Gov 2002: 9. Differences in Differences

Matthew Blackwell

October 30, 2015

1. Basic differences-in-differences model
2. Conditional DID
3. Standard error issues
4. Other DID approaches

Where are we? Where are we going?

- For every week on causal inference, we have identified a different source of exogenous variation:
- Message: simply having panel data does not identify an effect, but it does allow us to rely on different identifying assumptions.

Ladd/Lenz data

- Do newspaper endorsements affect reader's votes?
- Problem: people might read newspapers because of the underlying political positions.
  - Liberals read the NYT, conservatives read the WSJ
- Ladd and Lenz look at British newspapers that switched their endorsement to the Labour party between 1992 and 1997.
- Compare the voting trends for readers of endorsement switching papers vs. non-switching papers.

```
labour <- foreign::read.dta("LaddLenz.dta", convert.factors = FALSE)
head(labour[, c("tolabor", "vote_l_92", "vote_l_97")])
```

```
##   tolabor vote_l_92 vote_l_97
## 1       0         1         1
## 2       0         1         0
## 3       0         0         0
## 4       0         1         1
## 5       0         1         1
## 6       0         1         1
```

1/ Basic differences-in-differences model

Setup

- Basic setup: two groups, two time periods.
- At $t = 0$, neither group is treated and in period $t = 1$, one (and only one) of the groups is treated.
- Differences: changes in treated group from $t = 0$ to $t = 1$
  - Problem: Might be secular changes in the outcome
- Differences in differences (diff-in-diff, DID, DD): difference between $t = 0$ to $t = 1$ changes in treatment and control groups
  - Resolution: changes in the control group identifies the secular trend
- Examples:
  - Minimum wage changes in NJ with PA as control (Card and Krueger)
  - Effect of artillery shelling on insurgent attacks (Lyall)

Panel versus two cross sections

- Y_{igt} is the outcome for unit i in group g at time t.
- $G_i = 1$ are those that are treated at $t = 1$ and $G_i = 0$ for those that are always untreated
  - $G_i = 1$ for NJ since the minimum wage is enacted there.
- DID can be applied with panel data or two cross-sections.
- Panel:
  - Y_{igt} measured for all i at both t.
  - Could calculate individual changes: $Y_{ig1} - Y_{ig0}$
  - Ladd/Lenz data is of this variety.
- Cross-sections:
  - Y_{igt} means that unit i is only measured at t
  - Y_{ig1} means that Y_{ig0} is not observed.

Potential outcomes approach to DID

- $Y_{igt}(d)$ is the potential outcome under treatment d at time t.
- Again, the individual causal effect is just $Y_{igt}(1) - Y_{igt}(0)$.
- Treatment status in each period:
  - No treatment in the first period for either group: $D_{ig0} = 0$
  - In treated group, $G_i = 1 \rightsquigarrow D_{ig1} = 1$
  - In control group, $G_i = 0 \rightsquigarrow D_{ig1} = 0$
- Consistency: $Y_{it} = D_{igt}Y_{igt}(1) + (1 - D_{igt})Y_{igt}(0)$
  - All control p.o.s in first period: $Y_{ig0}(0) = Y_{ig0}$
  - In treated group: $G_i = 1 \rightsquigarrow Y_{i11} = Y_{i11}(1)$
  - In control group: $G_i = 0 \rightsquigarrow Y_{i01} = Y_{i01}(0)$

Constant effects linear DID model

- Start with constant effects linear model:

$$\mathbb{E}[Y_{it}(1) - Y_{it}(0)] = \tau$$

- Linear separable model:

$$\mathbb{E}[Y_{it}(0)] = \delta_t + \alpha_g$$

- Consistency plus these assumptions gives us:

$$Y_{igt} = \delta_t + \tau D_{igt} + \alpha_g + \eta_{it}$$

- Parameters:
  - Period effect: δ_t
  - Group effect α_g
  - Transitory shock/idiosyncratic error, η_{it}, with $\mathbb{E}[\eta_{it}] = 0$
- Without further assumptions, τ not identified because $G_i = D_{ig1}$ might be correlated with shocks:

$$\text{Cov}(G_i, \alpha_g) \neq 0$$

Baseline trends

- Baseline trend without treatment:

