Tutorial Week 8 and 9: Matching and Entropy Balancing

Problem Set - Propensity Scores, Matching, and Synthetic Control

Maximilian Birkle   Daniel Lehmann   Henry Lucas

2025-10-27

Contents

1 Load Packages **1**

2 Part I: Propensity Scores, Matching, and Robust Post-Matching Inference **3**
2.1 Background . . . 3
2.2 Load Data . . . 3
2.3 Define Treatment and Covariates . . . 3
2.4 Q0. First Check of the Data . . . 3
2.5 Q1. Estimating Propensity Scores . . . 4
2.6 Q2. Implement 1:1 Nearest-Neighbor Matching . . . 4
2.7 Q3. Standardized Mean Differences (SMDs) . . . 5
2.8 Q4. Overlap . . . 6
2.9 Q5. Matched-Pair ATT . . . 6
2.10 Q5.5. Bias–Variance Tradeoff in Matching Ratios . . . 7
2.11 Q6. Robust Post-Matching Inference (Abadie & Spiess, 2021) . . . 8
2.12 Q7. (Optional) Bootstrap Check . . . 8
2.13 Q8. Reflection . . . 9

3 Part II: Synthetic Control - German Reunification Study **10**
3.1 Background . . . 10
3.2 Load Data . . . 10
3.3 (a) Conceptual Questions . . . 10
3.4 (b) Mathematical/Optimization Questions . . . 10
3.5 (c) Estimation, Balance Before & After . . . 11
3.6 (d) Effect Size & Permutation Test . . . 11
3.7 (e) Placebo Test on Earlier Years . . . 12

4 Conclusion **13**

5 References **13**

1 Load Packages

```
library(tidyverse)      # Data manipulation and visualization
library(haven)          # Read Stata files
library(MatchIt)        # Matching methods
library(cobalt)         # Balance assessment
library(randomForest)   # Random forest for propensity scores
library(lmtest)         # Robust standard errors
```

```r
library(sandwich)       # Cluster-robust variance
library(knitr)          # Tables
library(kableExtra)     # Enhanced tables
library(broom)          # Tidy model output
library(Synth)          # Synthetic control method
```

2 Part I: Propensity Scores, Matching, and Robust Post-Matching Inference

2.1 Background

Research Question: Do **United Nations interventions** help shorten the duration of civil wars?

Gilligan and Sergenti (2008) use matching methods to re-evaluate earlier findings suggesting that UN interventions *prolong* conflict. They argue that this conclusion stems from **selection bias** — the UN tends to intervene in the *worst* conflicts.

Dataset: `war_pre_snapshots.dta`

Each row represents a **conflict episode** observed before a potential UN intervention. Our goal is to estimate the causal effect of UN involvement (`UN`) on the length of the conflict (`t1 - t0`), while balancing on key pre-treatment covariates.

2.2 Load Data

```r
# Read the UN intervention dataset


# Display structure and summary
```

2.3 Define Treatment and Covariates

```r
# Treatment variable: UN intervention (1 = Yes, 0 = No)


# Outcome variable: conflict duration (t1 - t0)


# Covariates for propensity score model
covar <- c(
  "inter", "deaths", "couprev", "sos", "drugs", "t0",
  "ethfrac", "pop", "lmtnest", "milper",
  "eeurop", "lamerica", "asia", "ssafrica"
)

# Create covariate matrix
```

2.4 Q0. First Check of the Data

2.4.1 Tasks:

1. Why might UN interventions **not** be randomly assigned across conflicts?
2. Which of the listed variables are most likely to confound the relationship between `UN` and conflict duration? Run a quick logistic regression and check.

```r
# Logistic regression: UN intervention as function of covariates
```

Discussion:

[Why might UN interventions not be randomly assigned? Which variables show strong associations with UN intervention?]

2.5 Q1. Estimating Propensity Scores

2.5.1 Theoretical Background

Let $T_i \in \{0, 1\}$ be the treatment indicator and $X_i = (X_{i1}, X_{i2}, \dots, X_{ip})$ the vector of pre-treatment covariates.

