Balancing, Regression, Difference-in-Differences, and Synthetic Control Methods: A Synthesis

Doudchenko and Imbens (2016)

Apoorva Lal - Panel Reading Group ; July 2021

Introduction

- Panel methods can be characterised into 3 broad groups (as of 2016):
  - Difference-in-differences : $\Delta Y^{\text{post}} - \Delta Y^{\text{pre}}$
  - Matching: on both pre-treatment outcomes and other covariates
  - Synthetic Control: For each treated unit, a 'synthetic control' is constructed as a weighted average of control units s.t. the weighted average matches pre-treatment outcomes and covariates
- This paper: framework to nest existing approaches + estimator that relaxes some assumptions.
  - Main contribution: framework to clarify assumptions
  - Resting WP; Cannibalised by later papers (esp. Arkhangelsky et al 2020)?

Notation

- $N + 1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards

Notation

- $N+1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards
- Treatment : $W_{i,t} = \mathbb{1}_{i=0 \,\wedge\, t \in T_0+1,\dots,T}$

Notation

- $N+1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards
- Treatment : $W_{i,t} = \mathbb{1}_{i=0 \,\wedge\, t \in T_0+1,\ldots,T}$
- Potential outcomes for unit 0 define the treatment effect: $\tau_{0,t} := Y_{0,t}(1) - Y_{0,t}(0)$ for $t = T_0 + 1, \ldots, T$

Notation

- $N + 1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards
- Treatment : $W_{i,t} = \mathbb{1}_{i=0 \,\wedge\, t \in T_0+1,\ldots,T}$
- Potential outcomes for unit 0 define the treatment effect:
  $\tau_{0,t} := Y_{0,t}(1) - Y_{0,t}(0)$ for $t = T_0 + 1, \ldots, T$
- Observed outcome: $Y_{i,t}^{obs} = Y_{i,t}(W_{i,t})$

Notation

- $N + 1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards
- Treatment : $W_{i,t} = \mathbb{1}_{i=0 \,\wedge\, t \in T_0+1,\dots,T}$
- Potential outcomes for unit 0 define the treatment effect: $\tau_{0,t} := Y_{0,t}(1) - Y_{0,t}(0)$ for $t = T_0 + 1, \dots, T$
- Observed outcome: $Y_{i,t}^{obs} = Y_{i,t}(W_{i,t})$
- Time-invariant characteristics $X_i := (X_{i,1}, \dots, X_{i,M})^\top$ for each unit, which may include lagged outcomes $Y_{i,t}^{obs}$ for $t \le T_0$

Notation

- $N + 1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards
- Treatment : $W_{i,t} = \mathbb{1}_{i=0 \, \wedge \, t \in T_0+1,\ldots,T}$
- Potential outcomes for unit 0 define the treatment effect: $\tau_{0,t} := Y_{0,t}(1) - Y_{0,t}(0)$ for $t = T_0 + 1, \ldots, T$
- Observed outcome: $Y_{i,t}^{obs} = Y_{i,t}(W_{i,t})$
- Time-invariant characteristics $X_i := (X_{i,1}, \ldots, X_{i,M})^\top$ for each unit, which may include lagged outcomes $Y_{i,t}^{obs}$ for $t \leq T_0$
    - $\mathbf{X}_c$ is $N \times M$ matrix that stacks Xs for control units

Notation

- $N + 1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards
- Treatment : $W_{i,t} = \mathbb{1}_{i=0 \, \wedge \, t \in T_0+1,\ldots,T}$
- Potential outcomes for unit 0 define the treatment effect: $\tau_{0,t} := Y_{0,t}(1) - Y_{0,t}(0)$ for $t = T_0 + 1, \ldots, T$
- Observed outcome: $Y_{i,t}^{obs} = Y_{i,t}(W_{i,t})$
- Time-invariant characteristics $X_i := (X_{i,1}, \ldots, X_{i,M})^\top$ for each unit, which may include lagged outcomes $Y_{i,t}^{obs}$ for $t \leq T_0$
  - $\mathbf{X}_c$ is $N \times M$ matrix that stacks Xs for control units
  - $\mathbf{X}_t$ is $M-$ row vector of covariates for control

