Week 6. Oaxaca-Blinder Decomposition

What is Oaxaca-Blinder Decomposition?

1. Mean difference between two groups (group 1 and group 2): $\bar{Y}_1$ and $\bar{Y}_2$
2. It is a technique to decompose the mean difference ($= \bar{Y}_1 - \bar{Y}_2$) into two components:
   (a) Coefficient effect
   = the rate effect,
   = the unexplained difference,
   which quantifies how much the mean difference between two groups can be attributable to the different treatment on the same characteristics. This effect is often regarded as "discrimination." However, I strongly recommend you not to call it "discrimination."
   (b) Endowment effect
   = the distributional effect,
   = the explained difference,
   which quantifies how much the mean difference between two groups can be attributable to the different distributions between two groups.
3. For example, the mean earnings gap between whites and blacks can be attributed to (1) the difference in the proportion of highly educated workers between two races, and (2) the differentiated rewards to the same level of education between races.
4. It is always about the difference between "two" groups.
5. Detailed decomposition is possible but you should be cautious. There are identification problems.

Statistical Explanation

Suppose you estimated the following model. One for white and the other for minority.

$$Y_i = a_i + \sum_J \sum_K b_{ijk} X_{ijk} + \epsilon \tag{1}$$

We can predict mean wages for groups w and m based on estimations from two regression models. The mean wages of the groups at time t can be expressed as follows:

$$\bar{Y}_i = a_{it} + \sum_J \sum_K b_{ijk} \bar{X}_{ijk} \tag{2}$$

where i indexes groups w (say, White) and m (say, minority) and t indexes time points. Y_{it} refers to the mean wage for group i. The a term is an estimated constant, and b represents the estimated coefficients. $\bar{X}$ refers to the mean values of the explanatory variables. Since the error term for a mean value is zero by the definition of OLS regression, I drop it from the equation. In case of b and $\bar{X}$, j indexes a jth category containing a set of relevant dummy variables and k indexes a kth dummy variable in each j. For simplicity, I drop j and k subscripts henceforth in the remaining part. The wage gap between group w and group m, $(Y_w - Y_m)$, can be calculated by substituting

the estimates for the right side of Equation 1, $([a_w + \sum\sum b_w \bar{X}_w] - [a_m + \sum\sum b_m \bar{X}_m])$. A Blinder-Oaxaca method breaks down this wage gap as follows:

$$\bar{Y}_w - \bar{Y}_m = \underbrace{[a_w - a_m]}_{\text{A1. Intercept Effect}} + \underbrace{\sum\sum \left[(b_w - b_m) \left(\frac{\bar{X}_w + \bar{X}_m}{2} \right) \right]}_{\text{A2. Coefficient Effect}} + \underbrace{\sum\sum \left[(\bar{X}_w - \bar{X}_m) \left(\frac{b_w + b_m}{2} \right) \right]}_{\text{A3. Endowment Effect}} \quad (3)$$

Note that instead of (a) $\left(\frac{\bar{X}_w + \bar{X}_m}{2}\right)$ and $\left(\frac{b_w + b_m}{2}\right)$, you can use (b) $\bar{X}_w$ and b_m, or (c) $\bar{X}_m$ and b_w as shown below. In most cases, your result will be only slightly different by the choice of (a), (b), or (c). But sometimes the difference is substantial. Sensitivity test is usually recommended.

$$\bar{Y}_w - \bar{Y}_m = \underbrace{[a_w - a_m]}_{\text{A1. Intercept Effect}} + \underbrace{\sum\sum (b_w - b_m)\bar{X}_w}_{\text{A2. Coefficient Effect}} + \underbrace{\sum\sum (\bar{X}_w - \bar{X}_m) b_m}_{\text{A3. Endowment Effect}} \quad (4)$$

From this result you report the size of two effects:

1. Coefficient Effect = A1 + A2
2. Endowment Effect = A3

You can compute this with Excel. Fortunately there is a Stata command.

