Causal Inference II

Mixtape Session

Roadmap

- Conditional Parallel Trends
  - Introducing Covariates
  - Choosing Covariates
  - Checking for Imbalance
- Estimators
  - Inverse probability weighting
  - Outcome Regression and Canonical OLS with Additive Controls
  - Double Robust
  - Canonical TWFE with Additive Covariates
- Applications
  - Simulations
  - LaLonde dataset
- Concluding Remarks

Review

- Diff in Diff is a quasi-experimental method that identifies the aggregate causal parameter, ATT, if:
  1. Parallel trends holds
  2. And also no anticipation and if we use comparisons that aren't treated
- But we worked with "unconditional parallel trends" which means this:

$$\Big(E[Y_k^0|Post] - E[Y_k^0|Pre]\Big) - \Big(E[Y_U^0|Post] - E[Y_U^0|Pre]\Big)$$

- This means that the *average* treatment group potential outcome trends for our treatment group is the same as our control group

Is Unconditional Parallel Trends Plausible?

- But what if these two groups are different in ways that predict Y^0?
- Let's say for instance that your study looks at an intervention that primarily effects cities, but your only control group are rural counties
- These groups differ in a ways that might, depending on your study, not be plausibly suggesting they would've been on the same trends

Conditional parallel trends

- While it isn't required that two groups – group k and group U – be similar on observables for parallel trends to hold:
  1. Remember, if treatment had been random, then they would have the average covariates
  2. And if they aren't, then you may need to provide more evidence that these differences are irrelevant
  3. So you should still check and see and if they aren't, you either adjust or have a good justification for why they're the same on trends in $E[Y^0]$
- Including covariates weakens the unconditional parallel trends assumption by only requiring that units with similar covariate values, as opposed to everyone, follow similar counterfactual trends in Y^0
- This is means switching out the *unconditional parallel trends* assumption for the *conditional parallel trends* assumption

Why covariates?

- The inclusion of covariates in diff-in-diff models is not about trying to find random variation in the treatment within values of the dimension of X
- It is based on the claim that the inclusion of covariates is necessary to re-establish parallel trends
- This is itself different than how covariates will be used in synthetic control, too

Correcting the missingness problem

$$
\begin{aligned}
\text{ATT} &= E[\delta|D=1] \\
&= E[Y^1 - Y^0|D=1] \\
&= E[Y^1|D=1] - E[Y^0|D=1] \\
&= E[Y|D=1] - E[Y^0|D=1]
\end{aligned}
$$

We were always missing Y^0 values for the treatment group units, but parallel trends allowed us to impute it using the change in $[Y^0]|D=0]$ as a guide

But if that trend is not a good guide, then we cannot.

Conditional parallel trends

The DiD equation yields:

$$\begin{aligned}\widehat{\delta} &= \Big(E[Y_k|Post] - E[Y_k|Pre]\Big) - \Big(E[Y_U|Post] - E[Y_U|Pre]\Big) \\ &= \text{ATT} + \text{Non-parallel trends bias}\end{aligned}$$

If we believe that conditional on covariates, parallel trends holds, but only within values of X, then there are methods we can use that incorporate covariates into the DiD equation and unbiasedness returns

The inclusion of covariates has particular regression specifications, plus there are alternative methods too, and we will review them

Picking covariates

- You want to select confounders, which are baseline untreated covariates that predict both the treatment (and thus cause imbalance) but also the potential outcome, Y^0
- You can use your common sense, logic, graphical models (i.e., DAGs) and familiarity of the subject
  - $\rightarrow$ Our focus is on the covariates that are determinants of the outcome, and less so the treatment itself, because note our concern is the untreated potential outcome Y^0
- But you can also use some simple approaches that I'll suggest now

Three ideas

- Note that we are missing $E[Y^0|D=1]$, and so we want covariates that are highly predictive of that missing potential outcome – problem is, it's missing
- But what if we did this:
  1. Drop all treated (post treatment treatment group) so that all outcomes are now Y^0 and regress that against candidate covariates, and select from there
  2. Drop post treatment treatment group units and drop the control units and regress Y^0 for the treated unit against baseline covariates and select from there
  3. Use $t-2$ and $t-1$ period, calculate $\Delta E[Y^0|D=1]$ and regress that against $t-1$ covariates and select from there
- This is a more data driven procedure and is helping you figure out which covariates might be candidate confounders – but note this still is not checking for imbalance

Covariate Balance and Parallel Trends

- The parallel trends assumption is untestable but can be evaluated using observed covariates.
- Balance in key covariates helps assess whether treated and control groups would follow similar trends.
- Differences in demographics or economic conditions between groups can indicate violations of parallel trends.
- Use covariates that are unlikely to be affected by treatment to check balance.

Baseline Covariates and Normalized Difference

- Baseline covariates are measured before treatment ($t-1$). Check for balance between treatment and control groups.
- Report the averages of covariates for both groups in a table.
- The normalized difference is calculated as:

$$\text{Norm. Diff}_{\omega} = \frac{\overline{X}_{\omega,T} - \overline{X}_{\omega,C}}{\sqrt{(S^2_{\omega,T} + S^2_{\omega,C})/2}}$$

- The normalized difference measures imbalance; it should be less than 0.25 to avoid problematic imbalance Imbens and Rubin (2015).

Normalized Difference vs. Z-Score

- **Normalized Difference:**
  - → Measures difference in group means (e.g., treatment vs. control)
  - → Interprets difference relative to pooled variance
  - → Used for comparing balance between groups
- **Z-Score:**
  - → Measures distance of a single observation from the mean
  - → Expresses this distance in standard deviation units
  - → Used for assessing how unusual a single observation is
- Both metrics standardize differences using standard deviation, making them unit-free and comparable across variables, but in our case we need to focus on differences between two groups, so we use the normalized difference and focus on a threshold of 0.25.

