

## Problem Set 2 (SOLUTIONS)

This problem set will take you through some Stata commands to estimate simple regression equations with dummy variables. You will learn how to interpret the estimated coefficients and test some linear hypotheses. Interpretation of these coefficients will be useful when we do treatment evaluation models later in term 1.

The hypothesis tests discussed in this problem set include standard T-tests and F-tests, which is assumed undergraduate knowledge for this module.

You will need to download the dataset `problemset2.dta`, which is available on Moodle.

### Conditional Expectation Function

Consider the Conditional Expectation Function (CEF),  $E[Y_i|X_i]$ . If  $X$  takes on discrete values:  $X_i \in \{x_1, x_2, \dots, x_m\}$ , then

$$E[Y_i|X_i] = E[Y_i|X_i = x_1] \cdot \mathbf{1}\{X_i = x_1\} + \dots + E[Y_i|X_i = x_m] \cdot \mathbf{1}\{X_i = x_m\}$$

where  $\mathbf{1}\{X_i = x_m\}$  is a dummy variable,  $= 1$  when  $X_i = x_m$ . Since the values of  $X_i$  are mutually exclusive there is no overlap of these dummy variables.

Note, we do not need to assume that  $X$  is a single random variable. It can be a vector of random variables that takes on discrete values.

We can re-arrange this expression using anyone of the values of  $X$ . The natural option is to choose the first, but this is arbitrary.

$$\begin{aligned} E[Y_i|X_i] &= E[Y_i|X_i = x_1] + (E[Y_i|X_i = x_2] - E[Y_i|X_i = x_1]) \cdot \mathbf{1}\{X_i = x_2\} + \dots \\ &\quad + (E[Y_i|X_i = x_m] - E[Y_i|X_i = x_1]) \cdot \mathbf{1}\{X_i = x_m\} \end{aligned}$$

Since,  $E[Y_i|X_i = x_m]$  is a constant ( $X_i$  is set to a specific value), we can express the CEF as a function that is linear in parameters.

$$E[Y_i|X_i] = \beta_1 + \beta_2 D_{i2} + \dots + \beta_m D_{im}$$

where  $D_{im} = \mathbf{1}\{X_i = x_m\}$ .

## Preamble

<IPython.core.display.HTML object>

Create a do-file for this problem set and include a preamble that sets the directory and opens the data. For example,

```
clear
//or, to remove all stored values (including macros, matrices, scalars, etc.)
*clear all

* Replace $rootdir with the relevant path to on your local harddrive.
cd "$rootdir/problem-sets/ps-2"

cap log close
log using problem-set-2-log.txt, replace

use problem-set-2-data.dta
```

## Questions

1. Consider the  $E[\ln(Wage_i)|Gender_i]$ , where  $Gender_i \in \{1\text{"Male"}, 2\text{"Female"}\}$ . Show that this CEF implies a linear model,

$$\ln(Wage_i) = \beta_1 + \beta_2 D_{i2} + \varepsilon_i$$

What do the parameters  $\beta_1$  and  $\beta_2$  imply?

$$\begin{aligned} E[\ln(Wage_i)|G_i] &= E[\ln(Wage_i)|G_i = 1] + (E[\ln(Wage_i)|G_i = 2] - E[\ln(Wage_i)|G_i = 1]) \cdot \mathbf{1}\{G_i = 2\} \\ &= \beta_1 + \beta_2 D_{i2} \end{aligned}$$

$$\Rightarrow E[\varepsilon_i|D_{i2}] = 0.$$

2. Regress `lwage` (log wage) on just a set of binary indicators that will enable you to test the hypothesis that males and females are on average, paid the same wage, ceteris paribus. Test this hypothesis.

```
reg lwage female
```

```
* or
```

```
gen male=1-female
```

```
reg lwage male
```

| Source   | SS         | df    | MS         | Number of obs | = | 4,165  |
|----------|------------|-------|------------|---------------|---|--------|
|          |            |       |            | F(1, 4163)    | = | 491.72 |
| Model    | 93.6914807 | 1     | 93.6914807 | Prob > F      | = | 0.0000 |
| Residual | 793.213421 | 4,163 | .190538895 | R-squared     | = | 0.1056 |
|          |            |       |            | Adj R-squared | = | 0.1054 |
| Total    | 886.904902 | 4,164 | .212993492 | Root MSE      | = | .43651 |

| lwage  | Coefficient | Std. err. | t      | P> t  | [95% conf. interval] |           |
|--------|-------------|-----------|--------|-------|----------------------|-----------|
| female | -.4744661   | .0213967  | -22.17 | 0.000 | -.516415             | -.4325171 |
| _cons  | 6.729774    | .00718    | 937.29 | 0.000 | 6.715697             | 6.74385   |

