REGRESSION ANALYSIS AND CAUSALITY WITH R

Difference-in-Differences

João Cerejeira[1] Miguel Portela[1,2,3]

[1]NIPE – UMinho
[2]IZA, Bonn
[3]Banco de Portugal

October 26, 2021

Introduction: John Snow study

Cholera in 19th C in England and Wales

1831-32: severe outbreak (around 20,000 dead) across many British towns and cities including London. England's Cholera Prevention Act followed the flawed and later repealed Quarantine Act of 1825.

1848-49: another severe outbreak (around 10,000 dead in 3 months in London; around 53,000 dead in England and Wales).

1853-54: there were a few cases in 1853 and the first half of 1854. Then, after August 31, 1854 there was the "most terrible outbreak of cholera which ever occurred in this kingdom" (Dr. John Snow, 1855).

1865-66: the fourth pandemic (1863-1879) only affected areas served by the East London Waterworks Company.

Prevailing thinking

Miasma theory was the prevailing 19th C dogma of public and medical community alike.

The theory of indirect and airborne transmission held that cholera was caused by the smell of the bad air, miasmata, a poisonous vapour with suspended particles of decaying matter and a foul smell.

At the time miasma theory made sense to most as disease and epidemics were concentrated in poor, filthy and foul-smelling city neighborhoods.

Physicians tended to believe that cholera was a condition of the blood.

Some believed that cholera was related to altitude.

Most believed it was not contagious.

Introduction: John Snow study

By the arrival of the 19th century, the River Thames had become the most contaminated river in the world. Water supplied to households by competing private companies.

The water companies did not filter or treat their water in 1848-49 or 1853-54.

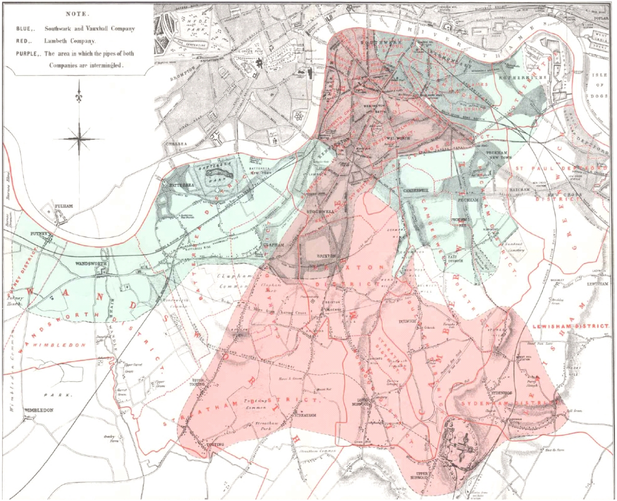
Companies competed for customers house by house, resulting in overlap between the areas supplied by the different companies, so sometimes different companies supplied households in same street

In south London two main companies:

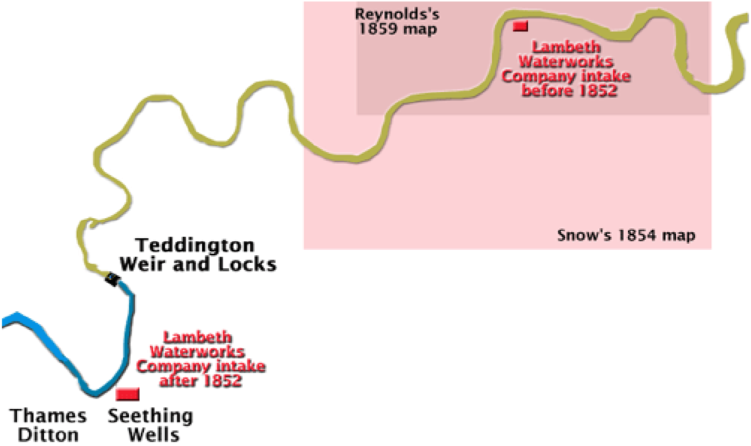
- Lambeth Company
- Southwark and Vauxhall Company

