

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

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? To answer this we proceed as follows.

1. 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 more education and more job experience.

```
. use http://www.nd.edu/~rwilliam/stats2/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
. estimates store blackmodel
. quietly reg income educ jobexp if !black
. estimates store whitemodel
. estimates table blackmodel whitemodel
```

Variable	blackmodel	whitemodel
educ	1.6779491	1.8933377
jobexp	.421975	.72225495
_cons	-3.0512005	-6.4611885

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 next do the following.

2. Regress income on education and job experience for blacks only.

```
. quietly reg income educ jobexp if black
```

3. Use the predict command, but select whites only:

```
. predict whcompblcoef if !black  
(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.

4. 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

5. Or, as an alternative to steps 3 and 4 above, we can use the adjust command:

```
. quietly reg income educ jobexp if black  
. adjust educ jobexp if !black
```

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

```
-----  
All |      xb  
-----+-----  
    | 26.2221  
-----
```

```
Key:  xb = Linear Prediction
```

The adjust command is very useful for “what if” scenarios, and we will use it more later on. In the simple form presented above, it plugs in the mean for each variable, and tells you what the predicted value is. Since the if qualifier limited the sample to those who were not black, it plugged in the means for whites. Specifically, adjust computed the following, using the coefficient values for blacks and the variable means for whites:

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

6. Here is how we interpret the results:

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.

7. 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.

Appendix: 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, \bar{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 ($P(A | E_i)^F$)

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

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

Using the marginal probability theorem, we get

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

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

```
. use http://www.nd.edu/~rwilliam/stats2/statafiles/goodpay.dta, clear
. tab female, sum(goodpay)
```

female	Summary of goodpay		
	Mean	Std. Dev.	Freq.
0	.77	.42295258	100
1	.37	.48523659	100
Total	.57	.4963181	200

This shows us that 77% of the men and 37% of the women receive good pay, a 40% difference.

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

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

```
. predict mcompfcoef if !female
(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  
. adjust occ2 occ3 if !female
```

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

```
-----  
All |          xb  
-----+-----  
    |      .579524  
-----
```

Key: xb = Linear Prediction

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.