| Source   | SS         | df    | MS         | Number of obs | = | 4,165  |
|----------|------------|-------|------------|---------------|---|--------|
|          |            |       |            | F(1, 4163)    | = | 491.72 |
| Model    | 93.6914807 | 1     | 93.6914807 | Prob > F      | = | 0.0000 |
| Residual | 793.213421 | 4,163 | .190538895 | R-squared     | = | 0.1056 |
|          |            |       |            | Adj R-squared | = | 0.1054 |
| Total    | 886.904902 | 4,164 | .212993492 | Root MSE      | = | .43651 |

| lwage | Coefficient | Std. err. | t      | P> t  | [95% conf. interval] |          |
|-------|-------------|-----------|--------|-------|----------------------|----------|
| male  | .4744661    | .0213967  | 22.17  | 0.000 | .4325171             | .516415  |
| _cons | 6.255308    | .020156   | 310.34 | 0.000 | 6.215791             | 6.294824 |

Alternatively, you could use Stata's factor notation:

```
reg lwage i.female
```

```
//note: defaults to smallest value as base category. This can be changed as follows.
```

```
reg lwage ib1.female
```

| Source   | SS         | df    | MS         | Number of obs | = | 4,165  |
|----------|------------|-------|------------|---------------|---|--------|
| Model    | 93.6914807 | 1     | 93.6914807 | F(1, 4163)    | = | 491.72 |
| Residual | 793.213421 | 4,163 | .190538895 | Prob > F      | = | 0.0000 |
|          |            |       |            | R-squared     | = | 0.1056 |
|          |            |       |            | Adj R-squared | = | 0.1054 |
| Total    | 886.904902 | 4,164 | .212993492 | Root MSE      | = | .43651 |

| lwage    | Coefficient | Std. err. | t      | P> t  | [95% conf. interval] |           |
|----------|-------------|-----------|--------|-------|----------------------|-----------|
| 1.female | -.4744661   | .0213967  | -22.17 | 0.000 | -.516415             | -.4325171 |
| _cons    | 6.729774    | .00718    | 937.29 | 0.000 | 6.715697             | 6.74385   |

| Source   | SS         | df    | MS         | Number of obs | = | 4,165  |
|----------|------------|-------|------------|---------------|---|--------|
| Model    | 93.6914807 | 1     | 93.6914807 | F(1, 4163)    | = | 491.72 |
| Residual | 793.213421 | 4,163 | .190538895 | Prob > F      | = | 0.0000 |
|          |            |       |            | R-squared     | = | 0.1056 |
|          |            |       |            | Adj R-squared | = | 0.1054 |
| Total    | 886.904902 | 4,164 | .212993492 | Root MSE      | = | .43651 |

| lwage    | Coefficient | Std. err. | t      | P> t  | [95% conf. interval] |          |
|----------|-------------|-----------|--------|-------|----------------------|----------|
| 0.female | .4744661    | .0213967  | 22.17  | 0.000 | .4325171             | .516415  |
| _cons    | 6.255308    | .020156   | 310.34 | 0.000 | 6.215791             | 6.294824 |

It is evident from the test p-value that the difference is statistically significant. Note, the standard `reg` command assumes homoskedastic SEs. If we believe that the variance of (log of) wages varies with gender, we should estimate heteroskedastic SEs. However, in this instance it will not make a difference to the conclusion.

```
reg lwage female, r
```

Linear regression

Number of obs = 4,165

|            |   |        |
|------------|---|--------|
| F(1, 4163) | = | 520.64 |
| Prob > F   | = | 0.0000 |
| R-squared  | = | 0.1056 |
| Root MSE   | = | .43651 |

|        |             | Robust    |        |       |                      |           |
|--------|-------------|-----------|--------|-------|----------------------|-----------|
| lwage  | Coefficient | std. err. | t      | P> t  | [95% conf. interval] |           |
| female | -.4744661   | .0207939  | -22.82 | 0.000 | -.5152332            | -.4336989 |
| _cons  | 6.729774    | .0072089  | 933.53 | 0.000 | 6.71564              | 6.743907  |

3. Extend the specification in (2) that will enable you to test the hypothesis that there is no difference in the wages between the following gender-ethnicity groups. Begin by defining the following dummy variables:

- $\text{female\_black} = \text{female} \times \text{black}$
- $\text{male\_black} = (1 - \text{female}) \times \text{black}$
- $\text{female\_nonblack} = \text{female} \times (1 - \text{black})$
- $\text{male\_nonblack} = (1 - \text{female}) \times (1 - \text{black})$

```
gen female_black=female*black
gen female_nonblack=female*(1-black)
gen male_black=(1-female)*black
gen male_nonblack=(1-female)*(1-black)
```

Then estimate the following regressions:

- lwage on female\_black, female\_nonblack, male\_black, male\_nonblack (without a constant: option nocons)
- lwage on female, black, female\_black
- lwage on female\_black, female\_nonblack, male\_black

For some of these exercises you may be able to use Stata's factor notation. However, in some instances you will need to manually create the above dummy-variable interactions.

