Applied Methods for PhD, Week 08

Difference-in-Differences

Michael E. Kummer
Theoretical Slide Set, inspired by Stephen Kastoryano

Nova SBE, OTIM

DIFFERENCE-IN-DIFFERENCES

- Previous two lectures about panel data: data on multiple entities (individuals, firms, etc.) observed over multiple time periods

$$Y_{it} = X_{it}\beta + \eta_i + U_{it} \quad i = 1, \ldots, N \quad t = 1, \ldots, T$$

- Policies/Treatment may take place at the aggregate level affecting only some municipalities/ states/ countries but not others.
- Source of omitted variable bias is at aggregate level.
- we observe individual outcomes both before and after receiving treatment then we can sometimes account for omitted variable bias and apply difference-in-differences (DiD).

DIFFERENCE-IN-DIFFERENCES

- Let Y_{i0} denote the outcome of individual i in year $t = 0$ before the treatment and Y_{i1} the outcome in year $t = 1$.
- The variable D_i indicates if individual i received treatment between year $t = 0$ and $t = 1$.
- DiD estimator assumes common trend

$$\widehat{\Delta} = (\overline{Y}_{treatment,after} - \overline{Y}_{treatment,before}) - (\overline{Y}_{control,after} - \overline{Y}_{control,before})$$

where

$$\overline{Y}_{treatment,after} = \frac{\sum_{i=1}^{N} D_i Y_{i1}}{\sum_{i=1}^{N} D_i} \quad \text{and} \quad \overline{Y}_{treatment,before} = \frac{\sum_{i=1}^{N} D_i Y_{i0}}{\sum_{i=1}^{N} D_i}$$

$$\overline{Y}_{control,after} = \frac{\sum_{i=1}^{N} (1 - D_i) Y_{i1}}{\sum_{i=1}^{N} (1 - D_i)} \quad \text{and} \quad \overline{Y}_{control,before} = \frac{\sum_{i=1}^{N} (1 - D_i) Y_{i0}}{\sum_{i=1}^{N} (1 - D_i)}$$

Table of Contents

TWO GROUPS & TWO TIME PERIODS

- Write outcomes for two groups $g = d, c$ as $Y_{ict} = Y_{it}|D_i = 0$ for the control group and $Y_{idt} = Y_{it}|D_i = 1$ for the treated group.
- In year $t = 0$ the outcomes are described by

$$\mathrm{E}[Y_{ic0}] = \alpha_c + \lambda_0 \quad \text{and} \quad \mathrm{E}[Y_{id0}] = \alpha_d + \lambda_0$$

- Assume homogenous treatment effect $\mathrm{E}[Y_{i1} - Y_{i0}] = \delta$.
- Outcomes in year $t = 1$ given by

$$\mathrm{E}[Y_{ic1}] = \alpha_c + \lambda_1 \quad \text{and} \quad \mathrm{E}[Y_{id1}] = \alpha_d + \delta + \lambda_1$$

- Comparing outcomes for treated and controls after treatment $\mathrm{E}[Y_{id1}] - \mathrm{E}[Y_{ic1}] = \delta + (\alpha_d - \alpha_c)$
- Comparing outcomes for treated before and after treatment $\mathrm{E}[Y_{id1}] - \mathrm{E}[Y_{id0}] = \delta + (\lambda_1 - \lambda_0)$
- DiD controls the additional change in trend

$$\begin{aligned} DiD &= (\mathrm{E}[Y_{id1}] - \mathrm{E}[Y_{id0}]) - (\mathrm{E}[Y_{ic1}] - \mathrm{E}[Y_{ic0}]) \\ &= ((\delta + \alpha_d + \lambda_1) - (\alpha_d + \delta_0)) - ((\alpha_c + \lambda_1) - (\alpha_c + \lambda_0)) \\ &= (\delta + \lambda_1 - \lambda_0) - (\lambda_1 - \lambda_0) \\ &= \delta \end{aligned}$$

CARD & KRUEGER (1994)

- What is the effect of an increase in minimum wage on employment?
- Prediction economic theory: rise in minimum wage leads perfectly competitive employers to cut employment.
- Card and Krueger investigate effect of increase in minimum wage from \$4.25 to \$5.05 in New Jersey (NJ) on April 1, 1992.
- Data on 410 fast-food restaurants also in neighboring state of Pennsylvania (PA):
  - individuals in NJ (treatment group)
  - individuals in Penn (control group)
  - in February/March 1992 (before)
  - in November/December 1992 (after)

