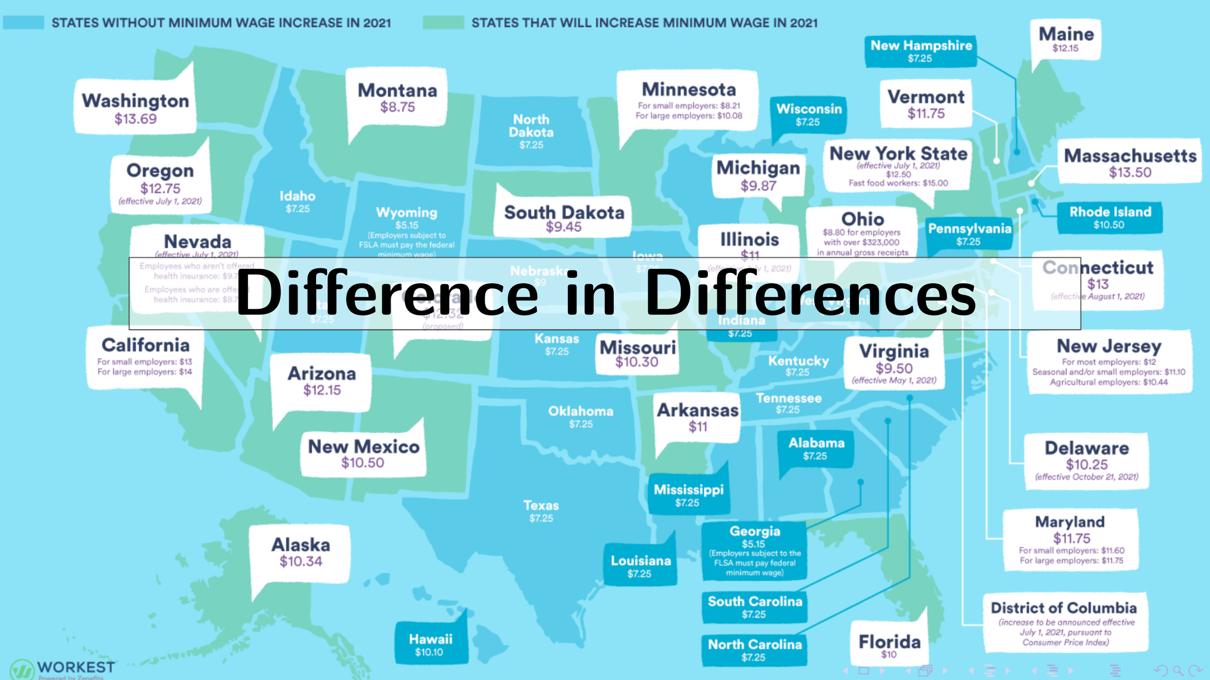
Difference in Differences



Today:

- Extend linear regression to make the most of natural experiment type data;
- Work through a case study applying difference in differences.

So far, we've used linear regression to:

1. Model a linear relationship between dependent variable y and independent variable x:

$$y = \beta_0 + \beta_1 x + \varepsilon;$$

2. Model a linear relationship between dependent variable y and many independent variables $x_1, x_2, x_3, \ldots$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \varepsilon;$$

3. Model a NON-linear relationship between dependent variable y and many independent variables x:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \ldots + \varepsilon;$$

4. Model relationships between independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon;$$

Motivation...

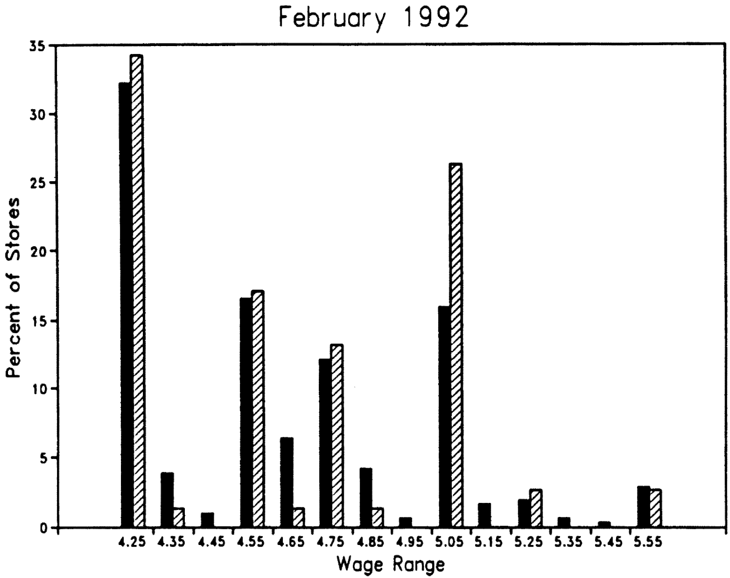
Does increasing the minimum wage decrease employment?