In each case, identify the base category and write down the parameters of the (implied) model in terms of conditional expectations.

```
* (a)
reg lwage female_black female_nonblack male_black male_nonblack

// note, Stata has dropped one variable due to perfect collinearity

reg lwage female_black female_nonblack male_black male_nonblack, nocons
```

note: male\_black omitted because of collinearity.

|             |  |            |       |            |               |   |        |
|-------------|--|------------|-------|------------|---------------|---|--------|
| Source      |  | SS         | df    | MS         | Number of obs | = | 4,165  |
| -----+----- |  |            |       |            |               |   |        |
| Model       |  | 107.436063 | 3     | 35.8120209 | F(3, 4161)    | = | 191.17 |
| Residual    |  | 779.468839 | 4,161 | .187327287 | Prob > F      | = | 0.0000 |
| -----+----- |  |            |       |            |               |   |        |
| Total       |  | 886.904902 | 4,164 | .212993492 | R-squared     | = | 0.1211 |
|             |  |            |       |            | Adj R-squared | = | 0.1205 |
|             |  |            |       |            | Root MSE      | = | .43281 |

|              |  |             |           |        |       |                      |           |
|--------------|--|-------------|-----------|--------|-------|----------------------|-----------|
| lwage        |  | Coefficient | Std. err. | t      | P> t  | [95% conf. interval] |           |
| -----+-----  |  |             |           |        |       |                      |           |
| female_black |  | -.4434409   | .0523433  | -8.47  | 0.000 | -.5460617            | -.34082   |
| female_non~k |  | -.2101553   | .0383456  | -5.48  | 0.000 | -.2853332            | -.1349774 |
| male_black   |  | 0           | (omitted) |        |       |                      |           |
| male_nonbl~k |  | .2239593    | .0317691  | 7.05   | 0.000 | .1616749             | .2862436  |
| _cons        |  | 6.517691    | .0309152  | 210.82 | 0.000 | 6.457081             | 6.578301  |
| -----+-----  |  |             |           |        |       |                      |           |

|             |  |            |       |            |               |   |          |
|-------------|--|------------|-------|------------|---------------|---|----------|
| Source      |  | SS         | df    | MS         | Number of obs | = | 4,165    |
| -----+----- |  |            |       |            |               |   |          |
| Model       |  | 185756.485 | 4     | 46439.1213 | F(4, 4161)    | > | 99999.00 |
| Residual    |  | 779.468839 | 4,161 | .187327287 | Prob > F      | = | 0.0000   |
| -----+----- |  |            |       |            |               |   |          |
| Total       |  | 186535.954 | 4,165 | 44.7865436 | R-squared     | = | 0.9958   |
|             |  |            |       |            | Adj R-squared | = | 0.9958   |
|             |  |            |       |            | Root MSE      | = | .43281   |

|              |  |             |           |        |       |                      |          |
|--------------|--|-------------|-----------|--------|-------|----------------------|----------|
| lwage        |  | Coefficient | Std. err. | t      | P> t  | [95% conf. interval] |          |
| -----+-----  |  |             |           |        |       |                      |          |
| female_black |  | 6.07425     | .0422382  | 143.81 | 0.000 | 5.991441             | 6.15706  |
| female_non~k |  | 6.307536    | .0226856  | 278.04 | 0.000 | 6.26306              | 6.352012 |
| male_black   |  | 6.517691    | .0309152  | 210.82 | 0.000 | 6.457081             | 6.578301 |
| male_nonbl~k |  | 6.74165     | .0073159  | 921.51 | 0.000 | 6.727307             | 6.755993 |
| -----+-----  |  |             |           |        |       |                      |          |

This model returns four parameter-estimates, each corresponding to the four gender-ethnicity groups. These are essentially conditional mean estimates.

```
* (b)
reg lwage female black female_black

// or, using factor notation:

reg lwage i.female##i.black
```

| Source   | SS         | df    | MS         | Number of obs | = | 4,165  |
|----------|------------|-------|------------|---------------|---|--------|
|          |            |       |            | F(3, 4161)    | = | 191.17 |
| Model    | 107.436063 | 3     | 35.8120209 | Prob > F      | = | 0.0000 |
| Residual | 779.468839 | 4,161 | .187327287 | R-squared     | = | 0.1211 |
|          |            |       |            | Adj R-squared | = | 0.1205 |
| Total    | 886.904902 | 4,164 | .212993492 | Root MSE      | = | .43281 |

| lwage        | Coefficient | Std. err. | t      | P> t  | [95% conf. interval] |           |
|--------------|-------------|-----------|--------|-------|----------------------|-----------|
| female       | -.4341146   | .0238361  | -18.21 | 0.000 | -.480846             | -.3873832 |
| black        | -.2239593   | .0317691  | -7.05  | 0.000 | -.2862436            | -.1616749 |
| female_black | -.0093263   | .057515   | -0.16  | 0.871 | -.1220865            | .1034339  |
| _cons        | 6.74165     | .0073159  | 921.51 | 0.000 | 6.727307             | 6.755993  |