Table 4: Covariate Balance Statistics

	Unweighted			Weighted		
Variable	Non-Adopt	Adopt	Norm. Diff.	Non-Adopt	Adopt	Norm. Diff.
2013 Covariate Levels						
% Female	27.94	28.32	0.19	29.89	30.13	0.14
% White	47.43	52.50	0.62	46.07	47.76	0.21
% Hispanic	5.86	4.75	-0.14	10.06	11.35	0.13
Unemployment Rate	7.10	7.77	0.25	6.98	8.00	0.50
Poverty Rate	18.54	16.22	-0.35	17.22	15.29	-0.37
Median Income	43.38	48.00	0.41	49.28	57.83	0.68
2014 - 2013 Covariate Differences						
% Female	-0.11	-0.13	-0.06	-0.05	-0.04	0.10
% White	-0.29	-0.35	-0.14	-0.29	-0.27	0.07
% Hispanic	0.11	0.11	-0.01	0.14	0.20	0.33
Unemployment Rate	-1.09	-1.26	-0.24	-1.08	-1.36	-0.54
Poverty Rate	-0.51	-0.28	0.13	-0.41	-0.35	0.04
Median Income	1.13	1.04	-0.04	1.11	1.73	0.32

This table reports the covariate balance between adopting and non-adopting states. In the top panel, we report the averages and standardized differences of each variable, measured in 2013, by adoption status. In the bottom panel we report the average and standardized differences of the county-level long differences between 2014 and 2013 of each variable. We report both weighted and unweighted measures of the averages to correspond to the different estimation methods of including covariates in a 2 $\times$ 2 setting.

Intuition for Checking Covariate Balance

- Remember that confounders are variables that do two things
  1. $X \rightarrow D$. Confounders have different distributions in treatment than control and these balance checks are about seeing if that's true
  2. $X \rightarrow \Delta E[Y^0]$. But it's only a problem if the imbalance is on variables that also predict potential outcome trends, and we choose covariates based on things we think that's true
- So you select covariates that you think are theoretically predictive of Y^0 trends, but then you check imbalance to see if they are differentially distributed

Example: Murder

- Say that there a set of cities pass gun law ordinances but smaller towns don't and you want to know its effect on homicides
- Maybe as much as 80% of homicides are in cities – the homicides in small towns are rare, and the trends are probably pretty noisy and different
- So you might think population size, could be a very important covariate to include or urbanization

Three covariate DiD papers

Three papers (though sometimes you see others) about covariate adjustment in DiD:

1. Abadie (2005) on semiparametric DiD – reweights the comparison group part of the DID equation using a propensity score based on X
2. Heckman, Ichimura and Todd (1997) on outcome regression uses baseline X and control group only to impute the missing counterfactual Y^0 for treatment group units in a DiD equation
3. Sant'Anna and Zhou (2020) is double robust which means the method does both of these at the same time so that you don't have to choose between them

We will discuss both of them and then compare their performance with the more straightforward fixed effects model

Roadmap

- Conditional Parallel Trends
  - Introducing Covariates
  - Choosing Covariates
  - Checking for Imbalance
- Estimators
  - Inverse probability weighting
  - Outcome Regression and Canonical OLS with Additive Controls
  - Double Robust
  - Canonical TWFE with Additive Covariates
- Applications
  - Simulations
  - LaLonde dataset
- Concluding Remarks

Semiparametric DiD

Abadie (2005) proposed a model that simply reweights the control group in the DiD equation using a particular specification ("semiparametric") of the propensity score on pretreatment covariates

1. Calculate each unit's "after minus before" (DiD equation)
2. Estimate the conditional probability of treatment based on baseline covariates (propensity score estimation)
3. Weight the comparison group's DiD equation with the propensity score

Remember – ATT is only missing Y^0 for treatment, so we only have to apply weights to the comparison group units

Novel elements of time in Abadie's model

- There is only one treatment group so therefore there is only one relevant treatment date, t
- The period prior to treatment is called the baseline, or b, period and it is when treated units were not treated
- X_b are "baseline" covariates meaning the value of X in the pre-treatment period for either the treated or comparison group units
- Propensity scores are estimated off the b period *only*
- Abadie "throws away" covariates after treatment because this is all about re-establishing parallel trends which is a *baseline* concept recall

Assumptions

Five main assumptions

1. No anticipation
2. SUTVA
3. Conditional parallel trends

$$E[Y_t^0 - Y_b^0 | D = 1, X_b] = E[Y_t^0 - Y_t^0 | D = 0, X_b]$$

4. Common support

$$Pr(D = 1) > 0; Pr(D = 1|X) < 1$$

5. Propensity score model is properly specified

Propensity scores as dimension reduction

- Propensity scores are ways of dealing with a conditioning set X that has large dimensions
- Dimensions are not the same as covariates – if you have continuous X, then it has infinite dimensions
- Common support means that *within* all combinations of the covariates (e.g., white male 47yo versus whites, males, age) there are units in treatment and control

Common support example

Think of common support like “exact matches” but on the propensity score

I’m a white male 47 years old with a PhD; can I find a white male 47 years old without a PhD

If I can, that’s common support; if I cannot that’s off support

Propensity scores as dimension reduction

- Propensity score theorem (Rosenbaum and Rubin 1983) showed that if you need X to satisfy some assumption, the propensity score will satisfy too
- Propensity scores essentially transform your large dimensional problem into a single scalar called the propensity score, which is the conditional probability of treatment (conditional on X)
- But we need to estimate the propensity score because we don't usually know it (only an experimentalist "knows" the true propensity score)

Common support and the propensity score

- Exact matches mean you have people who are identical on covariate values in both treatment and control
- Common support and the propensity score means you have people nearly identical on their probability of treatment
- I am 47yo white male with a PhD with a propensity score of 0.75, but you are an Asian female 27yo without a PhD and have a propensity score of 0.75
- Same idea, but for this to work, we need to have “matches” like that (just on the propensity score)

