Lecture 7

using regression models to "decompose" differences in means

Very widely used in applied studies e.g.

−differences in wages between men and women

- differences in college completion rates between 2 groups

Originally invented by Ron Oaxaca – "Oaxaca decomposition"

Assume to start that there is a “population model”:

$$y_i = x_i'\beta^* + u_i.$$

In our example for today, y_i is log wage of person i , x_i is $i's$ education (and possibly other characteristics).

We are going to be working with the OLS estimator $\widehat{\beta}$. We recall the critical FOC:

$$\sum_{i=1}^{N} x_i(y_i - x_i'\widehat{\beta}) = 0$$

We will assume $x_i' = (1, x_{2i}, ...x_{Ki})$ – the first regressor is a constant.

We will assume there are two groups, a and b. Let $\overline{y}^a$ represent the mean of y_i for group a, and let $\overline{x}^a$ represent the mean of the vector of $x's$ for group a. Suppose we fit a model

$$y_i = \beta_1 + \sum_{j=2}^{K} x_{ji}\beta_j + u_i$$

just for group a (with sample size N^a). Then we know:

$$\overline{y}^a = \widehat{\beta}_1 + \sum_{j=2}^{K} \overline{x}_j^a \widehat{\beta}_j = \overline{x}^a \widehat{\beta}$$

Why?

Refresher. The FOC require:

$$\sum_{i=1}^{N^a} x_i(y_i - x_i'\widehat{\beta}) = 0$$

But one of the elements of x_i is 1 so (dividing by N^a) :

$$\frac{1}{N^a}\sum_{i=1}^{N^a} 1(y_i - x_i'\widehat{\beta}) = 0$$

$$\Rightarrow \overline{y}^a = \frac{1}{N^a}\sum_{i=1}^{N^a} y_i = \frac{1}{N^a}\sum_{i=1}^{N^a} x_i'\widehat{\beta} = \overline{x}^a\widehat{\beta}$$

Intuitively, the constant can always be selected so that, given the other $\widehat{\beta}'s$ the predictions from the regression model fit the mean – so that is what the regression does!

Now let's add the second group to the sample, and add a dummy variable indicating observations from group b: $x_{iK+1} = D_i = 1[i \in b]$. For this new model (with total sample size $N = N^a + N^b$) it will be true that:

$$\begin{aligned} \overline{y}^a &= \overline{x}^{a\prime}\widehat{\beta} \\ \overline{y}^b &= \overline{x}^{b\prime}\widehat{\beta} \end{aligned}$$

Why? From the row of the FOC corresponding to β_{K+1} :

$$\sum_{i=1}^{N} D_i(y_i - x_i'\widehat{\beta}) = 0$$

But this requires that

$$\sum_{i \in b}(y_i - x_i'\widehat{\beta}) = 0$$

FOC dummy for group b requires:

$$\sum_{i \in b} (y_i - x_i' \widehat{\beta}) = 0$$

$$\Rightarrow \overline{y}^b = \frac{1}{N^b} \sum_{i \in b} y_i = \frac{1}{N^b} \sum_{i \in b} x_i' \widehat{\beta} = \left(\overline{x}^b \right)' \widehat{\beta}$$

And the FOC for the constant terms requires

$$\sum_i (y_i - x_i' \widehat{\beta}) = \sum_{i \in a} (y_i - x_i' \widehat{\beta}) + \sum_{i \in b} (y_i - x_i' \widehat{\beta})$$

$$\begin{array}{rcl} \Rightarrow & \sum_{i \in a} (y_i - x_i' \widehat{\beta}) & = 0 \\ \Rightarrow & \overline{y}^a - (\overline{x}^a)' \widehat{\beta} & = 0 \end{array}$$

As a result, in the pooled model with a constant and a dummy for group b we know:

$$\begin{aligned}
\overline{y}^a &= \widehat{\beta}_1 + \sum_{j=2}^{K} \overline{x}_j^a \widehat{\beta}_j \\
\overline{y}^b &= \widehat{\beta}_1 + \sum_{j=2}^{K} \overline{x}_j^b \widehat{\beta}_j + \widehat{\beta}_{K+1}
\end{aligned}$$

Thus:

$$\overline{y}^b - \overline{y}^a = \sum_{j=2}^{K} (\overline{x}_j^b - \overline{x}_j^a)\widehat{\beta}_j + \widehat{\beta}_{K+1}$$