Motivation...

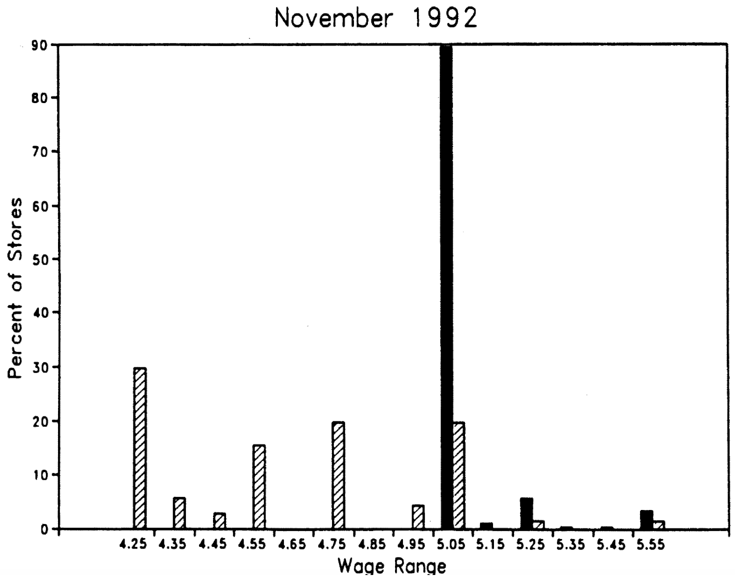
- In early 1990 NJ raised the minimum wage from \$4.25 to \$5.05 effective 1 April 1992 – highest minimum wage in the country;
- Notably, nearby eastern PA maintained a lower minimum wage;
- A **natural** within-subjects experiment;
  - NJ employers before/after;
  - PA employers before/after;
- Card and Krueger surveyed 473 fast food restaurants before and after law;
  - Low wage workers;
  - High response survey response rate.

TABLE 1—SAMPLE DESIGN AND RESPONSE RATES

		Stores in:	
	All	NJ	PA
Wave 1, February 15–March 4, 1992:			
Number of stores in sample frame:[a]	473	364	109
Number of refusals:	63	33	30
Number interviewed:	410	331	79
Response rate (percentage):	86.7	90.9	72.5
Wave 2, November 5–December 31, 1992:			
Number of stores in sample frame:	410	331	79
Number closed:	6	5	1
Number under rennovation:	2	2	0
Number temporarily closed:[b]	2	2	0
Number of refusals:	1	1	0
Number interviewed:[c]	399	321	78



February 1992
Percent of Stores
0
5
10
15
20
25
30
35
4.25
4.35
4.45
4.55
4.65
4.75
4.85
4.95
5.05
5.15
5.25
5.35
5.45
5.55
Wage Range


November 1992
Percent of Stores
90
80
70
60
50
40
30
20
10
0
4.25
4.35
4.45
4.55
4.65
4.75
4.85
4.95
5.05
5.15
5.25
5.35
5.45
5.55
Wage Range

Recall...

- If the researcher controls assignment of independent variables to observations (e.g. selection of treatment and control groups) then the data is experimental;

- A natural experiment is a situation in which:
  - The researcher does NOT control assignment of independent variables to observations...
  - ...but whatever does ends up creating a psuedo-randomly assigned treatment and control group anyway.

- So, can we extend linear regression to deal with observational data that has the characteristics of an experiment?

