Group Comparisons:
Using “What If” Scenarios to Decompose Differences Across Groups

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/

Last revised February 15, 2015

We saw that the effects of education and job experience were smaller for blacks than for whites. However, we also saw that blacks had lower levels of education and less job experience than whites. Both of these contribute to the differences in income between whites and blacks. How can we disentangle the relative importance of these differences?

One way to address this is via a “what if” question: Suppose that blacks had as much education and job experience as whites, but the effects of education and job experience were the same for blacks as they are now. What would the gap be between whites and blacks then? In other words, if you control for compositional differences on the independent variables, how much difference remains between whites and blacks on the dependent variable? I’ll show you a variety of approaches.

First, remember that in the real world (the real world we made up in this example, anyway) the difference between the average black and white income is $11,250. Further, on average whites also have 3.7 more years of education and 2.9 more years of job experience than do blacks.

```
. use https://www3.nd.edu/~rwilliam/statafiles/blwh.dta, clear
. tabstat income educ jobexp, by(black) columns(variables)

Summary statistics: mean
  by categories of: black

black |    income      educ    jobexp
------+------------------------------
white |     30.04      13.9      14.1
black |     18.79      10.2      11.2
------+------------------------------
Total |     27.79     13.16     13.52
-------------------------------------
```

Remember too that the effects of education and job experience are greater for whites (for variety I’ll use Stata’s `estimates` command to make side by side comparisons easier):

```
. quietly reg income educ jobexp if black==1
. estimates store blackmodel
. quietly reg income educ jobexp if black==0
. estimates store whitemodel
. quietly reg income educ jobexp
. estimates store bothmodel
. estimates table blackmodel whitemodel bothmodel

-----------------------------------------------------
    Variable | blackmodel   whitemodel   bothmodel   
-------------+---------------------------------------
        educ |  1.6779491    1.8933377    1.9451204  
      jobexp |    .421975    .72225495    .70822118  
       _cons | -3.0512005   -6.4611885   -7.3829348  
-----------------------------------------------------
```

To find out how much difference there would be in our hypothetical world where blacks have the same average levels of education and job experience as whites, while education and job experience continue to have the same effect on blacks as they do now, we can now use Stata's `margins` command.

```
. quietly reg income educ jobexp if black == 1
. margins, at(educ = 13.9 jobexp = 14.1)

Adjusted predictions                              Number of obs   =        100
Model VCE    : OLS

Expression   : Linear prediction, predict()
at           : educ            =        13.9
               jobexp          =        14.1

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   26.22214   .4666424    56.19   0.000     25.29598    27.14829
------------------------------------------------------------------------------
```

Here is what we did. We estimated the model for blacks only. Then, with the `margins` command, we asked Stata to compute the predicted income for a black who had the white average values for educ (13.9) and jobexp (14.1). The predicted value was $26, 222.14. While this is much more than the average black makes ($18,790) it is still much less than the average white makes ($30,040).

Put another way, `margins` computed the following, using the coefficient values for blacks (i.e. b_{educ} = 1.6779491, b_{jobexp} = .421975, constant = -3.0512005) and the variable means for whites (mean of educ = 13.9, mean of educ = 14.1):

$$E(Y) = -3.0512005 + (1.6779491 * 13.9) + (.421975 * 14.1) = \mathbf{26.22214}$$

Interpretation. In the "real" world, there is a difference of $11,250 in black and white income. Blacks average $18,790, whites average $30,040 dollars. HOWEVER, if blacks had the same average levels of education (13.9 years) and job experience (14.1 years) as whites, while the effects of education and job experience stayed the same for them as they currently are, their average income would increase to $26,222.14. This would be an increase of $26,222.14 - $18,790 = $7,432.14. However, they would still trail whites by $30,040 – $26,222.14 = $3,817.86.

So, of the original difference between black and white income, 66% (7432.14/11250) of the difference is due to blacks having lower levels of education and job experience than do whites; 34% (3817.86/11250) is due to the effects of education and job experience being different for whites than they are for blacks.