We can use this expression to “decompose” the difference in means. For example, if regressor number 2 is “education”, and $\widehat{\beta}_2 = 0.10$, $\overline{x}_2^b = 12$ and $\overline{x}_2^a = 14$ then differences in education explain a gap of $(\overline{x}_2^b - \overline{x}_2^a)\widehat{\beta}_2 = (12 - 14) \times 0.10 = -0.20$

We have shown that

$$\overline{y}^b - \overline{y}^a = \sum_{j=2}^{K} (\overline{x}_j^b - \overline{x}_j^a)\widehat{\beta}_j + \widehat{\beta}_{K+1}$$

Let's think of the special case where our model only has 2 explanatory variables: a constant, and a dummy for group b. In this case,

$$\begin{aligned} \overline{y}^a &= \widehat{\beta}_1 \\ \overline{y}^b &= \widehat{\beta}_1 + \widehat{\beta}_{K+1} \end{aligned}$$

which implies that the coefficient on the dummy for group b is

$$\widehat{\beta}_{K+1} = \overline{y}^b - \overline{y}^a$$

which we knew from Lecture 3! When we add other explanatory variables, however, the estimate will (in general) change.

Example: let's look at our 2012 sample from the CPS. Here we will focus on men, age 30-35, and consider group a = natives and group b = immigrants. Some relevant information:

Natives:

mean log wage = 3.0129

mean education = 14.092 years

Immigrants:

mean log wage = 2.7660

mean education = 12.409 years

	Pooled Model: Fit to Natives and Immigrants	
	(1)	(2)
Constant	3.013	1.546
	(0.006)	(0.025)
Immigrant	-0.247	-0.072
	(0.013)	(0.013)
Education (yrs)	--	0.104
		(0.002)
MSE	0.757	0.695
Adj. R-sq	0.018	0.173
Sample Size	19,092	19,092

Notes: Fit to data for males age 30-45 in March 2012 CPS. Dependent variable is log average hourly wage. Mean and standard deviation are 2.959 (0.764). Standard errors in parentheses.

Difference in mean wages

Difference in mean wages after "controlling" for education

Let's perform the decomposition. We have $K = 2$, with the second variable being education.

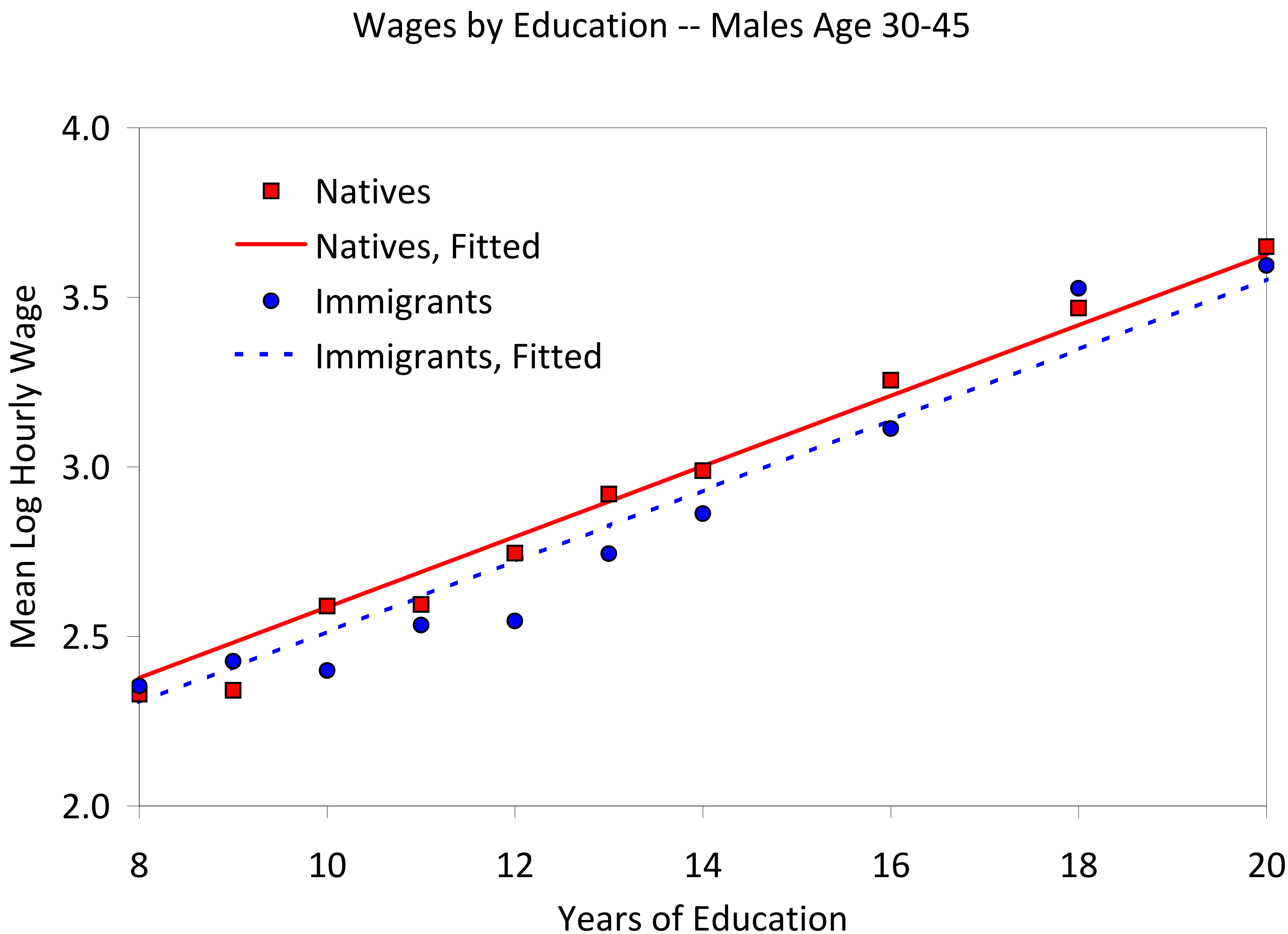
$$\overline{y}^b - \overline{y}^a = 2.766 - 3.013 = -0.247$$

