Randomized Controlled Trial and Difference-in-Differences

Causal Inference

- Focus on **research design**
  - Methods are important but secondary
- Good design ➔ credible results
- Assumptions are your enemies
  - They undermine credibility of results
  - How can they be minimized?
  - Carefully defend the ones you can't avoid
- Emphasis on intuition
  - Much of the math will be easy
  - the intuition sometimes less so

Notation

- "Dependent" or "outcome" variable Y → eg. Recovery time in F B.
- Main "Independent" or "predictive" variable X_1
- (Maybe) some "control" variables or "covariates" $\boldsymbol{X}_{-1} = (X_2, X_3, \ldots X_K)$
- **boldface** = vector or matrix
- Sample size N, observations indexed by i

→ Everything else apart from X_1

Notation, Cont'd

- **Major** simplification:

  Replace X_1 with binary W
  - Some units are "treated" ($w_i = 1$)
  - Others are "control" ($w_i = 0$)
  - We will call W the treatment indicator (or dummy)
- What about multivalued and "continuous" treatments?
  - Multi-valued = straightforward extension, just clunky → More than 2 groups
  - Continuous = at research frontier

    ↳ Treatment scale is varying

    eg size changes from 8 → 8.11

    vs. " " " 8 → 12

Major conceptual move: Potential outcomes

- Define: Every unit i has two "potential outcomes"
  - $y_i(w = 1)$:= outcome if treated [shorthand y_{i1}]
  - $y_i(w = 0)$:= outcome if control [shorthand y_{i0}]
- One of these is observed; one is not — for the same customer !!
  - Missing outcome is often called "counterfactual"
- But more useful to think of it as "real", just not observed
  - Much as in ordinary regression, we observe the **sample**
  - But the superpopulation from which we take the sample is not observed (and may not exist)

Observed y_i is a mix of y_{i1}, y_{i0}

- Define: $y_i^{obs} := w_i y_i(1) + (1 - w_i) y_i(0)$
- And: $y_i^{mis} := w_i y_i(0) + (1 - w_i) y_i(1)$
- Part of why regression is often misleading:
  - Tempts you to treat y_i^{obs} as a real quantity
  - It's not; it's only the mixture of y_{i0} and y_{i1} that you happen to observe
- Regression is really:

$$Y^{obs} = \alpha + \beta w + \boldsymbol{\gamma X}_{-1} + \epsilon$$

- Mixture in; mess out
  - Except in special cases

You either observe the individual if he is in treatment/control group

The groups will be "counter-factual"

This data do NOT there

Simple Interpretation: A person will only appear in one group at any point.

If w_i is NOT randomized, then β is just correlation.

With no missing data, this is easy

- Want to know: will treatment affect outcome?
  - Will ΔW cause ΔY? That is: is $y_{i1} \neq y_{i0}$?
  - **Define:** Treatment effect: $\tau_i = (y_{1i} - y_{0i})$

You only observe 1 of them

- Rubin's central insight: Causal inference is a missing data problem:
  - Need to credibly estimate the missing potential outcomes
  - The "fundamental problem of causal inference" [Holland, 1986]
  - Do that and you're done

BUT the missing data is NOT random.

- OK, so maybe not that easy . . .
  - But we have a clear goal
  - And a centrally nonparametric core research design

Second **major** conceptual move, and complication

- Heterogeneous treatment effects
  - Treatment effect: $\tau_i = (y_{1i} - y_{0i})$ depends on characteristics of unit i
  - Some characteristics are observed: $\mathbf{x}_i$
  - Some are not observed (omitted): $\mathbf{u}_i$

The (partly missing) design matrix is. . .

Known variables about the customer

Outcome if treated	Outcome if control	Treatment effect	Treatment dummy	First covariate		Last Covariate	Unobserved covariates
y_{11}	y_{10}	τ_1	w_1	x_{12}	. . .	x_{1K}	$\mathbf{u}_K$
y_{21}	y_{20}	τ_2	w_2	x_{22}	. . .	x_{1K}	$\mathbf{u}_{1K}$
y_{31}	y_{30}	τ_3	w_3	x_{32}	. . .	x_{3K}	$\mathbf{u}_{3K}$
y_{41}	y_{40}	τ_4	w_4	x_{42}	. . .	x_{4K}	$\mathbf{u}_{4K}$
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
y_{N1}	y_{N0}	τ_{N0}	w_N	x_{N2}	. . .	x_{NK}	$\mathbf{u}_{NK}$

Lots of them

All treatment data is missing because one of control / treatment is always missing

red = not observed

Want to know: are the τ_i's ≠ 0?

This is a hard problem

- Regression, applied to the partial data we observe, won't get us there
  - Except in special cases
- We need research designs that let us:
  - credibly estimate the missing potential outcomes
  - not worry about the omitted covariates

Core assumption 1: manipulation

- w_i is manipulable
- Counterexample: Effect of gender on income
  - Observe y_{i1} = income if male
  - Want to impute y_{i0} = income if female
  - All else about you is the same (*ceteris paribus*)
- Not achievable
  - "no causation without manipulation" [Holland, 1986]
  - If you were dictator, with infinite resources [and no morals], could you design an experiment to answer the question you have in mind? [Dorn, 1953]

Core Assumption 2 (& 3): SUTVA

- "Stable Unit Treatment Value Assumption" (SUTVA)
- Really two separate assumptions:
  - Only one kind of treatment (w = 0 or 1)
    - Can be relaxed (multivalued treatments)
  - responses of different units are **independent:**

$$\tau_i \perp\!\!\!\perp (\tau_j, w_j) \; \forall \; j \neq i$$

This is SUTVA

- If not satisfied, no easy answers
  - Can sometimes aggregate to higher level
    - E.g., study classrooms, not students
  - Or model spillovers

No spillover effect -
ie Me staying more on FB should not cause another friend to stay more (maybe because I chat with them)

Some common estimands and estimates

Estimand	Estimator (if know τ_i)
$ATE = E[\tau]$	$\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N} \tau_i$
$ATT = E[\tau \mid w = 1]$	$\widehat{ATT} = \frac{1}{N_t} \sum_{i:w_i=1} \tau_i$
$ATC = E[\tau \mid w = 0]$	$\widehat{ATC} = \frac{1}{N_c} \sum_{i:w_i=0} \tau_i$
$\tau_{0.5}$ = median treatment effect	$\widehat{\tau_{0.5}} = \alpha$: 50% $< \alpha$ 50% $\geq \alpha$
$\tau_{0.25}$ = 25th percentile (0.25 quantile)	$\widehat{\tau_{0.25}} = \alpha$: 25% $< \alpha$ 75% $\geq \alpha$
Conditional: $ATT_X(x) = E[\tau \mid w = 1, X = x]$	$\widehat{ATT_X}(x) = \frac{1}{N_{tx}} \sum_{i:\, w_i=1, X_i=x} \tau_i$

Avg. treatment effect on treatment group

Control group

All are equal if randomized control trial is implemented

Treatment effect on multiple groups

Toy example (N = 4, no covariates)

Unit i	y_i^{obs}	w_i	y_{i1}	y_{i0}	τ_i
1	3	1	3	?	?
2	1	1	1	?	?
3	0	0	?	0	?
4	1	0	?	1	?

What are:

$$\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N}\tau_i = \frac{1}{N}\sum_{i=1}^{N} y_i(1) - y_i(0)$$

$$\widehat{ATT} = \frac{1}{N_t}\sum_{i:w_i=1}\tau_i = \frac{1}{N_t}\sum_{i:w_i=1} y_i(1) - y_i(0)$$

$$\widehat{ATC} = \frac{1}{N_c}\sum_{i:w_i=0}\tau_i = \frac{1}{N_c}\sum_{i:w_i=0} y_i(1) - y_i(0)$$

Without more information, we don't know.

Apply magic (insert missing potential outcomes)

Unit i	y_i^{obs}	w_i	y_{i1}	y_{i0}	τ_i
1	3	1	3	0	3
2	1	1	1	0	1
3	0	0	0	0	0
4	1	0	1	1	0

Not necessarily randomized

Never really happens

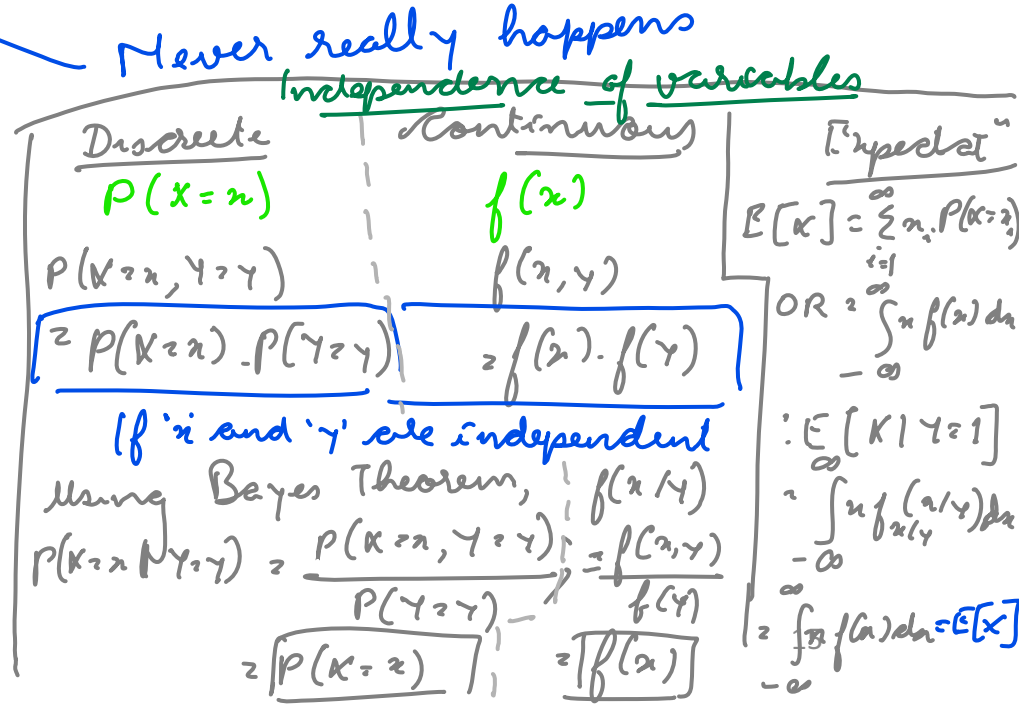
Can now compute:

$$\widehat{ATE} = (3 + 1 + 0 + 0)/4 = 1$$

$$\widehat{ATT} = (3 + 1)/2 = 2$$

$$\widehat{ATC} = (0 + 0)/2 = 0$$

Not the same (and in general, won't be)



Next major concept: "Assignment mechanism"

- Process (perhaps unknown) for determining which units are treated
- For our example, is assignment ***random***?
- Doesn't look that way! Units are:
  - treated if treatment "helps"
  - control if treatment is ineffective
- Still, assignment in superpopulation could be random
  - Our toy sample could be non-representative

Compare naïve estimator using observed values

Unit i	y_i^{obs}	w_i	y_{i1}	y_{i0}	τ_i
1	3	1	3	?	?
2	1	1	1	?	?
3	0	0	?	0	?
4	1	0	?	1	?

$$\overline{y_1^{obs}} = (3+1)/2 = 2$$

$$\overline{y_0^{obs}} = (0+1)/2 = 0.5$$

Estimator: $\hat{\tau}_{naive} := \overline{y_1^{obs}} - \overline{y_0^{obs}}$

Estimate: $\hat{\tau}_{naive} = 2 - 0.5 = 1.5$

But what's the **estimand?** Not ATT, ATC, or ATE

NOT a correct way Why?
Let's take an example:
If w_i is when people drink coffee and y_i is energy level, the problem here is that ONLY PEOPLE WHO HAVE POSITIVE EFFECT OF COFFEE will drink it!!!