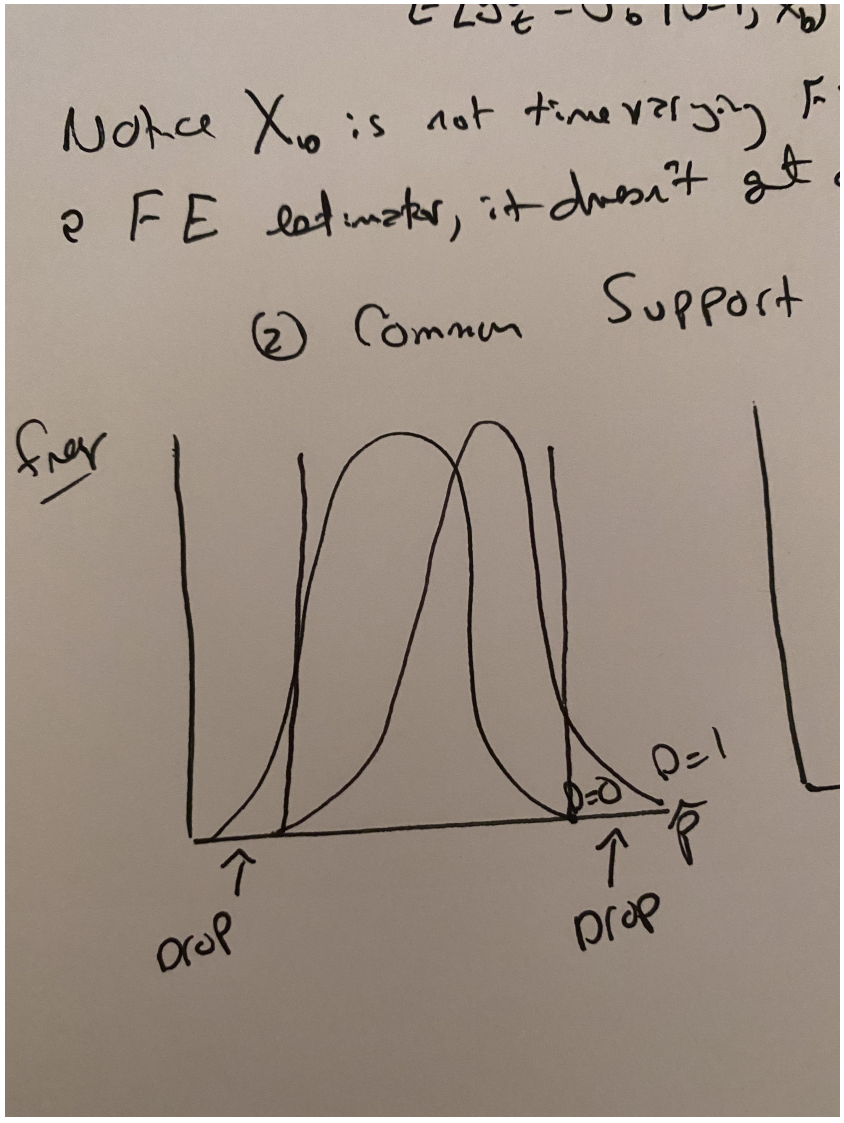
How do these work together?

Since we are identifying the ATT, and the ATT is missing Y^0 for the treated group, we are using the control group Y^0 in its place

Under conditional parallel trends and common support, some of the comparison group units are recovering the parallel trends because of their X values creating projections that in their differences perfectly aligned in expectation with the missing $\Delta E[Y^0|D=1]$

But we have to have all three for it to work

Visualizing propensity score to get common support



Definition and estimation

Defining the ATT parameter of interest

$$\begin{aligned} ATT &= E[Y_t^1 - Y_t^0 | D = 1] \\ &= E[Y_t^1 | D = 1] - E[Y_t^0 | D = 1] \end{aligned}$$

Abadie's inverse probability weighting (IPW) estimator

$$E\left[\frac{Y_t - Y_b}{Pr(D_t = 1)} \times \frac{D_t - Pr(D = 1|X_b)}{1 - Pr(D = 1|X_b)}\right]$$

The first is our causal parameter; the second is our reweighted DiD equation that *estimates* our causal parameter, but we need to estimate that propensity score

Abadie's IPW estimator

Look closely; what happens mathematically when you substitute $D = 1$ vs $D = 0$?

$$E\left[\frac{Y_t - Y_b}{Pr(D_t = 1)} \times \frac{D_t - Pr(D = 1|X_b)}{1 - Pr(D = 1|X_b)}\right]$$

The reweighting with the propensity only happens to the comparison group's first differences – not the treatment groups! Why? Because it's the Y^0 that is missing, not the Y^1

Propensity scores

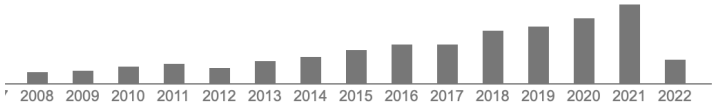
- It's common to hear people say that we don't know the propensity score; we can only estimate it. Same here – we approximate it with regressions
- Paper is titled "Semi-parametric DiD" because Abadie imposes structure on the polynomials used to construct the propensity score ("series logit")

Abadie 2005 influence

Alberto Abadie

Semiparametric difference-in-differences estimators

Authors	Alberto Abadie
Publication date	2005/1/1
Journal	The Review of Economic Studies
Volume	72
Issue	1
Pages	1-19
Publisher	Wiley-Blackwell
Description	The difference-in-differences (DID) estimator is one of the most popular tools for applied research in economics to evaluate the effects of public interventions and other treatments of interest on some relevant outcome variables. However, it is well known that the DID estimator is based on strong identifying assumptions. In particular, the conventional DID estimator requires that, in the absence of the treatment, the average outcomes for the treated and control groups would have followed parallel paths over time. This assumption may be implausible if pre-treatment characteristics that are thought to be associated with the dynamics of the outcome variable are unbalanced between the treated and the untreated. That would be the case, for example, if selection for treatment is influenced by individual-transitory shocks on past outcomes (Ashenfelter's dip). This article considers the case in which differences in observed …
Total citations	Cited by 2330



