Chapter 7, Dummy Variable

1. A dummy variable takes on 1 and 0 only. The number 1 and 0 have no numerical (quantitative) meaning. The two numbers are used to represent groups. In short dummy variable is categorical (qualitative).

   (a) For instance, we may have a sample (or population) that includes both female and male. Then a dummy variable can be defined as $D = 1$ for female and $D = 0$ for male. Such a dummy variable divides the sample into two subsamples (or two sub-populations): one for female and one for male.

   (b) Dummy variable follows Bernoulli distribution. The distribution is characterized by the parameter p

$$D = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases} \tag{1}$$

2. Consider using dummy variable as regressor

$$Y = \beta_0 + \beta_1 D + u \tag{2}$$

   Regression (2) can be broken into two separate regressions as

$$Y = \begin{cases} \beta_0 + u, & \text{when } D = 0 \\ (\beta_0 + \beta_1) + u, & \text{when } D = 1 \end{cases} \tag{3}$$

   Taking expectation of (3) leads to

$$E(Y|D=0) = \beta_0 \tag{4}$$
$$E(Y|D=1) = \beta_0 + \beta_1 \tag{5}$$

   and

$$\beta_0 = E(Y|D=0) \tag{6}$$
$$\beta_1 = E(Y|D=1) - E(Y|D=0) \tag{7}$$

   Therefore β_0 is the mean of Y conditional on $D = 0$ (or mean of Y in the subpopulation with $D = 0$), β_1 is the <u>difference</u> in mean Y between the two sub-populations.

3. Sample mean is the estimate for population mean, so we have the following interpretation for the estimated coefficients in (2)

$$\hat{\beta}_0 = \bar{y}_{D=0} \tag{8}$$
$$\hat{\beta}_1 = \bar{y}_{D=1} - \bar{y}_{D=0} \tag{9}$$

where $\bar{y}_{D=0}$ denotes the average Y in the sub-sample for which $D = 0$, $\bar{y}_{D=1}$ denotes the average Y in the sub-sample for which $D = 1$. Equation (2) provides a simple way to carry out a comparison of means test (or two sample t test) between the two groups. The null hypothesis of two-sample t test says that there is no difference between two groups:

$$H_0 : \beta_1 = 0$$

This hypothesis is rejected when the p-value for $\hat{\beta}_1$ is less than 0.05.

4. For example, let Y be wage, and $D = 1$ for female, and $D = 0$ for male. Then consider the regression

$$wage = \beta_0 + \beta_1 D + u,$$

and we know $\hat{\beta}_0$ is the average wage for male, and $\hat{\beta}_1$ equals average female wage minus average male wage. The two wages are significantly different if $\hat{\beta}_1$ is significant.

5. Now consider a regression with regressor X

$$Y = \beta_0 + \beta_1 D + \beta_2 X + u \tag{10}$$

which can be rewritten as

$$Y = \begin{cases} \beta_0 + \beta_2 X + u, & \text{when } D = 0 \\ (\beta_0 + \beta_1) + \beta_2 X + u, & \text{when } D = 1 \end{cases} \tag{11}$$

It follows that

$$E(Y|X, D = 0) = \beta_0 + \beta_2 X \tag{12}$$
$$E(Y|X, D = 1) = (\beta_0 + \beta_1) + \beta_2 X \tag{13}$$
$$\beta_1 = E(Y|X, D = 1) - E(Y|X, D = 0) \tag{14}$$

so β_1 measures the change in mean Y across two groups, holding X constant (or given

the same level of X). For instance, if X is edu(cation), in the regression

$$wage = \beta_0 + \beta_1 D + \beta_2 edu + u,$$

β_1 equals the average female wage minus average male wage, given the same level of education.