What went wrong? Selection bias

- Units not randomly chosen for treatment.
- Let's see what $\hat{\tau}^{naive}$ converges to:

Add and subtract $E[y_0|w=1]$

$$\hat{\tau}_{naive} := \overline{y_1^{obs}} - \overline{y_0^{obs}} \xrightarrow{p}$$

$$E[y_1|w=1] - E[y_0|w=0]$$

$$= E[y_1 - y_0|w=1] + \{E[y_0|w=1] - E[y_0|w=0]\}$$

$$= ATT \quad + \quad \text{Baseline bias}$$

What is the difference when people are treated?

Individuals in control group when they are in control

This should be 0. How? If $w=0$ and $w=1$ are the same.

Randomized control trial

Baseline bias := diff. between treated and controls if neither were treated

Often called "**selection bias**" (when units self-select into treatment)

In our example: $\widehat{ATT} = 2$

Baseline Bias $= -0.5$

$\hat{\tau}_{naive} = 2 - 0.5 = 1.5$

Would regression help? No.

Regression uses only observed values, regress y on w:

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
w	**1.5**	1.118034	1.34	0.312	-3.3105	6.3105
_cons	.5	.7905694	0.63	0.592	-2.901546	3.901546

Regression coefficient $\hat{\beta}$ estimates τ_{naive}!

Regression:

Separate concepts of ATE, ATT, ATC have no meaning:

We (silently) assume **homogeneous treatment effects:**

same τ for all units (ATE = ATT = ATC)

Also (silently) assume: **no baseline bias**

Can also measure bias of τ_{naive} relative to ATC

$$\hat{\tau}_{naive} = \overline{y_1^{obs}} - \overline{y_0^{obs}} \xrightarrow{p}$$

$$E[y_1|w=1] - E[y_0|w=0]$$

$$= \{E[y_1|w=1] - E[y_1|w=0]\} + E[y_1 - y_0|w=0]$$

$$= \quad \text{Outcome bias} \quad + \text{ATC}$$

This should be 0.

Add and subtract $E[y_1|w=0]$

Outcome bias = difference between treated and controls if both were treated. In our example:

$$\widehat{ATC} = 0$$

$$\text{Outcome bias} = 1.5$$

$$\hat{\tau}_{naive} = 1.5 + 0 = 1.5$$

Can also decompose outcome bias

Intuition:

Outcome bias = Baseline bias + "Treatment heterogeneity"

Some algebra:

Outcome bias $-$ Baseline bias $=$

$$= \{E[y_1|w=1] - E[y_1|w=0]\} - \{E[y_0|w=1] - E[y_0|w=0]\}$$

$$= \{E[y_1|w=1] - E[y_0|w=1]\} + \{E[y_0|w=0] - E[y_1|w=0]\}$$

$$= E[y_1 - y_0|w=1] - E[y_1 - y_0|w=0] =$$

$$= \text{ATT} - \text{ATC} := \textbf{Treatment heterogeneity}$$

In our example: Treatment heterogeneity $= 2.0$

$$\tau_{naive} = 0.0 + (-0.5) + 2.0 = 1.5$$

Summary: causal inference as missing data problem

- We have a partly observed "design matrix"
  - And treatment heterogeneity (ATT ≠ ATC ≠ ATE)
- Running a simple regression won't work
  - In expectation: gives τ_{naive}

$$= ATT + \text{baseline bias}$$

$$= ATC + \text{baseline bias} + \text{treatment heterogeneity}$$

$$\neq \text{ATE either}$$

Randomized Controlled Trial as Gold Standard

- Suppose we have **random** assignment across whole sample*:*

$$w_i \perp \left(j, y_{j1}, y_{j0}, \boldsymbol{x}_j, \boldsymbol{u}_j\right) \forall\, j \Leftrightarrow P[w_i = 1] = p \;\forall\, i$$

[If truly random, w $\perp$ (everything incl. **u)**]

- Often choose $p = 0.5$ (but don't need to)
  - Need probabilistic assignment: $0 < p < 1$
  - Higher variance if p near 0 (few treated units) or 1 (few control units)

Why Does Randomization Help?

Baseline bias $= E[y_0|w=1] - E[y_0|w=0]$

$= E[y_0] - E[y_0] =$ **0, by randomization**

Randomizing w does NOT predict anything for $E(y_0)$ i.e. conditional w is actually unconditional

Outcome bias $= E[y_1|w=1] - E[y_1|w=0]$

$= E[y_1] - E[y_1] =$ **0, by randomization**

Actually by independence !!

E[Treatment heterogeneity] = 0 **(by randomization)**

- treated and controls are similar in expectation

ATE = ATT = ATC **(by randomization)**

- treated and controls are similar in expectation

Regression now works: $\widehat{\tau_{naive}}$ is unbiased for ATE!

Estimator given randomization

- Estimator for ATE, ATT, ATC using analogy principle:
  - Unbiased (see prior slide) and consistent

$$\hat{\tau}_{naive} = \overline{y_1^{obs}} - \overline{y_0^{obs}}$$
$$= [\frac{1}{N_t} \sum_{i:w_i=1}^{N} y_i^{obs}] - [\frac{1}{N_c} \sum_{i:w_i=0}^{N} y_i^{obs}]$$

Intuition: Why does randomization work?

- Random assignment → treated and controls are similar on average.
  - So in estimating $ATE = E[y_1 - y_0]$, we introduce no bias by estimating y_1 using only treated units and y_0 using only control units
- Randomization → ATE = ATT = ATC
  - Treatment effects can still be heterogenous
  - Without covariates, poor estimate of τ_i for particular unit *i*:
    - For treated units, crude estimate

$$\hat{\tau}_i = y_{i1} - \overline{y_0}$$

    - Similarly for control units

$$\hat{\tau}_i = \overline{y_1} - y_{i0}$$

Heterogeneous treatment effects

- For better estimates, estimate "**response surfaces**":
  - $\widehat{y_1}(\boldsymbol{x})$ using treated units
  - $\widehat{y_0}(\boldsymbol{x})$ using controls.
  - Then estimate "**effect surface**": $\widehat{\tau(\boldsymbol{x})} = \widehat{y_1}(\boldsymbol{x}) - \widehat{y_0}(\boldsymbol{x})$
- In practice, harder than it sounds:
  - For some binary covariates (e.g., men vs. women), can use block randomization
    - Estimate ATE_{men} and $\text{ATE}_{\text{women}}$
  - Otherwise, quickly run into "curse of dimensionality"

So you have a randomized experiment:

- No: You **think** you have one
- **Always:** check for "covariate balance"
- For each control variable, can check:
  - Normalized difference in means
  - *t*-statistic for difference in means
  - Difference in (normalized) standard deviations
  - Kolmogorov-Smirnov statistic
  - Kernel density plots

→ 'Propensity score matching' can be used if these tests show that mean of control ≠ mean of treatment of control variables

Covariate balance tests

- Are fairly standard (and should be) for:
  - Randomized experiments
  - Regression discontinuity (RD)
  - Pure observational studies
- Are not (**but should be**) for:
  - Difference-in-differences
  - Binary instrumental variables
    - Can “dichotomize” non-binary instruments
- These other methods all seek to approach randomized experiments
  - covariate balance: one test for how well they succeed

“Combined” designs

- If covariate balance is imperfect, **fix it!**
  - Or, more realistically, improve it
  - Often feasible, if balance is not too bad.
  - Variety of “balancing methods”
    - Trimming
    - Matching
    - Inverse propensity weighting

Estimands and statistical significance for randomized experiments

Fisher's sharp null: H_0: $\tau_i = 0$ for each unit i

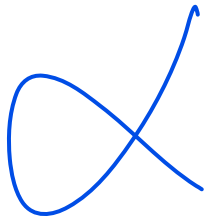
- Implication if true: we know **both** potential outcomes!
  - We can complete the "design matrix" under the null
  - And test for every difference between treated and controls you can think of
    - Using randomization methods
    - Not just differences in means or medians (for which we can compute standard errors)

Under Fisher's H_0: (imputed) data look like . . .

Outcome if treated	Outcome if control	Treatment dummy	Treatment effect	Control var. 1		Control var. k-1
y_{11}	y_{11}	w_1	$\tau_1 = 0$	x_{21}	. . .	x_{k1}
y_{12}	y_{12}	w_2	$\tau_2 = 0$	x_{22}	. . .	x_{k2}
y_{03}	y_{03}	w_3	$\tau_3 = 0$	x_{23}	. . .	x_{k3}
Y_{04}	y_{04}	w_4	$\tau_4 = 0$	x_{24}	. . .	x_{k4}
. . .	. . .	. . .		. . .	. . .	. . .
y_{0n}	y_{0n}	w_n	$\tau_n = 0$	X_{2n}	. . .	x_{kn}

Purple = not observed, but imputed under H_0

Assume: **unobserved** potential outcome = observed outcome



Statistical significance: ATE $\neq 0$

- Fisher's sharp null is extreme
- We could ask instead: is **average** effect $\neq 0$
  - Neyman's null
  - Intuition: Weaker null ➔ higher standard errors (lower *t*-stats)
- By how much?
  - Empirical answer, not much.
- Construct a *t*-test (Jerzy Neyman's approach)

Neyman's *t*-test for ATE ≠ 0

- Upper bound

$$V_{Neyman} = \frac{s_t^2}{N_t} + \frac{s_c^2}{N_c}$$

variances

$H_0: ATE = 0$

$H_A: ATE \neq 0$ (or $>$, $<$)

- Leads to standard two-sample t-test

$$t_{Neyman} = \frac{ATE}{\sqrt{\frac{s_t^2}{N_t} + \frac{s_c^2}{N_c}}}$$

; (t_n = ATE − 0 / $\sqrt{var}$)

means

- In practice, only slightly conservative

Major themes for randomized trials

- Key nature of "assignment mechanism"

  prob(w=1|y_0, y_1, **x**)

- Random assignment:

  w $\perp\!\!\!\perp$ (y_0, y_1, **x, u**)

- Regression: model for the (**observed**) data
- Causal inference: model for assignment mechanism
  - Model for the data not needed!
    - Can help if unsure about assignment mechanism

“Block” or “Stratified” Randomized Trials

- Simple core idea.
- Imagine drug trial:
  - Drug might work for men but not women (or vice versa)
  - Might have stronger side effects for old than for young
- Can use “important” covariates to create blocks:
  - male vs. female [2 blocks]
  - male vs. female and old vs. young [4 blocks]
- Randomize within each block

Overall and within-block effects

- Assume two blocks (male = m; female = f)
- Estimate ATE_m, ATE_f within each block

$$\widehat{ATE} = \frac{(N_m \times \widehat{ATE_m}) + (N_f \times \widehat{ATE_f})}{N}$$

In terms of probability, this is the expectation

$N = N_m + N_f$

- More generally, create *J* blocks B_j (j = 1, J):

$$\widehat{ATE} = \sum_{j=1}^{J} \frac{N_j}{N} \times \widehat{ATE_j}$$

Does block randomization create bias?

Errors in the slide!!