From the model in column 2 of the table, we have that

$$\overline{y}^b - \overline{y}^a = \sum_{j=2}^{K} (\overline{x}_j^b - \overline{x}_j^a)\widehat{\beta}_j + \widehat{\beta}_{K+1}$$

$$-0.247 = (12.409 - 14.092) \times 0.1041 - 0.0718$$

So the "effect of education" is $-1.683 \times 0.1041 = -0.175$ which is 70.9% of the wage gap. The remainder (29.1%) is "unexplained"



Wages by Education -- Males Age 30-45
Natives
Natives, Fitted
Immigrants
Immigrants, Fitted
Mean Log Hourly Wage
4.0
3.5
3.0
2.5
2.0
8
10
12
14
16
18
20
Years of Education

A common problem that arises in applying this idea is that the coefficients of the explanatory variables are not the same in the two samples. This gives rise to a more general version, which is "the" Oaxaca decomposition. Suppose we fit our model separately in the two subsamples. For group a we obtain OLS estimates $\widehat{\beta}^a$, and (since we have a constant in the model) we know

$$\overline{y}^a = (\overline{x}^a)'\widehat{\beta}^a.$$

For group b we obtain OLS estimates $\widehat{\beta}^b$, and we know

$$\overline{y}^b = (\overline{x}^b)'\widehat{\beta}^b.$$

So we can construct

$$\overline{y}^b - \overline{y}^a = (\overline{x}^b)'\widehat{\beta}^b - (\overline{x}^a)'\widehat{\beta}^a$$

Now let's manipulate this expression:

$$\begin{aligned}
\overline{y}^b - \overline{y}^a &= (\overline{x}^b)'\widehat{\beta}^b - (\overline{x}^a)'\widehat{\beta}^a \\
&= (\overline{x}^b - \overline{x}^a)'\widehat{\beta}^a + (\overline{x}^b)'(\widehat{\beta}^b - \widehat{\beta}^a) \\
&= (\overline{x}^b - \overline{x}^a)'\widehat{\beta}^b + (\overline{x}^a)'(\widehat{\beta}^b - \widehat{\beta}^a)
\end{aligned}$$

These are both algebraically true. The first says that the difference is the difference in mean $x's$, weighted by the estimated coefficients from group a, plus the difference in the coefficients, weighted by the mean from group b. The second reverses the groups. Note that if $\widehat{\beta}^a = \widehat{\beta}^b = \widehat{\beta}$ we get our previous method, taking account of the fact that our previous model had a dummy for group b included as one of the $x's$.