$$Y_{ig1}(0) - Y_{ig0}(0) = (\delta_1 - \delta_0) + (\eta_{i1} - \eta_{i0})$$

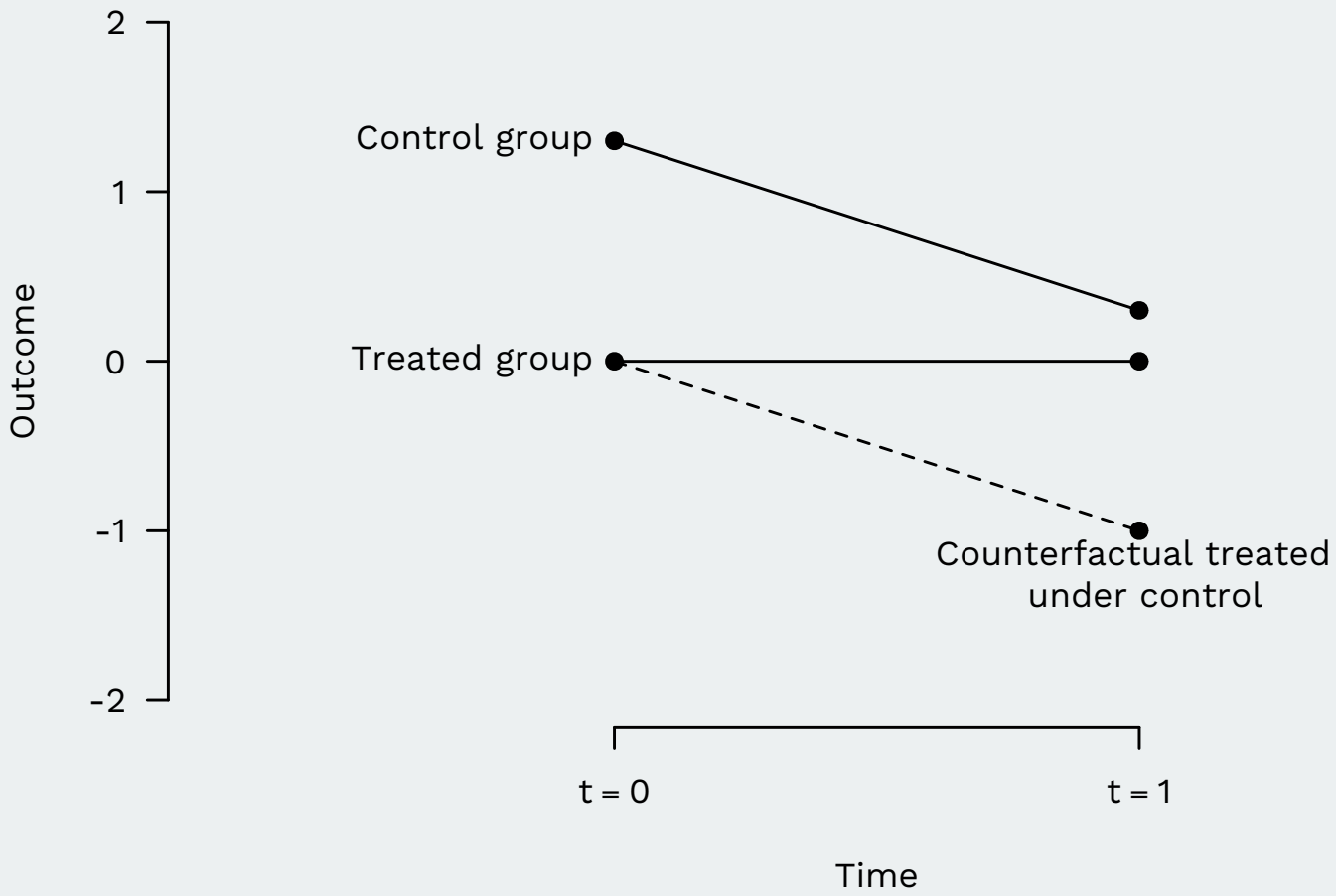
- Assumption Idiosyncratic errors are also independent of the treatment:

$$\mathbb{E}[\eta_{i1}|G_i] = \mathbb{E}[\eta_{i0}|G_i] = 0$$

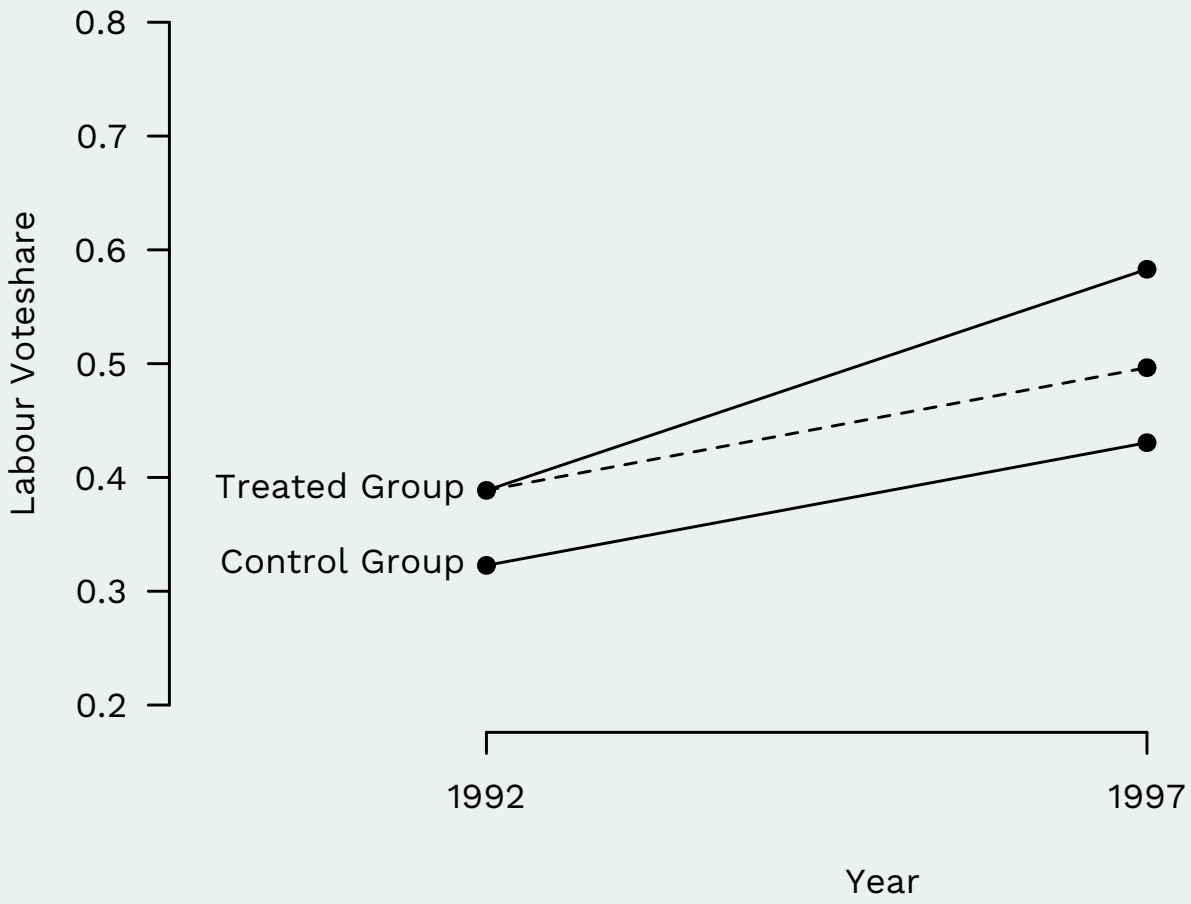
- Baseline trends are independent of G_i:

$$\begin{aligned}
\mathbb{E}[Y_{ig1}(0) - Y_{ig0}(0)|G_i = 1] &= (\delta_1 - \delta_0) + \mathbb{E}[(\eta_{i1} - \eta_{i0})|G_i] \\
&= (\delta_1 - \delta_0) + \mathbb{E}[(\eta_{i1} - \eta_{i0})] \\
&= (\delta_1 - \delta_0) \\
&= \mathbb{E}[Y_{ig1}(0) - Y_{ig0}(0)]
\end{aligned}$$

Common trends in a graph



Ladd/Lenz plot



Identification

- Remember that we are comparing $t = 1$ to $t = 0$.
- With this assumption, we can rewrite the above model as the following:

$$Y_{igt} = \mu + \delta \mathbb{1}(t = 1) + \gamma G_i + \tau(\mathbb{1}(t = 1) \times G_i) + \varepsilon_{igt}$$

- The parameters are the following:
  - Baseline trend: $\delta = \mathbb{E}[Y_{ig1}(0) - Y_{ig0}(0)] = (\delta_1 - \delta_0)$
  - Control start: $\mu = \mathbb{E}[Y_{ig0}(0)] = \mathbb{E}[\alpha_g | G_i = 0] + \delta_0$
  - Baseline differences: γ

$$\begin{aligned}\gamma &= \mathbb{E}[Y_{ig0}(0)|G_i = 1] - \mathbb{E}[Y_{ig0}(0)|G_i = 0] \\ &= \mathbb{E}[\alpha_g|G_i = 1] - \mathbb{E}[\alpha_g|G_i = 0]\end{aligned}$$

  - New error: $\varepsilon_{igt} = Y_{igt}(0) - \mathbb{E}[Y_{igt}(0)|D_{igt}] = \alpha_g - \mathbb{E}[\alpha_g|G_i] + \eta_{it}$