- No. Unbiased estimate *within each group j:*

$$[\mathrm{E}[y_1|w = 1, \mathrm{group} = j] = [\mathrm{E}[y_1|w = 0, \mathrm{group} = j]$$

$$[\mathrm{E}[y_0|w = 1, \mathrm{group} = j] = [\mathrm{E}[y_0|w = 0, \mathrm{group} = j]$$

So $\mathrm{ATE}_j =: \mathrm{E}[y_1 - y_0|\mathrm{group} = j]$

$$= \mathrm{E}[y_1|\mathrm{group} = j] - \mathrm{E}[y_0|\mathrm{group} = j]$$

$$= [\mathrm{E}[y_1|w = 1, \mathrm{group} = j] = [\mathrm{E}[y_0|w = 0, \mathrm{group} = j]$$

Covariates, omitted variables

- Should have covariate balance *within each group j:*

$$[E[\boldsymbol{x}|w = 1, \text{group} = j] = E[\boldsymbol{x}|w = 0, \text{group} = j]$$

Measure covariate balance within groups → test for within-group randomization

- Again no worries about omitted variables:

$$[E[\boldsymbol{u}|w = 1, \text{group} = j] = E[\boldsymbol{u}|w = 0, \text{group} = j]$$

Sum across groups using LIE

- Across groups *j*, apply Law of Iterated Expectations (LIE):

  Estim**and**: $\mathrm{E}[\mathrm{y}_1 - \mathrm{y}_0] = \mathrm{E}_\mathrm{j}[\mathrm{E}[\mathrm{y}_1 - \mathrm{y}_0|\mathrm{group} = j]]$

  $= \mathrm{E}_\mathrm{j}\big[\mathrm{E}[\mathrm{y}_1|\mathrm{w} = 1, \mathrm{group}\, j] - \mathrm{E}[\mathrm{y}_0|\mathrm{w} = 0, \mathrm{group} = j]\big]$

  Estim**ate**:

$$\widehat{ATE} = \sum_{j=1}^{J} \frac{N_j}{N} \widehat{ATE_j} = \sum_{j=1}^{J} \frac{N_j}{N} \Big(\sum_{i \in j: \cdot w_i = 1} \frac{y_{i1}}{N_{tj}} - \sum_{i \in j:\, w_i = 0} \frac{y_{i0}}{N_{tj}} \Big)$$

Not same as average over sample

- If treatment effects **and** proportion of treated N_{tj}/N_j both vary across blocks, the "global estimate" below is biased:

$$\widehat{ATE} \neq \sum_{i:w_i=1} \frac{y_{1i}}{N_t} - \sum_{i:w_i=0} \frac{y_{0i}}{N_c}$$

- If you use a block design, you have to use it consistently!
  - Estimate ATE within blocks first, them sum across blocks

Block randomized trial example

- Tennessee STAR experiment
- Study of value of smaller class sizes (for K)
  - STAR = Student/teacher achievement ratio
  - first convincing evidence that smaller classes ➔ higher test performance
  - Chetty et al, (2011): later-life performance too!
- Eligible schools: 3+ kindergarten classes
- Three groups of classes:
  - Small = small class (13-17 students)
  - Regular = regular class (22-25 students)
  - Reg + Aide = regular class w teacher's aide

What are the blocks?

- Randomly assign:
  - class types within schools (at least 1 of each type)
  - students and teachers to classes
- What are the “blocks”?

STAR experiment and SUTVA independence

- Is SUTVA "independence" satisfied for **students**?
  - Yes, under Fisher's sharp null (no effect on anyone)
  - No, if $\tau \neq 0$ [students could influence each other]
- We'll study results at class level
  - Is SUTVA independence satisfied for **classes**?
- Will regression still work
  - Tables below are from Angrist & Pischke (2009), who adapt them from Kreuger (1999)

Multivariate results

(n = 5,681, s.e., clustered on class in parentheses)

Dependent variable	Avg. percentile score			
Explanatory Variable	(1)	(2)	(3)	(4)
Small class	4.82**	5.37***	5.36***	5.37***
	(2.19)	(1.26)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian	-	-	8.35***	8.44***
			(1.35)	(1.36)
Girl	-	-	4.48***	4.39***
			(.63)	(.63)
Free Lunch	-	-	-13.15***	-13.07***
			(.77)	(.77)
White teacher	-	-	-	-.57
				(2.10)
Teacher experience	-	-	-	.26
				(.10)
Teacher Master's degree	-	-	-	-0.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

Value of rich covariates

- Estimated value of small class is stable as add covariates
  - As it should be, for randomized trial
- Not (too) worried about omitted variables
  - They should matter only by accident
- But suppose this was a pure observational study
- Then we worry a lot about omitted variables
- What can we do about OVB risk?
  - If a variable is included as a covariate, it isn't omitted ☺
  - If many included covariates, we worry less
  - If many included covariates, and estimate **insensitive** as we add them, we worry still less
    - Logic: If the covariates we **can** measure do not affect estimate then more likely that the omitted covariates won't either

Regression as weighted average causal effect

- So, is the school FE estimate unbiased (or close enough)?
- OLS *assumes* constant treatment effect
  - seeks most precise estimate given this assumption
    - the "B[est]" in BLUE
  - implicitly weights block *j* by conditional variance:
    - $wgt_j = s_j = p_{tj} \cdot (1 - p_{tj})$
- So (with r_j = fraction of sample in school *j*):

$$\hat{\tau}_{OLS} \xrightarrow[p]{} \tau_{wgt} = \sum_{j=1}^{J} r_j \cdot p_{tj} \cdot (1 - p_{tj})\, \tau_j \Big/ \sum_{j=1}^{J} p_{tj} \cdot (1 - p_{tj})$$

Regression and conditional variance weighting

- Example: HRS dataset (Black et al., Does Health Insurance Affect Mortality, WP 2015)
  - treat ***as if*** block randomized experiment; four blocks:
    - Hispanic
    - non-Hispanic black
    - non-Hispanic white
    - non-Hispanic other
  - estimate treatment effect: effect of insurance at wave 1 (1992) on mortality in 10 years (wave 6, 2002)

True treatment (uninsurance) effect estimates

Group	Sample size	$p_{insured}$	block ATE
Hispanic	880	0.613	-0.0539
non-Hispanic Black	1,619	0.794	-0.0037
non-Hispanic White	6,583	0.869	0.0575
non-Hispanic Other	197	0.746	0.0376
For full sample			
ATE			**0.0358**
ATC (for insured)	7,691		**0.0391**
ATT (for uninsured)	1,588		**0.0201**

So a situation where conditional variance weights can matter:

- Heterogeneous treatment effects
- Differing probabilities of treatment across blocks

By how much?

Regression vs. block treatment effect estimates

Stata:

```
. regress rdead6 noins rahispan nhispblack nhispwhite nhispother, robust
note: rahispan omitted because of collinearity

Linear regression                               Number of obs =    9279
                                                F(  4,  9274) =   17.44
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0097
                                                Root MSE      =  .32897

------------------------------------------------------------------------------
             |               Robust
      rdead6 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       noins |   .0256571   .0100564     2.55   0.011     .0059443    .0453699
    rahispan |          0  (omitted)
  nhispblack |   .0933497   .0144385     6.47   0.000      .065047    .1216524
  nhispwhite |   .0140573   .0115748     1.21   0.225    -.0086319    .0367465
  nhispother |   .0320008   .0263053     1.22   0.224    -.0195634    .0835649
       _cons |    .093467   .0112984     8.27   0.000     .0713196    .1156143
------------------------------------------------------------------------------
```

In this example, regression → mess

- Regression estimate is not close to ATE
  - 0.0257 (regression) vs. 0.0358 (true estimate)

- Why?
  - Large positive treatment effect for whites
    - Close to zero for Blacks
    - Large **negative** effect for Hispanics
  - Whites are more likely to be insured
    - Regression downweights whites (wt. = 0.114)
    - Versus Hispanics (wt. = 0.237); Blacks (wt. = 0.163)

Internal vs. external validity

- Internal validity = valid results ***for this sample***
  - Or larger population from which sample was drawn at random
- External validity = valid for larger population, not directly studied
  - Much harder, rarely achievable from single study

External validity of STAR experiment

- Let's explore our confidence in extrapolating from Tennessee STAR experiment, to:
  - smaller schools, not eligible for study
  - eligible schools, which decided not to participate
  - public schools in other states
  - private schools (secular, religious)
  - smaller class sizes than "small" STAR classes
  - larger class sizes than "regular" STAR classes
  - intermediate class sizes (17-22)
  - public schools in other countries

Randomized experiments: When to block?

- Always, if you can
  - "block what you can and randomize what you cannot" [Box, Hunter, and Hunter (1978, p.103]
  - Intuition: Get exact balance on important covariate instead of balance only in expectation
    - Still get benefits of randomization for other variables
- What to block on:
  - "science", not statistics
- The (minor) cost of blocking
  - Higher variance for the estimate of the variance

Experiments with one-sided noncompliance

- People often don't agree to be randomized
- Can have one-sided or two-sided noncompliance
  - treatment is *offered* at random
    - Some offerees accept = **compliers**
    - Some offerees decline = **noncompliers**
  - If non-offerees can't get the treatment, we have one-sided noncompliance
  - If some non-offerees figure out how to be treated, we have two-sided noncompliance
- Start with easier, one-sided case

Example: Sommer-Zeger (1991) Vit. A experiment

- Vitamin A shots offered for kids age 2-3m, again 6m later
  - Indonesian villages chosen at random
    - people in 225 villages received offer; 225 didn't
    - z = **offered** treatment
    - w = **received** treatment
- Treated villages: 12,094 kids (z=1)
    - 9,675 compliers (80.0%) (w=1)
    - 2,419 noncompliers (20.0%) (w=0)
- Control villages: 11,588 kids (z=0)
    - ?? compliers (w=0)
    - ?? noncompliers (w=0)

Can analyze as "Intent-to-Treat (ITT)"

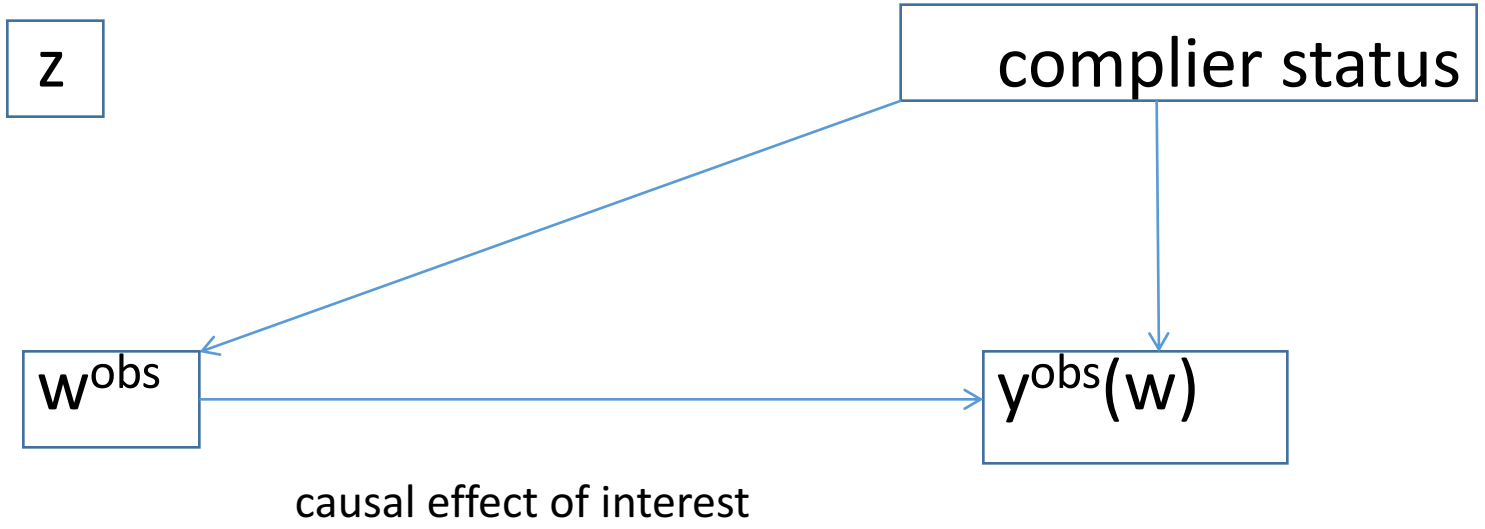
- **Estimand**: child death ($d_i = 1$)
- Treated kids: 12,094
  - 46 deaths (0.380%)
    - 12 among compliers; 34 among noncompliers
- Control kids: 11,588
  - 74 deaths (0.638%)
- Treated vs. controls
  - $\hat{\tau}_{ITT} = 0.00380 - .00638 = -.00258$

  ratio: 0.0038/0.00638 = 0.595 (40% drop in mortality)

Link to (classic and causal) IV

- One-sided non-compliance: first example of “causal IV”
- z is an instrument for w
  - satisfies usual IV assumptions
  - unlike traditional IV, causal IV allows for heterogeneous (“local”) treatment effects
- z addresses endogeneity of w
  - w depends on (unobserved) complier status
  - So do $(y_i(w_i = 0), y_i(w_i = 1))$
    - Noncompliers have higher mortality rates when not treated $y_i(w_i=0)$
    - Could have different treatment effects, if (forcibly) treated?