Notation

- $N+1$ units observed for T periods, with a subset of treated units (for simplicity - unit 0) treated from T_0 onwards
- Treatment : $W_{i,t} = \mathbb{1}_{i=0 \,\wedge\, t \in T_0+1,\ldots,T}$
- Potential outcomes for unit 0 define the treatment effect: $\tau_{0,t} := Y_{0,t}(1) - Y_{0,t}(0)$ for $t = T_0+1,\ldots,T$
- Observed outcome: $Y_{i,t}^{obs} = Y_{i,t}(W_{i,t})$
- Time-invariant characteristics $X_i := (X_{i,1},\ldots,X_{i,M})^\top$ for each unit, which may include lagged outcomes $Y_{i,t}^{obs}$ for $t \leq T_0$
  - $\mathbf{X}_c$ is $N \times M$ matrix that stacks Xs for control units
  - $\mathbf{X}_t$ is $M-$ row vector of covariates for control
  - stack them to get $\mathbf{X}$

Outcome Matrices

$$\mathbf{Y}^{\text{obs}} = \begin{bmatrix} \mathbf{Y}^{\text{obs}}_{\text{t, post}} & \mathbf{Y}^{\text{obs}}_{\text{c, post}} \\ \mathbf{Y}^{\text{obs}}_{\text{t, pre}} & \mathbf{Y}^{\text{obs}}_{\text{c, pre}} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{\text{t, post}}(1) & \mathbf{Y}_{\text{c, post}}(0) \\ \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{c, pre}}(0) \end{bmatrix} \qquad T \times (N+1)$$

$$\mathbf{Y}(0) = \begin{bmatrix} ? & \mathbf{Y}_{\text{c, post}}(0) \\ \color{red}{\mathbf{Y}_{\text{t, pre}}(0)} & \color{red}{\mathbf{Y}_{\text{c, pre}}(0)} \end{bmatrix} = \begin{bmatrix} ? & \color{blue}{\mathbf{Y}_{\text{c, post}}(0)} \\ \mathbf{Y}_{\text{t, pre}}(0) & \color{blue}{\mathbf{Y}_{\text{c, pre}}(0)} \end{bmatrix}$$

- relative magnitudes of T and N might dictate whether we impute the missing potential outcome $?$ using this or this comparison
  - **Many Units and Multiple Periods**: $N >> T_0$, $\mathbf{Y}(0)$ is 'fat', and red comparison becomes challenging relative to blue. So matching methods are attractive.

Outcome Matrices

$$\mathbf{Y}^{\text{obs}} = \begin{bmatrix} \mathbf{Y}^{\text{obs}}_{\text{t, post}} & \mathbf{Y}^{\text{obs}}_{\text{c, post}} \\ \mathbf{Y}^{\text{obs}}_{\text{t, pre}} & \mathbf{Y}^{\text{obs}}_{\text{c, pre}} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{\text{t, post}}(1) & \mathbf{Y}_{\text{c, post}}(0) \\ \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{c, pre}}(0) \end{bmatrix} \qquad T \times (N+1)$$

$$\mathbf{Y}(0) = \begin{bmatrix} ? & \mathbf{Y}_{\text{c, post}}(0) \\ \color{red}{\mathbf{Y}_{\text{t, pre}}(0)} & \color{red}{\mathbf{Y}_{\text{c, pre}}(0)} \end{bmatrix} = \begin{bmatrix} ? & \color{blue}{\mathbf{Y}_{\text{c, post}}(0)} \\ \mathbf{Y}_{\text{t, pre}}(0) & \color{blue}{\mathbf{Y}_{\text{c, pre}}(0)} \end{bmatrix}$$

- relative magnitudes of T and N might dictate whether we impute the missing potential outcome $?$ using this or this comparison
  - **Many Units and Multiple Periods**: $N >> T_0$, $\mathbf{Y}(0)$ is 'fat', and red comparison becomes challenging relative to blue. So matching methods are attractive.
  - $T_0 >> N$, $\mathbf{Y}(0)$ is 'tall', and matching becomes infeasible. So it might be easier to estimate blue dependence structure.

