Lab 10 - Binary Variables

Spring 2018

Contents

1 Introduction

In previous labs we briefly discussed how sometimes it is important to incorporate *qualitative* factors in our empirical analysis. We introduced the notion of *dummy* variable, and included in our empirical investigations, for instance, *gender* (male-female). Such characteristics are considered qualitative factors, and can be represented by dummy variables.

A dummy (or indicator, binary, boolean) variable is a variable which takes on the value of 0 or 1 depending on the presence of a specific characteristic. For example an individual is either male or female; a firm can be publicly or privately held. The characteristics above cannot be represented using a continuous variable, but a binary variable (or a set of binary variables) can be used to provide a categorical description. A dummy variable can appear both as a dependent or independent variable in a linear regression. Today we will consider only the case when dummies are used as independent explanatory variables.

Today we will use `wage1` dataset available with the `bcuse` command.

2 SLR on a Dummy

Let's look at the following model:

$$wage = \beta_0 + \beta_1 female + U$$

where $female$ is a dummy variable: $female = 1$ if the worker is female and $female = 0$ when the worker is male. Notice that the expected (or predicted) wage for a male:

$$E(wage|female = 0) = \beta_0$$

whereas the expected wage for a female is:

$$E(wage|female = 1) = \beta_0 + \beta_1$$

So that β_1 measures the expected wage gap between the two genders.

$$E(wage|female = 1) - E(wage|female = 0) = \beta_1$$

Now estimate the simple model above:

```
. reg wage female
```

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =   68.54
       Model |  828.220467     1  828.220467           Prob > F      =  0.0000
    Residual |  6332.19382   524  12.0843394           R-squared     =  0.1157
-------------+------------------------------           Adj R-squared =  0.1140
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.4763

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -2.51183   .3034092    -8.28   0.000    -3.107878   -1.915782
       _cons |   7.099489   .2100082    33.81   0.000     6.686928     7.51205
------------------------------------------------------------------------------
```

Following from above, 7.1 is the predicted wage for male workers, 4.59 is the predicted wage for females and -2.51 is the wage gap. We can immediately get the mean wage by gender by executing the following command:

```
margins, by(female)
```

```
Predictive margins                                 Number of obs   =        526
Model VCE    : OLS

Expression   : Linear prediction, predict()
over         : female

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
          0  |   7.099489   .2100082    33.81   0.000     6.687881    7.511098
          1  |   4.587659   .2189834    20.95   0.000     4.158459    5.016858
------------------------------------------------------------------------------
```

How do the summary statistics below relate to your estimates above?

```
. bysort female: sum wage
-------------------------------------------------------------------------------
-> female = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wage |       274    7.099489    4.160858        1.5      24.98

-------------------------------------------------------------------------------
-> female = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wage |       252    4.587659    2.529363        .53      21.63
```

3 MLR with binary independent variables

The SLR above does not control for important variables that affect workers wages, so the result by itself is not sufficient evidence of a wage gender gap. For example, as we have seen in the previous labs, education can explain part of the wage differentials across individuals: if women have less years of schooling than men, then the results found above might be due to differences in education.

Let's examine the difference in wages between the female and male workers controlling for education. Before we look at the regression estimates, let's plot the above relationship:

```
. twoway (lfit wage educ if female==1, clc(red)) (lfit wage educ
if female==0, clc(green))
```

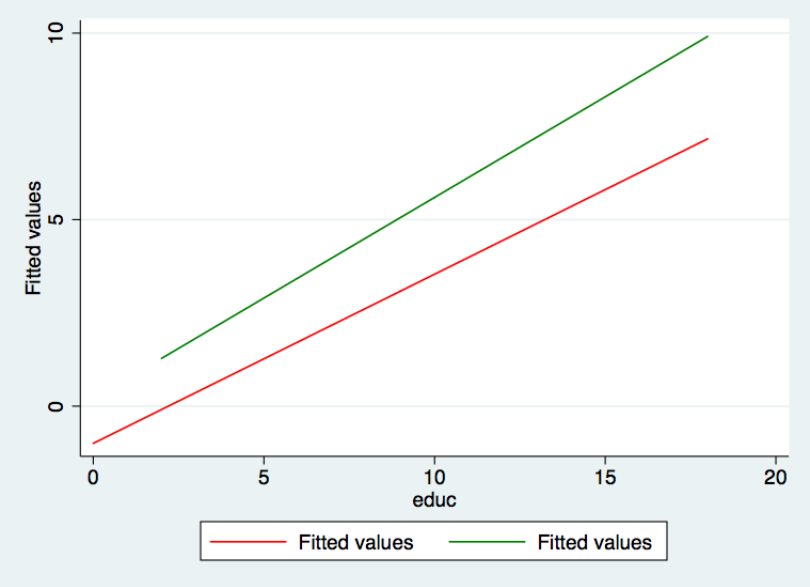

Figure 1: Relationship between wage and education

where the option `, clc()` defines the color (type `help colorstyle` for more colors). What can we say about the gender wage gap controlling for education? (red line for females, green for men).

- Notice that both the intercept and the slope seem to be different between the two regressions.
- This means that there is not only a systematic wage gap between the two genders (represented by the different intercept) but this gap *changes* with the level of education (represented by the different slope).
- Another interpretation of the different slope is that the effect of an extra year of education on wages is different for males and females.

Think hard about the previous bullet points and make sure you understand what they mean!

3.1 MLR with a Dummy: different intercepts, same slope

First, let's consider a model where we allow for a wage differential between female and male workers (different intercepts), but in which the effect of education is the same for both men and women (same slope).

$$wage = \beta_0 + \beta_1 female + \beta_2 educ + U$$

Here for a given level of education the predicted wage for a female is:

$$E(wage|female = 1, educ) = \beta_0 + \beta_1 + \beta_2 educ$$

whereas the expected wage for a male for a given level of education is:

$$E(wage|female = 0, educ) = \beta_0 + \beta_2 educ$$

So the gender gap for two people with the same level of education is

$$E(wage|female = 1, educ) - E(wage|female = 0, educ) = \beta_1$$

which is a constant as we previously argued. Now estimate the regression:

```
.  reg wage female educ

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   91.32
       Model |  1853.25304     2  926.626518           Prob > F      =  0.0000
    Residual |  5307.16125   523  10.1475359           R-squared     =  0.2588
-------------+------------------------------           Adj R-squared =  0.2560
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.1855

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.273362   .2790444    -8.15   0.000    -2.821547   -1.725176
        educ |   .5064521   .0503906    10.05   0.000     .4074592     .605445
       _cons |   .6228168   .6725334     0.93   0.355     -.698382    1.944016
------------------------------------------------------------------------------
```

- So, holding education fixed, the wage gap between men and women is 2.27.
- An extra year of education increases the wage by 50 cents, irrespective of whether you are a female or a male.
- Compare your results with those of the simple regression estimated above, are they the same? Did you expect the change?

Now try to graph the predicted values from this regression. One way to do it is the following:

```
. predict wagehat
(option xb assumed; fitted values)