Graphical depiction of role of z (as causal IV)

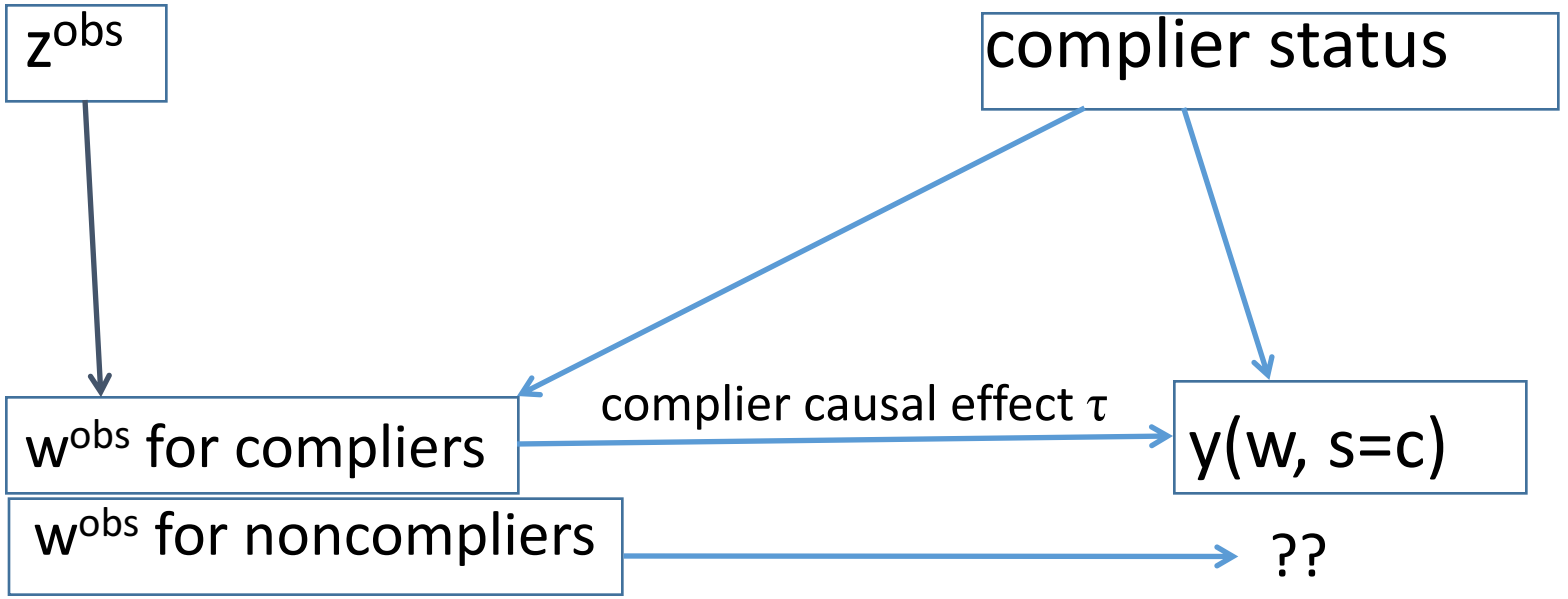


If z not used, w^{obs} is endogenous

Omitted variable, complier status, predicts both w^{obs} and y^{obs}

Note: An example of a Judea Pearl "directed acyclic graph (DAG)

Graphical depiction: IV for one-sided compliance



z affects y **only through** w, breaks endogeneity

Only affects compliers: Can only estimate τ for compliers

Only through assumption is crucial: z does not affect y either directly, or indirectly through **x, u**

Vitamin A experiment treated as classic IV

Stata:

```
ivregress 2sls outcome (treatment=instrument), robust

Instrumental variables (2SLS) regression              Number of obs =   23682
                                                      F(  1, 23680) =    7.75
                                                      Prob > F      =  0.0054
                                                      R-squared     =  0.0015
                                                      Root MSE      =  .07095
             |               Robust
     outcome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |   -.003228   .0011592    -2.78   0.005    -.0055002   -.0009559
       _cons |   .0063859     .00074     8.63   0.000     .0049355    .0078364
------------------------------------------------------------------------------
Instrumented:  treatment
Instruments:   instrument
------------------------------------------------------------------------------
```

- *t*-stat is slightly smaller than our lower bound estimate (2.78 vs. 2.80)
- Reflects uncertainty in proportion of compliers
- Should cluster on village, but data not available

Two-sided noncompliance

- Can also have two-sided noncompliance
  - Among those offered the treatment (z-treated):
    - Some offerees accept
    - Some offerees decline
  - Among the controls:
    - Some non-offerees get the treatment anyway
    - Some non-offerees don’t
- This can be handled by using IV as well

Recap on randomized experiments

- Gold standard for causal inference
  - Treated and controls: same in expectation *if not treated.*
  - Can (and should) test the randomization
  - Block randomization: important subclasses or controls
  - Naïve regression sometimes works
  - IV can address noncompliance
  - Naïve regression and simple 2sls sometimes work

Two-period Difference-in-Differences (DiD)

- Start simple, then add complexities
- Two time periods, before and after treatment.
  - treated and control groups
  - observe groups both before (t=b) and after (t=a)
  - no covariates

Near-random assignment (we hope)

- Assignment not random, but “close”
  - Comes from “shock” of some kind
    - often called “natural” or “quasi” experiment
    - Shock should be “exogenous”:
      - units don’t choose whether to be treated
      - division between treated and controls is unrelated to characteristics that affect response to treatment
      - no anticipation
      - shock expected to be permanent
  - **Assume**: Difference between treated and controls *would have been stable* ***but for the treatment***
    - Core, **untestable** “parallel changes” assumption
    - Can be plausible if assignment is close enough to random

Requirements for a "good shock"

(1) ***Shock Strength***: Strong enough to significantly change firm behavior.

(2) ***Exogenous Shock***. Came from "outside" the system. Firms did not choose to be treated, could not anticipate the shock, no reason to think unobservables predict potential outcomes or which firms were treated.

(3) ***"As If Random" Assignment****:* Separates firms into treated and controls in *close to random manner*. Exception for forcing variable which determines which firms are treated.

(4) ***Covariate balance***. Reasonable covariate balance between treated and control firms, including "common support". Somewhat imperfect balance can be address with balancing methods.

(5) ***Only-Through Condition(s)***: The effect of the shock on the outcome must come *only through* the shock. No other shock, at around the same time, could affect treated firms differently than control firms. For IV – the shock must affect the outcome only through the instrumented variable.

“Shock-based” design

- So to repeat (because this is central to good design)
- Common “requirements” for a “good shock” across shock-based designs (DiD, RD, IV, event study (ES))
  1. “strong” shock
  2. Exogenous: firms did not choose to be treated
     - No avoidance or anticipation
  3. “as if random” assignment to treatment
  4. Leading to covariate balance
     - Including reasonably thick “common support”
     - And parallel pre-treatment trends
  5. “only through” condition(s)

  Can improve design through “balancing methods”
- Commonalities are not well known
  - methods are studied separately, not together
  - some requirements are “soft” – credibility, not formal assumption

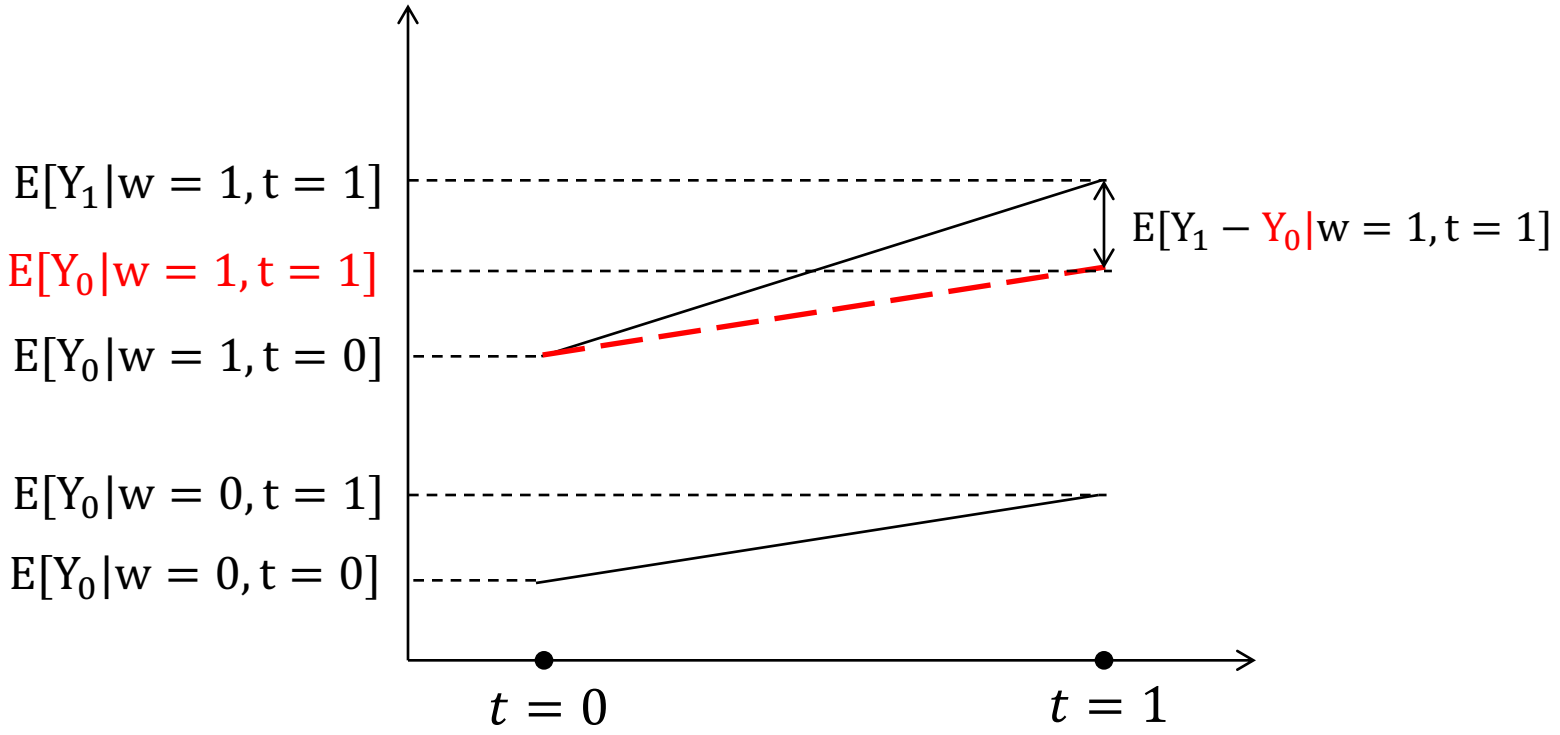
Implication of parallel changes

DiD setup	After	Before
Treated	$y_{t1,a}$	$y_{t0,b}$
Controls	$y_{c0,a}$	$y_{c0,b}$

- a = after; b = before, t=treatment group, c=controls
- 1-> treated (treatment group, **after**)
- 0-> not treated (treatment group **before**, controls **always**)
- *Change in (difference between treated and controls) is due to treatment*
  - ATT = after-minus-before difference in (difference between treated and controls)
- Hence the name: **difference-in-differences**

Graphical representation of DiD

Red = not observed



Those are the ideas behind DiD.
Now the (simple) math

- Not randomized trial ➔units **not** randomly chosen for treatment. Recall:

$$\tau_{naive} = E[y^{obs}|w = 1] - E[y^{obs}|w = 0]$$
$$= E[y_1 - y_0|w = 1] + \{E[y_0|w = 1] - E[y_0|w = 0]\}$$
$$= ATT + \text{baseline bias}$$