	Pooled Model: Fit to Natives and Immigrants		Model for Natives	Model for Immigrants
	(1)	(2)	(3)	(4)
Constant	3.013	1.546	1.365	1.676
	(0.006)	(0.025)	(0.033)	(0.035)
Immigrant	-0.247	-0.072	--	--
	(0.013)	(0.013)		
			Coefficients are NOT the same	
Education (yrs)	--	0.104	0.117	0.088
		(0.002)	(0.002)	(0.002)
MSE	0.757	0.695	0.689	0.707
Adj. R-sq	0.018	0.173	0.146	0.208
Sample Size	19,092	19,092	14,921	4,141

Notes: Fit to data for males age 30-45 in March 2012 CPS. Dependent variable is log average hourly wage. Mean and standard deviation are: for overall sample, 2.959 (0.764); for natives 3.013 (0.746); for immigrants 2.766 (0.795). Standard errors in parentheses.

Let's apply this to our example. Here we have

$$\begin{aligned}
\widehat{\beta}_2^a &= 0.117 \\
\widehat{\beta}_2^b &= 0.088 \\
(\overline{x}_2^b - \overline{x}_2^a) &= 12.409 - 14.092 = 1.683
\end{aligned}$$

And we know $\overline{y}^b - \overline{y}^a = -0.247$. So if we use the coefficient for natives we have:

$$(\overline{x}_2^b - \overline{x}_2^a)\widehat{\beta}_2^a = -0.197$$

$$\overline{x}_2^b(\widehat{\beta}_2^b - \widehat{\beta}_2^a) = -0.360$$

Whereas if we use the coefficient for immigrants we have

$$\begin{aligned}
(\overline{x}_2^b - \overline{x}_2^a)\widehat{\beta}_2^b &= -0.148 \\
\overline{x}_2^a(\widehat{\beta}_2^b - \widehat{\beta}_2^a) &= -0.409
\end{aligned}$$

This shows a couple of important things. First, we have 2 estimates of the contribution of the difference in mean education:

$$
\begin{aligned}
(\overline{x}_2^b - \overline{x}_2^a)\widehat{\beta}_2^a &= -0.197 \\
(\overline{x}_2^b - \overline{x}_2^a)\widehat{\beta}_2^b &= -0.148
\end{aligned}
$$

Usually people interpret this as meaning that the effect is somewhere between -0.15 and -0.20 out of the total -0.247 wage gap. But what do we make out of the other term?