Outcome Matrices

$$\mathbf{Y}^{\text{obs}} = \begin{bmatrix} \mathbf{Y}^{\text{obs}}_{\text{t, post}} & \mathbf{Y}^{\text{obs}}_{\text{c, post}} \\ \mathbf{Y}^{\text{obs}}_{\text{t, pre}} & \mathbf{Y}^{\text{obs}}_{\text{c, pre}} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{\text{t, post}}(1) & \mathbf{Y}_{\text{c, post}}(0) \\ \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{c, pre}}(0) \end{bmatrix} \qquad T \times (N+1)$$

$$\mathbf{Y}(0) = \begin{bmatrix} ? & \mathbf{Y}_{\text{c, post}}(0) \\ \color{red}{\mathbf{Y}_{\text{t, pre}}(0)} & \color{red}{\mathbf{Y}_{\text{c, pre}}(0)} \end{bmatrix} = \begin{bmatrix} ? & \color{blue}{\mathbf{Y}_{\text{c, post}}(0)} \\ \mathbf{Y}_{\text{t, pre}}(0) & \color{blue}{\mathbf{Y}_{\text{c, pre}}(0)} \end{bmatrix}$$

- relative magnitudes of T and N might dictate whether we impute the missing potential outcome $?$ using this or this comparison
  - **Many Units and Multiple Periods**: $N >> T_0$, $\mathbf{Y}(0)$ is ‘fat’, and red comparison becomes challenging relative to blue. So matching methods are attractive.
  - $T_0 >> N$, $\mathbf{Y}(0)$ is ‘tall’, and matching becomes infeasible. So it might be easier to estimate blue dependence structure.
  - Finally, if $T_0 \approx N$, regularization strategy for limiting the number of control units that enter into the estimation of $Y_{0,T_0+1}(0)$ may be important

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}, \mathbf{Y}_{\text{t, pre}}^{\text{obs}}, \mathbf{Y}_{\text{c, pre}}^{\text{obs}}$

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y^{\text{obs}}_{0,T} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y^{\text{obs}}_{i,T}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}^{\text{obs}}_{\text{c, post}}, \mathbf{Y}^{\text{obs}}_{\text{t, pre}}, \mathbf{Y}^{\text{obs}}_{\text{c, pre}}$
- Impose four constraints

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}, \mathbf{Y}_{\text{t, pre}}^{\text{obs}}, \mathbf{Y}_{\text{c, pre}}^{\text{obs}}$
- Impose four constraints
  1. **No Intercept**: $\mu = 0$. Stronger than Parallel trends in DiD.

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y^{\text{obs}}_{0,T} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y^{\text{obs}}_{i,T}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}^{\text{obs}}_{\text{c, post}}, \mathbf{Y}^{\text{obs}}_{\text{t, pre}}, \mathbf{Y}^{\text{obs}}_{\text{c, pre}}$
- Impose four constraints
  1. **No Intercept:** $\mu = 0$. Stronger than Parallel trends in DiD.
  2. **Adding up** : $\sum_{i=1}^{n} \omega_i = 1$. Common to DiD, SC.

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}, \mathbf{Y}_{\text{t, pre}}^{\text{obs}}, \mathbf{Y}_{\text{c, pre}}^{\text{obs}}$
- Impose four constraints
  1. **No Intercept**: $\mu = 0$. Stronger than Parallel trends in DiD.
  2. **Adding up** : $\sum_{i=1}^{n} \omega_i = 1$. Common to DiD, SC.
  3. **Non-negativity**: $\omega_i \geq 0 \; \forall \; i$. Ensures uniqueness via 'coarse' regularisation + precision control. Negative weights may improve out-of-sample prediction.

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}, \mathbf{Y}_{\text{t, pre}}^{\text{obs}}, \mathbf{Y}_{\text{c, pre}}^{\text{obs}}$
- Impose four constraints
  1. **No Intercept**: $\mu = 0$. Stronger than Parallel trends in DiD.
  2. **Adding up** : $\sum_{i=1}^{n} \omega_i = 1$. Common to DiD, SC.
  3. **Non-negativity**: $\omega_i \geq 0 \; \forall \; i$. Ensures uniqueness via 'coarse' regularisation + precision control. Negative weights may improve out-of-sample prediction.
  4. **Constant Weights**: $\omega_i = \overline{\omega} \; \forall \; i$

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}, \mathbf{Y}_{\text{t, pre}}^{\text{obs}}, \mathbf{Y}_{\text{c, pre}}^{\text{obs}}$
- Impose four constraints
  1. **No Intercept**: $\mu = 0$. Stronger than Parallel trends in DiD.
  2. **Adding up** : $\sum_{i=1}^{n} \omega_i = 1$. Common to DiD, SC.
  3. **Non-negativity**: $\omega_i \geq 0 \; \forall \; i$. Ensures uniqueness via 'coarse' regularisation + precision control. Negative weights may improve out-of-sample prediction.
  4. **Constant Weights**: $\omega_i = \overline{\omega} \; \forall \; i$
- DiD imposes 2-4.

Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}, \mathbf{Y}_{\text{t, pre}}^{\text{obs}}, \mathbf{Y}_{\text{c, pre}}^{\text{obs}}$
- Impose four constraints
  1. **No Intercept**: $\mu = 0$. Stronger than Parallel trends in DiD.
  2. **Adding up** : $\sum_{i=1}^{n} \omega_i = 1$. Common to DiD, SC.
  3. **Non-negativity**: $\omega_i \geq 0 \; \forall \; i$. Ensures uniqueness via 'coarse' regularisation + precision control. Negative weights may improve out-of-sample prediction.
  4. **Constant Weights**: $\omega_i = \overline{\omega} \; \forall \; i$
- DiD imposes 2-4.
- ADH(2010, 2014) impose 1-3

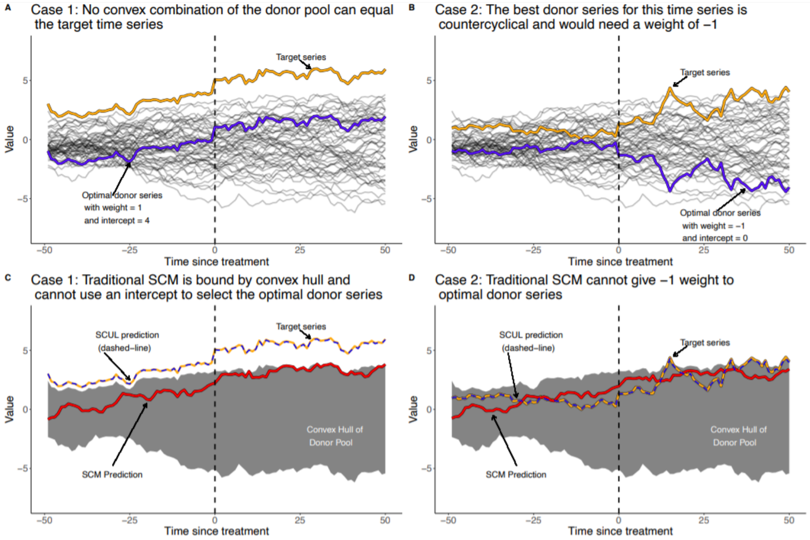
Common Structure: 4 assumptions

- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}, \mathbf{Y}_{\text{t, pre}}^{\text{obs}}, \mathbf{Y}_{\text{c, pre}}^{\text{obs}}$
- Impose four constraints
  1. **No Intercept**: $\mu = 0$. Stronger than Parallel trends in DiD.
  2. **Adding up** : $\sum_{i=1}^{n} \omega_i = 1$. Common to DiD, SC.
  3. **Non-negativity**: $\omega_i \geq 0 \; \forall \; i$. Ensures uniqueness via 'coarse' regularisation + precision control. Negative weights may improve out-of-sample prediction.
  4. **Constant Weights**: $\omega_i = \overline{\omega} \; \forall \; i$
- DiD imposes 2-4.
- ADH(2010, 2014) impose 1-3
  - 1 + 2 imply 'No Extrapolation'.

Relaxing the assumptions

- Negative weights
  - If treated units are outliers on important covariates, negative weights might improve fit
  - Bias reduction - negative weights increase bias-reduction rate
- When $N >> T_0$, (1-3) alone might not result in a unique solution. Choose by
  - Matching on pre-treatment outcomes : one good control unit is better than synthetic one comprised of disparate units
  - Constant weights - implicit in DiD
- Given many pairs of (μ, ω)
- prefer values s.t. synthetic control unit is similar to treated units in terms of lagged outcomes
- low dispersion of weights
- few control units with non-zero weights

Case for nonconvex or negative Weights : Hollingworth and Wing (2021)



