Causal Inference Part 2

Dr. Syed Badruddoza

Texas Tech

May 2, 2025

Synthetic Control Method: Motivation

- Goal: Estimate the causal effect of an intervention on a single treated aggregate unit (e.g., country, state, city).
- Challenge: No single untreated unit provides a valid counterfactual.
- Solution: Construct a "synthetic control" — a weighted average of untreated units that mimics the treated unit's pre-treatment outcomes and predictors.
- Especially useful when standard DiD is infeasible due to poor comparison groups.
- Read Abadie, Diamond, and Hainmueller (2010)

SCM Setup and Notation

- Units: 1 treated unit (e.g., USA), J untreated (e.g., donor pool of other developed countries).
- Outcome decomposition:

$$Y_{it} = Y_{it}(0) + \alpha_{it} D_{it}$$

- Treatment effect:

$$\alpha_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \quad \text{for } t > T_0$$

- Weights w_j^* are chosen to match pre-treatment outcomes and predictors.

Constructing the Synthetic Control

- Find weights $W = (w_2, ..., w_{J+1})'$ such that:

$$\sum_{j=2}^{J+1} w_j = 1, \quad w_j \geq 0$$

- Minimize pre-treatment loss:

$$\min_W (X_1 - X_0 W)' V (X_1 - X_0 W)$$

  where:

  - X_1 = Predictors for treated unit.
  - X_0 = Predictors for control units.
  - V = Weighting matrix reflecting predictor importance.

- V chosen to minimize pre-treatment MSPE (Mean Squared Prediction Error).

Key Assumptions of SCM

- **No Interference (SUTVA)**: Treatment affects only the treated unit.
- **Convex Hull**: Treated unit can be approximated as a convex combination of controls.
- **No Time-Varying Unobservables**: Unobserved factors evolve similarly across units.
- **Good Pre-Treatment Fit**: Synthetic control closely matches treated unit before intervention.

[Python Code for SCM]

Example: Tobacco Control in California

Abadie, Diamond, and Hainmueller (2010)

Objective: Estimate the causal effect of California's 1988 anti-smoking law (Proposition 99) on per-capita cigarette consumption.

- **Treated Unit:** California.
- **Synthetic Control:** Weighted average of U.S. states that match California's pre-1988 smoking trends and covariates (e.g., income, demographics, cigarette prices).
- **Outcome:** Annual per-capita cigarette sales (1970–2000).
- **Estimation:** Choose weights w_j to minimize the distance between treated and control units on pre-treatment predictors.

Findings:

- After 1988, California's smoking dropped significantly more than its synthetic control. By 2000, smoking in CA was 26 packs/person lower than the counterfactual. Placebo tests confirmed the estimated effect was unusually large relative to untreated states.

Advantages and Limitations

Advantages

- Transparent and data-driven; avoids arbitrary selection of controls.
- Handles aggregate interventions where only one unit is treated.
- Robust to functional form misspecification.
- Useful when treatment affects all units in the treated group (no within-group variation).

Limitations

- Needs a large and rich donor pool for a good synthetic match.
- Sensitive to extrapolation beyond the convex hull.
- Typically focuses on one treated unit; inference is non-standard.
- Results sensitive to choice of predictors and pre-treatment periods.

SCM Diagnostics

- **Pre-treatment Fit:**
  - Examine how well the synthetic control reproduces the treated unit's outcomes before intervention. Use pre-intervention Mean Squared Prediction Error (MSPE) as a metric.
- **Placebo Tests (Inference):**
  - Apply SCM to control units as if they were treated. Compare post-treatment gaps of treated vs. placebo units. Helps assess if the estimated treatment effect is unusually large.
- **MSPE Ratio:**
  - Compute ratio of post/pre-intervention MSPE for treated and placebo units. Large MSPE ratio for treated unit relative to control units suggests significance.
- **Balance Table:**
  - Compare predictor means for treated unit, synthetic control, and donor pool. Good balance supports the credibility of the synthetic counterfactual.
- **Graphical Inspection:**
  - Plot outcome paths for treated unit and synthetic control. Clear divergence post-treatment with good pre-treatment fit indicates strong evidence.

