
Industry 2 intercept coefficient for females = $\beta_0 + \delta_0 + \beta_4 + \delta_4$
Industry 3 intercept coefficient for females = $\beta_0 + \delta_0 + \beta_5 + \delta_5$
Industry 4 intercept coefficient for females = $\beta_0 + \delta_0 + \beta_6 + \delta_6$
Industry 5 intercept coefficient for females = $\beta_0 + \delta_0 + \beta_7 + \delta_7$
Industry 6 intercept coefficient for females = $\beta_0 + \delta_0 + \beta_8 + \delta_8$
Industry 7 intercept coefficient for females = $\beta_0 + \delta_0 + \beta_9 + \delta_9$

The set of **industry effects for female workers in Model 2** – i.e., the inter-industry differences in conditional mean ln-wages for female employees with given values of *ed* and *exp* – are given by the ***female*** **slope coefficients of the industry dummy variables** $\text{ind}2_i$, $\text{ind}3_i$, $\text{ind}4_i$, $\text{ind}5_i$, $\text{ind}6_i$, and $\text{ind}7_i$ that are included as regressors in Model 2. From the industry intercept coefficients for females given above, it follows that these female industry effects are:

$\beta_4 + \delta_4$ = the industry 2 – industry 1 difference in mean ln-wages for females;
$\beta_5 + \delta_5$ = the industry 3 – industry 1 difference in mean ln-wages for females;
$\beta_6 + \delta_6$ = the industry 4 – industry 1 difference in mean ln-wages for females;
$\beta_7 + \delta_7$ = the industry 5 – industry 1 difference in mean ln-wages for females;
$\beta_8 + \delta_8$ = the industry 6 – industry 1 difference in mean ln-wages for females;
$\beta_9 + \delta_9$ = the industry 7 – industry 1 difference in mean ln-wages for females.

- The ***female-male* differences in industry effects** implied by Model 2 are obtained by subtracting the male industry coefficients (β_4, β_5, β_6, β_7, β_8 and β_9) from the corresponding female industry coefficients ($\beta_4 + \delta_4$, $\beta_5 + \delta_5$, $\beta_6 + \delta_6$, $\beta_7 + \delta_7$, $\beta_8 + \delta_8$ and $\beta_9 + \delta_9$) – i.e., by the slope coefficients δ_4, δ_5, δ_6, δ_7, δ_8, and δ_9 of the female-industry interaction terms $f_i ind2_i$, $f_i ind3_i$, $f_i ind4_i$, $f_i ind5_i$, $f_i ind6_i$, and $f_i ind7_i$ in Model 2.

For industry 2:

$\beta_4 + \delta_4$ = the industry 2 – industry 1 difference in mean ln-wages for females

β_4 = the industry 2 – industry 1 difference in mean ln-wages for males

Therefore

δ_4 = the industry 2 – industry 1 difference in mean ln-wages for females
minus
the industry 2 – industry 1 difference in mean ln-wages for males

Similarly, ***for industries 3, 4, 5, 6 and 7:***

δ_5 = the industry 3 – industry 1 difference in mean ln-wages for females
minus
the industry 3 – industry 1 difference in mean ln-wages for males

δ_6 = the industry 4 – industry 1 difference in mean ln-wages for females
minus
the industry 4 – industry 1 difference in mean ln-wages for males

δ_7 = the industry 5 – industry 1 difference in mean ln-wages for females
minus
the industry 5 – industry 1 difference in mean ln-wages for males

δ_8 = the industry 6 – industry 1 difference in mean ln-wages for females
minus
the industry 6 – industry 1 difference in mean ln-wages for males

δ_9 = the industry 7 – industry 1 difference in mean ln-wages for females
minus
the industry 7 – industry 1 difference in mean ln-wages for males

Note that the conditional effects of industry on mean ln-wages are identical or equal for male and female workers if the regression coefficients δ_4, δ_5, δ_6, δ_7, δ_8 and δ_9 are jointly equal to zero: i.e., if $\boldsymbol{\delta_4 = \delta_5 = \delta_6 = \delta_7 = \delta_8 = \delta_9 = 0}$, or if $\boldsymbol{\delta_j = 0}$ **for all j = 4, 5, …, 9**.