Stata command

1. The personal earnings gap between whites and blacks is .2665 log dollars.

```
. tabstat lnpincome, by(black)

   black |      mean
---------+----------
       0 |   10.5726
       1 |  10.30613
---------+----------
   Total |  10.54621
--------------------
```

2. OLS Regression without Covariates

   If you estimate an OLS model without any control variables except `female`, you get the following result. Here is the coefficient of black is the mean difference in earnings between whites and blacks.

```
. reg lnpincome black

------------------------------------------------------------------------------
   lnpincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -.2664713   .0277216    -9.61   0.000    -.3208088   -.2121337
       _cons |    10.5726    .008724  1211.90   0.000      10.5555     10.5897
------------------------------------------------------------------------------
```

3. OLS Regression with Covariates

If you estimate an OLS model without an interaction, you get the following result. The coefficient of black becomes substantially reduced (= -.1056) compared to the basic model (-.2665). Usually we say that the control variables account for 60% of racial earnings gap (= (-.26647+.10558)/-.26647).

```
. reg lnpincome black female i.edu age married

      Source |       SS           df       MS      Number of obs   =    15,348
-------------+----------------------------------   F(8, 15339)     =    503.24
       Model |  3377.83676         8  422.229595   Prob > F        =    0.0000
    Residual |  12869.7402    15,339  .839020813   R-squared       =    0.2079
-------------+----------------------------------   Adj R-squared   =    0.2075
       Total |   16247.577    15,347  1.05868098   Root MSE        =    .91598

------------------------------------------------------------------------------
   lnpincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -.1018714   .0254998    -3.99   0.000    -.1518541   -.0518888
      female |    -.65128   .0148911   -43.74   0.000    -.6804683   -.6220916
             |
         edu |
           2 |   .2672748    .042582     6.28   0.000     .1838089    .3507406
           3 |   .4894489   .0416642    11.75   0.000     .4077821    .5711157
           4 |   .8708241   .0420922    20.69   0.000     .7883185    .9533297
           5 |   1.237103   .0435773    28.39   0.000     1.151686    1.322519
             |
         age |   .0103045   .0026002     3.96   0.000     .0052077    .0154012
     married |  -.0213489   .0222227    -0.96   0.337     -.064908    .0222102
       _cons |   9.820821   .1121696    87.55   0.000     9.600956    10.04069
------------------------------------------------------------------------------
```

4. Simple Decomposition

   Below is the Oaxaca-Blinder decomposition results. `Explained` is the endowment effect and `Unexplained` is the coefficient effect.

```
. oaxaca lnpincome female edu2 edu3 edu4 edu5 age married, by(black) w(1)

Blinder-Oaxaca decomposition                    Number of obs     =     15,348

           1: black = 0
           2: black = 1

------------------------------------------------------------------------------
    lnpincome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
Differential  |
 Prediction_1 |    10.5726   .0088174  1199.06   0.000     10.55532    10.58989
 Prediction_2 |   10.30613   .0236947   434.96   0.000     10.25969    10.35257
   Difference |   .2664713   .0252821    10.54   0.000     .2169193    .3160232
--------------+---------------------------------------------------------------
Decomposition |
    Explained |    .169037   .0142391    11.87   0.000     .1411288    .1969451
  Unexplained |   .0974343   .0247042     3.94   0.000     .0490151    .1458536
------------------------------------------------------------------------------
```