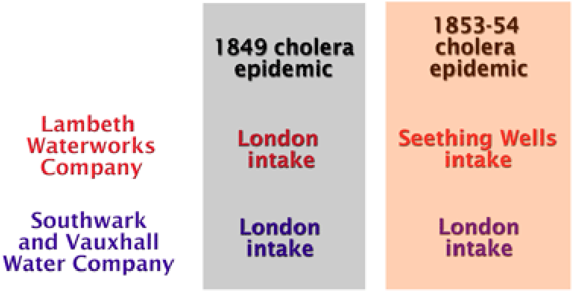
During 1845-52, both companies supplied people living in the same area with polluted water drawn from the River Thames.

During the 1848-49 cholera epidemic in London, the "water of the...Southwark [and] Vauxhall, and Lambeth [companies], is by far the worst of all those who take their supply from the Thames." - Snow, John. Communication of Cholera, 1855


NOTE.
BLUE,. Southwark and Vauxhall Company
RED,. Lambeth Company.
PURPLE,. The area in which the pipes of both Companies are intermingled.
HYDE PARK
BROMPTON
CHELSEA
SOUTHWARK
RIVER THAMES
ROTHERHITHE
ISLE OF DOGS
BERMONDSEY
WALWORTH
KENNINGTON
PECKHAM NEW TOWN
ST PAUL DEPTFORD
HATCHAM
NEW CROSS DISTRICT
BATTERSEA
BATTERSEA PARK
FULHAM
PUTNEY
WANDSWORTH
WANDSWORTH DISTRICT
STOCKWELL
CAMBERWELL
CAMBERWELL DISTRICT
PECKHAM
PECKHAM RYE
CLAPHAM
BRIXTON
BRIXTON DISTRICT
Clapham Common
EAST DULWICH
DULWICH
LEWISHAM
LEWISHAM DISTRICT
GREENWICH
WIMBLEDON
WIMBLEDON PARK
UPPER TOOTING
TOOTING
STREATHAM
STREATHAM DISTRICT
Streatham Common
LOWER NORWOOD
UPPER NORWOOD
SYDENHAM
SYDENHAM DISTRICT
Putney Heath
Wimbledon Common

Then in 1852 the Lambeth Waterworks Company move its water intake to a cleaner location upriver, while the Southwark and Vauxhall Water Company left its intake in the same contaminated location.


Reynolds's 1859 map
Lambeth Waterworks Company intake before 1852
Snow's 1854 map
Teddington Weir and Locks
Lambeth Waterworks Company intake after 1852
Thames Ditton
Seething Wells



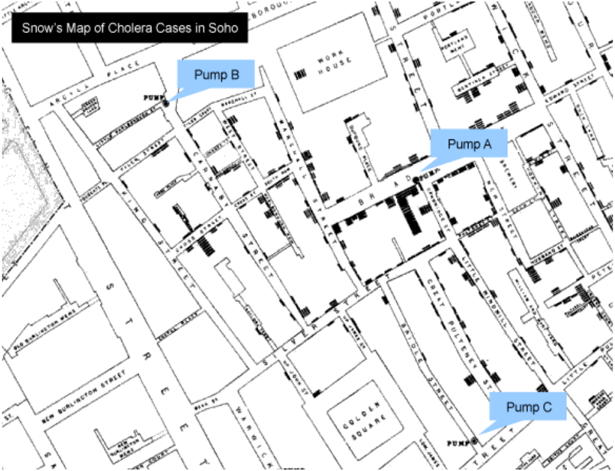
	1849 cholera epidemic	1853-54 cholera epidemic
Lambeth Waterworks Company	London intake	Seething Wells intake
Southwark and Vauxhall Water Company	London intake	London intake

When the next cholera epidemic appeared in 1853-54, some neighbors were unknowingly receiving cleaner water from the Lambeth company while others consumed more polluted water from the Southwark and Vauxhall company.