The optimisation problem: general case

Ingredients of objective function

- **Balance**: difference between pre-treatment outcomes for treated and linear-combination of pre-treatment outcomes for control
  - $\left\|\mathbf{Y}_{\text{t, pre}} - \mu - \omega^\top \mathbf{Y}_{\text{c, pre}}\right\|_2^2 = (\mathbf{Y}_{\text{t, pre}} - \mu - \omega^\top \mathbf{Y}_{\text{c, pre}})^\top (\mathbf{Y}_{\text{t, pre}} - \mu - \omega^\top \mathbf{Y}_{\text{c, pre}})$
- **Sparse and small weights**:
  - sparsity : $\|\omega\|_1$
  - magnitude: $\|\omega\|_2$

$$(\widehat{\mu}^{en}(\lambda,\alpha), \widehat{\omega}^{en}(\lambda,\alpha)) = \underset{\mu,\omega}{\operatorname{argmin}} \; Q(\mu,\omega | \mathbf{Y}_{\text{t, pre}}, \mathbf{Y}_{\text{c, pre}}; \lambda, \alpha)$$

where $$Q(\mu,\omega | \mathbf{Y}_{\text{t, pre}}, \mathbf{Y}_{\text{c, pre}}; \lambda, \alpha) = \left\|\mathbf{Y}_{\text{t, pre}} - \mu - \omega^\top \mathbf{Y}_{\text{c, pre}}\right\|_2^2 + \lambda \left(\frac{1-\alpha}{2} \|\omega\|_2^2 + \alpha \|\omega\|_1 \right)$$

Choosing α, λ: Tailored regularisation

- don't want to scale covariates $\mathbf{Y}_{\text{c, pre}}$ to preserve interpretability of weights
- Instead, treat each control unit as a 'pseudo-treated' unit and compute $\widehat{Y}_{j,T}(0) = \widehat{\mu}^{\text{en}}(j;\alpha,\lambda) + \sum_{i \neq j} \widehat{\omega}_i(j;\alpha,\lambda) \cdot Y_{i,T}^{\text{obs}}$ where

$$(\widehat{\mu}^{en}(j;\lambda,\alpha), \widehat{\omega}^{en}(j;\lambda,\alpha)) = \underset{\mu,\omega}{\operatorname{argmin}} \sum_{t=1}^{T_0} \left(Y_{j,t} - \mu - \sum_{i \neq 0,j} \omega_i Y_{i,t} \right)^2 + \lambda \left(\frac{1-\alpha}{2} \left\|\omega\right\|_2^2 + \alpha \left\|\omega\right\|_1 \right)$$

pick the value of the tuning parameters $(\alpha_{opt}^{en}, \lambda_{opt}^{en})$ that minimises

$$CV^{en}(\alpha,\lambda) = \frac{1}{N} \sum_{j=1}^{N} (Y_{j,T} - \overbrace{\widehat{\mu}^{en}(j;\alpha,\lambda) - \sum_{i \neq 0,j} \widehat{\omega}_i^{en}(j;\alpha,\lambda) \cdot Y_{i,T}}^{\widehat{Y}_{j,T}(0)})$$

Re-expressing Standard Methods

Difference in Differences

- assume (2-4)
- No unique μ, ω solution for $T = 2$, so fix $\omega = \frac{1}{N}$

$$\omega_i^{\text{did}} = \frac{1}{N} \quad \forall i \in \{1, \ldots N\}$$
$$\widehat{\mu}^{\text{did}} = \frac{1}{T_0} \sum_{s=1}^{T_0} Y_{0,s} - \frac{1}{NT_0} \sum_{s=1}^{T_0} \sum_{i=1}^{N} Y_{i,s}$$

Best Subset; One-to-one Matching
$(\widehat{\mu}^S, \widehat{\omega}^S) = \operatorname{argmin}_{\mu,\omega} Q(\cdot; \lambda = 0, \alpha)$ with
$\sum_{i=1}^N \mathbb{1}_{\omega_i \neq 0} \leq k$ (=1 for OtO)

Synthetic Control

- assume (1-3) (i.e. $\mu = 0$)
- For $M \times M$ PSD diagonal matrix $\mathbf{V}$

$$(\widehat{\omega}(\mathbf{V}), \widehat{\mu}(\mathbf{V})) = \operatorname*{argmin}_{\omega,\mu} \{(\mathbf{X}_t - \mu - \omega^\top \mathbf{X})^\top \mathbf{V} (\mathbf{X}_t - \mu - \omega^\top \mathbf{X})\}$$
$$\widehat{\mathbf{V}} = \operatorname*{argmin}_{\mathbf{V}=\operatorname{diag}(v_1,\ldots,v_M)} \{(\mathbf{Y}_{\text{t, pre}} - \widehat{\omega}(\mathbf{V})^\top \mathbf{Y}_{\text{c, pre}})^\top (\mathbf{Y}_{\text{t, pre}} - \widehat{\omega}(\mathbf{V})^\top \mathbf{Y}_{\text{c, pre}})\}$$