♦ ***OLS Estimation of Model 2***

- Estimate Model 2 by OLS on the full sample of 526 female and male employees. Enter *on one line* the **regress** command:

```
regress lnwage ed exp expsq ind2 ind3 ind4 ind5 ind6 ind7 female
find2 find3 find4 find5 find6 find7
```

- Use **lincom** commands to compute and test the ***female intercept*** **coefficient estimate** and the ***female*** **coefficient estimates for the *industry* dummy variables** $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$ in Model 2. Enter the following series of **lincom** commands:

```
lincom _b[_cons] + _b[female]
```
$= \hat{\beta}_0 + \hat{\delta}_0$

```
lincom _b[ind2] + _b[find2]
```
$= \hat{\beta}_4 + \hat{\delta}_4$

```
lincom _b[ind3] + _b[find3]
```
$= \hat{\beta}_5 + \hat{\delta}_5$

```
lincom _b[ind4] + _b[find4]
```
$= \hat{\beta}_6 + \hat{\delta}_6$

```
lincom _b[ind5] + _b[find5]
```
$= \hat{\beta}_7 + \hat{\delta}_7$

```
lincom _b[ind6] + _b[find6]
```
$= \hat{\beta}_8 + \hat{\delta}_8$

```
lincom _b[ind7] + _b[find7]
```
$= \hat{\beta}_9 + \hat{\delta}_9$

♦ ***Test for Industry Effects in Model 2***

We now wish to test for industry effects in Model 2. There are three such tests that should be performed.

Industry Effects Test 1: Test for industry effects for *male* employees

Test the null hypothesis of no industry effects for male employees in Model 2. Since industry effects for males in Model 2 are represented by the male slope coefficients β_4, β_5, β_6, β_7, β_8 and β_9 of the industry dummy variables $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$, the null and alternative hypotheses are specified as follows:

H_0: $\beta_j = 0$ for all j = 4, 5,…, 9; or

$\beta_4 = 0$ *and* $\beta_5 = 0$ *and* $\beta_6 = 0$ *and* $\beta_7 = 0$ *and* $\beta_8 = 0$ *and* $\beta_9 = 0$

H_1: $\beta_j \neq 0$ j = 4, 5,…, 9; or

$\beta_4 \neq 0$ *and/or* $\beta_5 \neq 0$ *and/or* $\beta_6 \neq 0$ *and/or* $\beta_7 \neq 0$ *and/or* $\beta_8 \neq 0$ *and/or* $\beta_9 \neq 0$

- Use the following **test** command to compute an F-test of the null hypothesis of **no industry effects for *male* employees**. Enter the commands:

```
test ind2 ind3 ind4 ind5 ind6 ind7
return list
```

Based on the computed outcome of this test, would you retain or reject the null hypothesis of no industry effects for male workers?

Industry Effects Test 2: Test for industry effects for *female* employees

Test the null hypothesis of no industry effects for female employees in Model 2. Since industry effects for females in Model 2 are represented by the female slope coefficients $\beta_4 + \delta_4$, $\beta_5 + \delta_5$, $\beta_6 + \delta_6$, $\beta_7 + \delta_7$, $\beta_8 + \delta_8$ and $\beta_9 + \delta_9$ of the industry dummy variables $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$ in the female ln-wage regression function, the null and alternative hypotheses are specified as follows:

H_0: $\beta_j + \delta_j = 0$ for all j = 4, 5,…, 9; or

$\beta_4 + \delta_4 = 0$ *and* $\beta_5 + \delta_5 = 0$ *and* $\beta_6 + \delta_6 = 0$ *and* $\beta_7 + \delta_7 = 0$ *and* $\beta_8 + \delta_8 = 0$ *and* $\beta_9 + \delta_9 = 0$

H_1: $\beta_j + \delta_j \neq 0$ j = 4, 5,…, 9; or

$\beta_4 + \delta_4 \neq 0$ *and/or* $\beta_5 + \delta_5 \neq 0$ *and/or* $\beta_6 + \delta_6 \neq 0$ *and/or* $\beta_7 + \delta_7 \neq 0$ *and/or* $\beta_8 + \delta_8 \neq 0$ *and/or* $\beta_9 + \delta_9 \neq 0$

- Use the following series of linked **test** commands to compute an F-test of the null hypothesis of **no industry effects for *female* employees**. Enter the commands:

```
test ind2 + find2 = 0, notest
test ind3 + find3 = 0, notest accumulate
test ind4 + find4 = 0, notest accumulate
test ind5 + find5 = 0, notest accumulate
test ind6 + find6 = 0, notest accumulate
test ind7 + find7 = 0, accumulate
return list
```

Note that only the results produced by the last of this sequence of ***six* test** commands correspond to the null hypothesis H_0 of no industry effects for female workers. That is why the **notest** option has been specified for the first five test commands in this sequence; the **notest** option simply suppresses the printing of the results of the test command to which it is attached.

Based on the computed outcome of this test, would you retain or reject the null hypothesis of no industry effects for female workers?