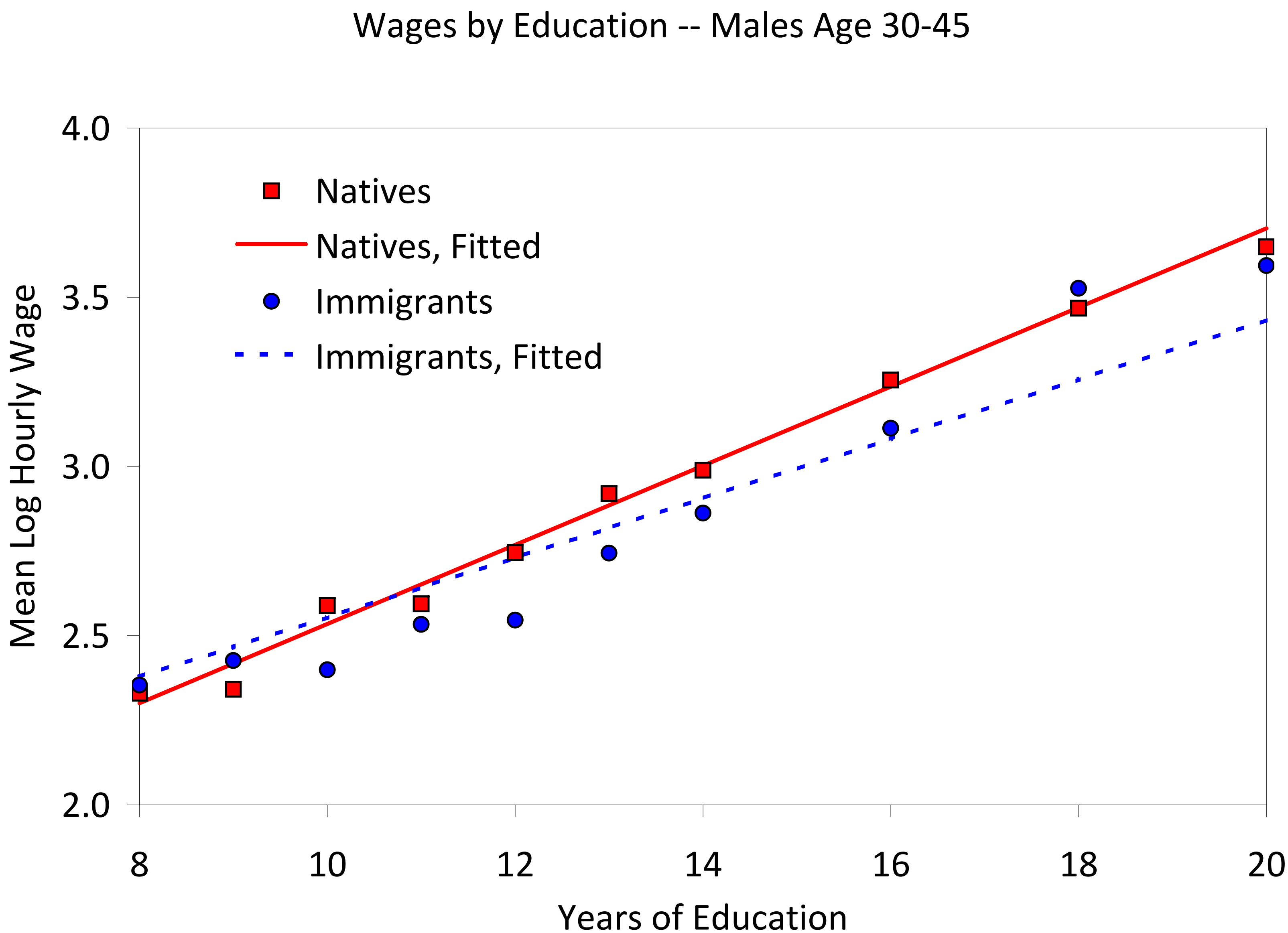
$$
\begin{aligned}
\overline{x}_2^b(\widehat{\beta}_2^b - \widehat{\beta}_2^a) &= -0.360 \\
\overline{x}_2^a(\widehat{\beta}_2^b - \widehat{\beta}_2^a) &= -0.409
\end{aligned}
$$

In either case we are "over-explaining" the wage gap (by a lot). If you look back at the fitted models you can see what is happening