Difference in Differences

- Say we have data in which we measure the dependent variable for four groups:
    1. ...a ‘treatment’ group before treatment
    2. ...a ‘treatment’ group after treatment
    3. ...a ‘control’ group before no treatment
    4. ...a ‘control’ group after no treatment

Difference in Differences

- Say we have data in which we measure the dependent variable for four groups:
  1. ...a ‘treatment’ group before treatment (fast food restaurants NJ early 1992);
  2. ...a ‘treatment’ group after treatment (fast food restaurants NJ late 1992);
  3. ...a ‘control’ group before no treatment (fast food restaurants PA early 1992);
  4. ...a ‘control’ group after no treatment (fast food restaurants PA late 1992);

Difference in Differences

- Say we have data in which we measure the dependent variable for four groups:
  1. ...a 'treatment' group before treatment (fast food restaurants NJ early 1992);
  2. ...a 'treatment' group after treatment (fast food restaurants NJ late 1992);
  3. ...a 'control' group before no treatment (fast food restaurants PA early 1992);
  4. ...a 'control' group after no treatment (fast food restaurants PA late 1992);

- This data could look something like:

Unit ID	T (time)	G (state)	y (# of employees)
fast food restaurant 1	0 (early 1992)	1 (NJ)	10
fast food restaurant 1	1 (late 1992)	1 (NJ)	11
fast food restaurant 2	0 (early 1992)	0 (PA)	15
fast food restaurant 2	1 (late 1992)	0 (PA)	12
⋮	⋮	⋮	⋮
fast food restaurant 473	0 (early 1992)	1 (NJ)	25
fast food restaurant 473	1 (late 1992)	1 (NJ)	32

Difference in Differences

- Let's run a simple linear regression to analyze this data:

$$y = \beta_0 + \beta_1 T + \beta_2 G + \beta_{12}(TG) + \varepsilon$$

- As usual, in this linear regression equation:
  - y is the dependent variable – it comes from data;
  - T is an indicator that is a 1 for after treatment and 0 for before treatment;
  - G is an indicator that is 1 for treatment group and 0 for control group;
  - the β's are the effects – linear regression will learn these from data;
  - ε is noise – we don't get to observe this;

Difference in Differences

- Let's run a simple linear regression to analyze this data:

$$y = \beta_0 + \beta_1 T + \beta_2 G + \beta_{12}(TG) + \varepsilon$$

- As usual, in this linear regression equation:
  - y is the dependent variable – it comes from data;
  - T is an indicator that is a 1 for late 1992 and 0 for early 1992;
  - G is an indicator that is 1 for NJ and 0 for PA;
  - the β's are the effects – linear regression will learn these from data;
  - ε is noise – we don't get to observe this;

Difference in Differences

- Let's run a simple linear regression to analyze this data:

$$y = \beta_0 + \beta_1 T + \beta_2 G + \beta_{12}(TG) + \varepsilon$$

- Check out β_{12} – it can be shown (though we won't) that:

$$\beta_{12} = (\overline{y}_{11} - \overline{y}_{01}) - (\overline{y}_{10} - \overline{y}_{00})$$

  - $\overline{y}_{TG}$ = average dependent variable for group G at time T;

Difference in Differences

- Let's run a simple linear regression to analyze this data:

$$y = \beta_0 + \beta_1 T + \beta_2 G + \beta_{12}(TG) + \varepsilon$$

- Check out β_{12} – it can be shown (though we won't) that:

$$\beta_{12} = (\bar{y}_{11} - \bar{y}_{01}) - (\bar{y}_{10} - \bar{y}_{00})$$

  - $\bar{y}_{TG}$ = average dependent variable for group G at time T;
  - $\bar{y}_{11}$ = average dependent variable for the treatment group after treatment;
  - $\bar{y}_{01}$ = average dependent variable for the treatment group before treatment;
  - $\bar{y}_{10}$ = average dependent variable for the control group after treatment;
  - $\bar{y}_{00}$ = average dependent variable for the control group before treatment;

Difference in Differences

- Let's run a simple linear regression to analyze this data:

$$y = \beta_0 + \beta_1 T + \beta_2 G + \beta_{12}(TG) + \varepsilon$$

- Check out β_{12} – it can be shown (though we won't) that:

$$\beta_{12} = (\overline{y}_{11} - \overline{y}_{01}) - (\overline{y}_{10} - \overline{y}_{00})$$