This change in location of the water intake was the basis for a natural experiment, which allowed Dr. Snow to compare mortality patterns by water source and strengthen his hypothesis regarding the transmission of cholera.

Snow started with descriptive epidemiology, which is a way of organising and summarising health-related data according to person, place and time (**who? where? when?**) with his questionnaire.

He obtained information on the number of cholera deaths (the numerator) and the number of households supplied by water (the denominator). Snow used death certificates for the number of deaths, company reports for the source of water as well as individual enquiry. This allowed him to describe the number of cases of cholera in different areas relative to the size of the population at risk.



When John Snow made this map of the Golden Square area, with a line showing where each person who had had a fatal case of cholera had lived, and the position of the public water pumps, he noticed clusters.

Snow found explanations for the exceptions within that radius that transformed the apparent inconsistencies into evidence supporting his theory. Outliers can supply important clues.

- None of the workers at the Broad Street brewery had cholera: they were very close to pump A, but tended to drink beer rather than water.

- Likewise the Poland Street Workhouse only recorded five deaths among its inmates.

- An elderly widow in West Hampstead (an area some distance away, which was free of cholera) liked the taste of Broad Street water, so she had a bottle brought to them every day from the pump. The fact that she and her visiting niece died of cholera was in Snow's view "the most conclusive".

Introduction: John Snow study (with numbers!)

In 1853/54 cholera outbreak:
Death Rates per 10000 people by water company
- Lambeth: 10
- Southwark and Vauxhall: 150.

Might be water but perhaps other factors. Snow compared death rates in **1849 epidemic**
- Lambeth: 150
- Southwark and Vauxhall: 125

A good estimate of effect of clean water

.

	1849	1853/54	Difference
Lambeth (treat)	**150**	**10**	**-140**
Vauxhall and Southwark (control)	*125*	*150*	*25*
Difference	-25	140	-165

Snow became convinced that the Broad Street pump was the source of the outbreak, and thus that transmission of cholera was indirect and carried by water (vehicle-borne rather than air-borne). Thus, the vehicle had to be decontaminated or eliminated. Indeed, later investigations showed that the superficial pump was probably contaminated by infected material, fecal matter.

Based on his detailed study which also noted the pump's proximity to a sewer, he persuaded the local authorities to implement a control measure immediately – and so the pump handle was removed on the 8th September 1854.

Basic idea of Differences-in-Differences

Have already seen idea of using differences to estimate causal effects;

Treatment/control groups in experimental data;

Often we would like to find ‘treatment’ and ‘control’ group who can be assumed to be similar in every way except receipt of treatment;

This may be very difficult to do.

Assumptions

Assume that, in absence of treatment, difference between ‘treatment’ and ‘control’ group is constant over time;

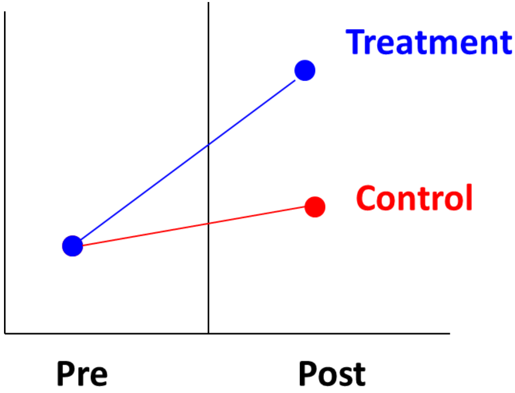
With this assumption can use observations on treatment and control group pre- and post-treatment to estimate causal effect.

$$\widehat{\delta} = (\overline{Y}_{T,A} - \overline{Y}_{C,A}) - (\overline{Y}_{T,B} - \overline{Y}_{C,B})$$

