Causal Effects in Policy - Lecture 3

Alice Wu

UW Madison Econ 695

This Lecture

- Recap: Regressions with group dummies
- Application: Wage Differentials between the Public Sector and the Private Sector
  - How much can the gap be explained by the composition of workers (across education groups) versus wage structure (differential return within educ group)?
- Decomposition Methods:
  - Oaxaca-Blinder
  - Reweighting Techniques (DiNardo, Fortin and Lemieux 1996)

Applications of Decomposition Methods

1. What are the most important explanations accounting for pay differences between men and women?
2. To what extent has wage inequality increased in the U.S. since the 1980s because of increasing returns to skill?
3. Which factors are behind most of the growth in U.S. GDP?

We will start with categorical covariates (e.g., educ, occupation...), decompose the pay gap between public-sector and private-sector workers into between-educ and within-educ differences, and then discuss the decomposition methods in general (continuous covariates).

Regressions with Multiple Dummies

- Suppose individuals belong to G mutually exclusive groups. For example, education level: less than HS, HS, College and above.
- Define $D_{gi} = 1[i \in g]$ for $g = 1, 2, ..., G$. Because of the mutual exclusivity, $D_{gi} \times D_{ki} = 0$ for any $g \neq k$.
- We posit a population regression:

$$y_i = x_i'\beta + u_i \tag{1}$$

$$x_i := \begin{bmatrix} D_{1i} \\ ... \\ D_{Gi} \end{bmatrix}$$

Given a random sample, let $N_g = \sum_i D_{gi}$ denote the num. individuals in group g. The fraction of individuals in g: $\bar{p}_g = \frac{1}{N}\sum_i D_{gi} = \frac{N_g}{N}$. Recall the OLS estimator:

$$\hat{\beta} = [\frac{1}{N}\sum_i x_i x_i']^{-1}\frac{1}{N}\sum_i x_i y_i = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ .. \\ \bar{y}_G \end{pmatrix}$$

Regressions with Multiple Dummies

$$\overline{y} = \overline{x}'\hat{\beta} = \sum_g \overline{p}_g \overline{y}_g \tag{2}$$

- So the OLS regression is just a way to get the group-specific means. The population version:

$$\begin{aligned} E^*[y_i|x_i] &= E[x_i]'\beta \\ &= \sum_g E[D_{gi}] \times E[y_i|D_{gi} = 1] \\ &= E[y_i|x_i] \end{aligned} \tag{3}$$

the linear projection is also the CEF!

Application of Dummy Regressions

- Now consider what happens with 2 groups, a and b. Let's denote:
  - $\overline{y}^a = \frac{1}{N^a}\sum_{i\in a} y_i$ = mean of outcome for group a
  - $\overline{y}^a_g = \frac{1}{N^a_g}\sum_{i\in g,a} y_i$ = mean of outcome for group a, category g
  - $N^a = \#obs$ in group a, $N^a_g = \#obs$ in group a, category g
  - $\overline{y}^a_g = \frac{1}{N^a_g}\sum_{i\in g,a} y_i$ = mean of outcome for group a, category g
  - $\overline{p}^a_g = \frac{1}{N^a}\sum_{i\in a} D_{gi}$ = fraction of group a in category g.
  - as before, $x_i = [D_{1i}, D_{2i}, ..., D_{Gi}]'$ indicate G mutually exclusive categories. Thus $\bar{x}^a = [\bar{p}^a_1, \bar{p}^a_2, ..., \bar{p}^a_G]'$

Clearly (by the Law of Iterated Expectations),

$$\overline{y}^a = \sum_g \overline{p}^a_g \overline{y}^a_g = (\overline{x}^a)'\hat{\beta}^a \tag{4}$$