Randomized trial: We focused on one time period = **after**

Baseline bias (aka selection bias) = (unobserved) difference between treated and controls if neither was treated

DiD: Dealing with baseline bias

- **Randomization** → baseline bias = 0
  - Treated and controls are same in expectation
- **DiD:** No randomization
  - Can't assume baseline bias = 0 [either before or after]
- But have both before and after
  - Estimate $(\text{baseline bias})_{\text{before}}$.
  - **Assume** $(\text{baseline bias})_{\text{after}} = (\text{baseline bias})_{\text{before}}$

ATT_{DiD} Estimand

- What we **can** estimate:

$$ATT_{DiD} = \tau_{naive,after} - baseline\ bias_{before}$$

- What we **want** to estimate:

$$ATT_{DiD} = \tau_{naive,after} - \left(baseline\ bias_{after}\right)$$

Not observed

- DiD "solves" this disconnect by **assuming** parallel changes:

$(\text{baseline bias})_{\text{after}} = (\text{baseline bias})_{\text{before}}$

Understanding parallel changes

- Ok, so we ***assume*** parallel changes
- What does this assumption mean? What might justify it?
- We're assuming:
  - Levels not randomly assigned → baseline bias ≠ 0
  - But **changes** are *as good as randomly assigned* →
    - $E[\delta(\text{baseline bias})] = 0$

DiD Assignment Mechanism

- This is an odd assignment mechanism
  - Usual assignment mechanism: Rule(s) determining who is treated
  - Here, there is a “sub-assignment mechanism”
    - applies to **changes** within each group
- Treated and controls must be similar enough to make this plausible
  - In pre-treatment covariates
  - In baseline $\text{bias}_{\text{before}}$

Another view of how DiD works

	After	Before	Unobserved potential outcomes	True treatment effects
Treated group	$y_{i1,a}$	$y_{i0,b}$	$y_{i0,a}$	ATT = $E[y_{1,a}]$-$E[y_{0,a}]$
Control group	$y_{i0,a}$	$y_{i0,b}$	$y_{i1,a}$	ATC = $E[y_{1,a}]$-$E[y_{0,a}]$

Data we need for ATT:

Top right cell: "after" outcomes for treated, if had not been treated

We assume:

$$E_{treated}[y_{0,a}] = E_{treated}[y_{0,b}] + E_{controls}[y_{0,a} - y_{0,b}]$$

Data we need for ATC:

Bottom right cell: "after outcomes for controls, ***if treated***

No good way to estimate $E_{controls}[y_{1,a}]$

So DiD let's us estimate ATT, but not ATC or ATE

First differences form of DiD

Alternate form of DiD **estimand:**

$$ATT_{DiD} = E\left[y_{t1,a} - y_{t0,a}\right] - E\left[y_{t0,b} - y_{c0,b}\right] =$$
$$E[\Delta(y_t)] - E[\Delta(y_c)]$$

ATT_{DiD} **estimator** relies on analogy principle, sample averages

True panel data: observe same units before and after:

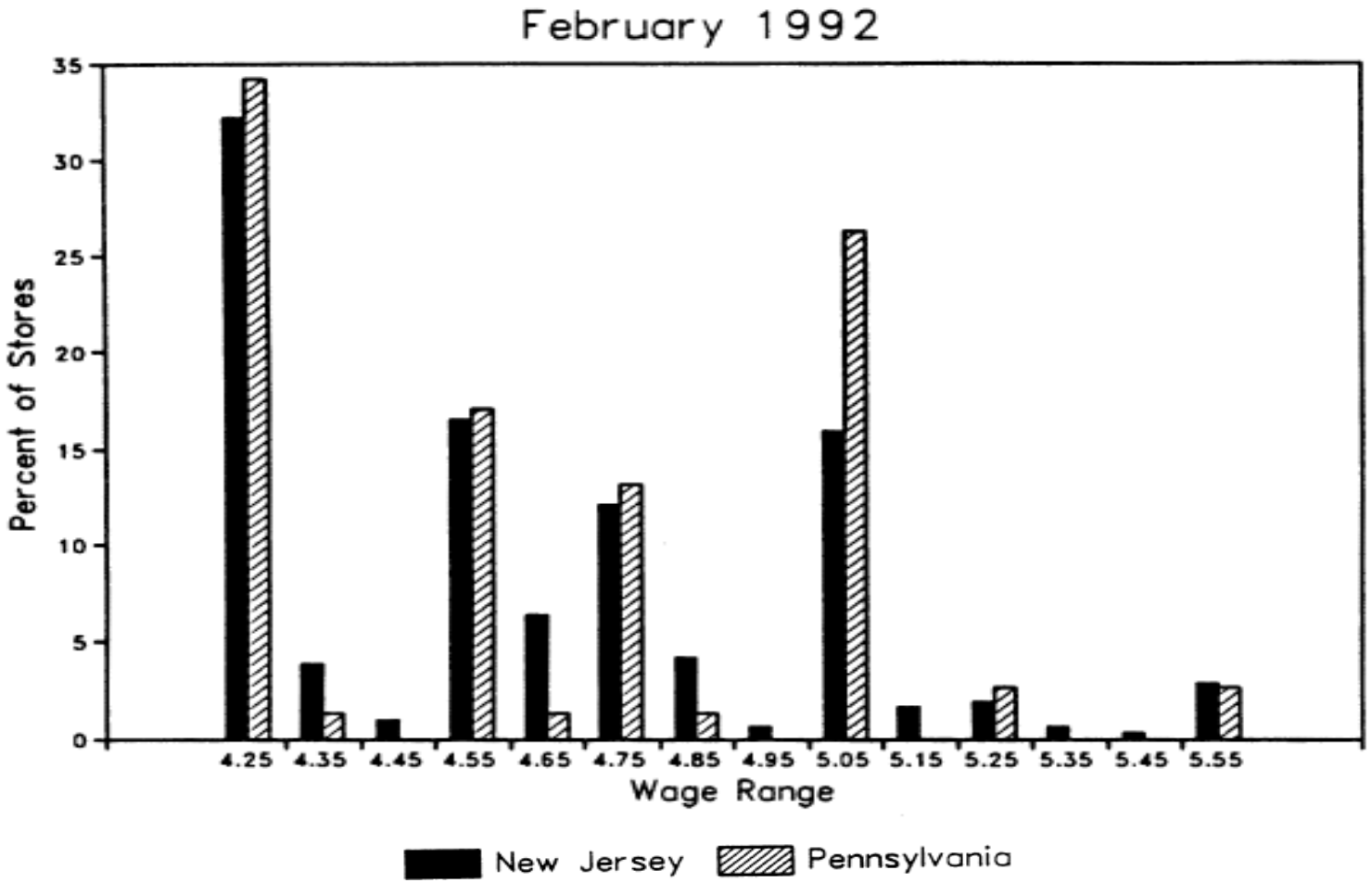
$$\widehat{ATT}_{DiD} = \frac{1}{N_t}\sum_{i\in t}\left(y_{i,a} - y_{i,b}\right) - \frac{1}{N_c}\sum_{i\in c}\left(y_{i,a} - y_{i,b}\right)$$
$$= \frac{1}{N_t}\sum_{i\in t}\Delta y_i - \frac{1}{N_c}\sum_{i\in c}\Delta y_i$$

Motivating example: Card & Krueger (1994)

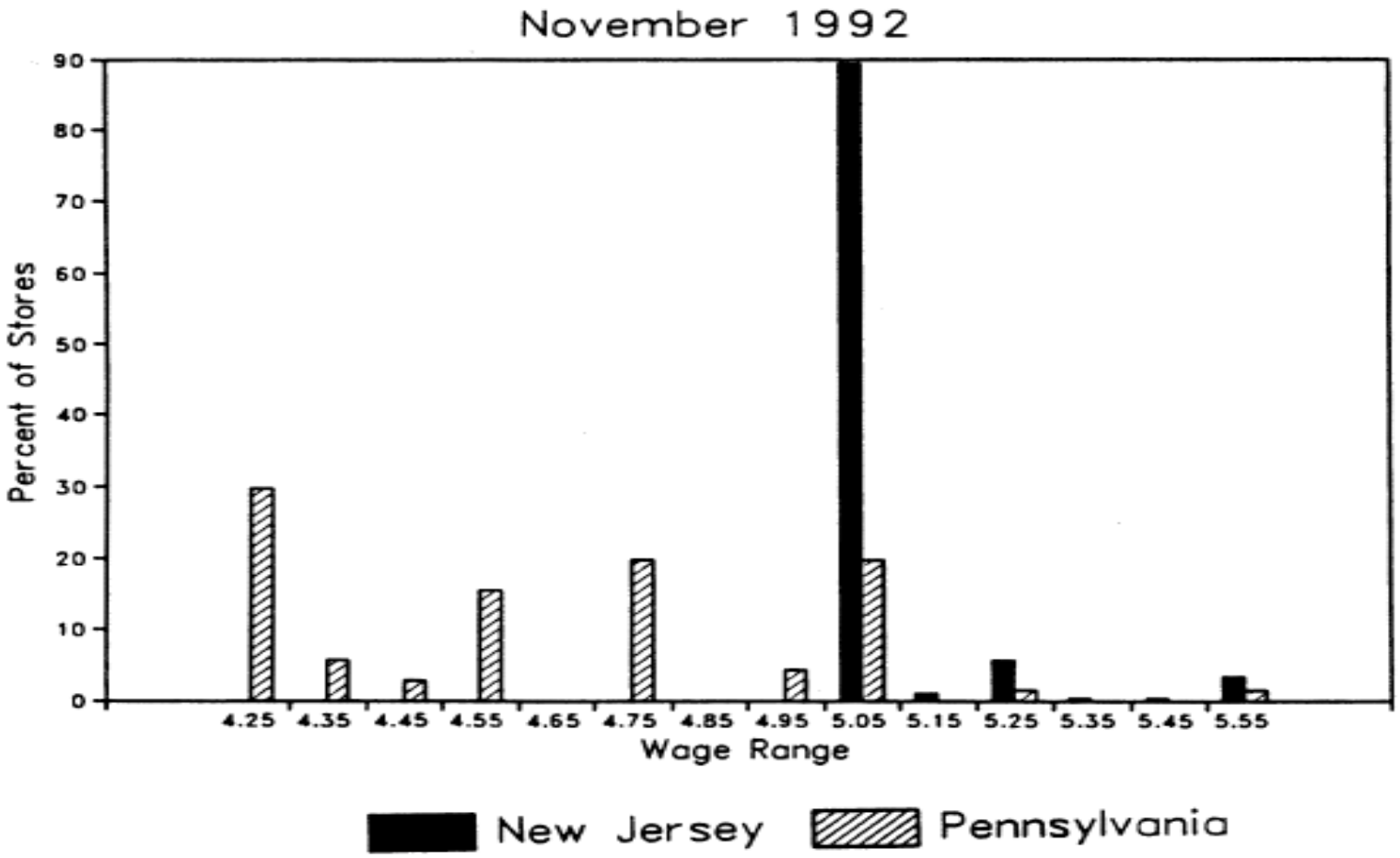
- Research question: Do (modestly) higher minimum wages reduce low-wage employment?
- Card and Krueger consider impact of New Jersey's 1992 minimum wage increase from $4.25 to $5.05 per hour
  - In 2013$: Equivalent to an Increase from $7.00 to $8.30
    - Compare US minimum wage [$7.25]; IL minimum wage: S8.25
- Compare 410 fast-food restaurants in New Jersey (treated) to eastern Pennsylvania (control) before and after the increase
- Data on wages and employment:
  - March & Dec 1992, one month before; 8 months after increase
- Note the "local" nature of the question:
  - Microeconomic theory: raise minimum **enough** ➔ lower employment
    - Higher prices ➔ lower equilibrium demand
    - Over time, higher labor cost ➔ substitute capital for labor

Wages before minimum wage increase

Note: National minimum = $4.25
New minimum is within range that some already pay.