. scatter wagehat educ
```

As you can see the predicted values lie on two separate lines, one for females and one for males. The two lines have the same slope. This is essentially what a dummy variable does.

3.2 MLR with a Dummy: different intercepts, different slopes

Above we assumed the slope coefficient on education is the same for both female and male workers - i.e. women and men had the same returns to education. Now, let's consider a model where we allow female and male workers to have different returns to education. Consider the following model:

$$wage = \beta_0 + \beta_1 female + \beta_2 educ + \beta_3 female * educ + U$$

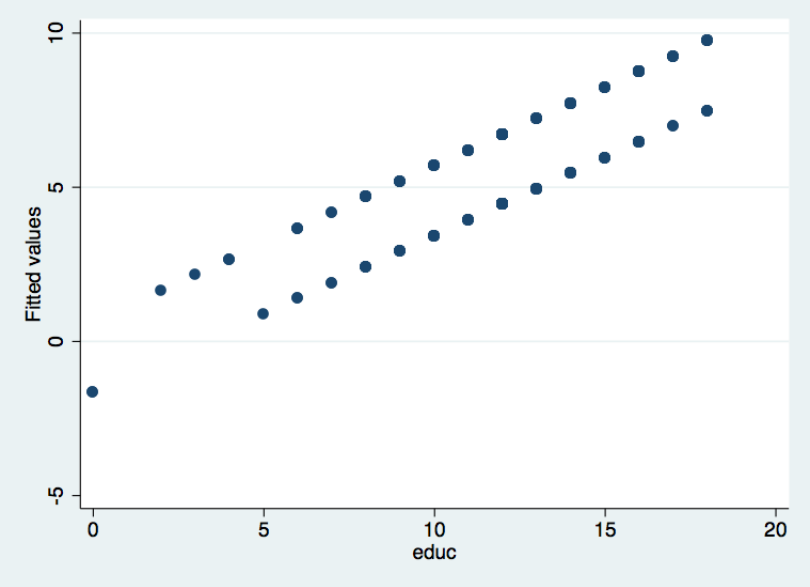

Figure 2: Plotted Predicted values

Note: The interaction term between the dummy variable ($female$) and the education variable allows us to estimate a slope coefficient of education separately for the female and male workers in the sample.

- The predicted wage for a male that has a certain number of years of schooling is:

$$E(wage|female = 0, educ) = \beta_0 + \beta_2 educ$$

- The predicted wage for a female that has a certain number of years of schooling is:

$$E(wage|female = 1, educ) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)educ$$

- and the predicted wage gap for two people with the same years of education will be:

$$E(wage|female = 1, educ) - E(wage|female = 0, educ) = \beta_1 + \beta_3 educ$$

  which **varies** with years of education.

Estimate the model by creating the interaction variable and then running the regression:

```
. gen feduc=female*educ

. reg wage female educ feduc
```

Or alternatively by typing the following without generating the interaction variable first (more on this syntax in the last section of this handout):

```
. reg wage i.female##c.educ
```

In any case the output will be:

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  3,   522) =   61.07
       Model |  1860.24439     3  620.081463            Prob > F      =  0.0000
    Residual |   5300.1699   522  10.1535822            R-squared     =  0.2598
-------------+------------------------------           Adj R-squared =  0.2555
       Total |  7160.41429   525  13.6388844            Root MSE      =  3.1865
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -1.198523    1.32504    -0.90   0.366    -3.801589    1.404543
        educ |    .539476   .0642229     8.40   0.000     .4133089    .6656432
       feduc |   -.085999   .1036388    -0.83   0.407    -.2895994    .1176014
       _cons |   .2004963   .8435616     0.24   0.812    -1.456696    1.857689
------------------------------------------------------------------------------
```

- Is the return to education greater for men or women?
- β_2 is the effect of an extra year of education on wages for a male whereas $\beta_2 + \beta_3$ is the effect for females.
- Clearly, returns for men are greater.

Again, plot the predicted values:

```
predict wagehat2
scatter wagehat2 educ
```