Industry Effects Test 3: Test for *female-male differences* in industry effects

Test the null hypothesis of no female-male differences in industry effects in Model 2. Since the female-male differences in industry effects in Model 2 are represented by the slope coefficients δ_4, δ_5, δ_6, δ_7, δ_8 and δ_9 of the female interactions with the industry dummy variables $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$ in Model 2, the null and alternative hypotheses are specified as follows:

H_0: $\delta_j = 0$ for all j = 4, 5,…, 9; or

$\delta_4 = 0$ *and* $\delta_5 = 0$ *and* $\delta_6 = 0$ *and* $\delta_7 = 0$ *and* $\delta_8 = 0$ *and* $\delta_9 = 0$

H_1: $\delta_j \neq 0$ j = 4, 5,…, 9; or

$\delta_4 \neq 0$ *and/or* $\delta_5 \neq 0$ *and/or* $\delta_6 \neq 0$ *and/or* $\delta_7 \neq 0$ *and/or* $\delta_8 \neq 0$ *and/or* $\delta_9 \neq 0$

- Use the following **test** command to compute an F-test of the null hypothesis of **no *female-male differences* in industry effects in Model 2**, i.e., that industry effects are equal, or identical, for male and female employees. Enter the commands:

  ```
  test find2 find3 find4 find5 find6 find7
  return list
  ```

  Based on the computed outcome of this test, would you retain or reject the null hypothesis of no industry effects for male workers?

  Note: If you have correctly computed the foregoing test of no female-male differences in industry effects in Model 2, you will have found that the null hypothesis H_0 is retained at all conventional significance levels. Despite this test outcome, we will, for pedagogical reasons, proceed with learning how *Stata* **graph bar** commands can be used to graphically illustrate the male and female industry effects in Model 2. But bear in mind that this test implies that industry effects in Model 2 do not differ significantly between male and female workers.

❑ Industry Effects in Model 2 for Male and Female Employees – margins

In this section, we demonstrate how the *Stata* **margins** command, to which you were introduced in **Stata 12 Tutorial 4**, can be used to compute for Model 2 industry effects for male and female employees and female-male differences in industry effects.

- First, re-estimate Model 2 by OLS using **factor-variable notation** to distinguish between *continuous* and *categorical* explanatory variables in the **regress** command. Recall from Stata 12 Tutorial 4 that the **c.** prefix must precede the name of each continuous explanatory variable, while the **i.** prefix must precede the name of each continuous explanatory variable. Enter *on one line* the command:

```
regress lnwage c.ed c.exp c.exp#c.exp i.industry i.female
i.female#i.industry
```

- Use the **margins** command to compute the conditional mean values of ln(wage$_i$) at the sample median values of **ed** and **exp** for male and female employees in the seven industry categories. Enter the following two **margins** commands, and observe that they yield identical results:

```
margins i.industry, at((median) ed exp female = (0 1))
margins i.female, at((median) ed exp industry = (1(1)7))
```

- Now use the **margins** command to compute female-male differences in the conditional mean values of ln(wage$_i$) at the sample median values of **ed** and **exp** for each of the seven industry categories. Enter the following two **margins** commands, and observe that they yield identical results:

```
margins r.female, at((median) ed exp industry = (1(1)7))
margins r.female, over(industry) at(ed = 12 exp = 13.5)
```

Note that the second **margins** command that uses the **over(industry)** option *must* specify the numerical values of **ed** and **exp** in order to produce the intended results.

- In the two previous **margins** commands, it is not necessary to specify the sample values of **ed** and **exp** in computing female-male differences in the conditional mean values of ln(wage$_i$) for each of the seven industry categories. This is so because Model 2 restricts the regression coefficients of the **ed** and **exp** regressors to be the same for male and female employees. To verify this, enter the following two **margins** commands and observe that they produce results identical to those of the two previous **margins** commands that specify the median values of **ed** and **exp**:

```
margins r.female, at(industry = (1(1)7))
margins r.female, over(industry)
```

- Next, use the **margins** command to compute separately for male and female employees inter-industry differences in the conditional mean values of ln(wage$_i$) relative to Industry 1 (the base group industry). Note that it is not necessary to specify the values of **ed** and **exp** to be used, because the pairwise industry differences in conditional mean ln(wage$_i$) values do not depend on, or vary with, the continuous explanatory variables **ed** and **exp**; they depend only on the categorical variable **female**. Enter the following four **margins** commands, and observe that they yield identical results:

```
margins r.industry, over(female)
margins r.industry, over(female) at(ed = 12 exp = 13.5)
margins r.industry, at(female = (0 1))
margins r.industry, at((median) ed exp female = (0 1))
```