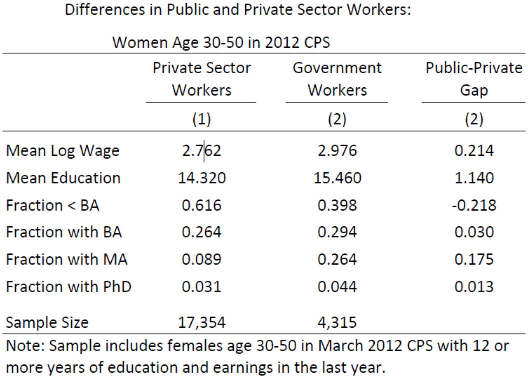
$$\overline{y}^b = \sum_g \overline{p}^b_g \overline{y}^b_g = (\overline{x}^b)'\hat{\beta}^b$$

So we can think about an **Oaxaca decomposition** of $\overline{y}^b - \overline{y}^a$: how much of the difference is explained by the difference between a and b in x_i?

- g : education. How much does the gap in earnings between b = public sector and a = private sector employees can be explained by differences in education levels?

Application: Public-Private Sector Pay Gap

- There is a lot of interest in the idea that government workers are "overpaid". Is that true? If so, by how much?
- To test this out, we'll use a March CPS sample. We'll focus on female workers age 30-50 with 12 or more years of education.
  - Government workers are paid more, but also have more education.
  - Define $\{D_{ig}\}$ as the indicators for education levels.



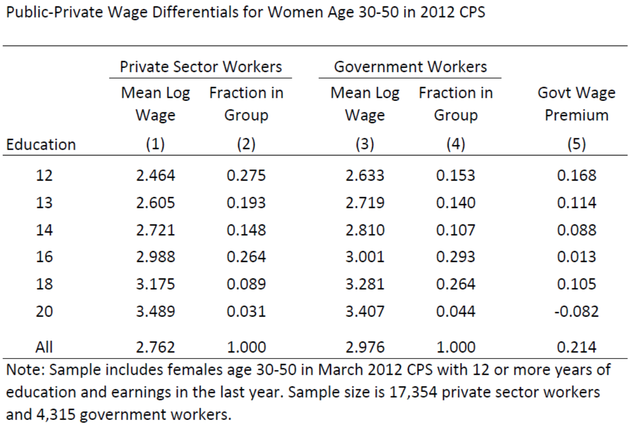
Differences in Public and Private Sector Workers:

Women Age 30-50 in 2012 CPS

	Private Sector Workers	Government Workers	Public-Private Gap
	(1)	(2)	(2)
Mean Log Wage	2.762	2.976	0.214
Mean Education	14.320	15.460	1.140
Fraction < BA	0.616	0.398	-0.218
Fraction with BA	0.264	0.294	0.030
Fraction with MA	0.089	0.264	0.175
Fraction with PhD	0.031	0.044	0.013
Sample Size	17,354	4,315	

Note: Sample includes females age 30-50 in March 2012 CPS with 12 or more years of education and earnings in the last year.

Application: Public-Private Sector Pay Gap



Public-Private Wage Differentials for Women Age 30-50 in 2012 CPS

	Private Sector Workers		Government Workers		
	Mean Log Wage	Fraction in Group	Mean Log Wage	Fraction in Group	Govt Wage Premium
Education	(1)	(2)	(3)	(4)	(5)
12	2.464	0.275	2.633	0.153	0.168
13	2.605	0.193	2.719	0.140	0.114
14	2.721	0.148	2.810	0.107	0.088
16	2.988	0.264	3.001	0.293	0.013
18	3.175	0.089	3.281	0.264	0.105
20	3.489	0.031	3.407	0.044	-0.082
All	2.762	1.000	2.976	1.000	0.214

Note: Sample includes females age 30-50 in March 2012 CPS with 12 or more years of education and earnings in the last year. Sample size is 17,354 private sector workers and 4,315 government workers.

