Causal Effects in Policy - Lecture 3b

Alice Wu

UW Madison Econ 695

More on Decomposition

- Previously, we apply decomposition methods to models with discrete covariates (group dummies).
- Today, we will present the more general versions of:
  - Oaxaca-Blinder

Oaxaca-Blinder Decomposition (General version)

- Let x_i denote a vector of covariates (including a constant, continuous/discrete covariates). Estimate an OLS regression of y_i on x_i separately for each group a, b:

$$\hat{\beta}^a = argmin_\beta \sum_{i \in a} (y_i - x_i' \beta)^2 \tag{1}$$
$$\hat{\beta}^b = argmin_\beta \sum_{i \in b} (y_i - x_i' \beta)^2$$

Recall with a constant in $x_i = (1, x_{2i}, ...x_{Ki})'$, the linear regression can fit the mean.

$$\overline{y}^a = (\overline{x}^a)' \hat{\beta}^a \tag{2}$$
$$\overline{y}^b = (\overline{x}^b)' \hat{\beta}^b$$

Oaxaca-Blinder Decomposition (General version)

- Oaxaca-Blinder Decomposition:

$$\begin{aligned}\bar{y}^b - \bar{y}^a &= (\bar{x}^b)'\hat{\beta}^b - (\bar{x}^a)'\hat{\beta}^a \\ &= \underbrace{(\bar{x}^b - \bar{x}^a)'\hat{\beta}^b}_{\text{between}} + \underbrace{(\bar{x}^a)'(\hat{\beta}^b - \hat{\beta}^a)}_{\text{within}} \\ &= \underbrace{(\bar{x}^b)'(\hat{\beta}^b - \hat{\beta}^a)}_{\text{within}} + \underbrace{(\bar{x}^b - \bar{x}^a)'\hat{\beta}^a}_{\text{between}}\end{aligned}$$

  - what's the difference between the last two lines? Hint: $\bar{y}^b_{counterf}$ vs. $\bar{y}^a_{counterf}$
  - "between": the composition effect that can be explained by differences in x's
  - "within": the effect that cannot be explained by diff in x's – in our application, recall this represents the differential returns to education in the public vs. private sectors.

Compare 3 regressions

1. Fit separate Models for groups $\{a, b\}$:

$$\forall i \in a : \hat{y}_i = \hat{\beta}_1^a + \sum_{k=2}^{K} x_{ki}\hat{\beta}_k^a; \quad \forall i \in b : \hat{y}_i \quad = \hat{\beta}_1^b + \sum_{k=2}^{K} x_{ki}\hat{\beta}_k^b \tag{3}$$

2. Pooled Regression: pool the observations from group a and group b, define $D_i = 1[i \in b]$

$$\hat{y}_i = \hat{\beta}_1 + \sum_{k=2}^{K} x_{ki}\hat{\beta}_k + \hat{\gamma}_1 D_i \tag{4}$$

with the inclusion of group dummy D_i, this regression fits the mean of the outcome for both group a and group b: (Hint: FOC w.r.t. β_1 and β_{K+1})

$$\bar{y}^a = (\bar{x}^a)'\hat{\beta}, \; \bar{y}^b = (\bar{x}^b)'\hat{\beta} + \hat{\gamma}_1$$

- what is $\hat{\gamma}_1$ if $x_i = 1$?
- so the diff in mean: $\bar{y}^b - \bar{y}^a = (\bar{x}^b - \bar{x}^a)'\hat{\beta} + \hat{\gamma}_1$. Compare with the decomposition from model #1.

Compare 3 regressions

- the pooled estimate, $\hat{\beta}$ is a weighted average of $\hat{\beta}^a$ and $\hat{\beta}^b$(Normally it lies between them, but not always.

3. Fully-interacted Pooled Regression: interact D_i fully with $x_i = (1, x_{2i}, ..., x_{Ki})'$,

$$\hat{y}_i = \hat{\beta}_1 + \sum_{k=2}^{K} x_{ki}\hat{\beta}_k + \hat{\gamma}_1 D_i + \sum_{k=2}^{K} (x_{ki} \times D_i)\,\hat{\gamma}_k \tag{5}$$

note this is a way to estimate $(\hat{\beta}_1^a, \hat{\beta}_1^b)$ in one regression:

$$\hat{\beta} = \hat{\beta}^a$$
$$\hat{\beta} + \hat{\gamma} = \hat{\beta}^b$$

Application: Immigrant-Native Pay Gap

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
- What's the difference in mean log wage? How would you estimate it in a regression?