Hence, compositional differences can account for about 2/3 of the income differences between blacks and whites; but a third of the difference still remains. One possible explanation might be that there is discrimination against blacks, but there are several other possible explanations as well, e.g. the quality of education and job experience may differ by race, or there could be other variables not currently included in the model that account for these differences.

Some other things to note. In this particular example, both compositional differences and differences in effects worked against blacks. There is nothing that says this always has to happen. For example, the effects of education and job experience could have been greater for blacks than whites; but if blacks were far behind in terms of years of education and job experience, they could still have lower levels of income than whites.

Also, how you phrase your "what if" question will affect your results. For example, you could have asked, what if whites had the same levels of education and job experience as blacks, while education and job experience continued to have the same effects on whites as they do now? Doing it that way would produce different numbers. To me, it is more intuitively appealing to "raise up" the group that is behind than it is to "bring down" the group that is ahead – but the main thing is to be clear as to how you are doing things. That is, I would usually use the coefficients for the disadvantaged group (in this case blacks) with the means for the advantaged group (in this case whites).

While I now prefer the above approach, Appendix A will demonstrate a few other alternatives. This is because

- If you don't understand one approach, a different approach may make things clearer to you
- These other approaches have been used in the past so you may encounter them being used on old exams
- In the above example, the means were exactly correct to 1 decimal place. Sometimes it may take several decimal places to report the value of a mean. Usually, rounding to a few decimal places will be sufficient. But, the most precise possible estimates can be obtained using methods outlined in Appendix A, which include some minor variations of the above `margins` command
- If there are several independent variables in the model, the methods in Appendix A may be a little easier to use since you don't have to specify the mean for each variable separately.

Appendix B presents a different detailed example using these techniques.

Appendix A: Alternative Approaches

In case the first approach isn’t clear, here are several other ways to do it.

Predict command. Regress income on education and job experience for blacks only. Then use the `predict` command, but select whites only:

```
. quietly reg income educ jobexp if black==1
. predict whcompblcoef if black==0
(option xb assumed; fitted values)
(100 missing values generated)
```

In other words, you estimate the regression using one group, blacks, but then compute the predicted values using the other group, whites. In effect, what this does is give us predicted values of income for a hypothetical group that has the same levels of education and job experience as whites, where the effects of education and job experience are the same as they currently are for blacks.

We then get the mean of this new variable:

```
. sum whcompblcoef

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
whcompblcoef |       400    26.22214    5.791108   19.23387   39.35931
```

Adjust command. The `adjust` command has largely been replaced by `margins`, but it still works. Again, you estimate the model for blacks, and then plug in the mean values for whites.

```
. quietly reg income educ jobexp if black == 1
. adjust educ jobexp if black == 0

-------------------------------------------------------------------------------------
     Dependent variable: income     Command: regress
 Covariates set to mean: educ = 13.9, jobexp = 14.1
-------------------------------------------------------------------------------------

----------------------
      All |         xb
----------+-----------
          |    26.2221
----------------------
     Key:  xb  =  Linear Prediction
```

Margins command using precise values for the means: In this example the means were exactly reported to 1 decimal place and hence were easy to specify in the `margins` command. You’ll probably be fine if you use values rounded to a few decimal places. But if, say, the means went to several decimal places, you can be super precise by doing something like the following.

Basically we compute scalars that are equal to the means (to several decimal places anyway) and we then use those scalar values in the `margins` command.

```
. * More precise approach
. sum educ if black == 0, meanonly
. scalar whitemeaneduc = r(mean)
. sum jobexp if black == 0, meanonly
. scalar whitemeanjobexp = r(mean)
. scalar list
whitemeanjobexp =       14.1
whitemeaneduc =       13.9

. quietly reg income educ jobexp if black == 1
. margins, at(educ = `=whitemeaneduc' jobexp = `=whitemeanjobexp')

Adjusted predictions                              Number of obs   =        100
Model VCE    : OLS

Expression   : Linear prediction, predict()
at           : educ            =        13.9
               jobexp          =        14.1

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   26.22214   .4666424    56.19   0.000     25.29598    27.14829
------------------------------------------------------------------------------
```