CARD & KRUEGER (1994)

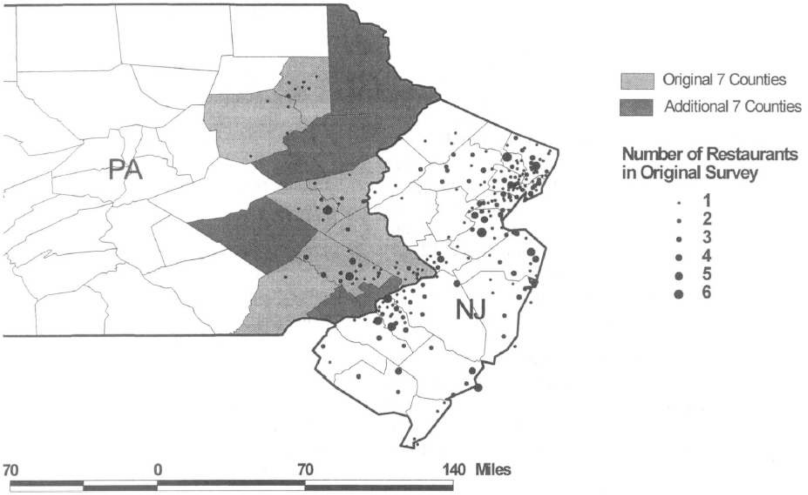


FIGURE 1. AREAS OF NEW JERSEY AND PENNSYLVANIA COVERED BY ORIGINAL SURVEY AND BLS DATA

CARD & KRUEGER (1994)

- Outcomes are employment in restaurant i in state g at time t

$$\begin{aligned} \text{NJ,F} : \mathrm{E}[Y_{id0}] = 20.44 \quad &\text{and} \quad \text{NJ,N} : \mathrm{E}[Y_{id1}] = 21.03 \\ \text{PA,F} : \mathrm{E}[Y_{ic0}] = 23.33 \quad &\text{and} \quad \text{PA,N} : \mathrm{E}[Y_{ic1}] = 21.17 \end{aligned}$$

- $\widehat{\beta}_{DiD} = (21.03 - 20.44) - (21.17 - 23.33) = 2.75.$
- Counter-intuitive result; Employment increased as consequence of increase in minimum wage (significant at 5%).
- When looking at trends by state: small change in NJ, but downward trend in Penn.
- Common trend assumption: In absence of treatment intervention in NJ, employment in NJ would have had same downward trend as Penn.

Table of Contents

DID REGRESSION FRAMEWORK

- We can translate DiD in a (panel) regression framework.
- In year $t = 0$ the outcomes are described by

$$Y_{igt} = \alpha_g + \lambda_t + \delta D_{gt} + \varepsilon_{igt}$$

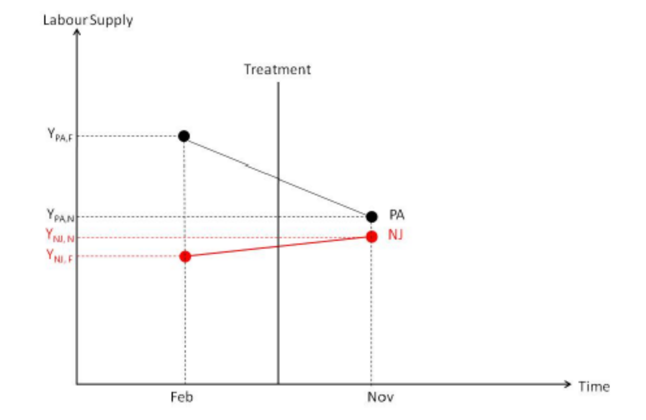
- α_g and λ_t are group and time dummies. $\varepsilon_{igt} = \eta_i + U_{igt}$.
- Can extend this framework to multiple groups and multiple time periods with $D_{gt} = 1$ if group g received treatment in period t and $D_{gt} = 0$ otherwise.
- Advantage of specifying difference-in-difference in regression equation:
  1. Convenient way to obtain standard errors.
  2. Easy to add additional time-varying regressors.
  3. Treatment variable can be continuous
- Common trend assumption: in the absence of the treatment intervention the treatment and control group would have followed a common trend.
- Stable unit treatment value assumption (SUTVA): treatment participation of one/some units does not affect the potential outcomes of other individuals (or themselves).