| Source   | SS         | df    | MS         | Number of obs | = | 4,165  |
|----------|------------|-------|------------|---------------|---|--------|
|          |            |       |            | F(3, 4161)    | = | 191.17 |
| Model    | 107.436063 | 3     | 35.8120209 | Prob > F      | = | 0.0000 |
| Residual | 779.468839 | 4,161 | .187327287 | R-squared     | = | 0.1211 |
|          |            |       |            | Adj R-squared | = | 0.1205 |
| Total    | 886.904902 | 4,164 | .212993492 | Root MSE      | = | .43281 |

| lwage        | Coefficient | Std. err. | t      | P> t  | [95% conf. interval] |           |
|--------------|-------------|-----------|--------|-------|----------------------|-----------|
| 1.female     | -.4341146   | .0238361  | -18.21 | 0.000 | -.480846             | -.3873832 |
| 1.black      | -.2239593   | .0317691  | -7.05  | 0.000 | -.2862436            | -.1616749 |
| female#black |             |           |        |       |                      |           |
| 1 1          | -.0093263   | .057515   | -0.16  | 0.871 | -.1220865            | .1034339  |

|       |  |         |          |        |       |          |          |
|-------|--|---------|----------|--------|-------|----------|----------|
|       |  |         |          |        |       |          |          |
| _cons |  | 6.74165 | .0073159 | 921.51 | 0.000 | 6.727307 | 6.755993 |

---

In this model, we have the following:

- $\beta_1 = E[\ln(Wage_i)|F = 0, B = 0]$
- $\beta_2 = E[\ln(Wage_i)|F = 1, B = 0] - E[\ln(Wage_i)|F = 0, B = 0]$
- $\beta_3 = E[\ln(Wage_i)|F = 0, B = 1] - E[\ln(Wage_i)|F = 0, B = 0]$
- $\beta_4 = (E[\ln(Wage_i)|F = 1, B = 1] - E[\ln(Wage_i)|F = 0, B = 1]) - (E[\ln(Wage_i)|F = 1, B = 0] - E[\ln(Wage_i)|F = 0, B = 0])$

```
* (c)
reg lwage female_black female_nonblack male_black

// note, this one is harder to replicate using factor notation.
```

|              |  |             |           |            |               |                      |           |
|--------------|--|-------------|-----------|------------|---------------|----------------------|-----------|
| Source       |  | SS          | df        | MS         | Number of obs | =                    | 4,165     |
| -----+       |  |             |           |            | F(3, 4161)    | =                    | 191.17    |
| Model        |  | 107.436063  | 3         | 35.8120209 | Prob > F      | =                    | 0.0000    |
| Residual     |  | 779.468839  | 4,161     | .187327287 | R-squared     | =                    | 0.1211    |
| -----+       |  |             |           |            | Adj R-squared | =                    | 0.1205    |
| Total        |  | 886.904902  | 4,164     | .212993492 | Root MSE      | =                    | .43281    |
| -----        |  |             |           |            |               |                      |           |
| lwage        |  | Coefficient | Std. err. | t          | P> t          | [95% conf. interval] |           |
| -----+       |  |             |           |            |               |                      |           |
| female_black |  | -.6674001   | .0428671  | -15.57     | 0.000         | -.7514426            | -.5833576 |
| female_non~k |  | -.4341146   | .0238361  | -18.21     | 0.000         | -.480846             | -.3873832 |
| male_black   |  | -.2239593   | .0317691  | -7.05      | 0.000         | -.2862436            | -.1616749 |
| _cons        |  | 6.74165     | .0073159  | 921.51     | 0.000         | 6.727307             | 6.755993  |

In this model, we have the following:

- $\beta_1 = E[\ln(Wage_i)|F = 0, B = 0]$
- $\beta_2 = E[\ln(Wage_i)|F = 1, B = 1] - E[\ln(Wage_i)|F = 0, B = 0]$
- $\beta_3 = E[\ln(Wage_i)|F = 1, B = 0] - E[\ln(Wage_i)|F = 0, B = 0]$



- $\beta_4 = E[\ln(Wage_i)|F = 0, B = 1] - E[\ln(Wage_i)|F = 0, B = 0]$

4. Compare the estimated coefficients with the sample average values for the lwage for the four subgroups. What do you see?

```
table female black, stat(mean lwage)
```

| -----          |  |          |          |          |
|----------------|--|----------|----------|----------|
|                |  | black    |          |          |
|                |  | 0        | 1        | Total    |
| -----          |  |          |          | -----    |
| female or male |  |          |          |          |
| 0              |  | 6.74165  | 6.517691 | 6.729774 |
| 1              |  | 6.307536 | 6.07425  | 6.255308 |
| Total          |  | 6.700755 | 6.363002 | 6.676346 |

You can compute the coefficients from each model simply using these averages.