Margins command using atmeans. Or, we can use the `margins` command, this time using the `atmeans` option. This would also be more precise if the means needed to be calculated to several decimal places, and be less tedious if there were several independent variables. As before, we estimate a model for blacks only, and then use the means for whites. By default, `margins` only does calculations using cases that were used in the estimation command. The `noesample` option overrides that.

```
. quietly reg income educ jobexp if black == 1
. margins if black == 0, noesample atmeans

Adjusted predictions                              Number of obs   =        400
Model VCE    : OLS

Expression   : Linear prediction, predict()
at           : educ            =        13.9 (mean)
               jobexp          =        14.1 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   26.22214   .4666424    56.19   0.000     25.29598    27.14829
------------------------------------------------------------------------------
```

Appendix B: A Blast from the Past

Here is a problem I use when I teach Grad Stats I. This problem addresses the same sorts of issues we are dealing with here. Men make more than women. However, men are also more likely to be in higher-paying occupations than are women. To what extent is difference in pay due to compositional differences in occupation distribution, and to what extent is it due to the fact that the effect of occupation on pay differs between men and women? We'll repeat how I did it originally, and then show how we can solve the problem using our current framework.

A researcher is doing a study of gender discrimination in the American labor force. She has come up with a 3-part classification of occupations (Occupation 1, Occupation 2, and Occupation 3) and a 2-part classification for wages ("good" and "bad"). She finds that, by gender, the distribution of occupation and wages is as follows:

	Women			Men		
Pay/Occ	Occ 1	Occ 2	Occ 3	Occ 1	Occ 2	Occ 3
Good Pay	20%	7%	10%	7%	10%	60%
Bad Pay	50%	8%	5%	8%	5%	10%

From the table, it is immediately apparent that 37% of all women receive good pay, compared to 77% of the men. At the same time, it is also very clear that the types of occupations are very different for men and women. For women, 70% are in occupation 1, which pays poorly, while 70% of men are in occupation 3, which pays very well. Therefore, the researcher wants to know whether differences in the types of occupations held by men and women account for the wage differential between them. How can she address this question?

SOLUTION. This problem is best addressed by asking a "what if" sort of question: Suppose women were distributed across occupations the same way men were, but within each occupation had the same wage structure that they do now. If differences in types of occupations alone account for the wage discrepancies, then this approach should control for those differences and wage differentials should disappear.

We will use the following terms:

Event A = Receives Good pay, $\overline{A}$ = Bad pay, E_i = Employed in occupation i.

Given these definitions, this problem requires that we combine the occupational distribution for men $(P(E_i))^M$ with the conditional probabilities that a woman receives good wages given the occupation she is in $(P(A \mid E_i))^F$

For men $P(E_1)^M = .15$, $P(E_2)^M = .15$, $P(E_3)^M = .70$

For women $P(A \mid E_1)^W = 2/7$, $P(A \mid E_2)^W = 7/15$, $P(A \mid E_3)^W = 10/15$

Using the marginal probability theorem, we get

$$P(A) = \sum P(E_i)^M P(A|E_i)^W = (.15 * \tfrac{2}{7}) + (.15 * \tfrac{7}{15}) + (.70 * \tfrac{10}{15}) = 0.5795238$$

Hence, if women had the same occupational distribution while continuing to make the same salaries within occupations that they do now, 58% of women would make good wages. This is much more than the 37% of women who currently make good wages, but still well short of the male figure of 77%. Differences in occupational structure account for much of the difference between men and women, but not all.

Alternative Solution. Let goodpay = 1 if pay is good, 0 otherwise; female = 1 if female, 0 if male; occ2 = 1 if in occupation 2, 0 otherwise; occ3 = 1 if in occupation 3, 0 otherwise. You then get

```
. clear all
. use https://www3.nd.edu/~rwilliam/statafiles/goodpay.dta, clear
. tabstat goodpay occ1 occ2 occ3, by(female) columns(variables)

Summary statistics: mean
  by categories of: female

female |   goodpay      occ1      occ2      occ3
-------+----------------------------------------
  Male |       .77       .15       .15        .7
Female |       .37        .7       .15       .15
-------+----------------------------------------
 Total |       .57      .425       .15      .425
------------------------------------------------
```