- Finally, suppose we wish to use Industry 4 rather than Industry 1 as the base group for purposes of computing the inter-industry differences in the conditional mean values of ln(wage$_i$) for male and female employees. That is, we wish to compute differences in the conditional mean values of ln(wage$_i$) for industries 1, 2, 3, 5, 6, and 7 relative to Industry 4. This can be easily done by simply changing the `r.` prefix on industry in the preceding margins commands to `rb4.`. The `rb4.` prefix instructs *Stata* to use the fourth smallest value of the categorical variable `industry` as the base group, rather than the smallest value, which is the default implied by the `r.` prefix. Enter the following four **margins** commands, and observe that they yield identical results:

```
margins rb4.industry, over(female)
margins rb4.industry, over(female) at(ed = 12 exp = 13.5)
margins rb4.industry, at(female = (0 1))
margins rb4.industry, at((median) ed exp female = (0 1))
```

❑ Model 2 -- Graphing Industry Effects for Male and Female Employees

This section demonstrates how to use the *Stata* **graph bar** command to graphically illustrate the male and female industry effects we have estimated for Model 2.

Before we can use the **graph bar** command, we must save the male and female industry coefficient estimates for Model 2 in a form that can be used in the **graph bar** command.

***Step 1*:** Generate a new variable that contains the values of the male industry coefficient estimates for Model 2; i.e., create a variable that contains the OLS estimates for Model 2 of the male coefficients for the industry dummy variables $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$ in Model 2.

- Use the following series of **generate** and **replace** commands to create a new variable named **`malindcoefs_model2`** that contains the OLS coefficient estimates for Model 2 of the male slope coefficients β_4, β_5, β_6, β_7, β_8 and β_9 of the industry dummy variables $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$. Enter the following series of **generate** and **replace** commands:

```
generate malindcoefs_model2 = _b[ind2] if ind2 == 1 & female ==
0
replace malindcoefs_model2 = _b[ind3] if ind3 == 1 & female == 0
replace malindcoefs_model2 = _b[ind4] if ind4 == 1 & female == 0
replace malindcoefs_model2 = _b[ind5] if ind5 == 1 & female == 0
replace malindcoefs_model2 = _b[ind6] if ind6 == 1 & female == 0
replace malindcoefs_model2 = _b[ind7] if ind7 == 1 & female == 0
```

- Next, we should verify that the variable **`malindcoefs_model2`** we have just created does indeed contain the OLS coefficient estimates for Model 2 of the male slope coefficients β_4, β_5, β_6, β_7, β_8 and β_9 of the industry dummy variables $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$. Enter the following series of **summarize** commands:

```
summarize malindcoefs_model2 if ind2 == 1 & female == 0
summarize malindcoefs_model2 if ind3 == 1 & female == 0
summarize malindcoefs_model2 if ind4 == 1 & female == 0
summarize malindcoefs_model2 if ind5 == 1 & female == 0
summarize malindcoefs_model2 if ind6 == 1 & female == 0
summarize malindcoefs_model2 if ind7 == 1 & female == 0
```

- Finally, enter the following **tab1** commands:

```
tab1 industry malindcoefs_model2, missing
tab1 industry malindcoefs_model2

table industry, contents(mean malindcoefs_model2)
```

Step 2: Generate a new variable that contains the values of the female industry coefficient estimates for Model 2; i.e., create a variable that contains the OLS estimates for Model 2 of the female coefficients for the industry dummy variables $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$ in Model 2.

- Use the following series of **generate** and **replace** commands to create a new variable named **femindcoefs_model2** that contains the OLS coefficient estimates for Model 2 of the female slope coefficients $\beta_4 + \delta_4$, $\beta_5 + \delta_5$, $\beta_6 + \delta_6$, $\beta_7 + \delta_7$, $\beta_8 + \delta_8$ and $\beta_9 + \delta_9$ of the industry dummy variables $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$. Enter the following series of **generate** and **replace** commands:

```
generate femindcoefs_model2 = _b[ind2] + _b[find2] if ind2 == 1
& female == 1

replace femindcoefs_model2 = _b[ind3] + _b[find3] if ind3 == 1 &
female == 1

replace femindcoefs_model2 = _b[ind4] + _b[find4] if ind4 == 1 &
female == 1

replace femindcoefs_model2 = _b[ind5] + _b[find5] if ind5 == 1 &
female == 1

replace femindcoefs_model2 = _b[ind6] + _b[find6] if ind6 == 1 &
female == 1

replace femindcoefs_model2 = _b[ind7] + _b[find7] if ind7 == 1 &
female == 1
```