5. Detailed Decomposition 1
(ref = edu1(LTHS) for a set of education dummy variables)

Below is the detailed decomposition results. The components of `Explained` and `Unexplained` are decomposed into each variable.

```
. oaxaca lnpincome female edu2 edu3 edu4 edu5 age married, by(black) w(1) detail(edu:edu*)

Blinder-Oaxaca decomposition                    Number of obs     =     15,348

           1: black = 0
           2: black = 1

------------------------------------------------------------------------------
   lnpincome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Differential |
Prediction_1 |    10.5726   .0088174  1199.06   0.000     10.55532    10.58989
Prediction_2 |   10.30613   .0236947   434.96   0.000     10.25969    10.35257
  Difference |   .2664713   .0252821    10.54   0.000     .2169193    .3160232
-------------+----------------------------------------------------------------
Explained    |
      female |   .0650001   .0093827     6.93   0.000     .0466105    .0833898
         edu |   .1067261   .0092775    11.50   0.000     .0885425    .1249098
         age |   .0014545   .0008633     1.68   0.092    -.0002376    .0031466
     married |  -.0041438    .006112    -0.68   0.498     -.016123    .0078354
       Total |    .169037   .0142391    11.87   0.000     .1411288    .1969451
-------------+----------------------------------------------------------------
Unexplained  |
      female |  -.2251413   .0271768    -8.28   0.000    -.2784069   -.1718758
         edu |  -.0098657   .1011814    -0.10   0.922    -.2081776    .1884462
         age |   -.199828   .3225986    -0.62   0.536    -.8321096    .4324537
     married |   -.049026   .0347166    -1.41   0.158    -.1170693    .0190173
       _cons |   .5812953   .3414981     1.70   0.089    -.0880287    1.250619
       Total |   .0974343   .0247042     3.94   0.000     .0490151    .1458536
------------------------------------------------------------------------------
edu: edu2 edu3 edu4 edu5
```

6. Detailed Decomposition 2
   (ref = edu4(BA) for a set of education dummy variables)

   As you see the result of Detailed Decompositions 1 and 2 are different. The unexplained effect of education was -.0099 in Decomposition 1 but it now became .0457222 in Decomposition 2. The unexplained effect of the constant has also been changed between two models. This is the identification problem of the detailed Oaxaca-Blinder Decomposition. As you change the reference group, the effect size of the unexplained part also changes.

```
. oaxaca lnpincome female edu1 edu2 edu3 edu5 age married, by(black) w(1) detail(edu:edu*)

Blinder-Oaxaca decomposition                    Number of obs     =     15,348

           1: black = 0
           2: black = 1

------------------------------------------------------------------------------
   lnpincome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Differential |
Prediction_1 |    10.5726   .0088174  1199.06   0.000     10.55532    10.58989
Prediction_2 |   10.30613   .0236947   434.96   0.000     10.25969    10.35257
  Difference |   .2664713   .0252821    10.54   0.000     .2169193    .3160232
-------------+----------------------------------------------------------------
Explained    |
      female |   .0650001   .0093827     6.93   0.000     .0466105    .0833898
         edu |   .1067261   .0092775    11.50   0.000     .0885425    .1249098
         age |   .0014545   .0008633     1.68   0.092    -.0002376    .0031466
     married |  -.0041438    .006112    -0.68   0.498     -.016123    .0078354
       Total |    .169037   .0142391    11.87   0.000     .1411288    .1969451
-------------+----------------------------------------------------------------
Unexplained  |
      female |  -.2251413   .0271768    -8.28   0.000    -.2784069   -.1718758
         edu |   .0457222   .0480695     0.95   0.342    -.0484923    .1399367
         age |   -.199828   .3225986    -0.62   0.536    -.8321096    .4324537
     married |   -.049026   .0347166    -1.41   0.158    -.1170693    .0190173
       _cons |   .5257074    .330021     1.59   0.111    -.1211219    1.172537
       Total |   .0974343   .0247042     3.94   0.000     .0490151    .1458536
------------------------------------------------------------------------------
edu: edu1 edu2 edu3 edu5
```

Identification Problem

See the PDF slides and read Kim's (2013) *Sociological Methodology* paper.

Decomposition of the change over time between two groups

See the PDF slides and read Kim's (2010) *Sociological Methods and Research* paper.