Propensity Score Matching

[Read Angrist and Pischke, Ch.3]

Propensity Score Matching: Setup

- We are interested in estimating the treatment effect of D:

$$Y_i = \tau D_i + \beta X_i + \varepsilon_i$$

- Problem: We never observe both $Y(D=1)$ and $Y(D=0)$ for the same unit. Treated and untreated units may differ systematically in X — selection bias.
- Idea: Instead of matching directly on high-dimensional X, match units based on their **propensity score**: probability of receiving the treatment.
- Goal: Use untreated units with similar X to estimate counterfactual $Y(0)$ for treated units.
- Assumption 1: Conditional Independence Assumption (CIA): Treatment assignment is independent of potential outcomes conditional on X.

$$Y(1), Y(0) \perp D \mid p(X)$$

- Assumption 2: Overlap (Common Support): Every unit has a positive probability of being both treated and untreated.

$$0 < \Pr(D = 1 \mid X) < 1$$

Propensity Score and Matching

- Propensity score: the probability of receiving treatment given covariates:

$$p(X) = f(D = 1|X)$$

  - Often estimated via logit, e.g., $\log\left(\frac{\Pr(D=1|X)}{1-\Pr(D=1|X)}\right) = X'\beta$
  - Machine learning (e.g., random forests, boosting) can also be used.
  - Prediction accuracy is prioritized in $p(X)$, not causal interpretation.
- Matching strategy:
  - Estimate $p(X)$ using logistic regression or machine learning.
  - Match each treated unit $D = 1$ with one or more control units $D = 0$ that have **similar** $p(X)$.
- Matching methods
  - Nearest Neighbor Matching: Match each treated unit to the nearest untreated unit based on $p(X)$.
  - Radius Matching: Match within a predefined caliper around $p(X)$.
  - Kernel Matching: Weighted average of all controls with weights decreasing with distance in $p(X)$.
  - Stratification (Subclassification): Divide data into strata based on $p(X)$ and compare outcomes within strata.

Implementation Steps of PSM

1. Estimate the propensity score
2. Match treated and untreated units based on their estimated $p(X)$.
3. Estimate the average treatment effect on the treated (ATT).
4. After matching, covariates X should be similar between treated and control groups.
   - Standardized mean differences.
   - Histograms/ density plots of $p(X)$ for treated and controls
   - Formal statistical tests (e.g., t-tests on covariate means).
   - Poor balance suggests poor matching. Re-specify or refine.
   - Be cautious about limited overlap (common support) and off-support matches.

[Python code on PSM]

Estimation

Matching-Based ATT Estimator

- For each treated unit i, find control units $j \in \mathcal{C}(i)$ with similar $e(X)$:

$$\hat{\tau}_{ATT}^{\text{match}} = \frac{1}{N_T} \sum_{i:D_i=1} \left(Y_i - \sum_{j \in \mathcal{C}(i)} w_{ij} Y_j \right)$$

where: $\mathcal{C}(i)$: matched control units and w_{ij}: weights summing to 1 (e.g., equal weights for nearest neighbors)

Another way: Inverse Probability Weighting (IPW)

- Use weights based on estimated propensity score $p(X)$:

$$\hat{\tau}_{ATT}^{\text{IPW}} = \frac{1}{N_T} \sum_{i=1}^{N} D_i Y_i - \sum_{i=1}^{N} \frac{(1-D_i)p(X_i)}{1-p(X_i)} Y_i$$

- Weights reweight control outcomes to resemble treated group.

Optional Regression Adjustment (after matching): Even after matching or weighting, residual differences in covariates might remain.

$$Y_i = \beta X_i + \tau D_i + \varepsilon_i$$

How Are Matching Weights w_{ij} Calculated?

1. Nearest Neighbor Matching

$$w_{ij} = \begin{cases} 1 & \text{if } j = \arg\min_{j \in \text{controls}} |p(X_i) - p(X_j)| \\ 0 & \text{otherwise} \end{cases}$$

Or for k nearest neighbors: $w_{ij} = \frac{1}{k}$ if $j \in \mathcal{C}(i)$

2. Caliper (Radius) Matching

$$w_{ij} = \begin{cases} \frac{1}{|\mathcal{C}(i)|} & \text{if } |p(X_i) - p(X_j)| \leq \delta \\ 0 & \text{otherwise} \end{cases}$$

3. Kernel Matching

$$w_{ij} = \frac{K_h(p(X_i) - p(X_j))}{\sum_{j'} K_h(p(X_i) - p(X_{j'}))}$$

- $K_h(\cdot)$: kernel function (e.g., Gaussian) with bandwidth h
- Weights decay with distance in $p(X)$

PSM Example

Does foreign aid causally affect recipient countries economic freedom? [Bologna Pavlik et al. 2022]

- *Treatment:* Sustained, large aid increases (AidData)
- *Outcome:* Change in Economic Freedom of the World (EFW) index
- *Covariates:* GDPpc, polity score, lagged EFW, etc.