Scholar articles
Semiparametric difference-in-differences estimators
A Abadie - The Review of Economic Studies, 2005
Cited by 2330 Related articles All 12 versions

Abadie (2005) is his fourth most cited paper

Doubly Robust Difference-in-differences

- DR models control for covariates twice – once using the propensity score, once using outcomes adjusted by regression – and are unbiased so long as:
  - → The regression specification for the outcome is correctly specified
  - → The propensity score specification is correctly specified
- Sant'Anna and Zhao (2020) incorporated DR into DiD by combining inverse probability weighting and outcome regression into a single DiD model
- It's in the engine of Callaway and Sant'Anna (2020) that we discuss later so it merits close study

Identification assumptions I: Data

Assumption 1: Assume panel data or repeated cross-sectional data

Handling repeated cross-sectional data is possible but assumes stationarity which is a kind of stability assumption, but I'll use panel representation.

Cross-sections will be potentially violated with changing sample compositions (e.g., the Napster example).

Identification assumptions II: Modification to parallel trends

Assumption 2: Conditional parallel trends

Counterfactual trends for the treatment group are the same as the control group for all values of X

$$E[Y_1^0 - Y_0^0 | X, D = 1] = E[Y_1^0 - Y_0^0 | X, D = 0]$$

Identification assumptions III: Common support

Assumption 3: Common support

For some $e > 0$, the probability of being in the treatment group is greater than e and the probability of being in the treatment group conditional on X is $\leq 1 - e$.

Heckman, et al doesn't use the propensity score so we need a more general expression of support

Estimating DD with Assumptions 1-3

- Assumptions 1-3 gives us a couple of options of estimating the DiD
- We can either use the outcome regression (OR) approach of Heckman, et al 1997 (will require correct model too)
- Or we can use the inverse probability weighting (IPW) approach of Abadie (2005) (will require correct model too)

Outcome regression

This is the Heckman, et al. (1997) approach where the potential outcome evolution for the treatment group is imputed with a regression based only on X_b for the control group *only*

$$\widehat{\delta}^{OR} = \overline{Y}_{1,1} - \left[\overline{Y}_{1,0} + \frac{1}{n^T} \sum_{i|D_i=1} (\widehat{\mu}_{0,1}(X_i) - \widehat{\mu}_{0,0}(X_i)) \right]$$

where $\overline{Y}$ is the sample average of Y among units in the treatment group at time t and $\widehat{\mu}(X)$ is an estimator of the true, but unknown, $m_{d,t}(X)$ which is by definition equal to $E[Y_t|D = d, X = x]$.

Outcome regression

$$\widehat{\delta}^{OR} = \overline{Y}_{1,1} - \left[\overline{Y}_{1,0} + \frac{1}{n^T} \sum_{i|D_i=1} (\widehat{\mu}_{0,1}(X_i) - \widehat{\mu}_{0,0}(X_i)) \right]$$

1. Regress changes ΔY on X among untreated groups using baseline covariates only
2. Get fitted values of the regression using all X from $D = 1$ only. Average those
3. Calculate change in this fitted Y among treated with the average fitted values

Inverse probability weighting

This is the Abadie (2005) approach where we use weighting

$$\widehat{\delta}^{ipw} = \frac{1}{E_N[D]} E\left[\frac{D - \widehat{p}(X)}{1 - \widehat{p}(X)}(Y_1 - Y_0)\right]$$

where $\widehat{p}(X)$ is an estimator for the true propensity score. Reduces the dimensionality of X into a single scalar.

These models cannot be ranked

- Outcome regression needs $\widehat{\mu}(X)$ to be correctly specified, whereas
- Inverse probability weighting needs $\widehat{p}(X)$ to be correctly specified
- It's hard to "rank" these two in practice with regards to model misspecification because each is inconsistent when their own models are misspecified
- But what if you could do both of them at the same time and not pay for it?

Double Robust DR

- Doubly robust combines them to give us insurance; we now get two chances to be wrong, as opposed to just one
- Two papers:
  1. Chang (2020) incorporates DR with double/debiased ML
  2. Sant'Anna and Zhau (2020) is based on the IPW (Abadie 2005) and OR (Heckman, Ichimura and Todd 1997)
- For now, I've prepped the latter, but will soon get Chang (2020) incorporated – I just have been relying on Brigham Frandsen to teach the DML material

Double Robust DiD

$$\delta^{dr} = E\left[\left(\frac{D}{E[D]} - \frac{\frac{p(X)(1-D)}{(1-p(X))}}{E\left[\frac{p(X)(1-D)}{(1-p(X))}\right]}\right)(\Delta Y - \mu_{0,\Delta}(X))\right]$$

$p(x)$: propensity score model

$\Delta Y = Y_1 - Y_0 = Y_{post} - Y_{pre}$

$\mu_{d,\Delta} = \mu_{d,1}(X) - \mu_{d,0}(X),$ where $\mu(X)$ is a model for

$m_{d,t} = E[Y_t|D = d, X = x]$

So that means $\mu_{0,\Delta}$ is just the control group's change in average Y for each $X = x$

Double Robust DiD

$$\delta^{dr} = E\left[\left(\frac{D}{E[D]} - \frac{\frac{p(X)(1-D)}{(1-p(X))}}{E\left[\frac{p(X)(1-D)}{(1-p(X))}\right]}\right)(\Delta Y - \mu_{0,\Delta}(X))\right]$$

Notice how the model controls for X: you're weighting the adjusted outcomes using the propensity score

The reason you control for X twice is because you don't know which model is right. DR DiD frees you from making a choice without making you pay too much for it

Efficiency

- Authors exploit all the restrictions implied by the assumptions to construct semiparametric bounds
- This is where the influence function comes in, which those who have studied the DID code closely may have noticed
- One of the main results of the paper is that the DR DiD estimator is also DR for inference

Standard TWFE Model

Consider our earlier TWFE specification:

$$Y_{it} = \alpha_1 + \alpha_2 T_t + \alpha_3 D_i + \delta(T_i \times D_t) + \varepsilon_{it}$$