Pooled Regression

- Model (1) fits a regression of mean wage on a constant and $D_i = 1[\text{Immigrant}]$. What is the coefficient on the dummy?

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

Pooled Regression

	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

Pooled Regression - Decomposition

- Let's perform the decomposition according to the pooled regression (column 2):

$$
\begin{aligned}
\overline{y}^b - \overline{y}^a &= (\overline{x}^b - \overline{x}^a)' \hat{\beta} + \hat{\gamma}_1 \\
&= (1-1)\hat{\beta}_1 + (\overline{x}^b - \overline{x}^a)\hat{\beta}_2 + \hat{\gamma}_1
\end{aligned}
$$

$$
-0.247 = \underbrace{(12.409 - 14.092)}_{\text{diff in educ}} \times \underbrace{0.104}_{\hat{\beta}_2} \underbrace{-0.072}_{\hat{\gamma}_1 \text{on Immig}}
$$

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" by diff in education (composition).

Separate Regressions

- Now we consider separate models: $\hat{\beta}_a$ estimated on natives, and $\hat{\beta}^b$ estimated on immigrants,

$$\text{Natives: } \hat{y}_i = \hat{\beta}_1^a + \hat{\beta}_2^a x_{2i}$$
$$\text{Immigrants: } \hat{y}_i = \hat{\beta}_1^b + \hat{\beta}_2^b x_{2i}$$

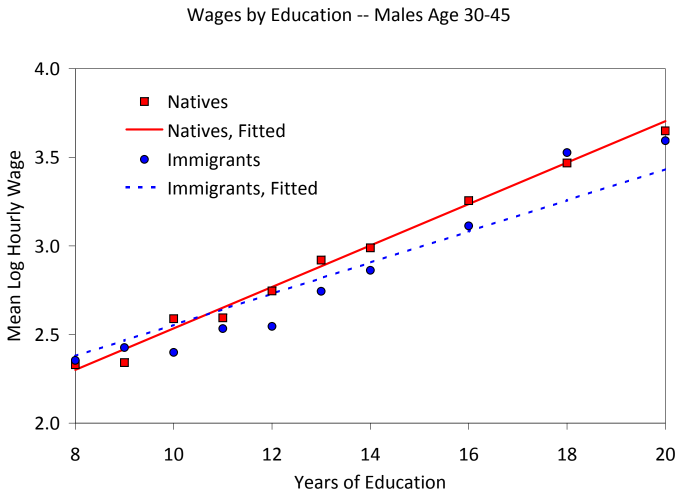
Separate Regressions

	Pooled Model: Fit to Natives and Immigrants		Model for Natives	Model for Immigrants
	(1)	(2)	(3)	(4)
Constant	3.013 (0.006)	1.546 (0.025)	1.365 (0.033)	1.676 (0.035)
Immigrant	-0.247 (0.013)	-0.072 (0.013)	--	--
Education (yrs)	--	0.104 (0.002)	0.117 (0.002)	0.088 (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

Coefficients are NOT the same

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.

Separate Regressions: Slopes may vary



Oaxaca Decomposition based on Separate Regressions

Evaluating terms:

$$\hat{\beta}_2^a = 0.117, \; \hat{\beta}_2^b = 0.088$$

And we know wage gap $\overline{y}^b - \overline{y}^a = -0.247$, and education gap $(\overline{x}_2^b - \overline{x}_2^a) = 12.409 - 14.092 = 1.683$. So if we use the coefficient for natives $(\hat{\beta}^a)$ we have:

$$\text{between: } (\overline{x}_2^b - \overline{x}_2^a)\hat{\beta}_2^a = -0.197$$
$$\text{within: } \overline{x}_2^b(\hat{\beta}_2^b - \hat{\beta}_2^a) = -0.360$$

Whereas if we use the coefficient for immigrants $(\hat{\beta}^b)$ we have

$$\begin{aligned} \text{between: } (\overline{x}_2^b - \overline{x}_2^a)\hat{\beta}_2^b &= -0.148 \\ \text{within: } \overline{x}_2^a(\hat{\beta}_2^b - \hat{\beta}_2^a) &= -0.409 \end{aligned}$$