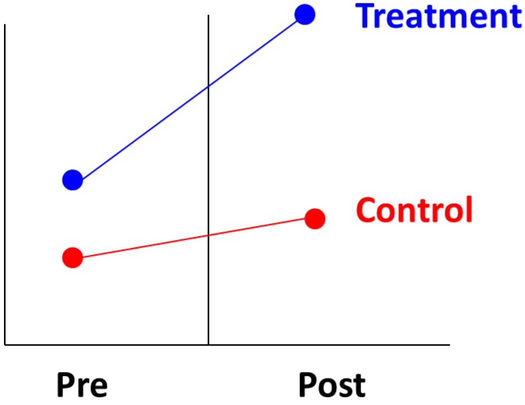
Idea:
- Difference pre-treatment is ‘normal’ difference $(\overline{Y}_{T,B} - \overline{Y}_{C,B})$;
- Difference post-treatment is ‘normal’ difference + causal effect $(\overline{Y}_{T,A} - \overline{Y}_{C,A})$;
- Difference-in-difference is causal effect.

Randomization, Graphically

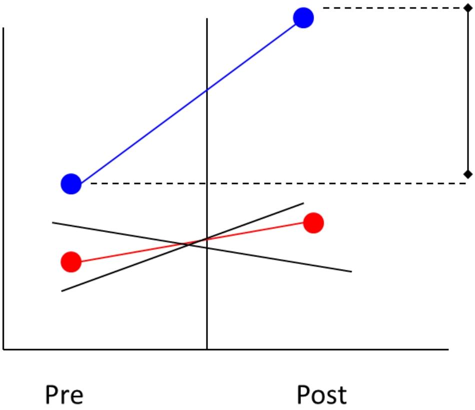
$$(\overline{Y}_{T,B} - \overline{Y}_{C,B}) = 0$$



Differences-in-Differences, Graphically

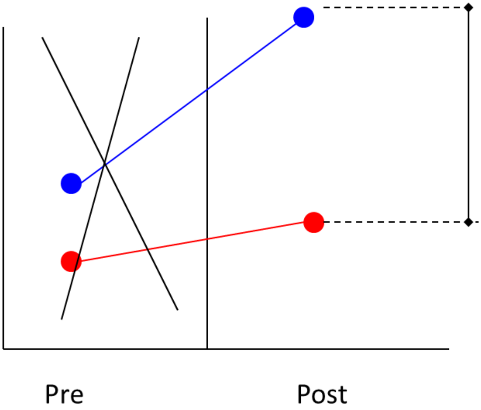


Differences-in-Differences, Graphically



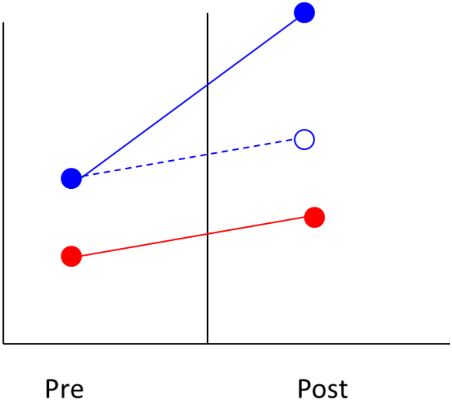
Effect of program using only pre- & post- data from T group (ignoring general time trend).

Differences-in-Differences, Graphically

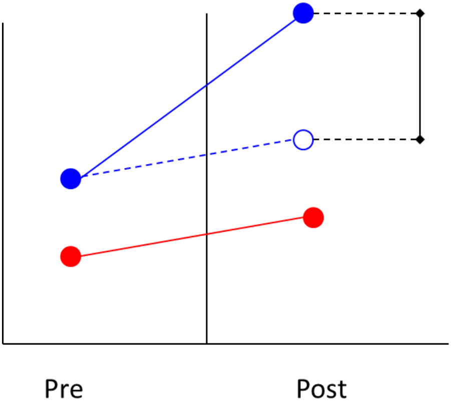


Effect of program using only T & C comparison from post-intervention (ignoring pre-existing differences between T & C groups).