CARD & KRUEGER (1994) DID

- In Card & Krueger case the regression model can be written for restaurant i in state g at time t:

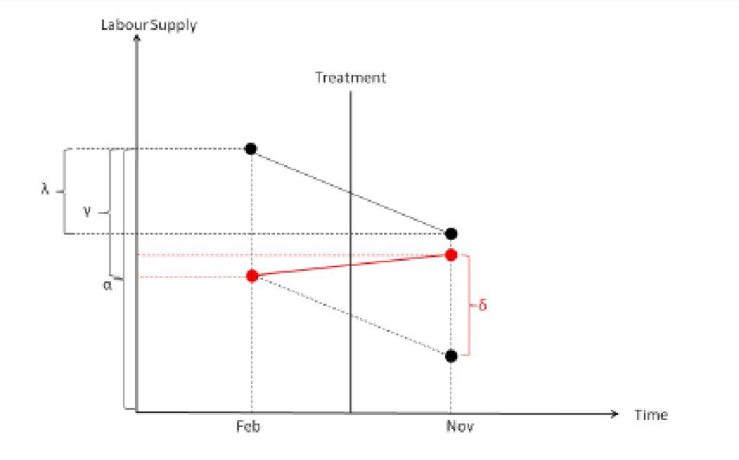
$$Y_{igt} = \alpha + \gamma NJ_s + \lambda \tau_t + \delta NJ_s \cdot \tau_t + \varepsilon_{igt}$$

- NJ is a dummy equal to 1 if the restaurant is in New Jersey. τ_t is a dummy equal to 1 if the observation is in November (after).



CARD & KRUEGER (1994) DID

$$Y_{igt} = \alpha + \gamma NJ_s + \lambda \tau_t + \delta NJ_s \cdot \tau_t + \varepsilon_{igt}$$



CARD & KRUEGER (2000) TREND

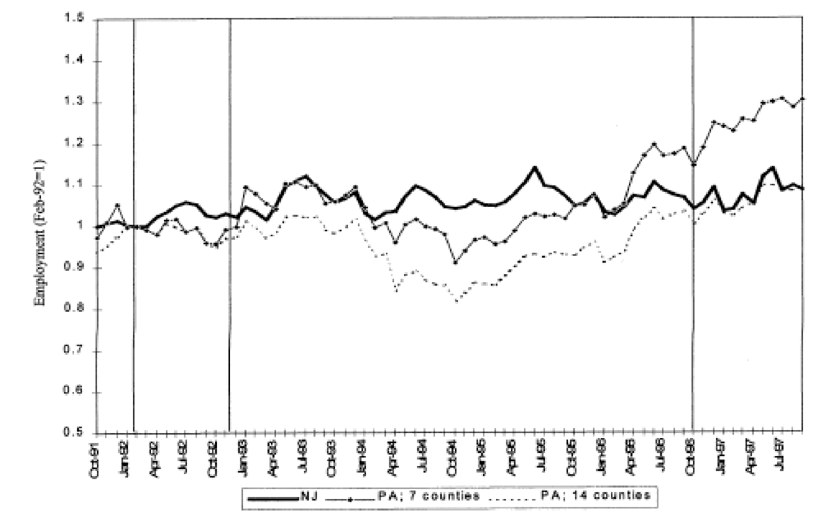


Table of Contents

DiD

	Kansas	NYC	Diff
before	2	1	*1.0*
after	3.2	1.3	*1.9*
Diff	1.2	0.3	**0.9**

- No longer assuming that the observations are counterfactuals (as in matching or regression)
- Works, if the **change over time** can be assumed to be equal absent treatment
- That's a higher level assumption, and sometimes it's even testable!

STEP 1:

- Explain and Defend the Experiment
  - Under which assumptions are treated and controlled assigned "like random?"
  - Does the quasi-experiment affect x in an interesting way?
  - Was the "shock" explicitly not meant to affect y?
  - "No proof is possible," but clarity is.
    - Allow your readers to understand, so they can judge.
  - Further support: Do steps 2-8

STEP 2:

- Present Raw Data in Terms of a Graph
  - Show outcome of interest:
    - before and after treatment
    - for treated and control separately
  - This communicates the basic variation and trends.