5. In each of the above models, describe the null hypothesis you would test to evaluate whether there is a significant earnings difference between the earnings of black and non-black females.

- $H_0 : \beta_1 = \beta_2$
- $H_0 : \beta_3 + \beta_4 = 0$
- $H_0 : \beta_2 - \beta_3 = 0$

6. Verify your solution to 4 by performing a test using the three set of regression output. You can use the post-estimation `test` command.

```
reg lwage female_black female_nonblack male_black male_nonblack, nocons

test female_black = female_nonblack

reg lwage female black female_black

test female_black + black = 0

reg lwage female_black female_nonblack male_black

test female_black - female_nonblack = 0
```

| Source   | SS         | df    | MS         | Number of obs | = | 4,165    |
|----------|------------|-------|------------|---------------|---|----------|
| Model    | 185756.485 | 4     | 46439.1213 | F(4, 4161)    | > | 99999.00 |
| Residual | 779.468839 | 4,161 | .187327287 | Prob > F      | = | 0.0000   |
|          |            |       |            | R-squared     | = | 0.9958   |
|          |            |       |            | Adj R-squared | = | 0.9958   |
| Total    | 186535.954 | 4,165 | 44.7865436 | Root MSE      | = | .43281   |

| lwage        | Coefficient | Std. err. | t      | P> t  | [95% conf. interval] |          |
|--------------|-------------|-----------|--------|-------|----------------------|----------|
| female_black | 6.07425     | .0422382  | 143.81 | 0.000 | 5.991441             | 6.15706  |
| female_non~k | 6.307536    | .0226856  | 278.04 | 0.000 | 6.26306              | 6.352012 |
| male_black   | 6.517691    | .0309152  | 210.82 | 0.000 | 6.457081             | 6.578301 |
| male_nonbl~k | 6.74165     | .0073159  | 921.51 | 0.000 | 6.727307             | 6.755993 |

( 1) female\_black - female\_nonblack = 0

F( 1, 4161) = 23.68  
 Prob > F = 0.0000

| Source   | SS         | df    | MS         | Number of obs | = | 4,165  |
|----------|------------|-------|------------|---------------|---|--------|
| Model    | 107.436063 | 3     | 35.8120209 | F(3, 4161)    | = | 191.17 |
| Residual | 779.468839 | 4,161 | .187327287 | Prob > F      | = | 0.0000 |
|          |            |       |            | R-squared     | = | 0.1211 |
|          |            |       |            | Adj R-squared | = | 0.1205 |
| Total    | 886.904902 | 4,164 | .212993492 | Root MSE      | = | .43281 |

| lwage        | Coefficient | Std. err. | t      | P> t  | [95% conf. interval] |           |
|--------------|-------------|-----------|--------|-------|----------------------|-----------|
| female       | -.4341146   | .0238361  | -18.21 | 0.000 | -.480846             | -.3873832 |
| black        | -.2239593   | .0317691  | -7.05  | 0.000 | -.2862436            | -.1616749 |
| female_black | -.0093263   | .057515   | -0.16  | 0.871 | -.1220865            | .1034339  |
| _cons        | 6.74165     | .0073159  | 921.51 | 0.000 | 6.727307             | 6.755993  |

( 1) black + female\_black = 0

F( 1, 4161) = 23.68  
 Prob > F = 0.0000

| Source   | SS         | df    | MS         | Number of obs | = | 4,165  |
|----------|------------|-------|------------|---------------|---|--------|
|          |            |       |            | F(3, 4161)    | = | 191.17 |
| Model    | 107.436063 | 3     | 35.8120209 | Prob > F      | = | 0.0000 |
| Residual | 779.468839 | 4,161 | .187327287 | R-squared     | = | 0.1211 |
|          |            |       |            | Adj R-squared | = | 0.1205 |
| Total    | 886.904902 | 4,164 | .212993492 | Root MSE      | = | .43281 |

| lwage        | Coefficient | Std. err. | t      | P> t  | [95% conf. interval] |           |
|--------------|-------------|-----------|--------|-------|----------------------|-----------|
| female_black | -.6674001   | .0428671  | -15.57 | 0.000 | -.7514426            | -.5833576 |
| female_non~k | -.4341146   | .0238361  | -18.21 | 0.000 | -.480846             | -.3873832 |
| male_black   | -.2239593   | .0317691  | -7.05  | 0.000 | -.2862436            | -.1616749 |
| _cons        | 6.74165     | .0073159  | 921.51 | 0.000 | 6.727307             | 6.755993  |