Just add in covariates then right?

$$Y_{it} = \alpha_1 + \alpha_2 T_t + \alpha_3 D_i + \delta(T_i \times D_t) + \theta \cdot X_{it} + \varepsilon_{it}$$

Sure! If you're willing to impose three *more* assumptions

Decomposing TWFE with covariates

TWFE places restrictions on the DGP. Previous TWFE regression under assumptions 1-3 implies the following:

$$E[Y_1^1|D=1,X] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X$$

Conditional parallel trends implies

$$E[Y_1^0 - Y_0^0|D=1,X] = E[Y_1^0 - Y_0^0|D=0,X]$$
$$E[Y_1^0|D=1,X] - E[Y_0^0|D=1,X] = E[Y_1^0|D=0,X] - E[Y_0^0|D=0,X]$$
$$E[Y_1^0|D=1,X] = E[Y_0^0|D=1,X] + E[Y_1^0|D=0,X] - E[Y_0^0|D=0,X]$$
$$E[Y_1^0|D=1,X] = E[Y_0|D=1,X] + E[Y_1|D=0,X] - E[Y_0|D=0,X]$$

Switching equation substitution

Last line from the switching equation. This gives us:

$$E[Y_1^0|D=1,X] = \alpha_1 + \alpha_2 + \alpha_3 + \theta X$$

Now compare this with our earlier Y^1 expression

$$E[Y_1^1|D=1,X] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X$$

We can define our target parameter, the ATT, now in terms of the fixed effects representation

Collecting terms

TWFE representation of our conditional expectations of the potential outcomes

$$E[Y_1^1|D=1,X] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta_1 X$$
$$E[Y_1^0|D=1,X] = \alpha_1 + \alpha_2 + \alpha_3 + \theta_2 X$$

Substitute these into our target parameter

$$\begin{aligned} ATT &= E[Y_1^1|D=1,X] - E[Y_1^0|D=1,X] \\ &= (\alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta_1 X) - (\alpha_1 + \alpha_2 + \alpha_3 + \theta_2 X) \\ &= \delta + (\theta_1 X - \theta_2 X) \end{aligned}$$

What if $\theta_1 X \neq \theta_2 X$?

Assumption 4: Homogeneous treatment effects in X

TWFE requires homogenous treatment effects in X (i.e., the treatment effect is the same for all X)

If X is sex, then effects are the same for males and females.

If X is continuous, like income, then the effect is the same whether someone makes $1 or $1 million.

X-specific trends

TWFE also places restrictions on covariate trends for the two groups too. Take conditional expectations of our TWFE equation.

$$
\begin{aligned}
E[Y_1|D=1] &= \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X_{11} \\
E[Y_0|D=1] &= \alpha_1 + \alpha_3 + \theta X_{10} \\
E[Y_1|D=0] &= \alpha_1 + \alpha_2 + \theta X_{01} \\
E[Y_0|D=0] &= \alpha_1 + \theta X_{00}
\end{aligned}
$$

X-specific trends

Now take the DiD formula:

$$\delta^{DD} = \Big((\alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X_{11}) - (\alpha_1 + \alpha_3 + \theta X_{10})\Big) - \Big((\alpha_1 + \alpha_2 + \theta X_{01}) - (\alpha_1 + \theta X_{00})\Big)$$

Eliminating terms, we get:

$$\delta^{DD} = \delta + (\theta X_{11} - \theta X_{10}) - (\theta X_{01} - \theta X_{00})$$

Second line requires that trends in X for treatment group equal trends in X for control group.

Assumption 5 and 6

We need “no X-specific trends” for the treatment group (assumption 5) and comparison group (assumption 6)

Intuition: No X-specific trends means the evolution of potential outcome Y^0 is the same regardless of X. This would mean you cannot allow rich people to be on a different trend than poor people, for instance.

Without these six, in general TWFE will not identify ATT.

Why not both?

- Let's review the problem. What if you claim you need X for conditional parallel trends?
- You have three options:
  1. Outcome regression (Heckman, et al. 1997) – needs Assumptions 1-3
  2. Inverse probability weighting (Abadie 2005) – needs Assumptions 1-3
  3. TWFE (everybody everywhere all the time) – needs Assumptions 1-6
- Problem is 1 and 2 need the models to be correctly specified
- Let's look at a couple of Monte Carlos – one by Pedro Sant'Anna, and then one by me

Monte Carlo details

- Compare DR with TWFE, OR and IPW
- Sample size is 1,000
- 10,000 Monte Carlo experiments
- Propensity score estimated with logit; OR estimated using linear specification

Table: Monte Carlo Simulations, DGP1, Both OR and Propensity score correct

	Bias	**RMSE**	**SE**	**Coverage**	**CI length**
TWFE	-20.9518	21.1227	2.5271	0.000	9.9061
OR	-0.0012	0.1005	0.1010	0.9500	0.3960
IPW	0.0257	2.7743	2.6636	0.9518	10.4412
DR	-0.0014	0.1059	0.1052	0.9473	0.4124

Figure 1: Monte Carlo for DID estimators, DGP1: Both pscore and OR are correctly specified