Constrained regression: When
$X_i = Y_{i,t};\ 1 \leq t \leq T_0$ (Lagged Outcomes only)
$\mathbf{V} = \mathbf{I}_N$ and $\lambda = 0$

Inference

- Need to be explicit about what is random in repeated-sampling
- Do not want to argue that controls have positive probability of treatment
- Since $\tau = Y_{0,T}^{obs} - Y_{0,T}(0)$, estimation error arises from imputation error
- $(\widehat{\tau} - \tau)^2 = (Y_{0,T}(0) - \widehat{Y}_{0,T}(0))^2$

define matrices $\mathbf{Y}_{i,s}^{j,t}(0)$, for $i \leq j \quad s \leq t$

$$\mathbf{Y}_{i,s}^{j,t} := \begin{bmatrix} Y_{i,t}(0) & \cdots & Y_{j,t}(0) \\ \vdots & \ddots & \vdots \\ Y_{i,s}(0) & \cdots & Y_{j,s}(0) \end{bmatrix}$$

$\mathbf{Y}_{(i),s}^{(i),t}$ is the same with unit i's column left out.

Estimators for the missing $Y_{0,T}(0)$

$$\widehat{Y}_{0,T}(0) = g\left(\mathbf{Y}_{0,1}^{0,T-1}, \mathbf{Y}_{(0),T}^{(0),T}, \mathbf{Y}_{(0),1}^{(0),T-1}\right)$$

which produces variance estimators based on assignment assumptions.

Random Assignment of Unit

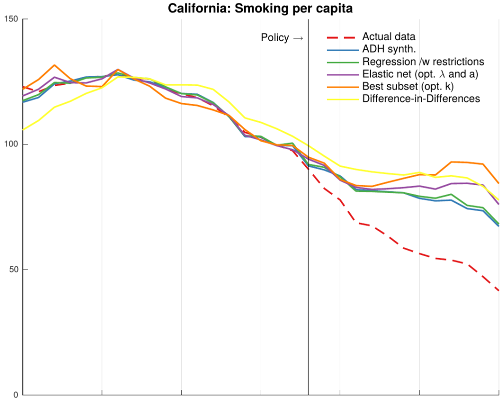
$$\widehat{\mathbb{V}}_c = \frac{1}{N}\sum_{i}^{N}(Y_{i,T}(0) - g\left(\mathbf{Y}_{i,1}^{i,T-1}, \mathbf{Y}_{(0,i),T}^{(0,i),T}, \mathbf{Y}_{(0,i),1}^{(0,i),T-1}\right)$$

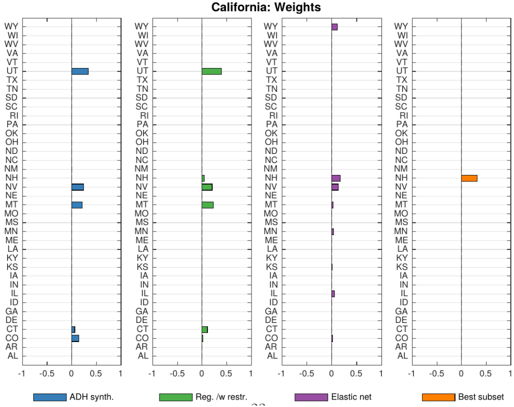
Random Timing of Treatment

$$\widehat{\mathbb{V}}_t = \frac{1}{s}\sum_{t=T_0-s+1}^{T_0}(Y_{i,T}(0) - g\left(\mathbf{Y}_{i,1}^{0,t-1}, \mathbf{Y}_{(0),t}^{(0),t}, \mathbf{Y}_{(0),1}^{(0),t-1}\right)$$

Combination : double-sum

Revisiting ADH California smoking example





Model	$\sum_i \omega_i$	μ	$\widehat{\tau}$	s.e.
Original Synth	1	0	-22.1	16.1
Constrained	1	0	-22.9	12.8
Elastic Net	.55	18.5	-26.9	16.8
Best Subset	.32	37.6	-31.9	20.3
Diff-in-Diff	1	-14.4	-32.4	18.9