STEP 3:

- Show treatment and control are similar pre-treatment.
  - Present separate summary statistics.
    - should be similar on observable dimensions.
    - if not: unlikely they are good control observations.
  - $\rightarrow$ you can attempt additional matching

STEP 4:

- Present Baseline Estimates.
  - Usually as regression.
  - Start without controls and add them. (Altonji et al.(2005))
  - Cluster Standard Errors on the relevant level.

STEP 5:

- Investigate pre-treatment patterns:
  - Examine whether behaviors were similar pre-treatment.
  - Contrast of pre- and post-shock differences can be more powerful.
  - e.g. replace crossterm of interest with controls before and after treatment.

STEP 6:

- Run many Robustness checks - Does the effect survive:
  - Various sets of controls?
  - Different functional forms?
  - Different choices of the placebo time-period or starting points?
  - Other dependent Variables?
  - Different choices of the size of the control group?
- Not always will the effect survive!
- Ideally it goes away when it should and stays when it should not!

STEP 7:

- Discuss the External Validity/Interest of the Treated Group
  - Can we assume homogeneous treatment?
  - UNDER WHICH ASSUMPTIONS!
  - if not, is the group we're studying interesting?
    - or are they geeks?
  - What else to do, if not: cf. Goldfarb and Tucker (2014)
  - To reiterate: be clear about your assumptions!

STEP 8:

- Apologize!!
  - Many things remain unproven and caveats cannot be eliminated!
  - Relate the assumptions once again, say what can be tested and what not.
  - Clarify what happens, if your untestable assumptions DO NOT hold!

Table of Contents

DID ESTIMATOR: THREATS TO COMMON TREND

- DiD estimator is more efficient than social experiment estimator even if treatment is randomly assigned (estimator balances covariates X).
- DiD estimator is consistent even if treatment assignment is correlated to (time-invariant) individual specific effect but must use fixed effects approach.
- However, Intervention should be random conditional on time and group fixed effects
- Problems arise if treatment depends on outcomes in previous and current period. Such feedback violates strict exogeneity which causes DiD estimator to be inconsistent.
- One potential concern is "Ashenfelter Dip".
  - Ashenfelter (1978) was the first to note that enrollment in a training program is more likely if temporary dip in earnings occurs just before start of program.
  - So treated and untreated have different pre-treatment trends.
  - Consequence is that earnings growth after enrollment likely different for participants even without treatment.

POTENTIAL SOLUTION FOR THREATS TO COMMON TREND

Potential solutions if common trend assumption is unlikely to hold?

1. Include time-varying covariates and/or group specific time trends.
2. Difference-in-difference-in-difference.
3. Difference-in-difference + instrumental variable approach.
4. Test model with *placebo treatments*: randomly place a treatment intervention at some (earlier) moment in time or/and on other units. Estimated treatment effect should be insignificant.
5. Synthetic control group approach.

1. TIME TREND REGRESSOR

- Common trend assumption can be relaxed by including time varying covariates (X_{igt}) and group specific (linear) time trends ($\mu_g \cdot t$)

$$Y_{igt} = \alpha_g + \lambda_t + \mu_g t + X_{igt}\beta + \delta D_{gt} + \varepsilon_{igt}$$

- We need at lest three time periods to estimate model with group specific time trends (but more would be better).
- Besley & Burgess (2004) study effect of labor market regulation on manufacturing performance in Indian states between 1958 amd 1992.
- Find that negative effects of labor regulation are lost after controlling for unit specific trends.
- We will get back to this model in the lecture on correlated random coefficient models.

1. REGRESSION INCLUDING LEADS AND LAGS

- Including lead variables allows inspecting pre-treatment trends.
- Lags can be included to analyze variation in post-treatment time varying effect.
- Write out a regression with leads and lags

$$Y_{igt} = \alpha_g + \lambda_t + X_{igt}\beta + ??? + \varepsilon_{igt}$$

- How to organize data when treatment occurs at different times t?