- Next, verify that the variable **femindcoefs_model2** we have just created does indeed contain the OLS coefficient estimates for Model 2 of the female slope coefficients $\beta_4 + \delta_4$, $\beta_5 + \delta_5$, $\beta_6 + \delta_6$, $\beta_7 + \delta_7$, $\beta_8 + \delta_8$ and $\beta_9 + \delta_9$ of the industry dummy variables $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$. Enter the following series of **summarize** commands:

```
summarize femindcoefs_model2 if ind2 == 1 & female == 1
summarize femindcoefs_model2 if ind3 == 1 & female == 1
summarize femindcoefs_model2 if ind4 == 1 & female == 1
summarize femindcoefs_model2 if ind5 == 1 & female == 1
summarize femindcoefs_model2 if ind6 == 1 & female == 1
summarize femindcoefs_model2 if ind7 == 1 & female == 1
```

- Finally, enter the following **tab1** commands:

```
tab1 industry femindcoefs_model2, missing
tab1 industry femindcoefs_model2

table industry, contents(mean femindcoefs_model2)
table industry, contents(mean malindcoefs_model2 mean
femindcoefs_model2)
```

Step 3: We can now use the newly created variables **`malindcoefs_model2`** and **`femindcoefs_model2`** to draw a bar graph of the estimated male and female industry effects in Model 2.

- Use the following basic **graph bar** command to create a first bar graph of the estimated male and female industry effects in Model 2. Enter *on one line* the **graph bar** command:

```
graph bar (mean) malindcoefs_model2 femindcoefs_model2 if
industry > 1, over(industry)
```

- We can produce a more complete and informative bar graph by adding to the above **graph bar** command some additional options that label the bars as Male or Female, that provide a title for the vertical y-axis, and that provide a title for the entire graph. Enter *on one line* the following expanded **graph bar** command:

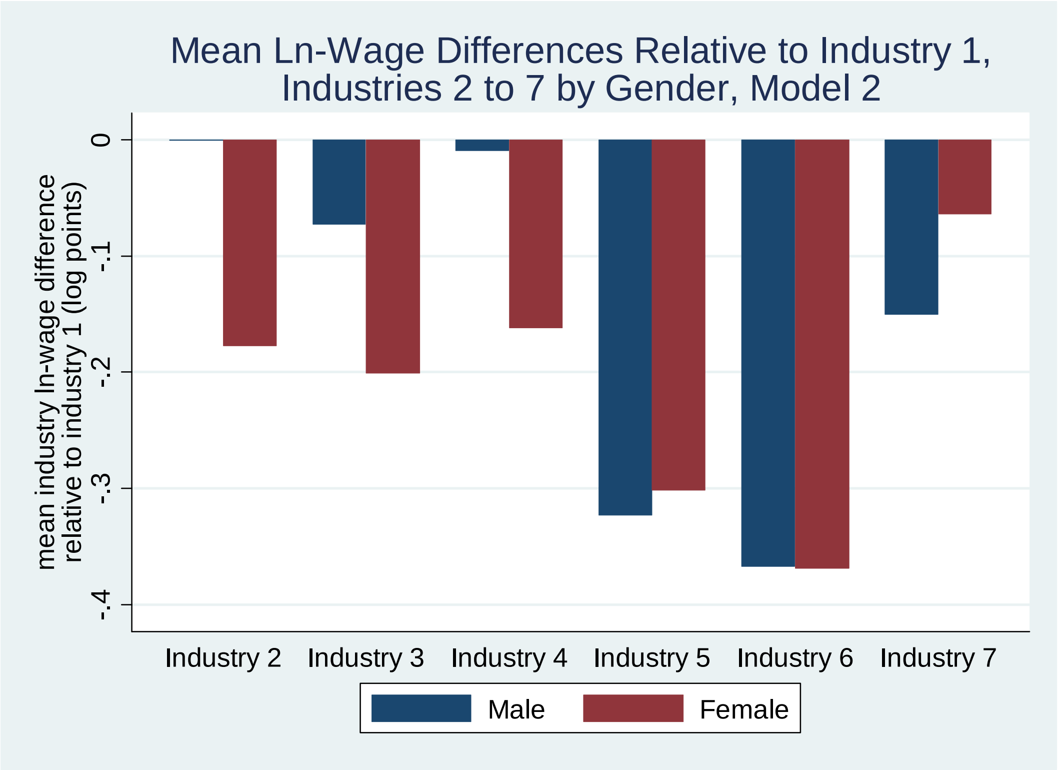
```
graph bar (mean) malindcoefs_model2 femindcoefs_model2 if
industry > 1, over(industry) legend ( label(1 "Male") label(2
"Female") ) ytitle("mean industry ln-wage difference" "relative
to industry 1 (log points)") title("Mean Ln-Wage Differences
Relative to Industry 1," "Industries 2 to 7 by Gender -- Model
2")
```

Carefully observe how the **graph bar** options **legend**, **ytitle** and **title** have been used to provide a more finished and complete bar graph.