Independence of new errors

- Using the above assumption, we can show that the treatment is independent of the error in this model:

$$\begin{aligned}
\mathbb{E}[\varepsilon_{igt}|D_{ig1}, D_{ig0}] &= \mathbb{E}[\varepsilon_{igt}|G_i] \\
&= \mathbb{E}[(\alpha_g - \mathbb{E}[\alpha_g|G_i] + \eta_{it})|G_i] \\
&= \mathbb{E}[\alpha_g|G_i] - \mathbb{E}[\mathbb{E}[\alpha_g|G_i]|G_i] + \mathbb{E}[\eta_{it}|G_i] \\
&= \mathbb{E}[\eta_{it}|G_i] \\
&= \mathbb{E}[\eta_{it}] = 0
\end{aligned}$$

- No assumptions about relationship between G_i and α_g.
- Just assumed independence of idiosyncratic errors: $\mathbb{E}[\eta_{it}|G_i] = 0$

Motivating DID

- Under common trends, control group identifies the baseline trend:

$$E[Y_{igt}|G_i = 0, t = 1] - E[Y_{igt}|G_i = 0, t = 0] = \delta$$

- The treated group is the baseline trend plus the treatment effect:

$$E[Y_{igt}|G_i = 1, t = 1] - E[Y_{igt}|G_i = 1, t = 0] = \delta + \tau$$

- Differences-in-differences:

$$\begin{aligned}&\big(E[Y_{igt}|G_i = 1, t = 1] - E[Y_{igt}|G_i = 1, t = 0]\big)\\ &- \big(E[Y_{igt}|G_i = 0, t = 1] - E[Y_{igt}|G_i = 0, t = 0]\big) = \tau\end{aligned}$$

- We can estimate each of these CEFs from the data and compute their sample versions to get an estimate of τ.

Estimation

- For the two period, binary treatment case, a regression of the outcome on time (pre-treatment $t = 0$, post-treatment $t = 1$), treated group, and their interaction can estimate τ:

$$Y_{igt} = \mu + \delta \mathbb{1}(t = 1) + \gamma G_i + \tau(\mathbb{1}(t = 1) \times G_i) + \varepsilon_{igt}$$

- $\widehat{\tau}$ would be the coefficient on the interaction between time and the treatment.

More than 2 periods/groups

$$Y_{igt} = \mu_g + \delta_t + \tau(I_{it} \times G_i) + \varepsilon_{igt}$$

- I_{it} is an indicator for the intervention being applied at time s:
  - $I_{it} = 1$ when $t > s$.
  - $I_{it} = 0$ when $t \leq s$.
  - Also include period effects δ_t
- More than two groups: include fixed effects for each group, μ_g
- Commonly referred to as two-way fixed effects model.

Panel data

- If we have panel data, then we can estimate this in a different, more direct way. Note that:

$$\tau = E[Y_{i1} - Y_{i0}|G_i = 1] - E[Y_{i1} - Y_{i0}|G_i = 0]$$

- Thus, in the panel data case, we can estimate the effect by regressing the change for each unit, $Y_{i1} - Y_{i0}$, on the treatment.
- Notice that this is the same as a first difference approach since $D_{ig0} = 0$ for all g and so $G_i = \Delta D_{igt}$.

Ladd/Lenz example

```
summary(lm(I(vote_l_97 - vote_l_92) ~ tolabor, data = labour))
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1078     0.0110    9.84   <2e-16 ***
## tolabor       0.0865     0.0301    2.87   0.0041 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.41 on 1591 degrees of freedom
## Multiple R-squared:  0.00516,    Adjusted R-squared:  0.00454
## F-statistic: 8.25 on 1 and 1591 DF,  p-value: 0.00412
```

Threats to identification

- Obviously, the treatment needs to be independent of the idiosyncratic shocks so that the variation of the outcome is the same for the treated and control groups, but this might not be plausible.
- Ashenfelter's dip: which is a empirical finding that people who enroll in job training programs see their earnings decline prior to that training.
- In the Ladd/Lenz paper, perhaps Labour leaning people selected into reading newspapers that we Labour leaning and thus both the editors and readers were changing together.
- Thus, the independence of the treatment and idiosyncratic shocks might only hold conditional on covariates.

Robustness checks

- Lags and Leads
  - if D_{igt} causes Y_{igt}, and not the other way around, then current and lagged values of D_{igt} should have an effect on Y_{igt}, but future values of D_{igt} should not.
  - Re-estimate model by lagging or leading the intervention indicator: $T'_{it} = 1$ when $t > s + 1$.
- Time trends
  - With more than two time periods, we can add unit-specific linear trends to the regression DID model:

$$Y_{igt} = \delta_t + \tau G_i + \alpha_{0g} + \alpha_{1g} \cdot t + \varepsilon_{igt}$$

  - Helps detect if there really are varying trends, if estimated from pre-treatment data.