1. REGRESSION INCLUDING LEADS AND LAGS

- Including lead variables allows inspecting pre-treatment trends.
- Lags can be included to analyze variation in post-treatment time varying effect.
- Regression with leads and lags

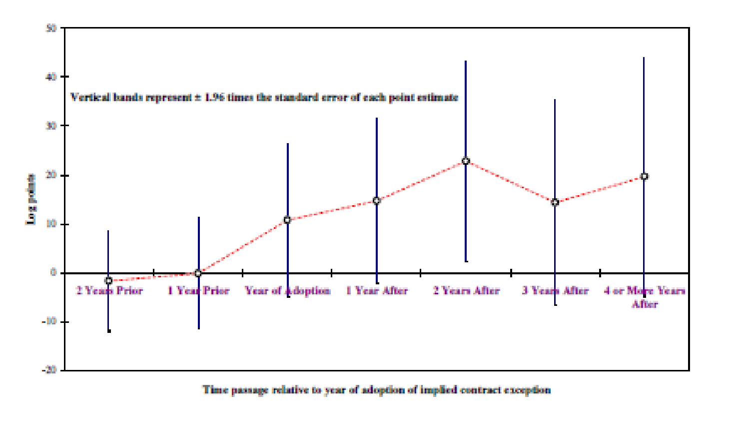
$$Y_{igt} = \alpha_g + \lambda_t + X_{igt}\beta + \sum_{\tau=1}^{q} \delta_{+\tau} D_{gt+\tau} + \sum_{\tau=0}^{m} \delta_{-\tau} D_{gt-\tau} + \varepsilon_{igt}$$

- Treatment occurs in year 0.
- Includes q leads (anticipation, ex-ante, or pre-treatment effects).
- Includes m lags (ex-post treatment effects).
- Remember: t is not necessarily calendar time, it can also be a cohort or elapsed time in a specific state (which lends itself more to duration DiD model).

1. REGRESSION INCLUDING LEADS AND LAGS: AUTOR (2003)

- Autor (2003) uses DiD model with leads and lags to analyze the effect of increased employment protection on a firm's use of temporary help workers.
- Less job security in US means easier to hire and fire workers.
- Some states courts have made some exceptions to this employment at will rule and have thus increased employment protection.
- Different states have passed laws to increase employment protection.
- These laws were implemented at different points in time in the different states.
- To fit this dynamic treatment into a static setting (DiD is static setting) usual approach is to normalize the treatment implementation year to 0.
- Autor then analyzes the effect of these employment protection laws on the use of temporary help workers.

1. REGRESSION INCLUDING LEADS AND LAGS: AUTOR (2003)



- No evidence of anticipation effects.
- The lags show time varying ex-post effects with increases in the first years then remains relatively constant.

2. DIFFERENCE-IN-DIFFERENCE-IN-DIFFERENCES

- When available, possible to use a third difference to adjust for common trend.
- Example: A state implements a change in health care policy for individuals aged 65 and older.
- DiD 1: data on health in treatment state before and after for people ≥ 65 and for people < 65 (control group)
  - If different trends between old and young people $\rightarrow$ possible violation of common trend.
- DiD 2: data on health before and after for people ≥ 65 in treatment state and in neighbouring state (control group)
  - If different trends in two states $\rightarrow$ possible violation of common trend.
- DiDiD:
  - DiD in treatment state with people < 65 as control group.
  - perform same DiD for control state.
  - DiDiD given by: $DiD_{treatment \quad state} - DiD_{control \quad state}$.

3. DID + INSTRUMENTAL VARIABLES

- If treatment variable is endogenous to group specific trends.
- Find variable that affects treatment but (arguably) only affect the outcome via it's effect through treatment.
- Given endogenous treatment D_{gt} in the model

$$Y_{igt} = \alpha_g + \lambda_t + \delta D_{gt} + \varepsilon_{igt}$$

- We need to find an instrument Z_{gt} which satisfies $Corr(Z_{gt}, X_{gt}) \neq 0$ and $Corr(Z_{gt}, \varepsilon_{igt}) = 0$.
- Beware that the interpretation of your treatment effect is no longer the same. You are now estimating a Local Average Treatment Effect (LATE).
- It is the effect on compliers. It is good to be able to say something about who these compliers might be. Who is affected by the effect found in your model?