Differences-in-Differences, Graphically



Differences-in-Differences, Graphically



Effect of program difference-in-difference (taking into account pre-existing differences between T & C and general time trend).

The D-in-D estimate

This is simply the difference in the change of treatment and control groups so can estimate as:

$$\Delta Y_i = c + dTreated_i + u_i \tag{1}$$

where *Treated* takes value one for treated individuals and zero for controls.

The D-in-D estimate

Alternatively, you can run:

$$Y_{it} = a + bTreat_i + cPost_t + d(Treated_{i,t} \times Post_{i,t}) + \varepsilon_{it} \qquad (2)$$

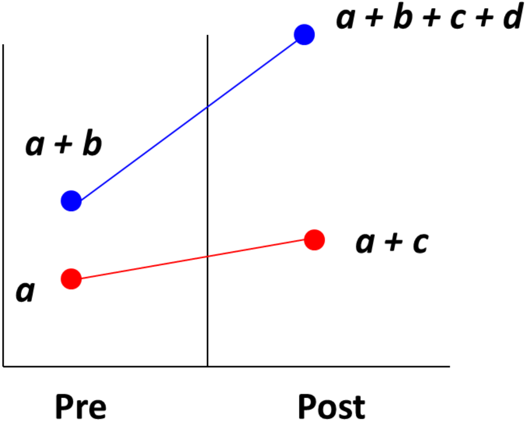
where t is equal to one for individuals observed after treatment and zero for individuals observed before treatment.

It will give different standard errors, because ‘levels’ version will assume residuals are independent – unlikely to be a good assumption.

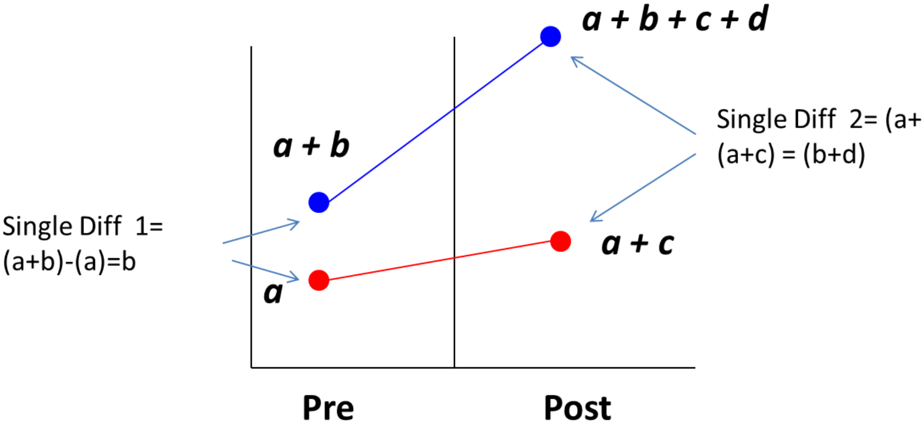
Can deal with this by using clustering option.

Putting Graph & Regression Together

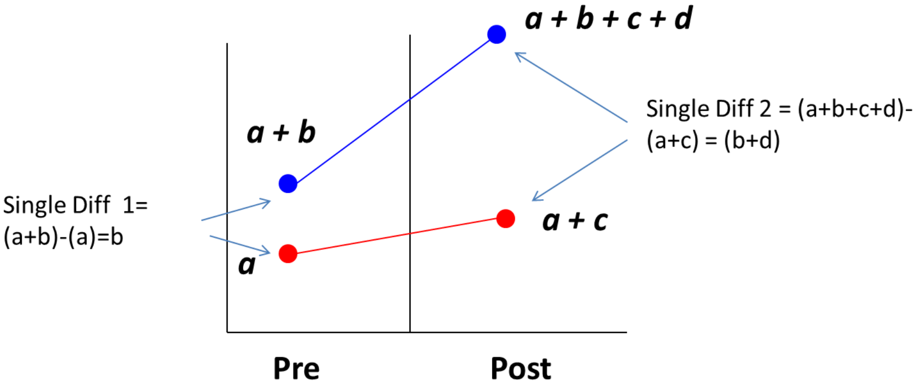
$$Y_{it} = a + bTreat_i + cPost_t + d(Treated_{i,t} \times Post_{i,t}) + \varepsilon_{it} \qquad (3)$$