Application: Public-Private Sector Pay Gap

- When we look at the “distribution table” we can think of many ways to explore the nature of the overall pay gap. One way is to imagine a “counterfactual world” where group *b* (the gov't workers) had the same distribution across categories as group *a* (the private sector workers) but within each category, the mean for group *b* remains unchanged. Under this counterfactual, the mean for group *b* would be:

$$\overline{y}^b_{counterf} = \sum_g \overline{p}^a_g \times \overline{y}^b_g = (\overline{x}^a)'\hat{\beta}^b \tag{5}$$

  This “reweighted counterfactual” allows us to decompose the wage gap as follows:

$$\begin{aligned} \overline{y}^b - \overline{y}^a &= \left(\overline{y}^b - \overline{y}^b_{counterf}\right) + \left(\overline{y}^b_{counterf} - \overline{y}^a\right) \\ &= \underbrace{\left(\overline{x}^b - \overline{x}^a\right)' \hat{\beta}_b}_{\text{between}} + \underbrace{(\overline{x}^a)' \left(\hat{\beta}^b - \hat{\beta}^a\right)}_{\text{within}} \end{aligned}$$

Application: Public-Private Sector Pay Gap

- Between: how much of the gap is explained by differences in the distribution across education levels ($\overline{p}_g^a$ vs. $\overline{p}_g^b$)

$$\begin{aligned} \overline{y}^b - \overline{y}^b_{counterf} &= \sum_g \overline{p}_g^b \overline{y}_g^b - \sum_g \overline{p}_g^a \overline{y}_g^b \\ &= \sum_g (\overline{p}_g^b - \overline{p}_g^a) \overline{y}_g^b \\ &= (\overline{\boldsymbol{x}}^{\boldsymbol{a}} - \overline{\boldsymbol{x}}^{\boldsymbol{b}})' \hat{\beta}^b \qquad (6) \end{aligned}$$

which is a weighted average of the difference in shares $(\overline{x}^a - \overline{x}^b)$ across the categories, using the average pay of the gov't workers $\hat{\beta}^b$ as weights.

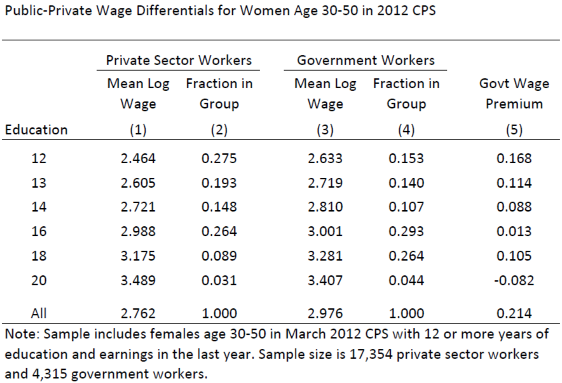
Application: Public-Private Sector Pay Gap

- Within: how much of the gap is explained by differences in returns to education ($\hat{\beta}^a$ vs. $\hat{\beta}^b$), holding fixed the distribution across educ levels:

$$\begin{aligned} \overline{y}^b_{counterf} - \overline{y}^a &= \sum_g \overline{p}^a_g \overline{y}^b_g - \sum_g \overline{p}^a_g \overline{y}^a_g \\ &= \sum_g \overline{p}^a_g \times (\overline{y}^b_g - \overline{y}^a_g) \\ &= (\overline{x}^a)'(\hat{\boldsymbol{\beta}}^{\boldsymbol{b}} - \hat{\boldsymbol{\beta}}^{\boldsymbol{a}}) \end{aligned} \tag{7}$$

which is a “group a - weighted” average of the pay gaps within each category. (column 2 × column 5)

Application: Public-Private Sector Pay Gap



Public-Private Wage Differentials for Women Age 30-50 in 2012 CPS

	Private Sector Workers		Government Workers				
	Mean Log Wage	Fraction in Group	Mean Log Wage	Fraction in Group	Govt Wage Premium	Col 2	Col 4
Education	(1)	(2)	(3)	(4)	(5)	× Col 5	× Col 5
12	2.464	0.275	2.633	0.153	0.168	0.046	0.026
13	2.605	0.193	2.719	0.140	0.114	0.022	0.016
14	2.721	0.148	2.810	0.107	0.088	0.013	0.009
16	2.988	0.264	3.001	0.293	0.013	0.003	0.004
18	3.175	0.089	3.281	0.264	0.105	0.009	0.028
20	3.489	0.031	3.407	0.044	-0.082	-0.003	-0.004
All	2.762	1.000	2.976	1.000	0.214	0.092	0.079

Note: Sample includes females age 30-50 in March 2012 CPS with 12 or more years of education and earnings in the last year. Sample size is 17,354 private sector workers and 4,315 government workers.