Table of Contents

CLUSTERING

- When computing standard errors be aware of possible grouped data (correct for clustering within groups).
- A cluster c is all observations within a group g at a given time t.
- Often many individuals within each cluster but few clusters (individuals in regions, students in schools).
- Consider DiD model:

$$Y_{igt} = \alpha_g + \delta D_{gt} + \lambda_t + U_{igt}$$

- Ignoring for ease of explanation observed individual covariates X_{igt}
- Clustering implies:

$$U_{igt} = V_{gt} + \varepsilon_{igt}$$

- with V_{gt} random effects and ε_{igt} error terms.
- In order to account for aggregate cluster specific variation V_{gt} we must adjust using cluster-robust standard errors. Rewrite DiD in cluster matrix notation:

$$y_c = x_c \gamma + u_c$$

- where x_c includes α_g, D_{gt} and λ_t

DONALD AND LANG (2007)

- Cluster-robust standard errors of OLS estimator

$$var(\hat{\gamma}) = \frac{N-1}{N-K}\frac{C}{C-1}\left(\sum_c x_c^{'} x_c\right)^{-1}\sum_c x_c^{'}\hat{u}_c\hat{u}_c^{'}x_c\left(\sum_c x_c^{'}x_c\right)^{-1}$$

- with N total number of observations, K number of parameters in γ and C number of clusters.
- The usual cluster-robust standard errors also account for heteroskedasticity.
- These standard errors depend on $C \rightarrow \infty$ not $N \rightarrow \infty$.
- Donald and Lang (REStat,2007) argue that small sample properties of usual cluster-robust standard errors may be poor.
- Propose two-step GLS estimator (see MHE chapter 8.2.1)
- This estimator has better small sample properties for standard errors
- Donald and Lang show standard errors are seriously underestimated in several classical DiD papers.

SERIAL CORRELATION

- Bertrand, Duflo, Mullainathan (QJE,2004) focus on serial correlation in U_{it}.

$$Y_{it} = \alpha_g + \beta D_{gt} + \lambda_t + U_{igt}$$

- Ignoring serial correlation in U_{it} seriously underestimates standard errors and thus over-rejects null hypothesis of no effect.
- This issue is only relevant if $T > 2$ (which is often the case).
- U_{it} arising from persistence in Y_{it} or D_{it}.
- Bertrand, Duflo, Mullainathan perform Monte Carlo placebo simulation experiments to investigate solutions.
- They find in their simulation that not accounting for serial correlation gives significant results in 45% of placebo interventions.

SERIAL CORRELATION: INTUITION

- Recall from linear panel lecture that Newey-West variance adjusts for cross-time correlation between $\widehat{\widetilde{U}}_{it}$ and $\widehat{\widetilde{U}}_{is}$ where in FE-within case $\widetilde{U}_{it} = U_{it} - \overline{U}_{it}$.
- When policy change is at the cluster level, the cross-time correlation must also be adjusted at the cluster level.
- Estimation of standard errors therefore requires large amount of clusters (groups) not individual units within clusters.

SOLUTIONS

Depending upon your study, it may be that $C = N$.

1. Cluster standard errors at group level (instead of group x time) (STATA "robust" or cluster option).
   - Performs well when C is reasonably large $C \geq 50$.
2. Use parametric specification for serial correlation (eg. AR(1) process $U_{it} = \rho U_{it-1} + \varepsilon_{it}$).
   - Even if correct specification is found, estimator does not perform well with small T.
3. Use empirical variance-covariance matrix (assuming homoskedasticity and common autocorrelation across states).
   - Does well when C is not small.
4. Block bootstrap (keep all observations of one group together).
   - Does well when C is reasonably large $C \geq 50$.
5. Average within group data before and after intervention (ie. ignore time series information).
   - Does well when C is small $C \sim 6$.
6. Wild Bootstrap (Cameron, Gelbach, Miller, 2008, 2011).
   - Does well when C is between 10 and 50.

Table of Contents

SYNTHETIC CONTROL GROUP

- In some cases, it may be better to choose a combination of more than one group rather than one control group.
- Similarly, if given many control groups it may be better to only use a subset.
- Synthetic control group method: weighted average of available controls to create a control group which emulates best the trends in the treatment group.
    - Synthetic control group built on pretreatment observations.
    - Suppose we observe T time periods and G groups.
    - the treatment group is designated by $g = d$ and is treated in the final period ($t = T$).
    - The available control groups are all groups g in G besides the group $g = d$ with outcome $\overline{Y}_{dt}$.
    - Estimated outcome for treatment group at $t = T$ in case of no treatment is $\sum_{G:g \neq d} \pi_g \overline{Y}_{gt}$ with weights π_g chosen to minimize

$$\left| \begin{array}{c} \overline{Y}_{d1} - \sum_{G:g \neq d} \pi_g \overline{Y}_{g1} \\ \vdots \\ \overline{Y}_{dT-1} - \sum_{G:g \neq d} \pi_g \overline{Y}_{gT-1} \end{array} \right|$$

    - Weights above minimize difference between treatment and control group based on outcome $\overline{Y}_{gt}$ prior to the intervention. Can also include covariates.
- Abadie, Diamond and Hainmueller (2010) apply synthetic control approach to estimate effect of a large scale tobacco control program implemented in California in 1988.