( 1) female\_black - female\_nonblack = 0

F( 1, 4161) = 23.68  
 Prob > F = 0.0000

7. In each case, test equality across all four gender-ethnicity groups. Again, you should get the same result.

```
reg lwage female_black female_nonblack male_black male_nonblack, nocons

test female_black = female_nonblack = male_black = male_nonblack

reg lwage female black female_black

test female_black = black = female = 0

reg lwage female_black female_nonblack male_black

test female_black = female_nonblack = male_black = 0
```

| Source   | SS         | df    | MS         | Number of obs | = | 4,165    |
|----------|------------|-------|------------|---------------|---|----------|
|          |            |       |            | F(4, 4161)    | > | 99999.00 |
| Model    | 185756.485 | 4     | 46439.1213 | Prob > F      | = | 0.0000   |
| Residual | 779.468839 | 4,161 | .187327287 | R-squared     | = | 0.9958   |
|          |            |       |            | Adj R-squared | = | 0.9958   |

Total | 186535.954      4,165    44.7865436    Root MSE      =      .43281

| lwage        | Coefficient | Std. err. | t      | P> t  | [95% conf. interval] |          |
|--------------|-------------|-----------|--------|-------|----------------------|----------|
| female_black | 6.07425     | .0422382  | 143.81 | 0.000 | 5.991441             | 6.15706  |
| female_non~k | 6.307536    | .0226856  | 278.04 | 0.000 | 6.26306              | 6.352012 |
| male_black   | 6.517691    | .0309152  | 210.82 | 0.000 | 6.457081             | 6.578301 |
| male_nonbl~k | 6.74165     | .0073159  | 921.51 | 0.000 | 6.727307             | 6.755993 |

- ( 1) female\_black - female\_nonblack = 0  
 ( 2) female\_black - male\_black = 0  
 ( 3) female\_black - male\_nonblack = 0

F( 3, 4161) = 191.17  
 Prob > F = 0.0000

| Source   | SS         | df    | MS         | Number of obs | = | 4,165  |
|----------|------------|-------|------------|---------------|---|--------|
| Model    | 107.436063 | 3     | 35.8120209 | F(3, 4161)    | = | 191.17 |
| Residual | 779.468839 | 4,161 | .187327287 | Prob > F      | = | 0.0000 |
| Total    | 886.904902 | 4,164 | .212993492 | R-squared     | = | 0.1211 |
|          |            |       |            | Adj R-squared | = | 0.1205 |
|          |            |       |            | Root MSE      | = | .43281 |

| lwage        | Coefficient | Std. err. | t      | P> t  | [95% conf. interval] |           |
|--------------|-------------|-----------|--------|-------|----------------------|-----------|
| female       | -.4341146   | .0238361  | -18.21 | 0.000 | -.480846             | -.3873832 |
| black        | -.2239593   | .0317691  | -7.05  | 0.000 | -.2862436            | -.1616749 |
| female_black | -.0093263   | .057515   | -0.16  | 0.871 | -.1220865            | .1034339  |
| _cons        | 6.74165     | .0073159  | 921.51 | 0.000 | 6.727307             | 6.755993  |

- ( 1) - black + female\_black = 0  
 ( 2) - female + female\_black = 0  
 ( 3) female\_black = 0

F( 3, 4161) = 191.17  
 Prob > F = 0.0000

| Source | SS | df | MS | Number of obs | = | 4,165  |
|--------|----|----|----|---------------|---|--------|
|        |    |    |    | F(3, 4161)    | = | 191.17 |

|             |  |            |       |            |               |   |        |
|-------------|--|------------|-------|------------|---------------|---|--------|
| Model       |  | 107.436063 | 3     | 35.8120209 | Prob > F      | = | 0.0000 |
| Residual    |  | 779.468839 | 4,161 | .187327287 | R-squared     | = | 0.1211 |
| -----+----- |  |            |       |            | Adj R-squared | = | 0.1205 |
| Total       |  | 886.904902 | 4,164 | .212993492 | Root MSE      | = | .43281 |

| lwage        | Coefficient | Std. err. | t      | P> t  | [95% conf. interval] |           |
|--------------|-------------|-----------|--------|-------|----------------------|-----------|
| female_black | -.6674001   | .0428671  | -15.57 | 0.000 | -.7514426            | -.5833576 |
| female_non~k | -.4341146   | .0238361  | -18.21 | 0.000 | -.480846             | -.3873832 |
| male_black   | -.2239593   | .0317691  | -7.05  | 0.000 | -.2862436            | -.1616749 |
| _cons        | 6.74165     | .0073159  | 921.51 | 0.000 | 6.727307             | 6.755993  |

( 1) female\_black - female\_nonblack = 0

( 2) female\_black - male\_black = 0

( 3) female\_black = 0

F( 3, 4161) = 191.17  
 Prob > F = 0.0000

8. Try to replicate the F-statistic for one of the above models. Hint, the F-stat for these models is the same as that of the whole model.

```
reg lwage female black female_black
ereturn list
scalar fstat = (e(r2)*e(df_r))/((1-e(r2))*e(df_m))
scalar list fstat
```

