Revised July 2018

Multiple Linear Regression Analysis with Indicator Variables

This set of notes discusses the use of Stata for multiple regression analysis involving indicator (dummy) variables. It assumes that you have set Stata up on your computer (see the "Getting Started with Stata" handout), and that you have read in the set of data that you want to analyze (see the "Reading in Stata Format (.dta) Data Files" handout).

In Stata, most tasks can be performed either by issuing commands within the "Stata command" window, **or** by using the menus. These notes illustrate both approaches, using the data file "GSS2016.DTA" (this data file is posted here: https://canvas.harvard.edu/courses/53958).

The illustration here considers sex and race differences in hours worked. First, create the indicator (dummy) variables that represent differences between categories of a nominal or ordinal variable.

The distribution of "sex" in the data set shows that there are 1,276 men and 1,591 women:

```
respondents |
        sex |      Freq.     Percent        Cum.
------------+-----------------------------------
       male |      1,276       44.51       44.51
     female |      1,591       55.49      100.00
------------+-----------------------------------
      Total |      2,867      100.00
```

In the command-line approach, one can create indicator variables identifying men and women via the following sequence of commands:

```
quietly tab sex,gen(sex)
rename sex1 male
rename sex2 female
```

The first command creates two indicator variables called "sex1" and "sex2"; these are 0/1 variables identifying the first and second categories of "sex", respectively. (The "quietly" tells Stata not to display the results of the "tab" command.) This command results in two indicator variables, one identifying men (reference category women) and the other identifying women (reference category men). (More on reference categories below.)

The two "rename" commands change the names of "sex1" and "sex2", to "male" and "female", respectively, making them a little more intuitive. Strictly speaking these name changes are not necessary. The distribution of "female" is as follows:

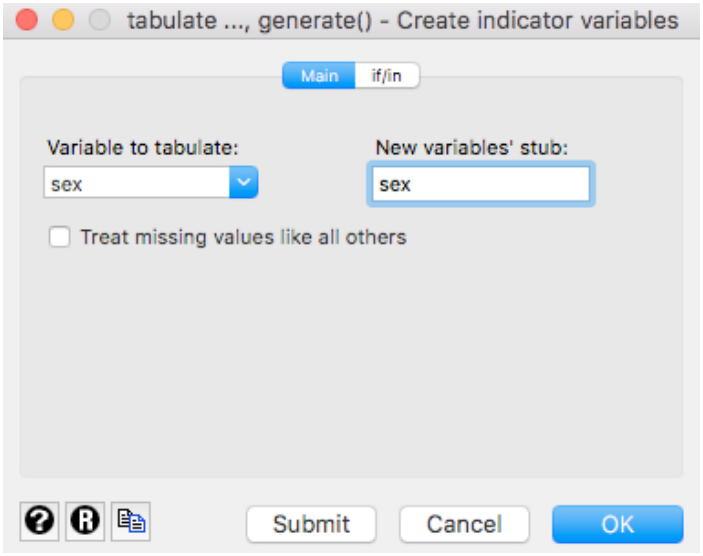
```
sex==female |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,276       44.51       44.51
          1 |      1,591       55.49      100.00
------------+-----------------------------------
      Total |      2,867      100.00
```

Note that the 1,591 women are now coded 1 while the 1,276 men are now coded 0. "Male" identifies men instead; for it, the 1,276 men are coded 1 and the 1,591 women are coded 0.

Via the menus, you can create indicator variables as follows:

click on "Data"
click on "Create or change data"
click on "Other variable creation commands"
click on "Create indicator variables"

A window like this will open up:



Fill in the name of the categorical variable in the "Variable to tabulate:" box and the "stub" name for the indicator variables in the "New variables' stub:" box. (The "stub" here is "sex"; Stata will add "1", "2" to it in order to name the indicator variables.) When you click "OK", Stata will create a 0/1 indicator variable that identifies each category of the categorical variable. Here, these are "sex1" and "sex2", and their distributions are

```
. tab sex1

  sex==male |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,591       55.49       55.49
          1 |      1,276       44.51      100.00
------------+-----------------------------------
      Total |      2,867      100.00

. tab sex2

sex==female |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,276       44.51       44.51
          1 |      1,591       55.49      100.00
------------+-----------------------------------
      Total |      2,867      100.00
```

One can "rename" sex1 and sex2, as shown above for the command-line approach (see the "Recoding…" handout for more information).