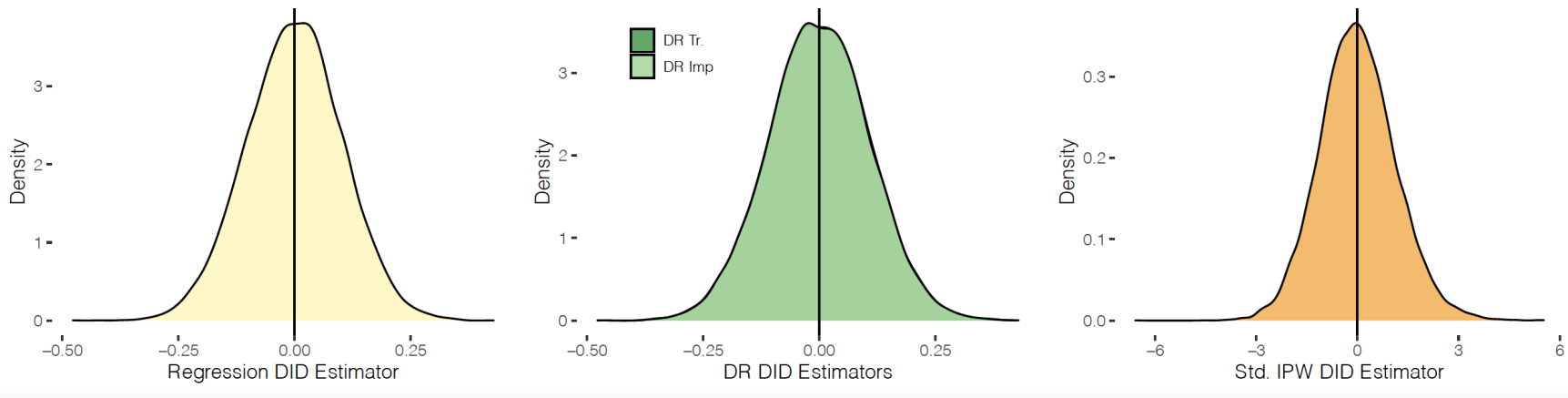
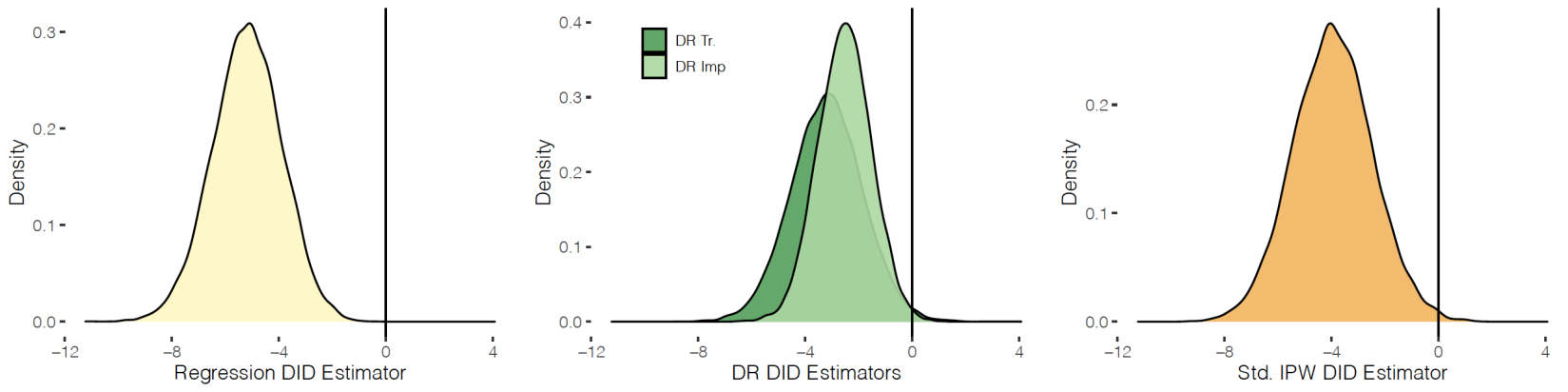


Table: Monte Carlo Simulations, DGP4, Neither OR and Propensity score correct

	Bias	**RMSE**	**SE**	**Coverage**	**CI length**
TWFE	-16.3846	16.5383	3.6268	0.000	14.2169
OR	-5.2045	5.3641	1.2890	0.0145	5.0531
IPW	-1.0846	2.6557	2.3746	0.9487	9.3084
DR	-3.1878	3.4544	1.2946	0.3076	5.0749

Figure 4: Monte Carlo for DID estimators, DGP4: Both OR and PS are misspecified



R and Stata Code

There is code in R and Stata (all DiD estimators are now beautifully arranged at a website hosted by Asjad Naqvi)

- Stata: **drdid**
- R: **drdid**

https://asjadnaqvi.github.io/DiD/docs/01_stata/

Remember – it's for 2x2 with covariates (i.e., one treatment group).