2/ Conditional DID

Nonparametric identification

- Up until now, we assumed a linear separable model and constant treatment effects. Can we identify things nonparametrically?
- Ease the notation: $Y_{it} = Y_{iG_i t}$
- Key assumption is parallel trends:

$$E[Y_{i1}(0) - Y_{i0}(0)|X_i, G_i = 1] = E[Y_{i1}(0) - Y_{i0}(0)|X_i, G_i = 0]$$

- What does this assumption say? It says that the potential trend under control is the same for the control and treated groups, conditional on covariates.

- We can show that this is the key assumption for identifying the ATT:

$$
\begin{aligned}
&\mathbb{E}[Y_{i1}(1) - Y_{i1}(0)|X_i, G_i = 1] \\
=&\mathbb{E}[Y_{i1}(1) - Y_{i0}(0) + Y_{i0}(0) - Y_{i1}(0)|X_i, G_i = 1] \\
=&(\mathbb{E}[Y_{i1}(1)|X_i, G_i = 1] - \mathbb{E}[Y_{i0}(0)|X_i, G_i = 1]) - (\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|X_i, G_i = 1]) \\
=&(\mathbb{E}[Y_{i1}(1)|X_i, G_i = 1] - \mathbb{E}[Y_{i0}|X_i, G_i = 1]) - (\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|X_i, G_i = 0]) \\
=&(\mathbb{E}[Y_{i1}|X_i, G_i = 1] - \mathbb{E}[Y_{i0}|X_i, G_i = 1]) - (\mathbb{E}[Y_{i1}(0)|X_i, G_i = 0] - \mathbb{E}[Y_{i0}(0)|X_i, G_i \\
=&\underbrace{(\mathbb{E}[Y_{i1}|X_i, G_i = 1] - \mathbb{E}[Y_{i0}|X_i, G_i = 1])}_{\text{differences for } G_i=1} - \underbrace{(\mathbb{E}[Y_{i1}|X_i, G_i = 0] - \mathbb{E}[Y_{i0}|X_i, G_i = 0])}_{\text{differences for } G_i=0}
\end{aligned}
$$

- $\rightsquigarrow$ the unconditional ATT:

$$\mathbb{E}[Y_{i1}(1) - Y_{i1}(0)|G_i = 1] = \mathbb{E}_X\left[\mathbb{E}[Y_{i1}(1) - Y_{i1}(0)|X_i, G_i = 1]\right]$$

- Each CEF could be estimated nonparametrically $\rightsquigarrow$ curse of dimensionality
- Can't identify the ATE because $\mathbb{E}[Y_{i1}(1)|X_i, G_i = 0]$ is unrestricted.

Nonparametric DID notes

- Note what is powerful here: no ignorability assumption. Relies only on parallel trends assumption.
- Sometimes we need higher order differences (diff-in-diff-in-diff).
- No obvious linearity assumption, but differences are a key part of the assumption:
  - Trends on one scale might not be parallel on another scale.
- With covariates, three general approaches (sound familiar?):
  - Regression DID, using linearity assumptions for X_i.
  - Matching on X_i, then using regular DID.
  - Weighting based on the propensity score.

Regression DD

- A Regression DID includes X_i in a linear, additive manner:

$$Y_{it} = \mu + X_i' \beta_t + \delta \mathbb{1}(t = 1) + \gamma G_i + \tau(\mathbb{1}(t = 1) \times G_i) + \varepsilon_{it}$$

- If we have repeated observations, we can take the differences between $t = 0$ and $t = 1$:

$$Y_{i1} - Y_{i0} = \delta + X_i' \beta + \tau G_i + (\varepsilon_{i1} - \varepsilon_{i0})$$

- Here, we have $\beta = \beta_1 - \beta_0$. Further note that because everyone is untreated in the first period, $D_{i1} - D_{i0} = D_{i1}$.
- As usual, for panel data, regress changes on treatment.
- This approach depends on constant effects and linearity in X_i. Could use matching to reduce model dependence here.