- To export and save this bar graph in Windows Enhanced Metafile format to a file named **`bargraph1_tutorial6.emf`** in the current *Stata* working directory, enter the **graph export** command:

```
graph export bargraph1_tutorial6.emf
```

- Here is what the bar graph you just exported in Windows Enhanced Metafile format to the file **`bargraph1_tutorial6.emf`** looks like when it is inserted into this MS Word document:



- You can produce a *horizontal* version of the bar graph you have just created by using the **graph hbar** command rather than the **graph bar** command. Enter the following expanded **graph hbar** command:

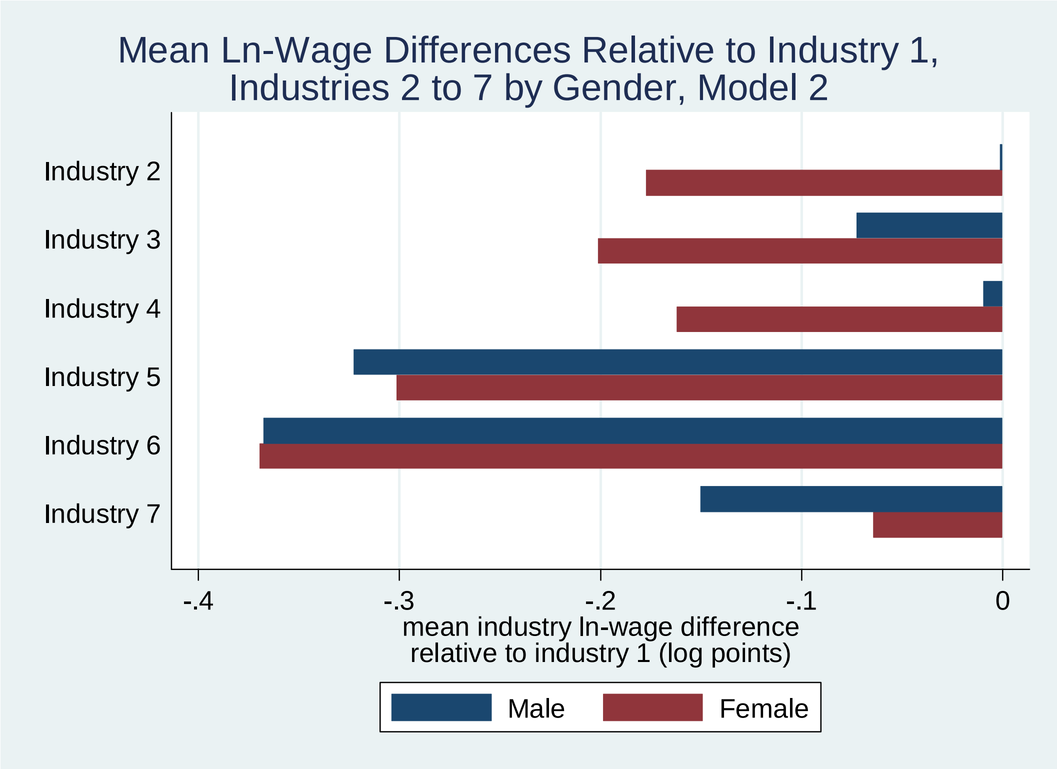
```
graph hbar (mean) malindcoefs_model2 femindcoefs_model2 if
industry > 1, over(industry) legend ( label(1 "Male") label(2
"Female") ) ytitle("mean industry ln-wage difference" "relative
to industry 1 (log points)") title("Mean Ln-Wage Differences
Relative to Industry 1," "Industries 2 to 7 by Gender -- Model
2")
```

Note that the above command is identical to the **graph bar** command on the previous page except for the use of the **graph hbar** command. Again observe how the **graph hbar** options **legend**, **ytitle** and **title** have been used to provide a more finished and complete bar graph.

- To export and save this bar graph in Windows Enhanced Metafile format to a file named **`bargraph2_tutorial6.emf`** in the current *Stata* working directory, enter the **graph export** command:

```
graph export bargraph2_tutorial6.emf
```

♦ Here is what the horizontal bar graph you just exported in Windows Enhanced Metafile format to the file **`bargraph2_tutorial6.emf`** looks like when it is inserted into this MS Word document:



A Third Bar Graph Depicting the Industry Effects for Males and Females

Finally, we can use an alternative **graph hbar** command to create a second horizontal bar graph that depicts the industry effects for male and female workers estimated in Model 2. But first we must save the male and female industry coefficient estimates for Model 2 in a form that can be used in the **graph hbar** command.