This shows us that 77% of the men and 37% of the women receive good pay, a 40% difference. It also shows the sharp differences in occupational distribution by gender. We next run a regression model for women only:

```
. reg  goodpay occ2 occ3 if female == 1

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =    4.45
       Model |  1.95761905     2  .978809524           Prob > F      =  0.0142
    Residual |   21.352381    97  .220127639           R-squared     =  0.0840
-------------+------------------------------           Adj R-squared =  0.0651
       Total |       23.31    99  .235454545           Root MSE      =  .46918

------------------------------------------------------------------------------
     goodpay |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        occ2 |   .1809524    .133491     1.36   0.178    -.0839904    .4458951
        occ3 |   .3809524    .133491     2.85   0.005     .1160096    .6458951
       _cons |   .2857143   .0560775     5.09   0.000     .1744161    .3970125
------------------------------------------------------------------------------
```

We regress goodpay on the two occupation dummies, for women only. This shows how occupation is related to women’s pay. The constant shows us that 28.57% of women in occupation 1 receive good pay. For occupation 2, 46.66% (.2857143 + .1809524) receive good pay; for occupation 3, 66.66% (.2857143 + .3809524) receive good pay. This is consistent with the original table. (As we’ll see, when the dependent variable is a dichotomy, logistic regression is preferable to OLS regression; but in this particular problem, where all the variables are dummy variables, both OLS regression and logistic regression will work equally well.)

We can now use the `margins` command. For occ2 and occ3, we plug in the mean values for men, i.e. 15% of men are in occupation 2 while 70% of men are in occupation 3:

```
. margins, at(occ2 = .15 occ3 = .7)

Adjusted predictions                              Number of obs   =        100
Model VCE    : OLS

Expression   : Linear prediction, predict()
at           : occ2            =         .15
               occ3            =          .7

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .5795238   .0871308     6.65   0.000     .4065932    .7524544
------------------------------------------------------------------------------
```

What this shows us is that, if women had the same occupational distribution as men while still receiving the same wages within occupations as they do now, 57.95% of women would receive good pay. Hence, about half the 40 point gap in men’s and women’s pay is due to differences in occupational structure, but about half is due to differences in pay within occupations. This is the exact same conclusion reached before.

We can get the same results using other approaches we have discussed.

```
. quietly reg goodpay occ2 occ3 if female == 1
. predict mcompfcoef if female == 0
(option xb assumed; fitted values)
(100 missing values generated)

. sum  mcompfcoef

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
  mcompfcoef |       100    .5795238    .1427537   .2857143   .6666667
```

Alternatively, using the `adjust` command,

```
. quietly reg goodpay occ2 occ3 if female == 1
. adjust occ2 occ3 if female == 0

--------------------------------------------------------------------------------------
     Dependent variable: goodpay     Command: regress
 Covariates set to mean: occ2 = .15, occ3 = .7
--------------------------------------------------------------------------------------

----------------------
      All |         xb
----------+-----------
          |    .579524
----------------------
     Key:  xb  =  Linear Prediction
```

Or, using `margins` with super-precise values for the means,

```
. sum occ2 if female==0, meanonly
. scalar malemeanocc2 = r(mean)
. sum occ3 if female==0, meanonly
. scalar malemeanocc3 = r(mean)
. scalar list
malemeanocc3 =         .7
malemeanocc2 =        .15

. quietly reg goodpay occ2 occ3 if female == 1
. margins, at(occ2 = `=malemeanocc2' occ3 = `=malemeanocc3')

Adjusted predictions                              Number of obs   =        100
Model VCE    : OLS

Expression   : Linear prediction, predict()
at           : occ2            =         .15
               occ3            =          .7

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .5795238   .0871308     6.65   0.000     .4065932    .7524544
------------------------------------------------------------------------------
```

Or, using `margins` with `atmeans`

```
. quietly reg goodpay occ2 occ3 if female == 1
. margins if female == 0, noesample atmeans

Adjusted predictions                              Number of obs   =        100
Model VCE    : OLS

Expression   : Linear prediction, predict()
at           : occ2            =         .15 (mean)
               occ3            =          .7 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .5795238   .0871308     6.65   0.000     .4065932    .7524544
------------------------------------------------------------------------------
```