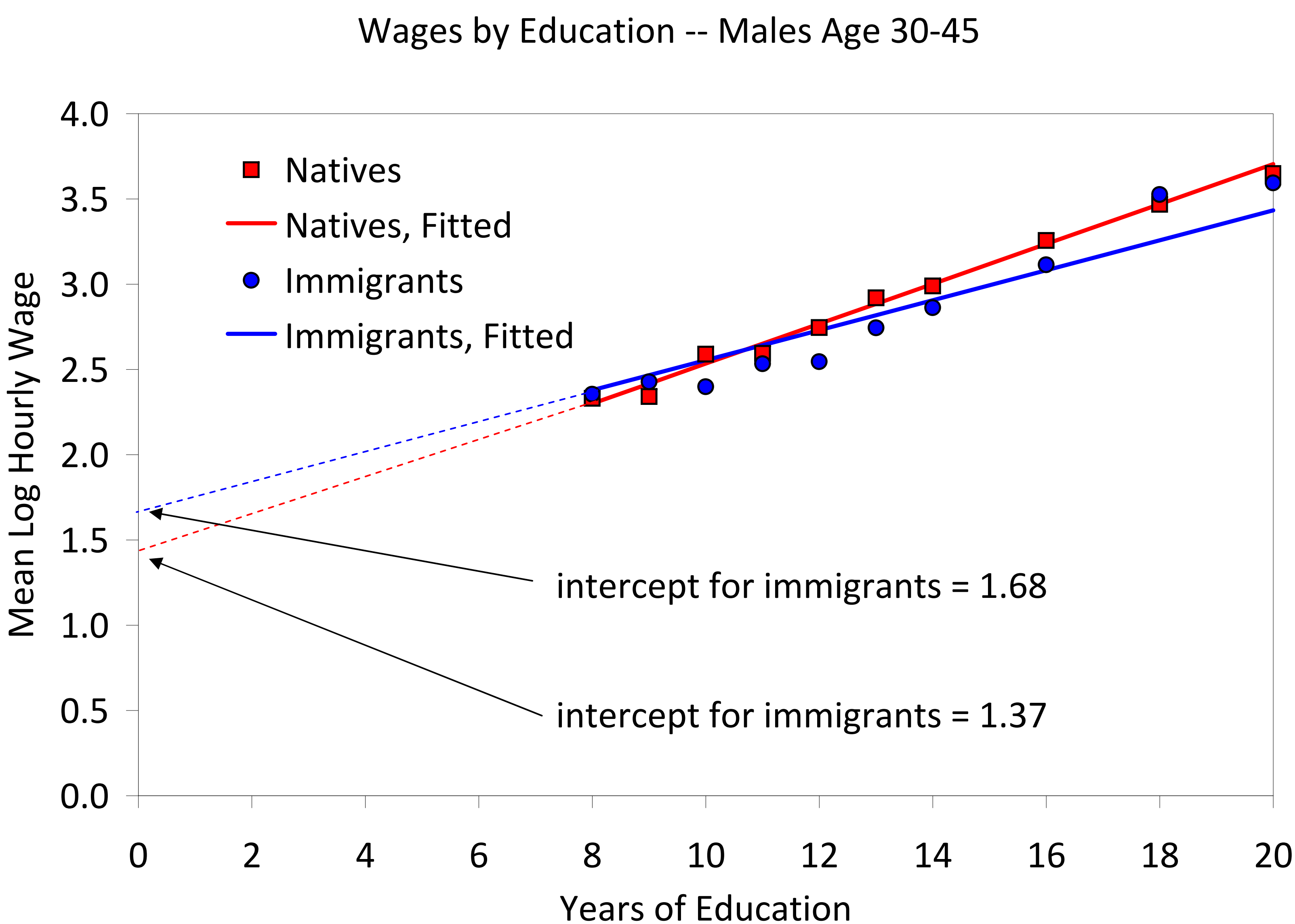
	Pooled Model: Fit to Natives and Immigrants		Model for Natives	Model for Immigrants
	(1)	(2)	(3)	(4)
Constant	3.013	1.546	1.365	1.676
	(0.006)	(0.025)	(0.033)	(0.035)
Immigrant	-0.247	-0.072	--	--
	(0.013)	(0.013)		
Education (yrs)	--	0.104	0.117	0.088
		(0.002)	(0.002)	(0.002)
MSE	0.757	0.695	0.689	0.707
Adj. R-sq	0.018	0.173	0.146	0.208
Sample Size	19,092	19,092	14,921	4,141

Estimated intercepts are much different

Notes: Fit to data for males age 30-45 in March 2012 CPS. Dependent variable is log average hourly wage. Mean and standard deviation are: for overall sample, 2.959 (0.764); for natives 3.013 (0.746); for immigrants 2.766 (0.795). Standard errors in parentheses.



Wages by Education -- Males Age 30-45
Natives
Natives, Fitted
Immigrants
Immigrants, Fitted
Mean Log Hourly Wage
4.0
3.5
3.0
2.5
2.0
8
10
12
14
16
18
20
Years of Education



Wages by Education -- Males Age 30-45
Natives
Natives, Fitted
Immigrants
Immigrants, Fitted
Mean Log Hourly Wage
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8
10
12
14
16
18
20
Years of Education
intercept for immigrants = 1.68
intercept for immigrants = 1.37

The decomposition is multiplying the difference in estimated "returns to education" – which is $0.088 - 0.117 = -0.029$ by numbers like 12 or 14, which "explains" a quite large difference in wages. The estimated constants are offsetting this so the total explained difference is always exactly -0.247.

We can see from this example that the part of the Oaxaca decomposition attributed to the difference in coefficients has to be interpreted carefully.

Let's probe this a little more. Suppose instead of measuring education in "years," we measured in in "years of high school or more" i.e., we subtracted 8 from all measures of education.

$$\begin{aligned}\overline{y}^a &= \widehat{\beta}_1^a + \widehat{\beta}_2^a \overline{x}_2^a \\ &= \widehat{\beta}_1^a + \widehat{\beta}_2^a(\overline{x}_2^a - 8) + 8\widehat{\beta}_2^a \\ &= (\widehat{\beta}_1^a + 8\widehat{\beta}_2^a) + \widehat{\beta}_2^a(\overline{x}_2^a - 8)\end{aligned}$$