Step 1: Generate a new variable that contains the values of the male industry coefficient estimates for Model 2; i.e., create a variable that contains the OLS estimates for Model 2 of the male coefficients for the industry dummy variables $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$ in Model 2.

- Use the following series of **generate** and **replace** commands to create a new variable named **`allindcoefs_model2`** that contains the OLS coefficient estimates for Model 2 of ***both*** the male slope coefficients β_4, β_5, β_6, β_7, β_8 and β_9 of the industry dummy variables $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$ and the female slope coefficients $\beta_4 + \delta_4$, $\beta_5 + \delta_5$, $\beta_6 + \delta_6$, $\beta_7 + \delta_7$, $\beta_8 + \delta_8$ and $\beta_9 + \delta_9$ of the industry dummy variables $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$. Enter the following series of **generate** and **replace** commands:

```
generate allindcoefs_model2 = _b[ind2] if ind2 == 1 & female ==
0
replace allindcoefs_model2 = _b[ind3] if ind3 == 1 & female == 0
replace allindcoefs_model2 = _b[ind4] if ind4 == 1 & female == 0
replace allindcoefs_model2 = _b[ind5] if ind5 == 1 & female == 0
replace allindcoefs_model2 = _b[ind6] if ind6 == 1 & female == 0
replace allindcoefs_model2 = _b[ind7] if ind7 == 1 & female == 0
replace allindcoefs_model2 = _b[ind2] + _b[find2] if ind2 == 1 &
female == 1
replace allindcoefs_model2 = _b[ind3] + _b[find3] if ind3 == 1 &
female == 1
replace allindcoefs_model2 = _b[ind4] + _b[find4] if ind4 == 1 &
female == 1
replace allindcoefs_model2 = _b[ind5] + _b[find5] if ind5 == 1 &
female == 1
```

```
replace allindcoefs_model2 = _b[ind6] + _b[find6] if ind6 == 1 &
female == 1
replace allindcoefs_model2 = _b[ind7] + _b[find7] if ind7 == 1 &
female == 1
```

- Next, verify that the variable `allindcoefs_model2` we have just created does indeed contain the OLS coefficient estimates for Model 2 of both the male slope coefficients β_4, β_5, β_6, β_7, β_8 and β_9 and the female slope coefficients $\beta_4 + \delta_4$, $\beta_5 + \delta_5$, $\beta_6 + \delta_6$, $\beta_7 + \delta_7$, $\beta_8 + \delta_8$ and $\beta_9 + \delta_9$ of the industry dummy variables $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$. Enter the following series of **summarize** commands:

```
summarize allindcoefs_model2 if ind2 == 1 & female == 0
summarize allindcoefs_model2 if ind3 == 1 & female == 0
summarize allindcoefs_model2 if ind4 == 1 & female == 0
summarize allindcoefs_model2 if ind5 == 1 & female == 0
summarize allindcoefs_model2 if ind6 == 1 & female == 0
summarize allindcoefs_model2 if ind7 == 1 & female == 0
summarize allindcoefs_model2 if ind2 == 1 & female == 1
summarize allindcoefs_model2 if ind3 == 1 & female == 1
summarize allindcoefs_model2 if ind4 == 1 & female == 1
summarize allindcoefs_model2 if ind5 == 1 & female == 1
summarize allindcoefs_model2 if ind6 == 1 & female == 1
summarize allindcoefs_model2 if ind7 == 1 & female == 1
```

- Finally, enter the following **tab1** and **tab2** commands:

```
tab1 industry allindcoefs_model2 if female == 0, missing
tab1 industry allindcoefs_model2 if female == 1, missing

tab2 allindcoefs_model2 female, missing
tab2 industry female, missing
```

Careful inspection of the results of these **tab1** and **tab2** commands will enable you to verify that the newly created variable `allindcoefs_model2` does indeed contain the OLS coefficient estimates for Model 2 of both the male slope coefficients β_4, β_5, β_6, β_7, β_8 and β_9 and the female slope coefficients $\beta_4 + \delta_4$, $\beta_5 + \delta_5$, $\beta_6 + \delta_6$, $\beta_7 + \delta_7$, $\beta_8 + \delta_8$ and $\beta_9 + \delta_9$ of the industry dummy variables $ind2_i$, $ind3_i$, $ind4_i$, $ind5_i$, $ind6_i$, and $ind7_i$.

Step 2: We can now use the newly created variable **`allindcoefs_model2`** to draw a horizontal bar graph of the estimated male and female industry effects in Model 2.

- Use the following basic **graph hbar** command to create a first horizontal bar graph of the estimated male and female industry effects in Model 2. Enter the **graph bar** command:

  ```
  graph hbar (mean) allindcoefs_model2 if industry > 1,
  over(female) over(industry)
  ```

  Note that the above **graph hbar** command includes both the **over(female)** and the **over(industry)** options.