|             |  |            |       |            |               |   |        |
|-------------|--|------------|-------|------------|---------------|---|--------|
| Source      |  | SS         | df    | MS         | Number of obs | = | 4,165  |
| -----+----- |  |            |       |            | F(3, 4161)    | = | 191.17 |
| Model       |  | 107.436063 | 3     | 35.8120209 | Prob > F      | = | 0.0000 |
| Residual    |  | 779.468839 | 4,161 | .187327287 | R-squared     | = | 0.1211 |
| -----+----- |  |            |       |            | Adj R-squared | = | 0.1205 |
| Total       |  | 886.904902 | 4,164 | .212993492 | Root MSE      | = | .43281 |

| lwage       |  | Coefficient | Std. err. | t      | P> t  | [95% conf. interval] |           |
|-------------|--|-------------|-----------|--------|-------|----------------------|-----------|
| -----+----- |  |             |           |        |       |                      |           |
| female      |  | -.4341146   | .0238361  | -18.21 | 0.000 | -.480846             | -.3873832 |
| black       |  | -.2239593   | .0317691  | -7.05  | 0.000 | -.2862436            | -.1616749 |

|              |           |          |        |       |           |          |
|--------------|-----------|----------|--------|-------|-----------|----------|
| female_black | -.0093263 | .057515  | -0.16  | 0.871 | -.1220865 | .1034339 |
| _cons        | 6.74165   | .0073159 | 921.51 | 0.000 | 6.727307  | 6.755993 |

---

scalars:

```

e(N) = 4165
e(df_m) = 3
e(df_r) = 4161
e(F) = 191.1735422330181
e(r2) = .1211359442526956
e(rmse) = .4328132235793649
e(mss) = 107.4360627542734
e(rss) = 779.4688391479763
e(r2_a) = .1205023003768864
e(ll) = -2419.902951629166
e(ll_0) = -2688.805870567022
e(rank) = 4

```

macros:

```

e(cmdline) : "regress lwage female black female_black"
e(title) : "Linear regression"
e(marginsok) : "XB default"
e(vce) : "ols"
e(depvar) : "lwage"
e(cmd) : "regress"
e(properties) : "b V"
e(predict) : "regres_p"
e(model) : "ols"
e(estat_cmd) : "regress_estat"

```

matrices:

```

e(b) : 1 x 4
e(V) : 4 x 4
e(beta) : 1 x 3

```

functions:

```

e(sample)
fstat = 191.17354

```

In the case where the F-test corresponds to the test of the entire model, you can write the F-statistic in terms of  $R^2$ .

9. Estimate the following model:

$$lwage = \beta_1 + \beta_2 F + \beta_3 B + \beta_4 F \times B + \beta_5 exp + \beta_6 exp^2 + \beta_7 educ + \varepsilon$$

- Interpret the estimated coefficients  $\hat{\beta}_7$ .
- Interpret the effect of experience variable **exp**. Use the median level of experience to make your calculation.

```
sum educ, det
reg lwage i.female##i.black exper expsq educ

di (exp(_b[educ])-1)*100
```

| years of education |             |           |             |               |                      |           |
|--------------------|-------------|-----------|-------------|---------------|----------------------|-----------|
| Percentiles        |             | Smallest  |             |               |                      |           |
| 1%                 | 6           | 4         |             |               |                      |           |
| 5%                 | 8           | 4         |             |               |                      |           |
| 10%                | 9           | 4         | Obs         |               | 4,165                |           |
| 25%                | 12          | 4         | Sum of wgt. |               | 4,165                |           |
| 50%                | 12          |           | Mean        |               | 12.84538             |           |
|                    |             | Largest   | Std. dev.   |               | 2.787995             |           |
| 75%                | 16          | 17        |             |               |                      |           |
| 90%                | 17          | 17        | Variance    |               | 7.772916             |           |
| 95%                | 17          | 17        | Skewness    |               | -.2581161            |           |
| 99%                | 17          | 17        | Kurtosis    |               | 2.71273              |           |
| Source             | SS          | df        | MS          | Number of obs | =                    | 4,165     |
|                    |             |           |             | F(6, 4158)    | =                    | 411.81    |
| Model              | 330.586     | 6         | 55.0976667  | Prob > F      | =                    | 0.0000    |
| Residual           | 556.318902  | 4,158     | .13379483   | R-squared     | =                    | 0.3727    |
|                    |             |           |             | Adj R-squared | =                    | 0.3718    |
| Total              | 886.904902  | 4,164     | .212993492  | Root MSE      | =                    | .36578    |
| lwage              | Coefficient | Std. err. | t           | P> t          | [95% conf. interval] |           |
| 1.female           | -.4032047   | .0203208  | -19.84      | 0.000         | -.4430444            | -.363365  |
| 1.black            | -.1551546   | .0269249  | -5.76       | 0.000         | -.2079417            | -.1023674 |
| female#black       |             |           |             |               |                      |           |

|       |  |           |          |        |       |           |           |
|-------|--|-----------|----------|--------|-------|-----------|-----------|
| 1 1   |  | -.002071  | .0488405 | -0.04  | 0.966 | -.0978246 | .0936826  |
|       |  |           |          |        |       |           |           |
| exper |  | .0427346  | .0022404 | 19.07  | 0.000 | .0383422  | .047127   |
| expsq |  | -.0006982 | .0000494 | -14.14 | 0.000 | -.0007951 | -.0006014 |
| educ  |  | .0731837  | .0020983 | 34.88  | 0.000 | .0690698  | .0772975  |
| _cons |  | 5.303667  | .0362462 | 146.32 | 0.000 | 5.232606  | 5.374729  |