Col 2*Col 5: $(\overline{x}^a)'(\hat{\beta}^b - \hat{\beta}^a)$

Application: Public-Private Sector Pay Gap

- We could also think of a different re-weighting counterfactual. Imagine that group *a* (the private sector workers) had the same distribution across categories as group *b* (the gov't workers) but within each category, the mean for group *a* remains unchanged. Under this counterfactual, the mean for group *a* would be:

  $$\overline{y}^a_{counterf} = \sum_g \overline{p}^b_g \overline{y}^a_g = (\overline{x}^b)'\hat{\beta}^a$$

  We can write:

  $$\begin{aligned} \overline{y}^b - \overline{y}^a &= (\overline{y}^b - \overline{y}^a_{counterf}) + (\overline{y}^a_{counterf} - \overline{y}^a) \\ &= \underbrace{(\overline{x}^b)'(\hat{\beta}^b - \hat{\beta}^a)}_{\text{within}} + \underbrace{(\overline{x}^b - \overline{x}^a)\hat{\beta}^a}_{\text{between}} \end{aligned} \quad (8)$$

  which is the alternative version of the Oaxaca decomposition.

- See column 4 × column 5 for the "within" difference based on this "group b - weighted" counterfactual.

Weighted Regression

- A nice thing about the "reweighted counterfactual" gap is that we can construct it very quickly using a *weighted regression*.
- Weights arise in applied problems where we have observations on y_i and x_i for a sample of individuals, and a set of "weights" $w_i \geq 0$ for each observation. The "weighted average" of y is

$$\overline{y}(w) = \frac{\sum_i w_i y_i}{\sum_i w_i}.$$

    - A lot of surveys have weights. For example, the March CPS sample has weights, which you can use to make the sample "nationally representative" of the US population. The sample has too many observations from small states and from certain cities to be representative without the weights.

Weighted Regression

- Many samples (including the CPS) are drawn from “sampling strata”, which are pre-determined sub-groups (or sub-areas) with known total populations. Observations from different strata are sampled with different probabilities - in this case the weight is just $w_i = 1/p_{s(i)}$, where $s(i)$ is the strata that i is drawn from, and $p_{s(i)}$ is the sampling probability in that strata:

$$p_{s(i)} = \frac{\text{num. sample units in strata } s(i)}{\text{known population in that strata}}$$

$$w_i = \frac{1}{p_{si}} : \text{ the num. people each sample unit represents}$$

Weighted Regression

The weighted OLS regression coefficient is:

$$\hat{\beta}(w) = \left(\sum_i w_i x_i x_i' \right)^{-1} \sum_i w_i x_i y_i$$

which can be computed in standard regression packages (Stata: pweight).

Reweighted Counterfactual

Back to the counterfactuals...

- We now show that we can construct the reweighted counterfactual mean for group b using a certain weighted average. For any observation from group b in category g, consider the weight:

  $$w_g = \frac{N_g^a}{N_g^b}$$

  Then

  $$\overline{y}^b_{counterf} = \sum_g \overline{p}^a_g \overline{y}^b_g = \frac{\sum_{i \in b} w_{g(i)} y_i}{\sum_{i \in b} w_{g(i)}}$$

  where $g(i)$ =the group that i belongs to. We'll prove this in two steps. First we look at the denominator. Then the numerator.