6. From (11) we can show

$$\frac{dE(Y|X)}{dX} = \begin{cases} \beta_2 & \text{when } D = 0 \\ \beta_2 & \text{when } D = 1 \end{cases} \tag{15}$$

So regression (10) is restrictive by assuming that the marginal effect of X on Y does not depend on D. Go back to the wage example. This restriction assumes that when education changes, wage changes at the same rate for female and male.

7. In chapter 6 we know interaction term can be used to allow the marginal effect of X to depend on another regressor. The regression with both dummy and interaction term of dummy and X is

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 (X * D) + u \tag{16}$$

which can be rewritten as

$$Y = \begin{cases} \beta_0 + \beta_2 X + u, & \text{when } D = 0 \\ (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X + u, & \text{when } D = 1 \end{cases} \tag{17}$$

The last equation makes it clear that

Dummy variable allows for different intercepts (or intercept shift)

Interaction term of dummy variable and X allows for different slopes

see Figure 7.2 in textbook.

8. Note regression (16) contains the same amount of information as two separate regressions of Y on X, one using subsample $D = 0$, and one using subsample $D = 1$.

9. Exercise: derive the marginal effect of X on Y implied by (16)

10. Suppose we have two subsamples, one for female and one for male. We want to estimate the effect of education on wage. We have two options. Option 1 is to run two separate regressions, one for female and one for male. Option two is pool (merge) the two subsamples together and just run one regression. Which option is better?

    (a) Essentially this problem is about whether the relationship between education and wage depends on gender

    (b) To answer this question, we just pool the two subsample, and run regression (16). The point is, we need to use dummy variable and interaction term. The null hypothesis is gender does not matter, so

    $$\beta_1 = \beta_3 = 0 \tag{18}$$

    We can use F test (called Chow test in this context) for this hypothesis.

    i. If p-value is less than 0.05, H_0 is rejected, so gender matters. We need to keep the dummy and interaction term in (16). That means, running two separate regressions, one for female and one for male, is better idea.

    ii. If p-value is greater than 0.05, H_0 is not rejected, so gender does not matter. We need to drop the dummy and interaction term from (16). That means, running one regression using both subsamples is better idea.

11. What if we have information about gender and marital status? Option one is to define two dummy variables as

    $$D_1 = \begin{cases} 1, & \text{female} \\ 0, & \text{male} \end{cases} \tag{19}$$

    $$D_2 = \begin{cases} 1, & \text{married} \\ 0, & \text{unmarried} \end{cases} \tag{20}$$

    and use them to run the regression of

    $$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + u \tag{21}$$

For this regression we can show

$$E(Y) = \begin{cases} \beta_0, & \text{if } D_1 = 0, D_2 = 0 \\ \beta_0 + \beta_1, & \text{if } D_1 = 1, D_2 = 0 \\ \beta_0 + \beta_2, & \text{if } D_1 = 0, D_2 = 1 \\ \beta_0 + \beta_1 + \beta_2, & \text{if } D_1 = 1, D_2 = 1 \end{cases}$$

Now we can see regression (22) is restrictive because it assumes

$$E(Y|D_1 = 1, D_2 = 1) - E(Y|D_1 = 1, D_2 = 0) = E(Y|D_1 = 0, D_2 = 1) - E(Y|D_1 = 0, D_2 = 0), \tag{22}$$

In words, when $D2$ changes from 0 to 1, the change in mean Y does not depend on $D1$. This is a kind of no-interaction restriction. Let Y be wage. Then no-interaction restriction says that when a person changes his/her marital status, the change in wage does not depend on the gender of the person.