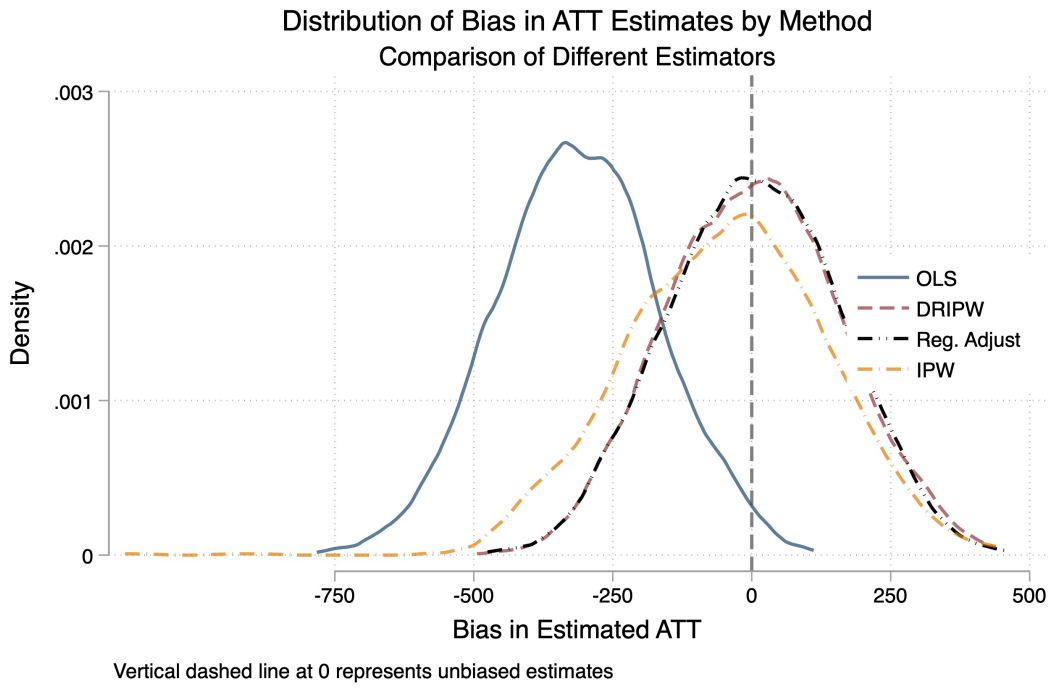
Roadmap

- Conditional Parallel Trends
  - Introducing Covariates
  - Choosing Covariates
  - Checking for Imbalance
- Estimators
  - Inverse probability weighting
  - Outcome Regression and Canonical OLS with Additive Controls
  - Double Robust
  - Canonical TWFE with Additive Covariates
- Applications
  - Simulations
  - LaLonde dataset
- Concluding Remarks

Fake Data

- First we will look at the use of these estimators using a simulation named `covariates.do` and `covariates.R`
- We will do it both with a single run, as that's faster, and then run a simulation of 1,000 simulated regenerated data (i.e., Monte Carlo simulation) to get a distribution
- We will examine all four estimators: (1) OLS, (2) IPW, (3) OR and (4) DR

Simulation


Distribution of Bias in ATT Estimates by Method
Comparison of Different Estimators
.003
.002
.001
0
Density
-750
-500
-250
0
250
500
Bias in Estimated ATT
OLS
DRIPW
Reg. Adjust
IPW
Vertical dashed line at 0 represents unbiased estimates

Application using real data

- Let's now use a real example with real data and see how well this does
- Famous paper in AER by Lalonde (1986), an Orley and Card student at Princeton
- Found that most program evaluation did badly, but let's revisit it with diff-in-diff

Description of NSW Job Trainings Program

The National Supported Work Demonstration (NSW), operated by Manpower Demonstration Research Corp in the mid-1970s:

- was a temporary employment program designed to help disadvantaged workers lacking basic job skills move into the labor market by giving them work experience and counseling in a sheltered environment
- was also unique in that it **randomly assigned** qualified applicants to training positions:
  - → **Treatment group**: received all the benefits of NSW program
  - → **Control group**: left to fend for themselves
- admitted AFDC females, ex-drug addicts, ex-criminal offenders, and high school dropouts of both sexes

NSW Program

- Treatment group members were:
  - → guaranteed a job for 9-18 months depending on the target group and site
  - → divided into crews of 3-5 participants who worked together and met frequently with an NSW counselor to discuss grievances and performance
  - → paid for their work
- Control group members were randomized so the same
- Note: the randomization balanced observables and unobservables across the two arms, thus enabling the estimation of an ATE for the people who self-selected into the program

NSW Program

- Other details about the NSW program:
  - → Wages: NSW offered the trainees lower wage rates than they would've received on a regular job, but allowed their earnings to increase for satisfactory performance and attendance
  - → Post-treatment: after their term expired, they were forced to find regular employment
  - → Job types: varied within sites – gas station attendant, working at a printer shop – and males and females were frequently performing different kinds of work

NSW Data

- NSW data collection:
  - → MDRC collected earnings and demographic information from both treatment and control at baseline and every 9 months thereafter
  - → Conducted up to 4 post-baseline interviews
  - → Different sample sizes from study to study can be confusing, but has simple explanations

NSW Data

- Estimation:
  - → NSW was a randomized job trainings program; therefore estimating the average treatment effect is straightforward:

$$\frac{1}{N_t}\sum_{D_i=1} Y_i - \frac{1}{N_c}\sum_{D_i=0} Y_i \approx E[Y^1 - Y^0]$$

    in large samples assuming treatment selection is independent of potential outcomes (randomization) – i.e., $(Y^0, Y^1) \perp\!\!\!\perp D$.

- NSW worked: Treatment group participants' real earnings post-treatment (1978) was positive and economically meaningful – $\approx$ \$900 (LaLonde 1986) to \$1,800 (Dehejia and Wahba 2002) depending on the sample used

LaLonde, Robert J. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". *American Economic Review*.

LaLonde's study was **not** an evaluation of the NSW program, as that had been done, but rather an evaluation of econometric models done by:

- replacing the experimental NSW control group with non-experimental control group drawn from two nationally representative survey datasets: Current Population Survey (CPS) and Panel Study of Income Dynamics (PSID)
- estimating the average effect using non-experimental workers as controls for the NSW trainees
- comparing his non-experimental estimates to the experimental estimates of $900

LaLonde (1986)

- LaLonde's conclusion: available econometric approaches were biased and inconsistent
  - → His estimates were way off and usually the wrong sign
  - → Conclusion was influential in policy circles and led to greater push for more experimental evaluations

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*[a,b]

Name of Comparison Group[d]	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Unadjusted (2)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Adjusted[c] (3)	NSW Treatment Earnings Less Comparison Group Earnings: Post-Training Year, 1978 Unadjusted (4)	NSW Treatment Earnings Less Comparison Group Earnings: Post-Training Year, 1978 Adjusted[c] (5)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons: Without Age (6)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons: With Age (7)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78: Unadjusted (8)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78: Adjusted[c] (9)	Controlling for All Observed Variables and Pre-Training Earnings (10)
Controls	$2,063	$39	$−21	$886	$798	$847	$856	$897	$802	$662
	(325)	(383)	(378)	(476)	(472)	(560)	(558)	(467)	(467)	(506)
PSID-1	$2,043	−$15,997	−$7,624	−$15,578	−$8,067	$425	−$749	−$2,380	−$2,119	−$1,228
	(237)	(795)	(851)	(913)	(990)	(650)	(692)	(680)	(746)	(896)
PSID-2	$6,071	−$4,503	−$3,669	−$4,020	−$3,482	$484	−$650	−$1,364	−$1,694	−$792
	(637)	(608)	(757)	(781)	(935)	(738)	(850)	(729)	(878)	(1024)
PSID-3	($3,322	($455	$455	$697	−$509	$242	−$1,325	$629	−$552	$397
	(780)	(539)	(704)	(760)	(967)	(884)	(1078)	(757)	(967)	(1103)
CPS-SSA-1	$1,196	−$10,585	−$4,654	−$8,870	−$4,416	$1,714	$195	−$1,543	−$1,102	−$805
	(61)	(539)	(509)	(562)	(557)	(452)	(441)	(426)	(450)	(484)
CPS-SSA-2	$2,684	−$4,321	−$1,824	−$4,095	−$1,675	$226	−$488	−$1,850	−$782	−$319
	(229)	(450)	(535)	(537)	(672)	(539)	(530)	(497)	(621)	(761)
CPS-SSA-3	$4,548	$337	$878	−$1,300	$224	−$1,637	−$1,388	−$1,396	$17	$1,466
	(409)	(343)	(447)	(590)	(766)	(631)	(655)	(582)	(761)	(984)