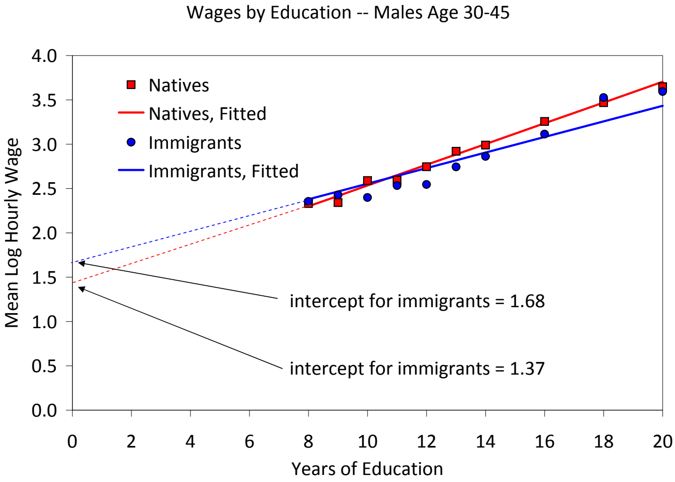
This shows a couple of important things.

1. We have 2 estimates of the contribution of the difference in mean education (composition effect - **"between"**): -0.197 or -0.148.
   - 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.
2. But what do we make of the **"within"** term? Between + Within = mean wage gap $\bar{y}^b - \bar{y}^a = -0.247$?

$$\bar{x}_2^b(\hat{\beta}_2^b - \hat{\beta}_2^a) \quad = -0.360 \quad , \bar{x}_2^a(\hat{\beta}_2^b - \hat{\beta}_2^a) = -0.409$$

   - Note that these are quite large – much larger than the immigrant/native wage gap. If you look back at the fitted models you can see what is happening.

Oaxaca Decomposition based on Separate Regressions



Renormalize X

- Let's probe the term "$\bar{x}^g \times \triangle\hat{\beta}$" 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.

$$\begin{aligned}
\bar{y}^a &= \hat{\beta}_1^a + \hat{\beta}_2^a \bar{x}_2^a \\
&= \hat{\beta}_1^a + \hat{\beta}_2^a(\bar{x}_2^a - 8) + 8\hat{\beta}_2^a \\
&= \underbrace{(\hat{\beta}_1^a + 8\hat{\beta}_2^a)}_{\text{constant}} + \underbrace{\hat{\beta}_2^a(\bar{x}_2^a - 8)}_{\text{effect of normalized x}}
\end{aligned}$$

  - 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 $8\hat{\beta}_2^a$).
  - Likewise for group b :

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

Renormalize X

- If we examined the "difference in $x's$" part of the Oaxaca decomposition, 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 $\hat{\beta}_2^a$ or $\hat{\beta}_2^b$ – so we would get the same answer as before.
- But for the "difference in coefficients" part of the decomposition, we would look at
$$(\hat{\beta}_2^b - \hat{\beta}_2^a) \times (\overline{x}_2^b - 8)$$
or$(\hat{\beta}_2^b - \hat{\beta}_2^a) \times (\overline{x}_2^a - 8)$.
- Returning to our example:
$$\begin{aligned} \overline{x}_2^a &= 14.09,\ \overline{x}_2^b = 12.41 \\ \hat{\beta}_2^a &= 0.117,\ \hat{\beta}_2^b = 0.088 \end{aligned}$$
So if we use the re-normalized mean for immigrants we have:
$$(\overline{x}_2^b - 8) \times (\hat{\beta}_2^b - \hat{\beta}_2^a) = 4.41 \times -0.029 = -0.128$$

Renormalize X

Whereas if we use renormalized mean for natives we have:

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

which still "over-explains" the immigrant-native wage gap because the constants have different coefficients in models for a and b $\hat{\beta}_1^b > \hat{\beta}_1^a$.

- Takeaway: we have to be careful in the interpretation of the "within" component (unexplained by composition) - $\overline{x}^g \times (\hat{\beta}^b - \hat{\beta}^a)$, because we can re-normalize the x variable and get different answers!!
- If the key x variable has a "natural scale" then it may be possible to compare the contribution of the coefficients. E.g., we could assert that everyone with education ≤ 8 years earns (more or less) the same, so the natural scale is to measure $Educ - 8$.