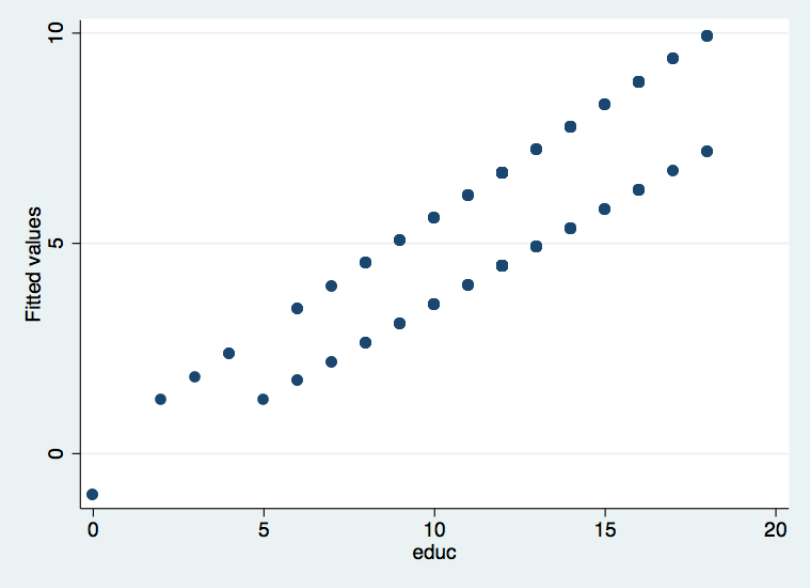

Figure 3: Plotted Predicted values

To conveniently interact multiple variables we can execute the following command:

```
reg wage i.female##c.(educ exper numdep)
```

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  7,   518) =   39.71
       Model |  2500.69654     7  357.242362           Prob > F      =  0.0000
    Residual |  4659.71776   518  8.99559412           R-squared     =  0.3492
-------------+------------------------------           Adj R-squared =  0.3404
       Total |  7160.41429   525  13.6388844           Root MSE      =  2.9993
------------------------------------------------------------------------------
           wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
       1.female |   3.963781   1.558856     2.54   0.011     .9013238    7.026238
           educ |   .7655612   .0665232    11.51   0.000     .6348728    .8962496
          exper |   .1145122   .0144667     7.92   0.000     .0860917    .1429327
         numdep |   .4007355   .1425077     2.81   0.005     .1207713    .6806997
                |
  female#c.educ |
             1  |  -.3081501   .1051549    -2.93   0.004    -.5147327   -.1015676
                |
 female#c.exper |
             1  |  -.0993157   .0205603    -4.83   0.000    -.1397074   -.0589239
                |
female#c.numdep |
             1  |  -.5445047   .2170939    -2.51   0.012    -.9709974    -.118012
                |
          _cons |  -5.103602   1.027289    -4.97   0.000    -7.121766   -3.085438
------------------------------------------------------------------------------
```

3.3 MLR with binary independent variables: *generalizing*

We can now generalize what we have seen to MLR models with more than one categorical variable. Suppose we want to estimate a model that allows for wage differences among four groups: married men, married women, single men and single women. Let's look at the following model:

$$wage = \beta_0 + \beta_1 female + \beta_2 married + \beta_3 female * married + \beta_4 educ + \beta_5 tenure + U$$

Note: In the above regression we allow for an interaction between gender and marital status. This way we have a different intercept for each of the four groups.

- What is the predicted wage for married females in terms of the population regression function's coefficients?
- How about for single females?

Now to estimate the model we use the following commands:

```
. gen fmarried=female*married

. reg wage female married fmarried educ tenure
```

Or by typing directly:

```
. reg wage i.female##i.married educ tenure
```

In any case the result is:

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  5,   520) =   66.29
       Model |  2787.43282     5  557.486564           Prob > F      =  0.0000
    Residual |  4372.98147   520  8.40957975           R-squared     =  0.3893
-------------+------------------------------           Adj R-squared =  0.3834
       Total |  7160.41429   525  13.6388844           Root MSE      =  2.8999
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2793782   .4102435    -0.68   0.496    -1.085317    .5265601
     married |   1.947501   .3879996     5.02   0.000     1.185262    2.709741
    fmarried |  -2.391728   .5285819    -4.52   0.000    -3.430146   -1.353309
        educ |   .5262234   .0461321    11.41   0.000     .4355952    .6168516
      tenure |   .1453963   .0184716     7.87   0.000     .1091081    .1816846
       _cons |  -1.907628   .6653354    -2.87   0.004    -3.214704   -.6005525
------------------------------------------------------------------------------
```

- Interpret each of the coefficients above. Which variables are statistically significant? What is the estimated difference in wages between single men and married men? How about between single men and single women?