The "quietly tab…" command will create an indicator variable for each category in a categorical variable. For example, "race" has three categories in this data set:

```
    race of |
 respondent |      Freq.     Percent        Cum.
------------+-----------------------------------
      white |      2,100       73.25       73.25
      black |        490       17.09       90.34
      other |        277        9.66      100.00
------------+-----------------------------------
      Total |      2,867      100.00
```

The following commands create three indicator variables identifying whites, blacks, and others, respectively:

```
quietly tab race,gen(race)
rename race1 white
rename race2 black
rename race3 othrace
```

Here are distributions for the indicator variables that are produced:

```
. tab white

race==white |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        767       26.75       26.75
          1 |      2,100       73.25      100.00
------------+-----------------------------------
      Total |      2,867      100.00

. tab black

race==black |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,377       82.91       82.91
          1 |        490       17.09      100.00
------------+-----------------------------------
      Total |      2,867      100.00

. tab othrace

race==other |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,590       90.34       90.34
          1 |        277        9.66      100.00
------------+-----------------------------------
      Total |      2,867      100.00
```

Once the indicator variables are created, they can be used like other explanatory (independent) variables in multiple regression. To use a categorical variable as a predictor, enter all but one of its indicator variables as explanatory variables. (If you try to enter all of them, one will be removed automatically by Stata and you will get an error message). The category identified by

the indicator variable that is not entered is the "reference category" for that particular categorical variable. It is the reference category for interpreting the coefficients of the other indicator variable(s). For example, here is a regression of hours worked (per week) on education and the indicator variable identifying females. The indicator variable for males is not entered. Males are the reference category for sex.

```
. regress hrs1 educ female

      Source |       SS           df       MS      Number of obs   =     1,645
-------------+----------------------------------   F(2, 1642)      =     42.52
       Model |  16811.1586         2  8405.57932   Prob > F        =    0.0000
    Residual |  324590.926     1,642  197.680223   R-squared       =    0.0492
-------------+----------------------------------   Adj R-squared   =    0.0481
       Total |  341402.085     1,644  207.665502   Root MSE        =     14.06

------------------------------------------------------------------------------
        hrs1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .2303655     .12042     1.91   0.056    -.0058276    .4665585
      female |  -6.339185   .6950595    -9.12   0.000    -7.702481   -4.975888
       _cons |   40.94105   1.755433    23.32   0.000     37.49793    44.38418
------------------------------------------------------------------------------
```

Controlling for education, women in the sample work about 6.34 hours less than do men (the reference category for sex). Controlling for sex, each additional year of education is linked to working about 0.23 hours more per week.

The t statistic for "female" is -9.12, with a p level lower than 0.001. So we are very confident in inferring that there is a difference between women and men in hours worked, controlling for education. The 95% confidence interval for the difference in hours worked between women and men (controlling for education) is (-7.7, -4.98), which does not include the value 0.

To examine race differences, one would add two of the three indicator variables for race. For example:

```
. regress hrs1 educ female black othrace

      Source |       SS           df       MS      Number of obs   =     1,645
-------------+----------------------------------   F(4, 1640)      =     22.39
       Model |  17681.3984         4  4420.34959   Prob > F        =    0.0000
    Residual |  323720.687     1,640  197.390663   R-squared       =    0.0518
-------------+----------------------------------   Adj R-squared   =    0.0495
       Total |  341402.085     1,644  207.665502   Root MSE        =     14.05

------------------------------------------------------------------------------
        hrs1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .2126625   .1214523     1.75   0.080    -.0255554    .4508805
      female |  -6.345217   .6962263    -9.11   0.000    -7.710804   -4.979631
       black |    .305603   .9308723     0.33   0.743    -1.520221    2.131427
     othrace |  -2.312312    1.16121    -1.99   0.047    -4.589923   -.0347009
       _cons |   41.37902    1.80553    22.92   0.000     37.83763     44.9204
------------------------------------------------------------------------------
```

Since "black" and "othrace" are the included indicator variables, the reference category for "race" is "white". The coefficients for "black" and "othrace" are interpreted as follows: Blacks work 0.31 hours more per week than do whites, controlling for education and sex. Non-white, non-black individuals (i.e., "other race" individuals) work 2.31 hours less per week than do whites, controlling for education and sex. Note that the partial regression coefficients for indicator variables are always interpreted as differences relative to the reference category.

Since the t-statistic for the "black" indicator variable has a p level greater than 0.05, we are not confident that the estimated difference between blacks and whites generalizes beyond the sample.

To test for whether there are differences (controlling for sex and education) among the three "race" categories, we can use the post-estimation "test" command to produce the "incremental F test" for removing "black" and "othrace" from the above regression (for more information on this command, see the "Multiple Linear Regression Analysis" handout):

```
. test black othrace

 ( 1)  black = 0
 ( 2)  othrace = 0

       F(  2,  1640) =    2.20
            Prob > F =    0.1106
```

This F statistic has a relatively high p level, indicating that controlling for sex and education, the differences in hours worked per week among whites, blacks, and others are not statistically significant.