- We can produce a more complete and informative bar graph by adding to the above **graph hbar** command some additional options that provide a title for the vertical y-axis, and that provide a title for the entire graph. Enter the following expanded **graph hbar** command:

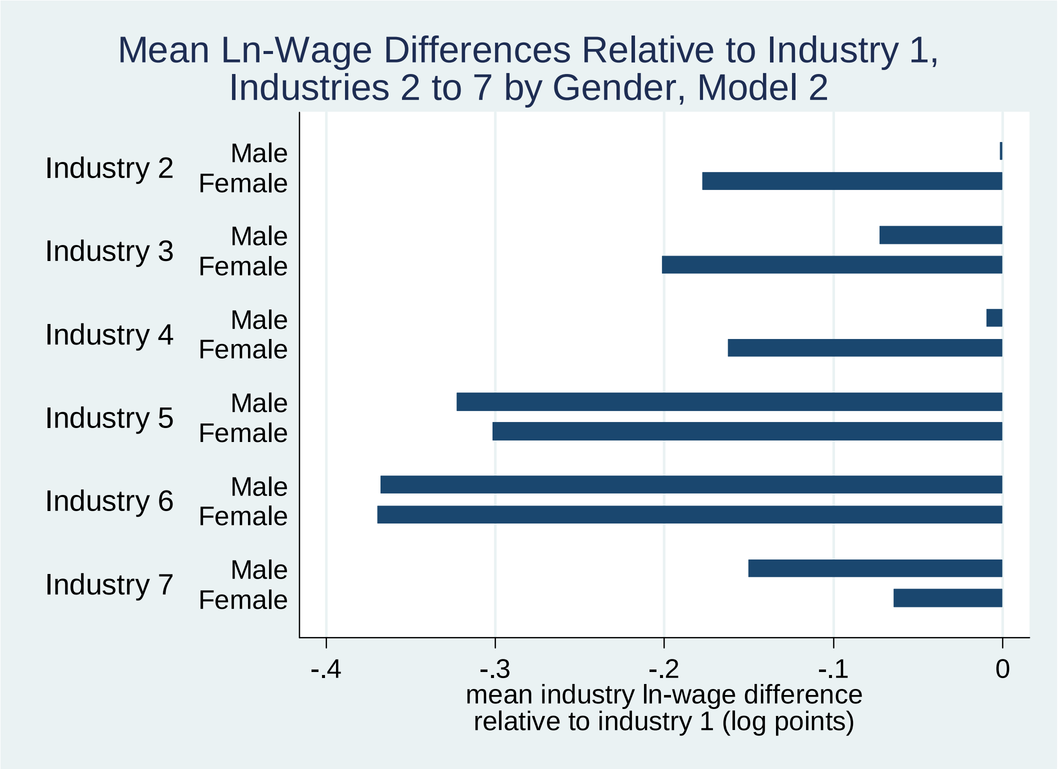
  ```
  graph hbar (mean) allindcoefs_model2 if industry > 1,
  over(female) over(industry) ytitle("mean industry ln-wage
  difference" "relative to industry 1 (log points)") title("Mean
  Ln-Wage Differences Relative to Industry 1," "Industries 2 to 7
  by Gender, Model 2", span)
  ```

  Note that the option **`span`** is specified in the **`title(   )`** portion of the above **graph hbar** command; it centers the title over the entire graph rather than over the plot region.

- To export and save this bar graph in Windows Enhanced Metafile format to a file named **`bargraph3_tutorial6.emf`** in the current *Stata* working directory, enter the **graph export** command:

  ```
  graph export bargraph3_tutorial6.emf
  ```

- Here is what the horizontal bar graph you just exported in Windows Enhanced Metafile format to the file **`bargraph3_tutorial6.emf`** looks like when it is inserted into this MS Word document:



❑ Preparing to End Your *Stata* Session

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

- First, you will probably want to **save the current dataset**. Enter the following **save** command with the **replace** option to save the current dataset as *Stata*-format dataset **wage1_econ452.dta**:

```
save wage1_econ452, replace
```

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

```
log close
```

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

```
cmdlog 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 portable electronic storage device such as a flash memory stick. These files will be found in the ***Stata* working directory**, which is usually **C:\data** on the computers in Dunning 350. There are three files you will want to be sure you have: the complete *Stata* log file **452tutorial6.log**; the *Stata* command log file **452tutorial6.txt**; and the changed *Stata*-format dataset **wage1_econ452.dta**. Use the Windows **copy** command to copy any files you want to keep to your own portable electronic storage device (e.g., a flash memory stick).

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.