Method: Propensity Score Matching (PSM)

- Logit model to estimate $p(X) = \Pr(D = 1|X)$
- Nearest neighbor matching (k = 1–4)
- Compared with Mahalanobis Distance Matching (MDM)

Findings:

- Overall aid had *no consistent effect* on EFW.
- Governance-targeted aid showed modest positive effects (10-year horizon).
- Effects concentrated in trade freedom and short-run government size.

Limitations and Best Practices

Strengths

- Reduces dimensionality problem: matching on a scalar score.
- Transparent and intuitive.
- Can be used in cross-sectional data.

Limitations

- Only adjusts for observed covariates X.
- Sensitive to model specification for propensity score.
- Poor overlap or limited support can severely limit estimation and bias estimates.

Best practices

- Matching quality should be assessed (e.g., standardized differences).
- Common practice: perform matching + regression adjustment on the matched sample.
- Always check for balance in covariates.

DML

Double Machine Learning

Why Double Machine Learning (DML)?

- Goal: Estimate a causal effect (e.g., of treatment D on outcome Y) in the presence of complex covariates X.
- Problem: Machine learning models are good at prediction, but biased for inference.
- DML addresses this by:
  - Separating prediction (nuisance estimation) from causal inference.
  - Using sample splitting (cross-fitting) to avoid overfitting.

Model Setup and Key Assumptions

Model: Partial Linear Form

$$Y = D\tau + g(X) + \varepsilon, \quad D = m(X) + \nu$$

Assumptions:

- **Unconfoundedness:** $Y(0), Y(1) \perp D \mid X$
- **Overlap:** $0 < \Pr(D = 1 \mid X) < 1$
- **Orthogonality:** Estimation of τ is robust to small errors in $\hat{g}(X)$, $\hat{m}(X)$

How DML Works (Algorithm Steps)

1. **Split** the data into K folds.
2. **Estimate** the nuisance functions $\hat{g}(X)$ and $\hat{m}(X)$ using machine learning (e.g., trees).
3. **Residualize**:
$$\tilde{Y}_i = Y_i - \hat{g}(X_i), \quad \tilde{D}_i = D_i - \hat{m}(X_i)$$
4. **Estimate causal effect**:
$$\tilde{Y}_i = \tau \tilde{D}_i + \eta_i$$

What Makes DML Work?

- **Cross-Fitting:** Prevents overfitting by estimating $\hat{g}(X)$, $\hat{m}(X)$ on separate subsamples.
- **Debiasing:** Residualization removes first-order ML bias, enabling valid estimation of τ.
- **Flexibility:** Any supervised ML algorithm can be used (e.g., Lasso, trees, boosting, neural nets).

[Python Code for DML]

Strengths and Limitations of DML

Strengths:

- Valid inference with high-dimensional or complex X
- Debiasing yields robust estimates
- Flexible nuisance function estimation using ML

Limitations:

- Requires strong overlap; sensitive to extreme $p(X)$
- Results depend on ML model quality
- More computationally intensive than standard regression

Choosing a Causal Inference Strategy

1. **Do you have repeated observations over time?**

- **Yes:**
  - *With a control group:*
    - Many pre/post observations → **Interrupted Time Series**, **Synthetic Control**.
    - Fewer time periods → **Difference-in-Differences (DiD)**.
  - *No control group:* **Interrupted Time Series (Single-group Pre/Post)**.
- **No:**
  - Treatment assigned based on a cutoff? → **Regression Discontinuity (RD)**.
  - Have a third variable that influences treatment but not directly outcome? → **Instrumental Variables (IV)**.
  - Have rich covariates measured pre-treatment?
    - → **Propensity Score Matching (PSM)** or **Double Machine Learning (DML)**.