  - $\overline{y}_{TG}$ = average dependent variable for group G at time T;
  - $\overline{y}_{11}$ = average employment for the NJ restaurants, late 92;
  - $\overline{y}_{01}$ = average employment for the NJ restaurants, early 92;
  - $\overline{y}_{10}$ = average employment for the eastern PA restaurants, late 92;
  - $\overline{y}_{00}$ = average employment for the eastern PA restaurants, early 92;

Difference in Differences

$$\beta_{12} = (\overbrace{\bar{y}_{11}}^{\text{NJ, late 92}} - \overbrace{\bar{y}_{01}}^{\text{NJ, early 92}}) - (\overbrace{\bar{y}_{10}}^{\text{PA, late 92}} - \overbrace{\bar{y}_{00}}^{\text{PA, early 92}})$$

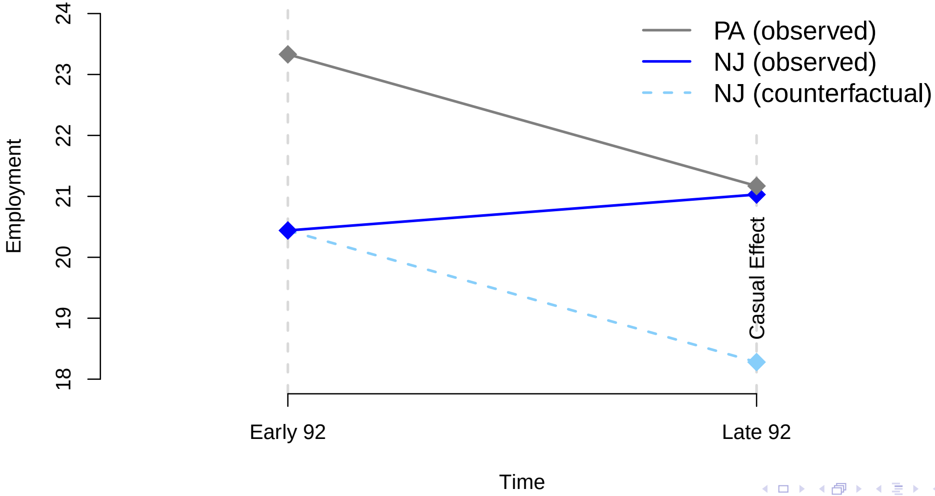
Difference in Differences

$$\beta_{12} = \underbrace{\Big(\overbrace{\bar{y}_{11}}^{\text{NJ, late 92}} - \overbrace{\bar{y}_{01}}^{\text{NJ, early 92}} \Big)}_{\text{Difference w/i Treatment}} - \underbrace{\Big(\overbrace{\bar{y}_{10}}^{\text{PA, late 92}} - \overbrace{\bar{y}_{00}}^{\text{PA, early 92}} \Big)}_{\text{Difference w/i Control}}$$

Difference in Differences

$$\beta_{12} = \underbrace{\underbrace{\Big(\overbrace{\bar{y}_{11}}^{\text{NJ, late 92}} - \overbrace{\bar{y}_{01}}^{\text{NJ, early 92}} \Big)}_{\text{Difference w/i Treatment}} - \underbrace{\Big(\overbrace{\bar{y}_{10}}^{\text{PA, late 92}} - \overbrace{\bar{y}_{00}}^{\text{PA, early 92}} \Big)}_{\text{Difference w/i Control}}}_{\text{Difference in Differences}}$$

Parallel Trend Assumption



So, does increasing minimum wage decrease employment?

	NJ	PA	Difference
Early 92	20.44	23.33	−2.89
Late 92	21.03	21.17	−0.14
Change	0.59	−2.16	**2.75**

$$\beta_{12} = (\overline{y}_{11} - \overline{y}_{01}) - (\overline{y}_{10} - \overline{y}_{00})$$

So, does increasing minimum wage decrease employment? Nope!

	NJ	PA	Difference
Early 92	20.44	23.33	−2.89
Late 92	21.03	21.17	−0.14
Change	0.59	−2.16	**2.75**

$$\beta_{12} = (\overline{y}_{11} - \overline{y}_{01}) - (\overline{y}_{10} - \overline{y}_{00})$$
$$\beta_{12} = (21.03 - 20.44) - (21.17 - 23.33) = 2.75.$$

Why should we care?

Experiments are the gold standard in establishing causality and we can use linear regression to model the data they generate.