Putting Graph & Regression Together



Diff-in-Diff=(Single Diff 2-Single Diff 1)=(b+d)-b=d



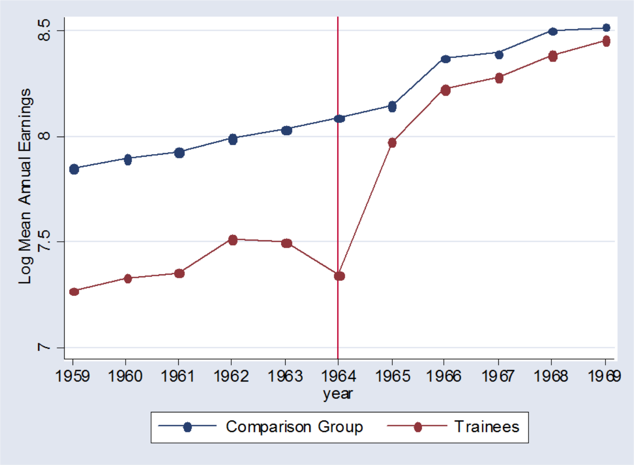
Single Diff 2 = (a+b+c+d)-(a+c) = (b+d)

Single Diff 1= (a+b)-(a)=b

The D-in-D estimate

- D-in-D estimate is d;
- Note: assumes trends in outcome variables the same for treatment and control groups;
- This is not testable;
- With two periods can get no idea of plausibility but can with more periods.

Ashenfelter's Dip



Example: Card & Krueger (1994)

Suppose you are interested in the effect of minimum wages on employment (a classic and controversial question in labour economics).

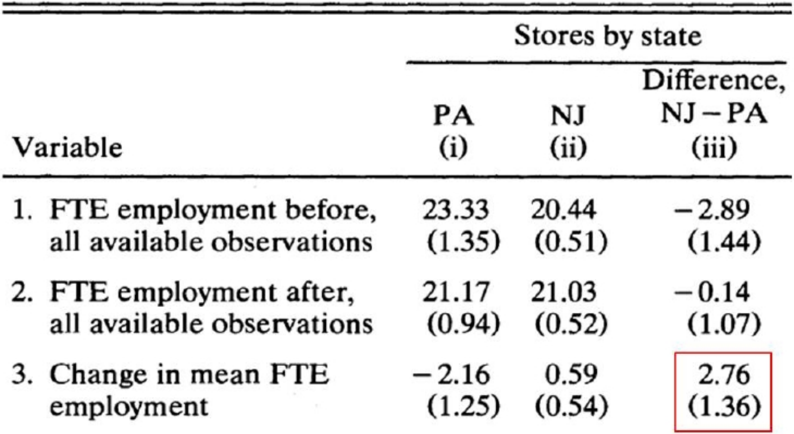
In a competitive labour market, increases in the minimum wage would move us up a downward-sloping labour demand curve and employment would fall.

Card & Krueger (1994) analyse the effect of a minimum wage increase in New Jersey using a dierences-in-dierences methodology.

In February 1992 NJ increased the state minimum wage from $4.25 to $5.05. Pennsylvania's minimum wage stayed at $4.25.

New York
Pennsylvania
Rhode Island
Connecticut
New Jersey

They surveyed about 400 fast food stores both in NJ and in PA both before and after the minimum wage increase in NJ.



	Stores by state		
Variable	PA (i)	NJ (ii)	Difference, NJ − PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Summary

A very useful and widespread approach;

Validity does depend on assumption that trends would have been the same in absence of treatment;

Can use other periods to see if this assumption is plausible or not;

Uses 2 observations on same individual – most rudimentary form of panel data.