Ladd/Lenz with covariates

```
summary(lm(I(vote_l_97-vote_l_92) ~ tolabor + parent_labor,
           data = labour))
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.1151     0.0133    8.66   <2e-16 ***
## tolabor        0.0882     0.0302    2.92   0.0035 **
## parent_labor  -0.0207     0.0212   -0.97   0.3309
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.41 on 1590 degrees of freedom
## Multiple R-squared:  0.00575,    Adjusted R-squared:  0.0045
## F-statistic:  4.6 on 2 and 1590 DF,  p-value: 0.0102
```

Semiparametric estimation with repeated outcomes

- How to estimate regression DID without strong linearity assumptions?
- Abadie (2005) on how to use weighting estimators to help with estimation.
- Basically, we are going to weight the treated and control groups so that they are balanced on the covariates.
- Abadie shows that:

$$\mathbb{E}[Y_{i1}(1) - Y_{i1}(0)|G_i = 1] = \mathbb{E}\left[\frac{(Y_{i1} - Y_{i0})}{\mathbb{P}(G_i = 1)} \cdot \frac{G_i - \mathbb{P}(G_i = 1|X_i)}{1 - \mathbb{P}(G_i = 1|X_i)}\right]$$

- Have to estimate the propensity score for being in the treated group $\mathbb{P}(G_i = 1|X_i)$
- Weights are slightly different here than with IPTW because we're interested in the ATT.

3/ Standard error issues

Serial correlation and placebo tests

$$Y_{igt} = \mu_g + \delta_t + \tau(I_{it} \times G_i) + \nu_{gt} + \varepsilon_{igt}$$

- Bertrand et al (2004) highlight the problem of serial correlation in ν_{gt}
- Placebo test:
  - Outcome: CPS data on state-level female wages from 1979 and 1999
  - Placebo treatment: randomly pick a fake intervention between 1985 and 1995.
  - Placebos significant 45% of time at the 5% level.
- Solutions:
  - Clustered SEs at the group-level
  - Block bootstrap on group (repeatedly sample entire g vectors rather than it rows)
  - Aggregate to g units with two time periods each: pre- and post-intervention.
- All solutions depend on large numbers of groups.

Cluster-robust SEs

- First, let's write the within-group regressions like so:

$$\mathbf{y}_g = \mathbf{X}_g \beta + \varepsilon_g$$

- $\mathbf{y}_g$ is the vector of responses for group g, let $\mathbf{X}$ be all the $\mathbf{X}_g$ stacked into one matrix.
- We assume that respondents are independent across units, but possibly dependent within clusters. Thus, we have

$$\text{Var}[\varepsilon_g | \mathbf{X}_g] = \Sigma_g$$

- Under clustered dependence, we can write the sandwich variance like so:

$$\text{Var}[\hat{\beta} | \mathbf{X}] = \mathbb{E}\left[\mathbf{X}_g \mathbf{X}_g'\right]^{-1} \mathbb{E}\left[\mathbf{X}_g' \Sigma_g \mathbf{X}_g\right] \mathbb{E}\left[\mathbf{X}_g \mathbf{X}_g'\right]^{-1}$$

- Using the plug-in principle:

$$\text{Var}[\hat{\beta} | \mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_g \mathbf{X}_g' \Sigma_g \mathbf{X}_g \right) (\mathbf{X}'\mathbf{X})^{-1}$$

Estimating CRSEs

- Way to estimate this matrix: replace Σ_g with an estimate based on the within-cluster residuals, $\hat{\varepsilon}_g$:

$$\widehat{\Sigma}_g = \hat{\varepsilon}_g \hat{\varepsilon}_g'$$

- Final expression for our cluster-robust covariance matrix estimate:

$$\widehat{\text{Var}}[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_g \mathbf{X}_g' \hat{\varepsilon}_g \hat{\varepsilon}_g' \mathbf{X}_g \right) (\mathbf{X}'\mathbf{X})^{-1}$$

- With small-sample adjustment (which is what most software packages report):

$$\widehat{\text{Var}}_a[\hat{\beta}|\mathbf{X}] = \frac{m}{m-1} \frac{n-1}{n-k-1} (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_g \mathbf{X}_g' \hat{\varepsilon}_g \hat{\varepsilon}_g' \mathbf{X}_g \right) (\mathbf{X}'\mathbf{X})^{-1}$$

Block bootstrap

- Block bootstrap is when we bootstrap whole groups (states, etc) instead of it pairs.
  - Accounts for correlations within the groups (serial correlation, etc).
- Let $g \in 1, \ldots, M$ be the group indices. Procedure:
  1. Randomly sample M indices with replacement: $g_b^* = (g_{b1}^*, \cdots, g_{bM}^*)$
  2. Grab the outcome vector and covariate matrix for each sampled index: $(\mathbf{y}_{g_{b1}^*}, \mathbf{X}_{g_{b1}^*})$
  3. Stack all of these together into one data matrix: $\mathbf{y}_b^*$, $\mathbf{X}_b^*$
  4. Estimate $\widehat{\tau}_b^*$ from a DID model with $\mathbf{y}_b^*$, $\mathbf{X}_b^*$.
  5. Repeat 1-4 B times to get the bootstrapped distribution of $\widehat{\tau}$
- Theoretically very simple: just bootstrap groups.
- Computationally tricky because you need to keep track of all the multilevel indices.