[a] The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is $886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

[b] Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

[c] The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

[d] See Table 3 for definitions of the comparison groups.

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*[a,b]

Name of Comparison Group[d]	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Unadjusted (2)	Pre-Training Year, 1975 Adjusted[c] (3)	Post-Training Year, 1978 Unadjusted (4)	Post-Training Year, 1978 Adjusted[c] (5)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons Without Age (6)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons With Age (7)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78 Unadjusted (8)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78 Adjusted[c] (9)	Controlling for All Observed Variables and Pre-Training Earnings (10)
Controls	$2,063	$39	$−21	$886	$798	$847	$856	$897	$802	$662
	(325)	(383)	(378)	(476)	(472)	(560)	(558)	(467)	(467)	(506)
PSID-1	$2,043	−$15,997	−$7,624	−$15,578	−$8,067	$425	−$749	−$2,380	−$2,119	−$1,228
	(237)	(795)	(851)	(913)	(990)	(650)	(692)	(680)	(746)	(896)
PSID-2	$6,071	−$4,503	−$3,669	−$4,020	−$3,482	$484	−$650	−$1,364	−$1,694	−$792
	(637)	(608)	(757)	(781)	(935)	(738)	(850)	(729)	(878)	(1024)
PSID-3	($3,322	($455	$455	$697	−$509	$242	−$1,325	$629	−$552	$397
	(780)	(539)	(704)	(760)	(967)	(884)	(1078)	(757)	(967)	(1103)
CPS-SSA-1	$1,196	−$10,585	−$4,654	−$8,870	−$4,416	$1,714	$195	−$1,543	−$1,102	−$805
	(61)	(539)	(509)	(562)	(557)	(452)	(441)	(426)	(450)	(484)
CPS-SSA-2	$2,684	−$4,321	−$1,824	−$4,095	−$1,675	$226	−$488	−$1,850	−$782	−$319
	(229)	(450)	(535)	(537)	(672)	(539)	(530)	(497)	(621)	(761)
CPS-SSA-3	$4,548	$337	$878	−$1,300	$224	−$1,637	−$1,388	−$1,396	$17	$1,466
	(409)	(343)	(447)	(590)	(766)	(631)	(655)	(582)	(761)	(984)

[a] The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is $886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

[b] Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

[c] The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

[d] See Table 3 for definitions of the comparison groups.

Imbalanced covariates for experimental and non-experimental samples

	All		CPS Controls $N_c = 15,992$	NSW Trainees $N_t = 297$		
covariate	mean	(s.d.)	mean	mean	t-stat	diff
Black	0.09	0.28	0.07	0.80	47.04	-0.73
Hispanic	0.07	0.26	0.07	0.94	1.47	-0.02
Age	33.07	11.04	33.2	24.63	13.37	8.6
Married	0.70	0.46	0.71	0.17	20.54	0.54
No degree	0.30	0.46	0.30	0.73	16.27	-0.43
Education	12.0	2.86	12.03	10.38	9.85	1.65
1975 Earnings	13.51	9.31	13.65	3.1	19.63	10.6
1975 Unemp	0.11	0.32	0.11	0.37	14.29	-0.26

Lab

https://github.com/Mixtape-Sessions/Causal-Inference-2/tree/main/Lab/Lalonde

Together let's do questions 1 and 2a-c

Roadmap

- Conditional Parallel Trends
  - Introducing Covariates
  - Choosing Covariates
  - Checking for Imbalance
- Estimators
  - Inverse probability weighting
  - Outcome Regression and Canonical OLS with Additive Controls
  - Double Robust
  - Canonical TWFE with Additive Covariates
- Applications
  - Simulations
  - LaLonde dataset
- Concluding Remarks

Concluding remarks

- Including covariates in a DiD design is done for reasons that are different than in regressions more generally – we are trying to address a parallel trends violation
- Typical regression modeling can only incorporate covariates, but that places restrictions on the model, whereas other methods will not
- We use baseline covariates because we cannot include covariates that are outcomes otherwise it introduces its own biases
- Doubly robust and IPW incorporate covariates through propensity scores and outcome regressions (or both) using baseline covariate means only

Suggestions

- Remember that the threat to validity comes from comparing aggregate groups of units for whom differences in observables, with different associations with Y^0, violate parallel trends
- Conditional parallel trends simply requires that groups which are comparable on covariates are more likely to have comparable Y^0 trends
- Check for imbalance at baseline using normalized differences
- Choose covariates that are highly predictive of the missing potential outcome, Y^0

Design stage

- Keep in your mind that you are attempting to reassemble the dataset with the RCT metaphor in mind
- Randomization distributes mean potential outcomes the same for treatment and control
- We only need similar trends in Y^0 to be the same, and we think that's more likely within covariate strata
- So, in the design stage, think carefully about this – this is still prior to looking at the outcome data as this can inadvertently cause p-hacking
- Remember – you're the expert so use your knowledge of the phenomena to choose judiciously the covariates and be careful about causing the curse of dimensionality