ABADIE, DIAMOND AND HAINMUELLER (2010)

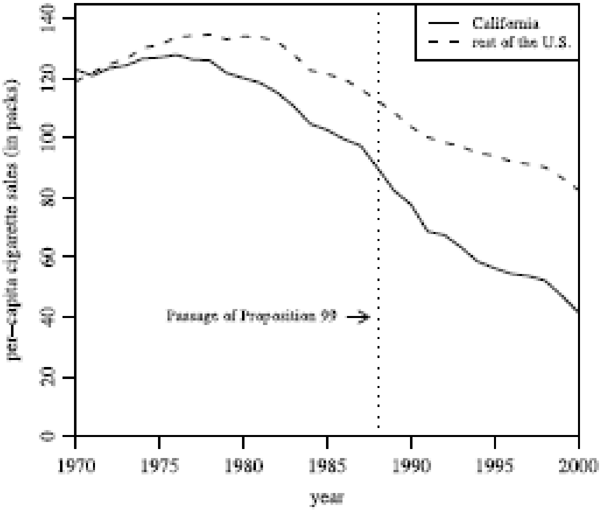


Figure 1. Trends in per-capita cigarette sales: California vs. the rest of the United States

ABADIE, DIAMOND AND HAINMUELLER (2010)

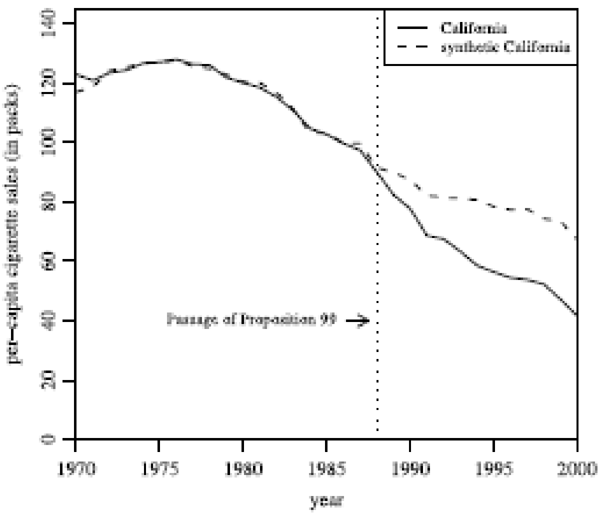


Figure 2. Trends in per-capita cigarette sales: California vs. synthetic California.

ABADIE, DIAMOND AND HAINMUELLER (2010)

- Placebo simulation approach to assess significance.
- For each iteration, assign treatment randomly to one of the control groups and re-estimate effect of 'placebo effect'.

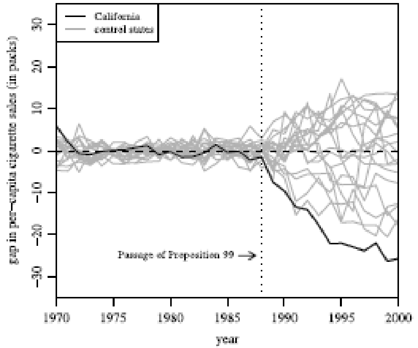


Figure 7. Per-capita cigarette sales gaps in California and placebo gaps in 19 control states (discards states with pre-Proposition 99 MSPE two times higher than California's).

- Probability of estimating effect as large as California-effect under random permutation of treatment is 5%.

CHANGES IN CHANGES

- Athey and Imbens (Econometrica 2006) propose generalization of standard DiD model.
- The model is called Changes-in-Changes (CiC)
- They weaken common trend assumption to allow effects of time to differ systematically across individuals.
- Effect of treatment is also allowed to differ systematically across individuals.
- Robust to rescaling of the outcome variable (for example levels vs. logarithms).
- With CiC it is possible to estimate entire counterfactual outcome distribution.
- And it is possible to estimate treatment effect on the non-treated.
- General idea:
  - Re-weight the control group such that at baseline their outcome distribution is identical to that of the treatment group.
  - Assume that individuals stay in the same quantile within group (regardless of the intervention)
  - Apply same re-weighting transformation for the controls after the intervention and compare to observed outcomes in treatment group.