Block bootstrap coding

- Toy data:

```
dat[c(1:2, 6:7), ]
```

```
##   groups d      y
## 1      A 0 -0.604
## 2      A 1 -0.074
## 6      B 1 -0.142
## 7      B 0 -1.332
```

- Trick to get the indices for each group:

```
lookup <- split(1:nrow(dat), dat$groups)
lookup[1]
```

```
## $A
##  [1]   1   2   3   4   5  26  27  28  29  30  51  52  53  54  55  76
## [17]  77  78  79  80 101 102 103 104 105 126 127 128 129 130 151 152
## [33] 153 154 155 176 177
```

Block bootstrap coding

- Take one sample of groups:

```
gnames <- names(lookup)
star <- sample(gnames, size = length(gnames), replace = TRUE)
head(lookup[star], n = 2)
```

```
## $C
##  [1]  11  12  13  14  15  36  37  38  39  40  61  62  63  64  65  86
## [17]  87  88  89  90 111 112 113 114 115 136 137 138 139 140 161 162
## [33] 163 164 165
##
## $C
##  [1]  11  12  13  14  15  36  37  38  39  40  61  62  63  64  65  86
## [17]  87  88  89  90 111 112 113 114 115 136 137 138 139 140 161 162
## [33] 163 164 165
```

- Use `unlist()` to get all of the indices from your sample:

```
dat.star <- dat[unlist(lookup[star]), ]
```

4/ Other DID approaches

Changes-in-changes

- Athey and Imbens (2006) generalize DID to handle looking at different changes in the distribution of Y_{it}
- Basic idea: relative distribution of units doesn't change across time.
  - Suppose someone went from 5th percentile to the median in the control group.
  - A treated unit with the same pretreatment outcome would have had the same change had they been the control group.
- Estimate the CDF and inverse CDF of the control group distributions at $t = 0$ and $t = 1$ to impute the counterfactual changes of the treated group over time.
- Can use these estimates to get ATT, or any change in the distribution (quantiles, variance, etc).
- Requires more data to estimate the CDFs.

Synthetic controls

- Abadie and Gardeazabal (2003) use a DID approach for "quantitative case studies."
- Application: effect of an intervention in a single country or state at one point in time.
- Basic idea: 1 treated group, many controls.
  - Compare the time-series of the outcome in the treated group to the control.
  - But which control group should you use? So many and they may not be comparable to the treated.
- Synthetic control: use a convex combination of the controls to create a synthetic control.
  - Choose the weights that minimize the pretreatment differences between treated and synthetic control.