---

7.5928119

- A one unit increase in years of educ is associated with an increase of 7.59% in expected wages, holding other regressors fixed.

```
sum exper, det
return list
di (exp(_b[exper]+2*r(p50)*_b[expsq])-1)*100
```

years of full-time work experience

---

| Percentiles |    | Smallest |             |          |
|-------------|----|----------|-------------|----------|
| 1%          | 3  | 1        |             |          |
| 5%          | 5  | 1        |             |          |
| 10%         | 7  | 1        | Obs         | 4,165    |
| 25%         | 11 | 1        | Sum of wgt. | 4,165    |
| 50%         | 18 |          | Mean        | 19.85378 |
|             |    | Largest  | Std. dev.   | 10.96637 |
| 75%         | 29 | 50       |             |          |
| 90%         | 36 | 50       | Variance    | 120.2613 |
| 95%         | 39 | 50       | Skewness    | .4000014 |
| 99%         | 44 | 51       | Kurtosis    | 2.072064 |

scalars:

```

r(N) = 4165
r(sum_w) = 4165
r(mean) = 19.85378151260504
r(Var) = 120.2612759224727
r(sd) = 10.96637022548814
r(skewness) = .4000013893781186
r(kurtosis) = 2.072064120792274
r(sum) = 82691
r(min) = 1

```



```

r(max) = 51
r(p1) = 3
r(p5) = 5
r(p10) = 7
r(p25) = 11
r(p50) = 18
r(p75) = 29
r(p90) = 36
r(p95) = 39
r(p99) = 44

```

1.7754427

- A one unit increase in years of experience is associated with an increase of 1.78% in expected wages, holding other regressors fixed.

**10.** Theoretically, how would you test the following restrictions for the model below?

- $\beta_2 = \beta_3$
- $\beta_4 + \beta_5 = 1$
- $\beta_2 = \beta_3$  **and**  $\beta_4 + \beta_5 = 1$

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

- Restrictions 1:  $H_0 : \beta_2 = \beta_3 \Rightarrow \beta_2 - \beta_3 = 0$ . This can be written as a simple T-test (or F-tests),

$$\text{T-stat} = \frac{\hat{\beta}_2 - \hat{\beta}_3}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_2 - \hat{\beta}_3)}}$$

where,

$$\text{Var}(\hat{\beta}_2 - \hat{\beta}_3) = \text{Var}(\hat{\beta}_2) + \text{Var}(\hat{\beta}_3) - 2 \cdot \text{Cov}(\hat{\beta}_2, \hat{\beta}_3)$$

Alternatively, rewrite the model, adding and subtracting  $\beta_3 X_2$  (or  $\beta_2 X_3$ ):

$$Y = \beta_1 + (\beta_2 - \beta_3)X_2 + \beta_3(X_2 + X_3) + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

Then test the hypothesis that the coefficient on  $X_2$  is  $= 0$ .

- Restrictions 1:  $H_0 : \beta_4 + \beta_5 = 1$ . This can be written as a simple T-test,

$$\text{T-stat} = \frac{\hat{\beta}_4 + \hat{\beta}_5 - 1}{\sqrt{\widehat{Var}(\hat{\beta}_4 + \hat{\beta}_5)}}$$

where,

$$\text{Var}(\hat{\beta}_4 + \hat{\beta}_5) = \text{Var}(\hat{\beta}_4) + \text{Var}(\hat{\beta}_5) + 2 \cdot \text{Cov}(\hat{\beta}_4, \hat{\beta}_5)$$

Alternatively, rewrite the model, adding and subtracting  $\beta_5 X_4 - X_5$  (or  $\beta_2 X_3$ ):

$$Y - X_4 = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + (\beta_4 + \beta_5 - 1)X_4 + (\beta_5)(X_5 - X_4) + \varepsilon$$

Then test the hypothesis that the coefficient on  $X_4$  is  $= 0$ .

c. To test both of the linear restrictions simultaneously, we would use an F-test.

Step 1: estimate the unrestricted model

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

Store the  $SSR_U$

Step 2: estimate the restricted model

$$(Y - X_4) = \gamma_1 + \gamma_2(X_2 + X_3) + \gamma_5(X_5 - X_4) + \varepsilon$$

Store the  $SSR_R$ .

Step 3: Compute the F-statistic

$$\text{F-stat} = \frac{(SSR_R - SSR_U)/(df_R - df_U)}{SSR_U/df_U}$$

**Postamble**

[log](#) [close](#)