Wages after minimum wage increase



Selected Card & Krueger results

Full-time equivalent employment, per restaurant.

Drop 6 restaurants which closed from pre to post; 4 which temporarily closed.

Time	PA	NJ	NJ - PA
Before	23.33	20.44	-2.89
	(1.35)	(0.51)	(1.44)
After	21.17	21.03	-0.14
	(0.94)	(0.52)	(1.07)
After - Before	-2.16	0.59	**2.76****
	(1.25)	(0.54)	**(1.36)**

NJ looks better after minimum wage increase

But effect is **entirely** because PA employment declines

Puts great stress on the parallel changes assumption.

Come back to this assumption . . .

Regression implementation of DiD

- Can use unit fixed effects regression
  - *post* = dummy for "after" (post-treatment) period
  - f_i = unit dummies
  - t=0 (before) or 1 (after)

$$y_{it} = \alpha + f_i + \beta \cdot t + \delta \cdot w_i \cdot t + \varepsilon_{it}$$

- Or first difference form:

$$\Delta y_{it} = \beta + \delta \cdot w_i + \varepsilon_i$$

- **Stata:** regress δy w, robust
  - Compare randomized experiment (**regress y w, robust**)

Why does regression work?

And what does it estimate?

- DiD: regress δy w, robust
- Randomized experiment: regress y w, robust

Randomized experiment	Difference-in-Differences
Regression actually estimates	Regression actually estimates:
$y_i^{obs} = y_{i1}*w_i + y_{i0}*(1-w_i) = \alpha + \beta*w_i + \varepsilon_i$	$\delta y_i^{obs} = \delta y_{i1}*w_i + \delta y_{i0}*(1-w_i) = \alpha + \beta*w_i + \varepsilon_i$
For $w_i = 1$: $y_{i1} = \alpha + \beta + \varepsilon_i$ For $w_i = 0$: $y_{i0} = \alpha + \varepsilon_i$	For $w_i = 1$: $\delta y_{i1} = \alpha + \beta + \varepsilon_i$ For $w_i = 0$: $\delta y_{i0} = \alpha + \varepsilon_i$
Random assignment of **units. For treated:**	Random assignment of **changes:**
$E[y_{i1} \mid w_i=1] = \alpha + \beta$ $E[y_{i0} \mid w_i=1] = E[y_{i0} \mid w_i=0] = \alpha$	$E[\delta y_{i1} \mid w_i=1] = \alpha + \beta$ $E[\delta y_{i0} \mid wi=1] = E[\delta y_{i0} \mid w_i=0] = \alpha$
Treated = controls in expectation	Treated ≠ controls in expectation
$\tau_{ATE/ATT/ATC} = E[y_{i1}] - E[y_{i0}] = [\alpha + \beta] - \alpha = \beta$	$\tau_{ATT} = E[\delta y_{i1}] - E[\delta y_{i0}] = [\alpha + \beta] - \alpha = \beta$

Regression: Minimum wage laws and employment

Method 1, pooled OLS, cluster on firm:

Stata:

```
. gen nj_post = nj*post
. regress emptot post nj nj_post, cluster(ID)
```

```
Linear regression                               Number of obs =     794
                                                F(  3,   409) =    1.80
                                                Prob > F      =  0.1462
                                                R-squared     =  0.0074
                                                Root MSE      =  9.4056
                                  (Std. Err. adjusted for 410 clusters in ID)
------------------------------------------------------------------------------
             |               Robust
      emptot |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        post |  -2.165584   1.218025    -1.78   0.076    -4.559954    .2287855
          nj |  -2.891761   1.439546    -2.01   0.045    -5.721593   -.0619281
     nj_post |   2.753606   1.306607     2.11** 0.036     .1851025    5.322109
       _cons |   23.33117   1.346536    17.33   0.000     20.68417    25.97816
```

Interaction term is positive and (barely) significant

Net change in NJ employ: -2.89 [coeff on nj] + 2.75 [coeff on nj*post] = -0.14

Note: if use "robust" instead of "cluster"; t = 1.53 instead of 2.11

Method 2: Restaurant FE

Stata:

```
. xtreg emptot post nj nj_post, fe robust
note: tsset already run; nj dropped due to collinearity

Fixed-effects (within) regression               Number of obs      =       794
Group variable: ID                              Number of groups   =       410

R-sq:  within  = 0.0147                         Obs per group: min =         1
       between = 0.0043                                        avg =       1.9
       overall = 0.0000                                        max =         2

                                                F(2,409)           =      2.14
corr(u_i, Xb)  = -0.0967                        Prob > F           =    0.1185

                                    (Std. Err. adjusted for 410 clusters in ID)
------------------------------------------------------------------------------
             |               Robust
      emptot |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        post |  -2.283333   1.247982    -1.83   0.068    -4.736592    .1699251
          nj |           0  (omitted)
     nj_post |        2.75   1.337555     2.06** 0.040     .1206598     5.37934
       _cons |   21.06045   .2281007    92.33   0.000     20.61206    21.50885
-------------+----------------------------------------------------------------
     sigma_u |   8.298003
     sigma_e |  6.3411612
         rho |  .63132515   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Note: 410 firms, but only 384 observed twice (794 - 410 = 384)
Fixed effects uses only the twice-observed firms.
Lose information on overall change in NJ employment

Comment on regression with interactions

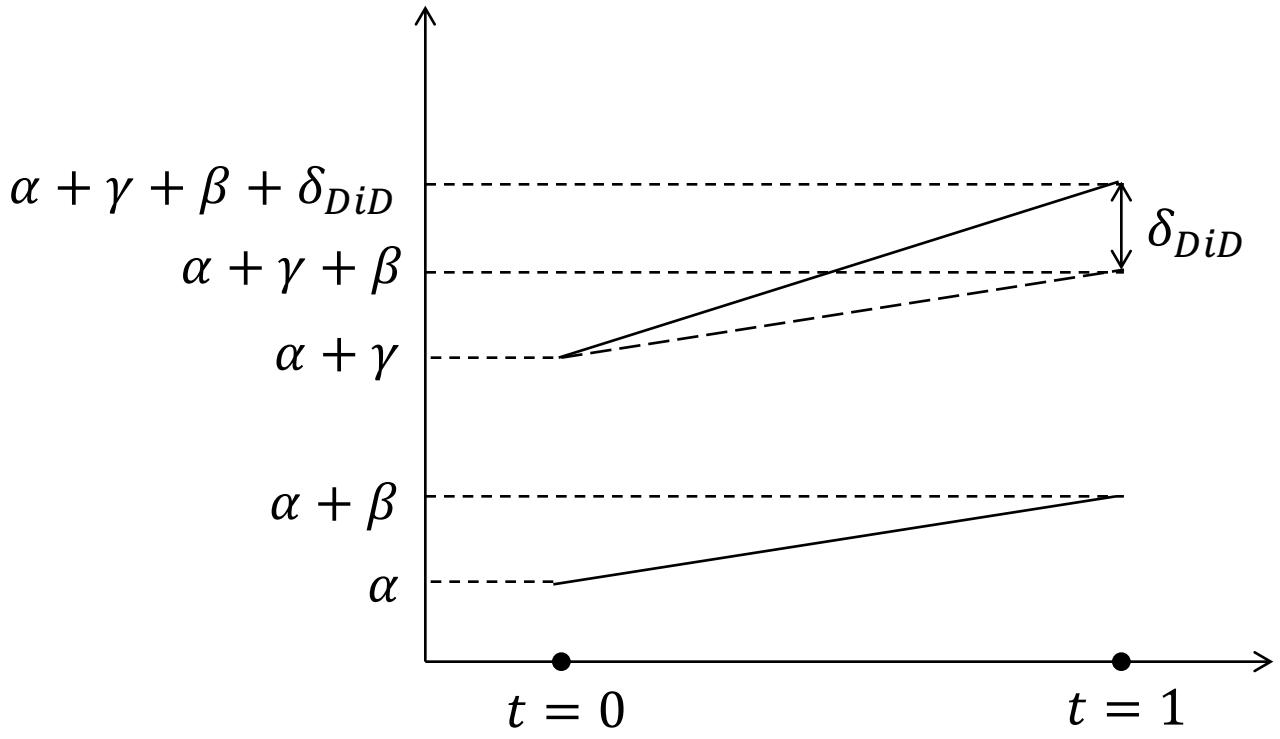
- Regression with interaction term, such as:

$$y_{it} = \alpha + \gamma \cdot w_i + \beta \cdot post + \delta_{DiD} \cdot w_i \cdot post + \varepsilon_i$$

- **Always** include non-interacted terms. Why?
- Suppose drop $\gamma * w_i$, what happens?
  - $\gamma * w_i$ will be absorbed into error term
  - Corr (w_i, $w_i * t$)≠ 0 (often large)
  - ➔ Corr (ε_i, $w_i * t$)≠ 0 (often large) ➔ omitted variable bias
  - δ_{DiD} will capture some of impact of (omitted) w_i
- Unit fixed effects will absorb treatment dummy
  - return to previous slide: n_j dummy is dropped

Meaning of coefficients on interaction terms

$$y = \alpha + \gamma \cdot w + \beta \cdot post + \delta_{DiD} \cdot w \cdot post + \varepsilon$$



Method 3. First-differences regression

Stata:

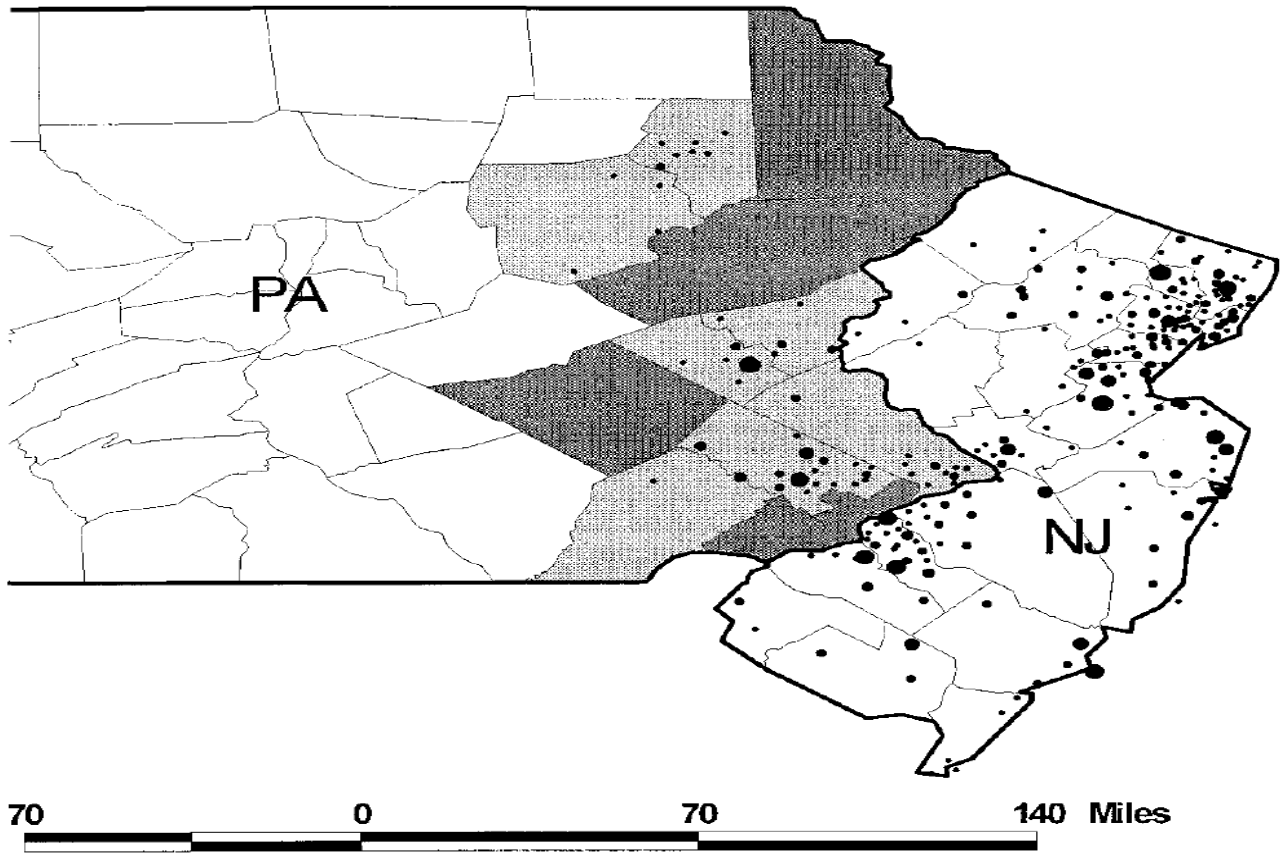
```
tsset ID post
       panel variable:  ID (strongly balanced)
        time variable:  post, 0 to 1
                delta:  1 unit