Reweighted Counterfactual

1. Denominator:

$$\begin{aligned}\sum_{i\in b} w_{g(i)} &= \sum_{i\in b}\frac{N^a_{g(i)}}{N^b_{g(i)}} = \sum_g \sum_{i\in g,b}\frac{N^a_g}{N^b_g} \\ &= \sum_g \frac{N^a_g}{N^b_g}\sum_{i\in g,b} 1 \\ &= \sum_g N^a_g = N^a \qquad (9)\end{aligned}$$

2. Numerator:

$$\begin{aligned}\sum_{i\in b} w_{g(i)} y_i &= \sum_{i\in b}\left(\frac{N^a_{g(i)}}{N^b_{g(i)}}\right) y_i = \sum_g \left(\frac{N^a_g}{N^b_g}\right)\sum_{i\in g,b} y_i \\ &= \sum_g N^a_g \overline{y}^b_g \\ &= N^a \sum_g \overline{p}^a_g \overline{y}^b_g \qquad (10)\end{aligned}$$

Reweighted Counterfactual

- So dividing by the denominator gets us

$$\frac{\sum_{i\in b} w_{g(i)} y_i}{\sum_{i\in b} w_{g(i)}} = \sum_g \overline{p}_g^a \overline{y}_g^b \tag{11}$$

  which means we can get the counterfactual mean for group b using a weighted mean. Note the weights "rebalance" the observations across categories. So if there are lots of $a's$ in category g, and not many $b's$, then we give more weight to the observations from the b group that we see in this category.

- We can also easily construct the difference

$$\overline{y}_{counterf}^b - \overline{y}^a = \frac{\sum_{i\in b} w_{g(i)} y_i}{\sum_{i\in b} w_{g(i)}} - \frac{\sum_{i\in a} y_i}{\sum_{i\in a} 1}$$

Reweighted Counterfactual

Estimate a weighted regression of y_i on a constant and a dummy group for group b:

$$y_i = \gamma_0 + \gamma_1 D_{bi} + \epsilon_i, \tag{12}$$

$$\text{weight } w_i = \begin{cases} w_{g(i)} & i \in b \\ 1 & i \in a \end{cases}$$

where $g(i)$: the category an observation i belongs to, and the weight for $i \in b$, $w_i = \frac{N^a_{g(i)}}{N^b_{g(i)}}$. The weight for all observations in group a is 1. The weighted OLS coefficient on the dummy for group b will equal the difference in the weighted means:

$$\hat{\gamma}_1 = \frac{\sum_{i \in b} w_i y_i}{\sum_{i \in b} w_i} - \frac{\sum_{i \in a} w_i y_i}{\sum_{i \in a} w_i} = \overline{y}^b_{counterf} - \overline{y}^a \tag{13}$$

Weights: DFL Method

- That then leaves us with the question: is there a “fast way” to get the weights $w_i = N^a_{g(i)}/N^b_{g(i)}$, $\forall i \in b$? The answer is yes. We will use a very important “trick” invented by John DiNardo, Thomas Lemieux, and Nicole Fortin – known by applied economists as the “DFL” method.
- The trick is this: if you considered a “crazy” regression on the combined sample of a and b:

$$m_i = x_i'\theta + \eta \tag{14}$$

where the dependent variable is $m_i = 1[i \in a]$ and $x_i' = (D_{1i}, D_{2i}...D_{Gi})$ what would you get for estimates of θ? From the first part of the lecture, we know that the OLS estimate of $\theta_g$will be:

$$\hat{\theta}_g = \frac{1}{N_g}\sum_{i\in g} m_i = \frac{N^a_g}{N^a_g + N^b_g}$$

Weights: DFL Method

So if we fit the “crazy” model predicting membership in group a, and took the predictions from this model, we’ll have:

$$\hat{m}_i = x_i'\hat{\theta} = \frac{N^a_{g(i)}}{N^a_{g(i)} + N^b_{g(i)}}$$

where $g(i)$ is the category that observation i fall into. Now form the weight for $i \in b$:

$$w_i = \frac{\hat{m}_i}{1 - \hat{m}_i} = \frac{N^a_{g(i)}}{N^b_{g(i)}} \tag{15}$$

This is the weight we want to form our counterfactual mean (and run the counterfactual gap regression).

- Since the object we are trying to predict in the “crazy regression” is the probability of membership in category g, most times people use a logit model rather than a linear probability model. (Will introduce logit regression later).