12. In order to relax the no-interaction restriction, we can define four dummy variables (because we have four groups of people) as

$$E_1 = \begin{cases} 1, & \text{female and married} \\ 0, & \text{otherwise} \end{cases}$$

$$E_2 = \begin{cases} 1, & \text{female and unmarried} \\ 0, & \text{otherwise} \end{cases}$$

$$E_3 = \begin{cases} 1, & \text{male and married} \\ 0, & \text{otherwise} \end{cases}$$

$$E_4 = \begin{cases} 1, & \text{male and unmarried} \\ 0, & \text{otherwise} \end{cases}$$

and run a regression using only <u>three</u> of them

$$Y = \beta_0 + \beta_1 E_1 + \beta_2 E_2 + \beta_3 E_3 + u \tag{23}$$

If we use all four dummies, then $E_1 + E_2 + E_3 + E_4 = 1$ so is perfectly correlated with the intercept term. This situation is called <u>dummy variable trap</u>. In order to avoid dummy variable trap, we leave out one dummy.

13. Exercise: Please show regression (23) does not impose no-interaction restriction.

14. Consider a special variable

$$X = \begin{cases} 1, & \text{using bus} \\ 2, & \text{using subway} \\ 3, & \text{driving car} \end{cases} \tag{24}$$

Note that X has no numerical meaning, so is qualitative. Numbers 1, 2 and 3 are used here to define three categories. Number 2 does not mean it is twice of 1. Because the variable is qualitative, we need to translate it into a set of dummy variables

$$F_1 = \begin{cases} 1, & \text{using bus} \\ 0, & \text{otherwise} \end{cases}$$

$$F_2 = \begin{cases} 1, & \text{using subway} \\ 0, & \text{otherwise} \end{cases}$$

$$F_3 = \begin{cases} 1, & \text{driving car} \\ 0, & \text{otherwise} \end{cases}$$

When running regression, we do not use X (since it has no numerical meaning). Instead we use two of the three dummy variables defined above.

15. The same idea can be applied to an ordinal variable such as

$$X = \begin{cases} 3, & \text{exceeds expecation} \\ 2, & \text{meets expecation} \\ 1, & \text{fails expecation} \end{cases} \tag{25}$$

For ordinal variable we only know ranking. The number has no numerical meaning. Actually we can replace number 3 with any number greater than 2 (to maintain the ordering). Because ordinal variable is qualitative, we need to translate it into a set of dummy variables. We cannot directly use ordinal variable in regression.

Example: Chapter 7

1. We use the data file 311_wage1.dta, downloadable at my webpage. See example 7.1 in textbook for detail.

2. We see for the first observation, wage = 3.1, educ = 11, female = 1 (so is female), and married = 0 (so is unmarried). Female and married are both dummy variables, for which the values 1 and 0 have no quantitative meaning.

3. Command `tab` is used to tabulate proportion (probability) for dummy variable. In this case 52.09 percent observations are male (female=0), and 47.91 percent are female.

4. Next we run regression (2), i.e., regress wage on dummy variable female. The estimated intercept $\hat{\beta}_0 = \bar{y}_{D=0} = 7.099489$ is the average wage for male. The estimated slope $\hat{\beta}_1 = \bar{y}_{D=1} - \bar{y}_{D=0} = -2.51183$ is average female wage minus average male wage. In this example female earns less than male since $\hat{\beta}_1$ is negative. The p-value for $\hat{\beta}_1$ is less than 0.05, so we reject the null hypothesis that female wage equals male wage. In other words, the two wages differ significantly.

5. Alternatively we can summarize wage separately for female and male. The command is

```
 sort female
by female: sum wage
```

   On average a male earns 7.099489, and a female earns 4.587659. The difference is $4.587659 - 7.099489 = -2.51183$, which is the same as $\hat{\beta}_1$ reported by regression (2). This finding confirms that

Regressing Y on dummy variable carries out the two sample t test.

6. Next we run regression (16) using X = educ:

$$wage = \beta_0 + \beta_1 female + \beta_2 educ + \beta_3 (educ * female) + u$$

   (a) The estimated intercept is $\hat{\beta}_0 = .2004963$. It measures the average male wage when $educ = 0$.

(b) $\hat{\beta}_1 = -1.198523$. It measures the average female wage when $educ = 0$ minus average male wage when $educ = 0$. In other words, when educ = 0, a female earns $.2004963 + (-1.198523) = -.9980267$. This number is not very meaningful since in this sample no female has zero education (two males have zero educ, and you can see them using command `list if educ==0`).