. gen d_emptot = D1.emptot
(436 missing values generated)

. regress d_emptot nj, robust

 Linear regression                               Number of obs =     384
                                                 F(  1,   382) =    4.23
                                                 Prob > F      =  0.0405
                                                 R-squared     =  0.0146
                                                 Root MSE      =  8.9678
------------------------------------------------------------------------------
             |               Robust
d_emptot |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          nj |        2.75   1.337725     2.06** 0.040     .1197732    5.380227
       _cons |  -2.283333    1.24814    -1.83   0.068    -4.737419    .1707523
```

Results: Identical to FE [coeff on constant = post – pre difference in overall means]
Almost same coefficient and *t*-statistic on NJ as in pooled OLS [OLS: on nj*post].

Dropped 26 restaurants with only "pre" or only "post" data

Restaurant Locations (Card and Krueger, 2000)

Are these locations similar enough?

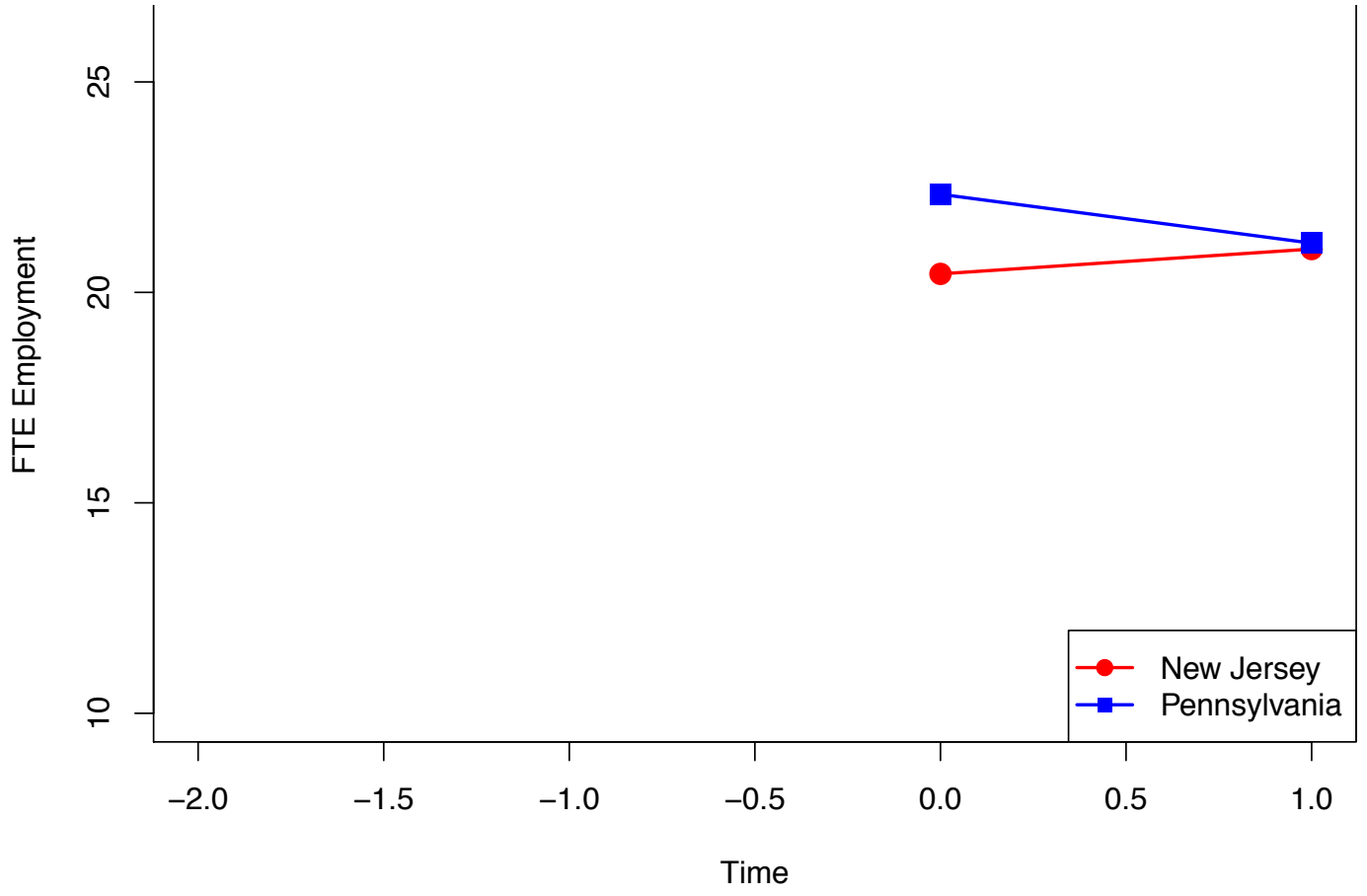


DiD threat: Non-parallel changes

- How can we test for parallel changes?
- Basic strategy: Use multiple pre-periods
    - See if (visually) parallel changes over t= [-n, 0]
    - If not parallel, assumption not justified over t = [0, 1]
    - Placebo shock: middle of pre-treatment period
        - Significant using only pre-treatment data?

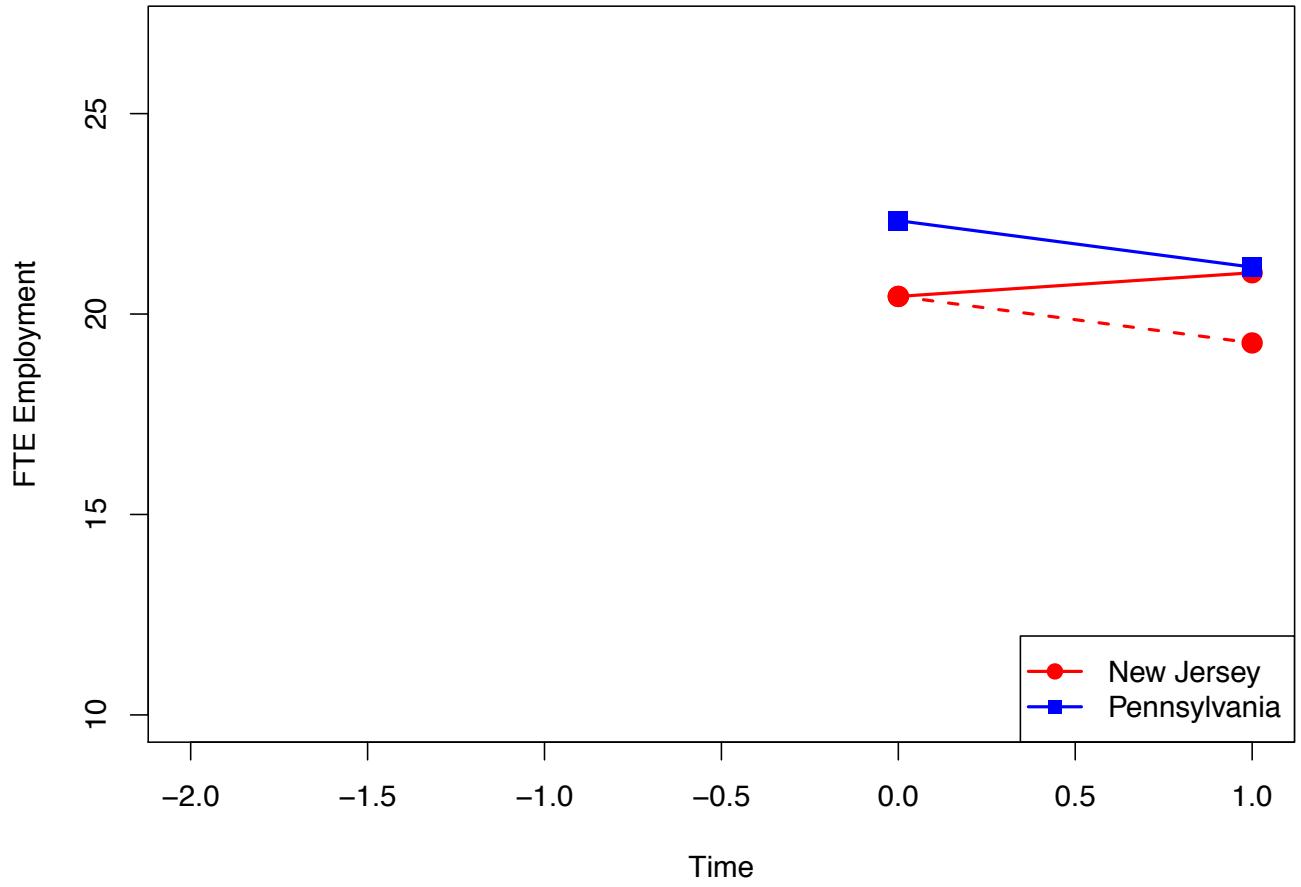
Falsification test: Use pre-period data 1

Card and Krueger (1994) observe:



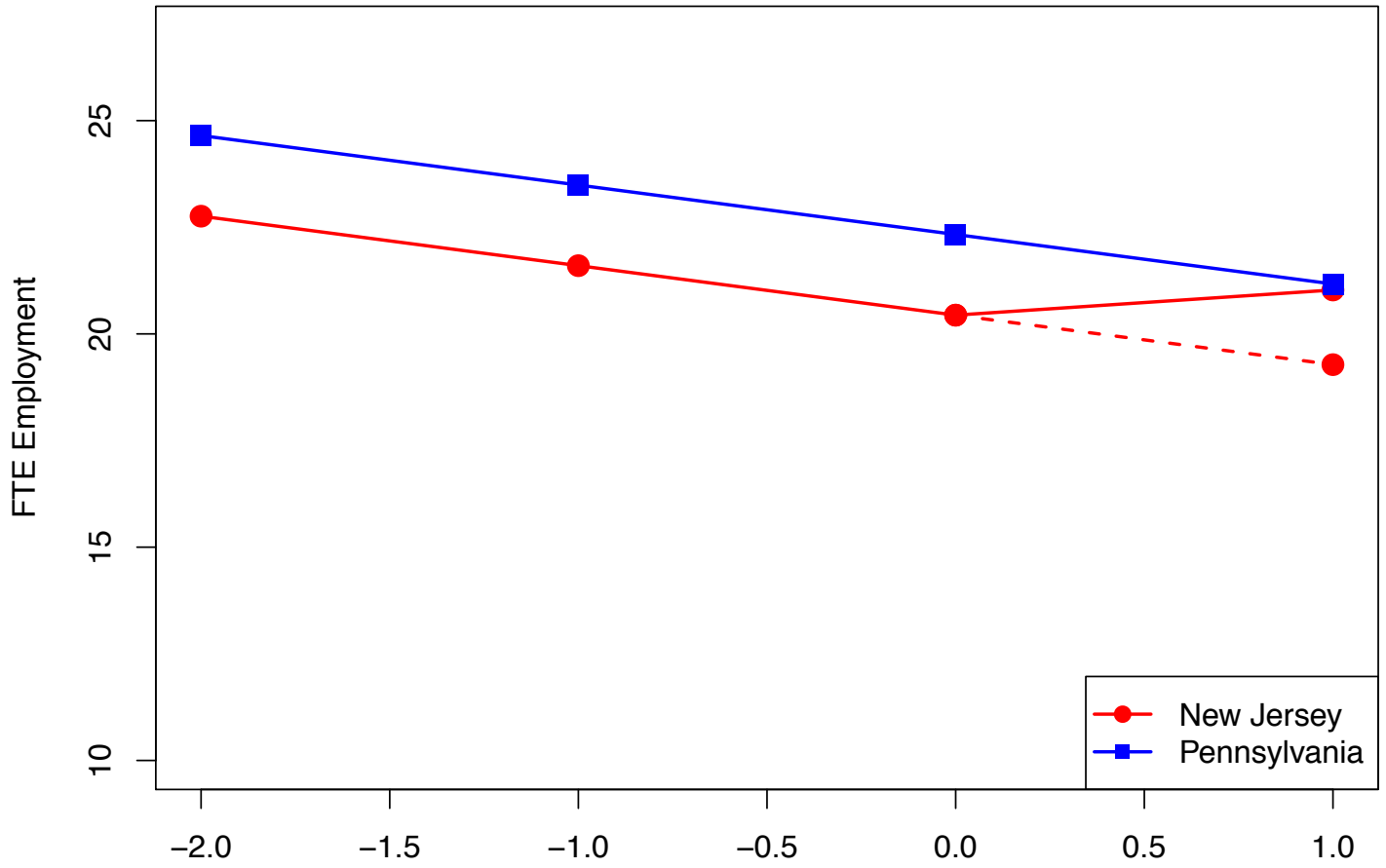
Falsification test: Use pre-period data 2

They ask us to **believe** the NJ counterfactual is this (decline with no min. wage increase)



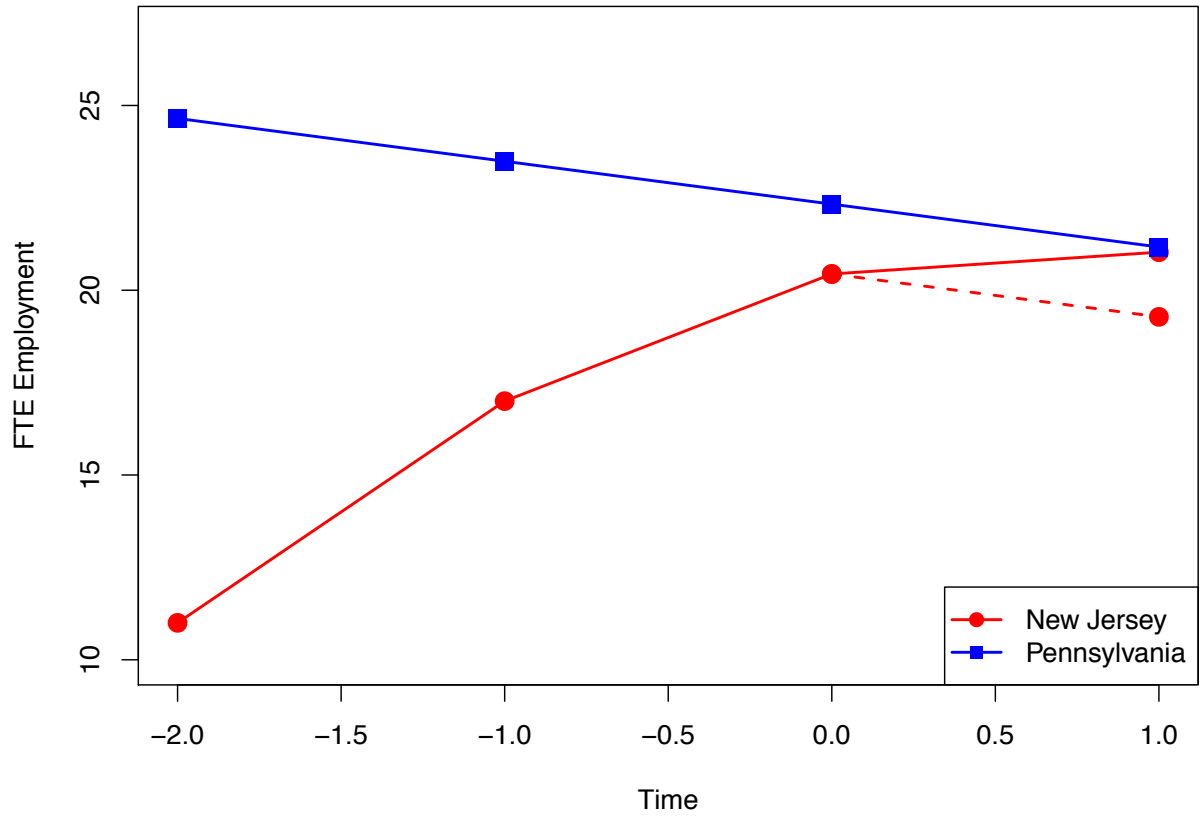
Falsification test: Use pre-period data 3

This would be credible if the pre-period looked like this:



Falsification test: Use pre-period data 4

But not if the pre period data looked like this:



Criticism of Card and Krueger (1994)

- Critics said: Why should we believe parallel trends?
- They also said: NJ and PA fast-food restaurants aren't similar enough
  - flash back to picture showing their locations
- Card and Krueger (2000) did more work
  - Presented the graph of locations shown above
  - And developed time series data

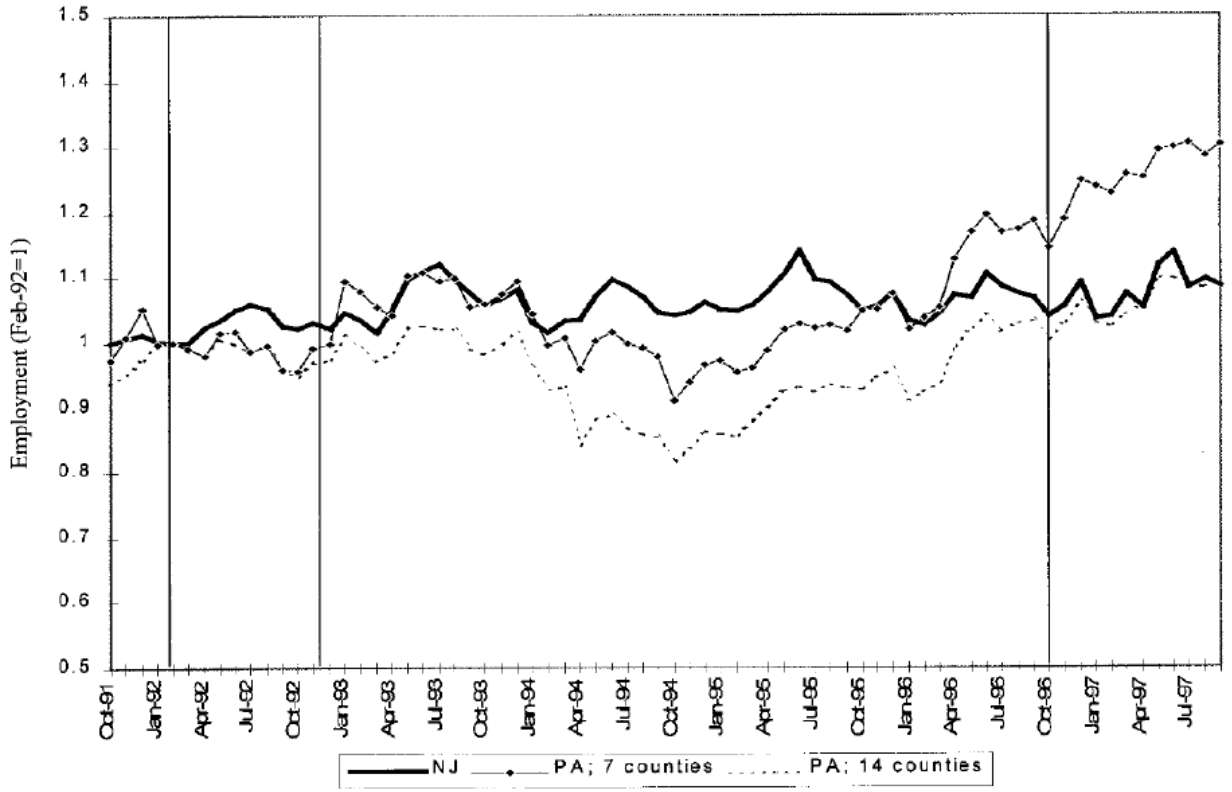
Longer trends in NJ v. PA fast-food employment

One surely wouldn't conclude that NJ employment ***rose***

Better research design: pre-period data for a longer time

Data not available

Still, no strong evidence of NJ employment **drop**