3.4 A Note on Dummies, Factor Variables and Interactions

In the sections above we have seen two ways to include in a regression an interaction term between a dummy (indicator) and a continuous variable and between two dummies.

One way to do it is to create the interaction variables first

```
. gen feduc=female*educ
. gen fmarried=female*married
```

and then running the appropriate regression.

The other way to do it is to use the "##" syntax within your regression command:

```
i.female##c.educ
```

```
i.female##i.married
```

The above command includes in the regression the original indicator variables and their interactions with either other indicator variables or continuous variables. When you use this syntax remember that indicator variables are prefixed with "i." whereas continuos variables are prefixed with "c.".

This notation is very useful if you have indicator variables with, say, n categories because stata creates $n-1$ dummy variables "on the fly". However, you may lose track of what you are actually doing so stick to generating the variables first if you are dealing with a small number of them.

Stata has several factor variables operators which can be used with most estimation commands to

- create indicator variables from categorical variables
- create interactions of indicators of categorical variables
- create interactions of continuous variables (polynomials)
- create interactions of categorical and continuous variables

Here are some tips on how to proceed:

- Prefix a variable with i. to specify indicators for each level (category) of the variable.
- Put # between two variables to create indicators for each interaction of (the categories of) the variables. This will not include the dummy variables on their own.
- Put # # to specify main effects for each variable and the interactions between them.
- To interact a continuous variable with a factor variable, just prefix the continuous variable with c..

For more examples and a complete description of factor variables operators in Stata, type:

```
help fvvarlist
```

4 Chow Test

- Looking for evidence of different intercepts and slopes can be thought of as testing joint significance of the dummy variable and its interactions with all other x variables. So, we could estimate the model with all interactions (unrestricted) and without interactions (restricted) to form the F-statistic. However, this approach requires generating a large number of interaction variables in order to calculate the SSR_{ur}, and this can be cumbersome when the number of interactions grows large. The Chow Test allows us to calculate the F-statistic without generating all of these interactions.
- To test whether the intercept and slopes of the model differ for male and female workers, run the restricted model for men to obtain SSR_1

  ```
  regress wage educ if female==0
  ```

  Note $\text{SSR}_1 = 4009.93077$
- Now run the restricted model for women to obtain SSR_2

  ```
  regress wage educ if female==1
  ```

  Note $\text{SSR}_2 = 1290.23913$

- Lastly, run the restricted model for both male and female workers

  ```
  regress wage educ
  ```

  Note $\text{SSR}_r = 5980.68225$

- To conduct the Chow Test, calculate the F-Statistic

$$F = \frac{[SSR_r - (SSR_1 + SSR_2)]/(k+1)}{(SSR_1 + SSR_2)/(n-2k-2)} \tag{1}$$

  Note $k = 1$ and $n = 526$. We get an F-Statistic of 33.51

- To compute the p-value write:

  ```
  display Ftail(2, 522, 33.51)
  ```

  where $k + 1 = 2$ are the numerator's degrees of freedom, $n_1 + n_2 - 2 \times (k + 1) = 522$ are the denominator's degrees of freedom, and 33.51 is the computed F-statistic.

- We get a p-value that is close to zero. We can therefore reject the null hypothesis that the slope and coefficient parameter estimates are the same for men and women.

- Alternatively, a Chow test can be conducted by introducing interaction terms and testing for their joint significance:

  ```
  . reg wage i.female##c.(educ exper)

  . testparm i.female i.female#c.*

   ( 1)  1.female = 0
   ( 2)  1.female#c.educ = 0
   ( 3)  1.female#c.exper = 0

         F(  3,   520) =   29.65
              Prob > F =    0.0000
  ```

  Note that we use `testparm` in place of `test` so that we can include certain Stata operators (e.g. * and #).