(c) $\hat{\beta}_2 = .539476$. So male wage rises by .539476 when his educ rises by 1 unit.

(d) $\hat{\beta}_3 = -.085999$. So female wage rises by $.539476 + (-.085999) = .453477$ when her educ rises by 1 unit.

(e) The null hypothesis that the relationship between wage and educ does not depend on gender (or there is NO difference in regression functions across female and male) can be formulated as

$$H_0 : \beta_1 = \beta_3 = 0.$$

The F test for difference in regression functions across groups is called Chow test

The stata command to conduct Chow test is `test female fe`. It is shown that $F = 33.51$, p-value < 0.05. So we reject the null hypothesis. That means there IS difference in regression functions across female and male. In other words, the relationship between wage and educ depends on gender.

(f) Note that $\hat{\beta}_1$ and $\hat{\beta}_3$ are individually insignificant (the p-values are 0.366 and 0.407, respectively), whereas the Chow test indicates that they are jointly significant. The lesson is, just focusing on individual coefficient can be misleading.

7. Because the relationship between wage and educ depends on gender, we can run two separate (group-wise) regressions, one using female and one using male. The stata command is `by female: reg wage educ`. We see the coefficients in the male regression are the same as $\hat{\beta}_0$ and $\hat{\beta}_2$ reported by the pooled regression (16). The female results can also be derived based on the pooled regression (16). In other words,

Regressing on dummy and interaction terms is as informative as groupwise regressions

The pooled regression (16) has one big advantage over groupwise regressions: we can run Chow test based on (16).

8. Finally you are shown how to define a set of dummy variables to represent multiple categories of gender and marital status. In theory we should define four dummies since

there are four groups. But, aware of dummy variable trap, we only define three. The group for which we do not define dummy is base group. In this example, the base group is unmarried male. The three dummy variable sare

$$\begin{aligned} D1 &= 1 \texttt{ for married male} \\ D2 &= 1 \texttt{ for unmarried female} \\ D3 &= 1 \texttt{ for married female} \end{aligned}$$

Consider the regression of

$$wage = \beta_0 + \beta_1 D1 + \beta_2 D2 + \beta_3 D3 + u$$

To facilitate interpreting coefficients, let break down the above regression to

$$Ewage = \begin{cases} \beta_0, & \text{when } D1 = D2 = D3 = 0 \\ \beta_0 + \beta_1, & \text{when } D1 = 1 \\ \beta_0 + \beta_2, & \text{when } D2 = 1 \\ \beta_0 + \beta_3, & \text{when } D3 = 1 \end{cases}$$

The interpretations of coefficients are

(a) $\hat{\beta}_0 = 5.168023$. It measures the average wage for unmarried male, the base group.

(b) $\hat{\beta}_1 = 2.815009$, So a married male earns 2.815009 more than an unmarried male. (So marriage enhances a male's market value)

(c) $\hat{\beta}_2 = -.5564399$, So an unmarried female earns .5564399 less than an unmarried male. (So there is is discrimination against female)

(d) $\hat{\beta}_3 = -.6021142$, So a married female earns .6021142 less than an unmarried male.

(e) Because $\hat{\beta}_3 - \hat{\beta}_2 = -.6021142 - (-.5564399) < 0$, marriage decreases a female's market value.