DiD within New Jersey

Restaurants	NJ low wage	NJ middle wage	NJ high wage	NJ all	PA all
wage level	$4.25	[$4.26, $4.99]	≥ $5.00		
FTE before	19.56	20.08	22.25	20.44	23.33
FTE after	20.88	20.96	20.21	21.03	21.17
change	+1.32	+0.87	-2.04	+0.59	-2.16
s.e.	(0.95)	(0.84)	(1.14)	(0.54)	(1.25)

Predict: No effect for high-wage NJ restaurants

They are alternate control group

Better control group than PA?: changes only in NJ

Placebo check: No difference between high-wage NJ and PA

If confirmed, supports validity of PA as source of controls

But . . . what would you worry about?

Other placebo tests?

- What else might we try?
  - Compare alternate control groups
    - just done
  - “Placebo shock” at a different time
    - Pre-treatment, or distant post-treatment
  - “Placebo outcome”: outcome that should not be affected by treatment
    - Placebo outcome = not supposed to be affected by the treatment
    - what would you suggest?

ATT_{DiD}: Repeated Cross-Section Data

- Observe different but similar units before and after = "**repeated cross-section**"
- Many survey datasets take this form

$$ATT_{DiD}^{repeated\ x-section}$$
$$= \left[\frac{1}{N_{Ta}}\left[\sum_{treated,a,i=1}^{N_{Ta}} y_{i1,\alpha}\right] - \frac{1}{N_{Tb}}\left[\sum_{treated,b,i=1}^{N_{Tb}} y_{i1,b}\right]\right]$$
$$- \left[\frac{1}{N_{Ca}}\left[\sum_{controls,a,i=1}^{N_{Ca}} y_{i0,\alpha}\right] - \frac{1}{N_{Cb}}\left[\sum_{treated,b,i=1}^{N_{Cb}} y_{i0,b}\right]\right]$$

DiD Estimators: repeated cross-section

Regression also works for repeated cross-section data:

$$y_{it} = \alpha + \gamma \cdot w_i + \beta \cdot t + \delta_{DiD} \cdot w_i \cdot t + \varepsilon_i$$

- t = 0 (before), 1 (after)
- Usual regression assumption E[ε|w, t] = 0

implicitly captures parallel changes assumption

Stata:

```
gen w_t = w*t
regress y  w  t  w_t, robust
```

Alternative (creates interaction term on the fly):

```
regress y w t w##t, robust
```

	After (t_i=1)	Before (t_i=0)	After – Before
Treated (w_i=1)	$\alpha + \gamma + \beta + \delta_{DiD}$	$\alpha + \gamma$	$\beta + \delta_{DiD}$
Control (w_i=0)	$\alpha + \beta$	α	β
Treated - Control	$\gamma + \delta_{DiD}$	γ	δ_{DiD}

How about covariates?

- Similar to randomized experiments, adding covariates can increase precision. But:
  - Fixed attributes will be absorbed by unit fixed effects
  - Simple time variation (e.g., everyone is a year older at t=1) will be absorbed by time fixed effects
  - For other time-varying covariates, we want to be confident that the treatment does **not** predict the covariate during the post-treatment period ($w_i \perp\!\!\!\perp x_i$)