Without synthetic controls

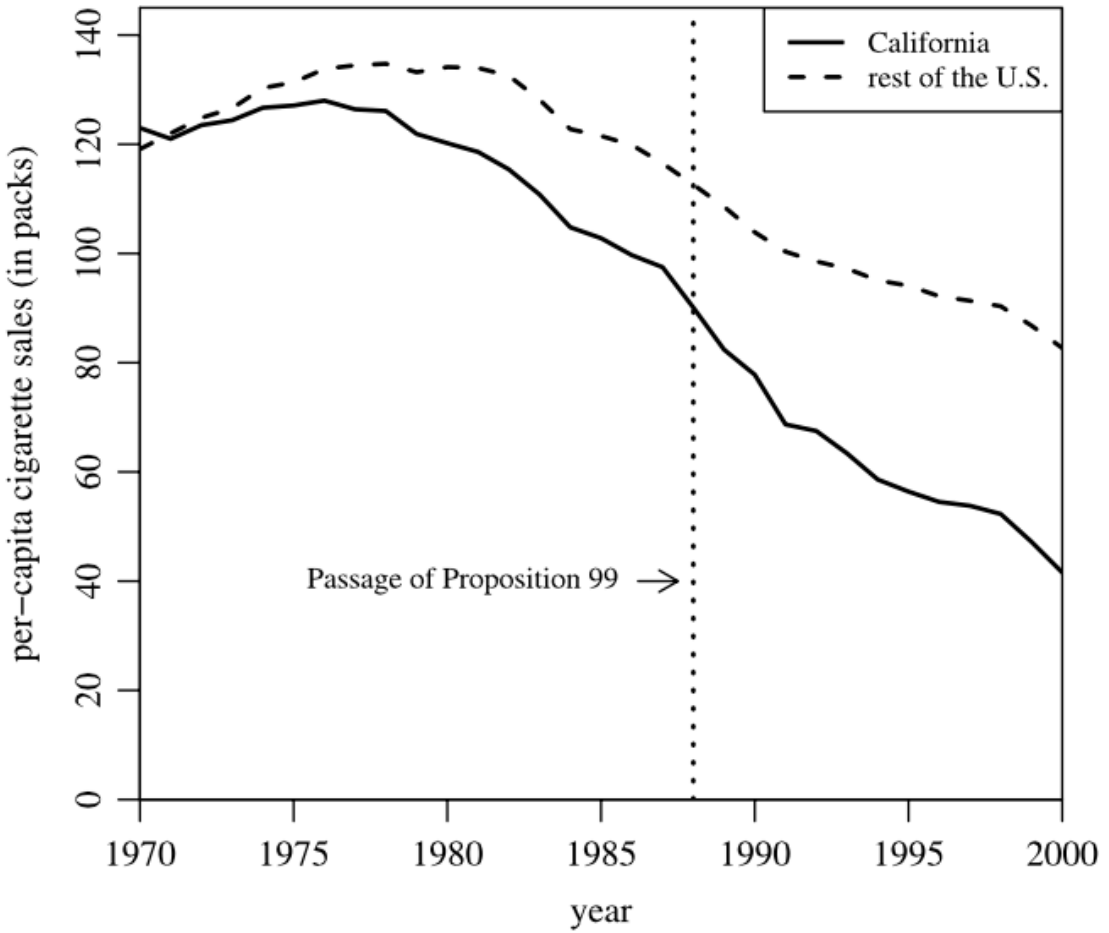


Figure 1. Trends in per-capita cigarette sales: California vs. the rest of the United States.

With synthetic controls

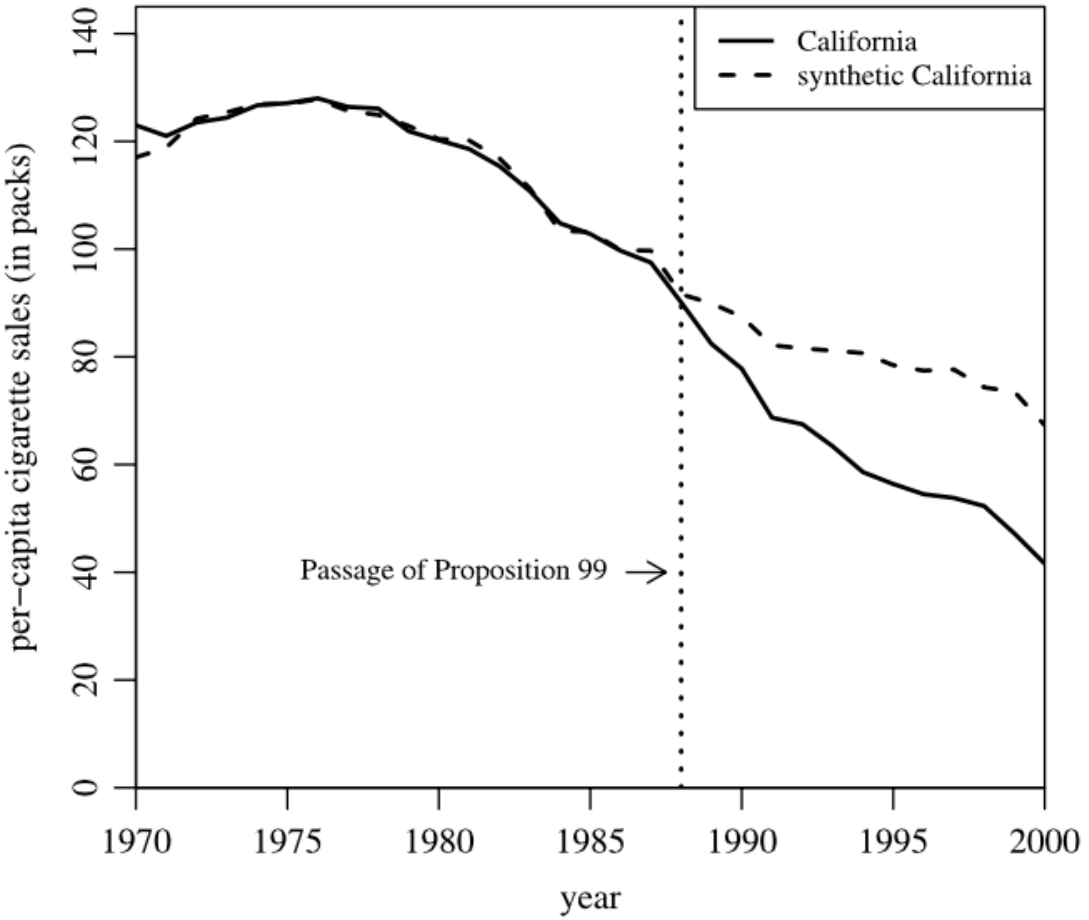


Figure 2. Trends in per-capita cigarette sales: California vs. synthetic California.