9. Exercise: Show that a female is discriminated more when she is married than when she is unmarried. Hint: compute $\hat{\beta}_3 - \hat{\beta}_1$

```
     +---------------------------------+
     | wage   educ   female   married  |
     |---------------------------------|
  1. |  3.1     11        1         0  |
  2. |  3.2     12        1         1  |
  3. |    3     11        0         0  |
  4. |    6      8        0         1  |
  5. |  5.3     12        0         1  |
     +---------------------------------+

. tab female

     =1 if |
    female |      Freq.     Percent        Cum.
-----------+-----------------------------------
         0 |        274       52.09       52.09
         1 |        252       47.91      100.00
-----------+-----------------------------------
     Total |        526      100.00

. 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
------------------------------------------------------------------------------

. by 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
```

```
. reg wage female educ fe

      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
          fe |   -.085999   .1036388    -0.83   0.407    -.2895994    .1176014
       _cons |   .2004963   .8435616     0.24   0.812    -1.456696    1.857689
------------------------------------------------------------------------------
. * chow test
. test female fe
 ( 1)  female = 0
 ( 2)  fe = 0
       F(  2,   522) =   33.51
            Prob > F =    0.0000
. by female: reg wage educ
-> female = 0

      Source |       SS       df       MS              Number of obs =     274
-------------+------------------------------           F(  1,   272) =   48.60
       Model |  716.445945     1  716.445945            Prob > F      =  0.0000
    Residual |  4009.93077   272  14.7423925            R-squared     =  0.1516
-------------+------------------------------           Adj R-squared =  0.1485
       Total |  4726.37672   273  17.3127352            Root MSE      =  3.8396

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .539476   .0773864     6.97   0.000     .3871237    .6918284
       _cons |   .2004963   1.016462     0.20   0.844    -1.800637    2.201629
------------------------------------------------------------------------------
-> female = 1

      Source |       SS       df       MS              Number of obs =     252
-------------+------------------------------           F(  1,   250) =   61.15
       Model |  315.577975     1  315.577975            Prob > F      =  0.0000
    Residual |  1290.23913   250  5.16095652            R-squared     =  0.1965
-------------+------------------------------           Adj R-squared =  0.1933
       Total |  1605.81711   251  6.39767771            Root MSE      =  2.2718

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .453477   .0579919     7.82   0.000     .3392622    .5676919
       _cons |  -.9980266   .7285069    -1.37   0.172     -2.43282    .4367665
```

```
. * multiple category
. gen d1 = 0

. replace d1 = 1 if female == 0 & married ==1
(188 real changes made)

. gen d2 = 0

. replace d2 = 1 if female == 1 & married ==0
(120 real changes made)

. gen d3 = 0

. replace d3 = 1 if female == 1 & married ==1
(132 real changes made)

. reg wage d1 d2 d3

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  3,   522) =   38.45
       Model |  1295.94158     3  431.980528            Prob > F      =  0.0000
    Residual |  5864.47271   522   11.234622            R-squared     =  0.1810
-------------+------------------------------           Adj R-squared =  0.1763
       Total |  7160.41429   525  13.6388844            Root MSE      =  3.3518

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          d1 |    2.815009   .4363413     6.45   0.000     1.957808    3.672209
          d2 |   -.5564399   .4735578    -1.18   0.241    -1.486753    .3738733
          d3 |   -.6021142   .4644846    -1.30   0.195    -1.514603    .3103745
       _cons |    5.168023   .3614348    14.30   0.000     4.457978    5.878069
------------------------------------------------------------------------------
```

Do File

```
* Do file for dummy variable (chapter 7)
set more off
clear
capture log close
cd "I:\311"
log using 311log.txt, text replace
use 311_wage1.dta, clear
* show first 5 observations
list wage educ female married in 1/5
* tabulate female
tab female
* run regression using dummy
reg wage female
* compare the means for male and female
sort female
by female: sum wage
* run regression using dummy and interaction term
gen fe = female*educ
reg wage female educ fe
* chow test
test female fe
* run separate regressions for male and female
by female: reg wage educ
* multiple category
gen d1 = 0
replace d1 = 1 if female == 0 & married ==1
gen d2 = 0
replace d2 = 1 if female == 1 & married ==0
gen d3 = 0
replace d3 = 1 if female == 1 & married ==1
reg wage d1 d2 d3
log close
```