If we were to measure education as years of high school or more, we would get *exactly the same coefficient* on education, but the constant would be bigger (by exactly $8\widehat{\beta}_2^a$). Likewise for group b :

$$\overline{y}^b = \widehat{\beta}_1^b + \widehat{\beta}_2^b \overline{x}_2^b = (\widehat{\beta}_1^b + 8\widehat{\beta}_2^b) + \widehat{\beta}_2^b(\overline{x}_2^b - 8)$$

If we examined the “difference in $x's$” part of the Oaxaca decompositiopn, we would compare differences in renormalized education:

$$(\overline{x}_2^b - 8) - (\overline{x}_2^a - 8) = \overline{x}_2^b - \overline{x}_2^a$$

multiplying by $\widehat{\beta}_2^a$ or $\widehat{\beta}_2^b$ – so we would get the same answer as before. But for the “difference in coefficients” part of the decomposition, we would look at

$$(\widehat{\beta}_2^b - \widehat{\beta}_2^a) \times (\overline{x}_2^b - 8)$$

or

$$(\widehat{\beta}_2^b - \widehat{\beta}_2^a) \times (\overline{x}_2^a - 8)$$

Returning to our example:

$$
\begin{aligned}
\overline{x}_2^a &= 14.09 \\
\overline{x}_2^a &= 12.41 \\
\widehat{\beta}_2^a &= 0.117 \\
\widehat{\beta}_2^b &= 0.088
\end{aligned}
$$

So if we use the renormalized mean for immigrants we have:

$$(\overline{x}_2^b - 8)(\widehat{\beta}_2^b - \widehat{\beta}_2^a) = 4.41 \times -0.029 = -0.128$$

Whereas if we use renormalized mean for natives we have:

$$(\overline{x}_2^a - 8)(\widehat{\beta}_2^b - \widehat{\beta}_2^a) = 6.09 \times -0.029 = -0.177$$

Which still "over-explains" the immigrant-native wage gap!

Bottom line:

1. we can always use a pooled model to evaluate the effect of differences in mean $x's$:

$$\overline{y}^b - \overline{y}^a = \sum_{j=2}^{K} (\overline{x}_j^b - \overline{x}_j^a)\widehat{\beta}_j + \widehat{\beta}_{K+1}$$

2. when the coefficients of the x variables are different for the two groups, we can evaluate two alternative terms:

$$\sum_{j=2}^{K}(\overline{x}_j^b - \overline{x}_j^a)\widehat{\beta}_j^a$$
$$\sum_{j=2}^{K}(\overline{x}_j^b - \overline{x}_j^a)\widehat{\beta}_j^b$$

3. when the coefficients are different there is also a "difference in coefficients" component

The two estimates of the difference in coefficients component are:

$$\sum_{j=2}^{K} \overline{x}_j^b(\widehat{\beta}_j^b - \widehat{\beta}_j^a)$$
$$\sum_{j=2}^{K} \overline{x}_j^a(\widehat{\beta}_j^b - \widehat{\beta}_j^a)$$

And can be evaluated. BUT – we have to be careful, because we can re-normalize the x variable and get different answers!!

Another example - gender gap in probability of entering a STEM (science-technology-engineering-math) program in university

Setting: Ontario (Canada) - entry directly to college programs (e.g., "economics" or "chemistry")

- entry based on grades in highest 6 courses in final year of HS (no SAT!)

- "STEM-ready" kids have at least 3 math/science courses

Table 5: Models For Probability of Registering in STEM-Related University Program

	Dependent Variable = 1 if Register in STEM-related Program		
	(1)	(2)	(3)
Female Indicator (× 100)	-5.0 (0.2)	-5.8 (0.2)	-1.7 (0.2)
Within-cohort Rank of Top 3 Grade 12 STEM courses	--	--	0.73 (0.01)
Within-cohort Rank of Top 6 Grade 12 Course	--	--	-0.52 (0.01)
Age, Year and High School Effects?	no	yes	yes
Gifted/special needs)?	no	yes	yes

Note: standard errors in parentheses. Table reports linear probability model coefficients for event of registering in STEM-related program. Sample is 170,288 STEM-ready students. Models in columns 2-4 include dummies for age, graduating cohort, student's main language, foreign-born status, and high school. See Table 2 for sample information.

Mean Ranks in Top 3 Grade 12 STEM courses: females 0.50, males 0.51

Mean Ranks in Top 6 Grade 12 courses: females 0.61, males 0.55