CHANGES IN CHANGES

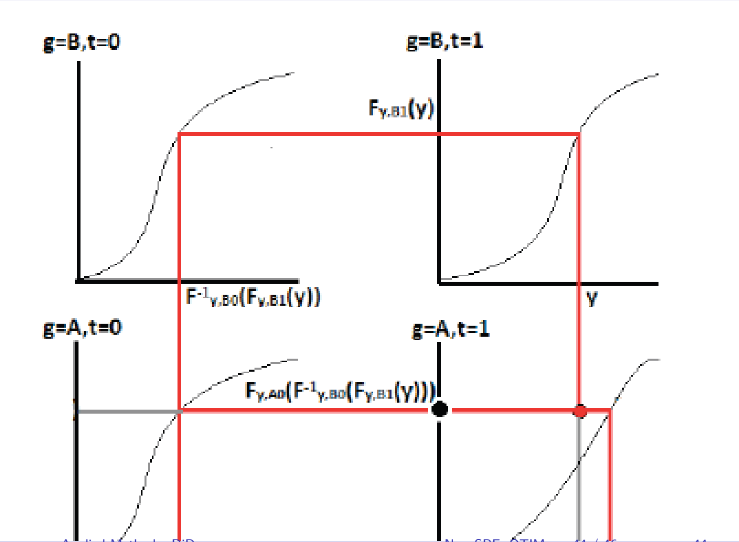
Assumptions in Changes in Changes Model (3. and 4. are standard DiD model assumptions)

- Y^I_{igt} potential outcome for individual i in group g in period t if treated.
- Y^N_{igt} potential outcome for individual i in group g in period t if not treated.

1. Outcome if not treated: $Y^N_{igt} = h(U_i, t_i)$ with $h(U_i, t_i)$ monotone increasing in U_i (unobservable characteristics).
    - Individual i outcome is the same in given time period irrespective of group membership g.
    - Outcome as function of unobservables can change over time $h(U_i, t = 0) \neq h(U_i, t = 1)$.
    - Trend in non-treated outcome should be the same when $U_i = U_j$ but trend can differ when $U_i \neq U_j$.
2. Distribution of unobservables can differ between groups but not within a group over time $U_i \perp U_j$.
3. $U_i = \alpha_g + \varepsilon_{igt}$ (additivity) with $\varepsilon_{igt} \perp (g_i, t_i)$.
4. $h(U_i, t_i) = U_i + \lambda_t = \alpha_g + \lambda_t + \varepsilon_{igt}$

CHANGES IN CHANGES

- Observed outcome distributions for treatment (A) and control (B) groups before (0) and after (1).



CHANGES IN CHANGES

- Distribution of counterfactual non-treatment outcomes ($F^N_{y,gt}(y)$) for the treatment group $g = d$ after treatment $t = 1$:

$$F^N_{y,d1}(y) = F_{y,d0}(F^{-1}_{y,c0}(F_{y,c1}(y)))$$

- Average treatment effect on the treated ATT:

$$\Delta CIC^{treated} = \mathrm{E}[Y^I_{d1}] - \mathrm{E}[Y^N_{d1}] = \mathrm{E}[Y^I_{d1}] - \mathrm{E}[F^{-1}_{y,c1}(F_{y,c0}(y_{d0}))]$$

- In rather similar way we can estimate average treatment effect on the non-treated.
- We do need the additional assumption

$$Y^I_{igt} = h^I(U_i, t_i)$$

- This says that effect of treatment at given time is identical for individuals with $U_i = U_j$ irrespective of group membership.

CONCLUDING REMARKS

- DiD estimators can potentially solve causal questions.
- DiD can be implemented on repeated cross section data.
- Fundamental identifying assumption is the common trend assumption.
  - Is common trend assumption plausible?
  - Can you give pre-treatment evidence of common trend?
  - Be aware of compositional effects
- You should also think about SUTVA.
- Choice of comparison group is crucial!
- Make sure to calculate correct standard errors. Or, in unclear situations, perform several robustness checks