The **propensity score** is defined as:

$$e(X_i) = P(T_i = 1 \mid X_i)$$

2.5.2 Tasks:

1. Define the propensity score
2. Estimate $\hat{e}(X_i)$ in two ways:
   - **(a)** Logistic regression: $\text{logit}(e(X_i)) = X_i'\beta$
   - **(b)** Random forest classifier
3. Report mean, SD, and range of $\hat{e}(X_i)$ for treated and control
4. Create histogram/density plot by treatment status

```
# (a) Logistic regression propensity score


# Extract predicted probabilities
```

```
# (b) Random forest propensity score


# Extract predicted probabilities
```

```
# Summary statistics of propensity scores
```

```
# Density plot of propensity scores by treatment status
```

Interpretation:

[Discuss the distribution of propensity scores. Are there regions of poor overlap?]

2.6 Q2. Implement 1:1 Nearest-Neighbor Matching

2.6.1 Matching Setup

For each estimated propensity score $\hat{e}(X_i)$, match each treated unit to the nearest control on the **logit of the propensity score**:

$$\ell_i = \log\left(\frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)}\right)$$

Use **replacement** and a **caliper** of $0.2 \times \text{SD}(\ell_i)$ to restrict poor matches.

2.6.2 Tasks:

1. Implement matching using both logit and RF propensity scores
2. Report how many treated units fail to find a match
3. How does this change the estimand?

```r
# Matching using logit propensity scores


# Number of matched treated units
```

```r
# Matching using RF propensity scores


# Number of matched treated units
```

Discussion:

[How many treated units were dropped? What does this mean for the target estimand (ATT)?]

2.7 Q3. Standardized Mean Differences (SMDs)

2.7.1 SMD Formulas

For each covariate X^k:

Before matching (ATT version):

$$\text{SMD}_{\text{raw}}(k) = \frac{\bar{X}^k_{T=1} - \bar{X}^k_{T=0}}{\sqrt{s^{2,k}_{T=1}}}$$

After matching:

$$\text{SMD}_{\text{match}}(k) = \frac{\bar{X}^{k,\text{treated}}_{\text{match}} - \bar{X}^{k,\text{control}}_{\text{match}}}{\sqrt{s^{2,k}_{T=1}}}$$

2.7.2 Tasks:

1. Compute SMDs before and after matching for all covariates
2. Create a Love plot showing balance before and after matching (both methods)
3. Add vertical line at 0.1 (acceptable threshold)
4. Comment on which design achieves better covariate balance
5. Create two additional Love plots including interactions and squared terms

```r
# Calculate SMDs before matching


# Calculate SMDs after matching (logit)


# Calculate SMDs after matching (RF)
```

```r
# Love plot: Base covariates
```

```r
# Love plot: Including interactions
```

```r
# Love plot: Including squared terms
```

Interpretation:

[Which method achieves better balance? Are all SMDs below 0.1?]

2.8 Q4. Overlap

2.8.1 Tasks:

1. For each method, report:
   - Min and max of $\hat{e}(X_i)$ for treated and controls
   - Proportion of treated units whose $\hat{e}(X_i)$ lies inside the support of controls (and vice versa)
2. Plot distributions of $\hat{e}(X_i)$ for treated and controls
3. Identify regions of poor overlap or extreme propensities
4. (Optional) Trim observations outside common support and re-compute ATT
5. Examine matched subsets - do matches seem like fair counterfactuals?

```
# Min/max propensity scores by treatment group


# Proportion in common support
```

```
# Plot propensity score distributions
```

```
# Optional: Trim and re-estimate
```

Discussion:

[Is there good overlap? Which observations are on the edge of common support?]

2.9 Q5. Matched-Pair ATT

2.9.1 ATT Estimator

Let each matched pair be denoted by $(i, j(i))$ where i is treated and $j(i)$ is its matched control.

The **average treatment effect on the treated** is:

$$\hat{\tau}_{\text{ATT}} = \frac{1}{N_T^*} \sum_{i \in \mathcal{T}^*} \left(Y_i - Y_{j(i)} \right)$$

where $\mathcal{T}^*$ is the set of treated units with a valid match.

2.9.2 Task:

Compute the ATT for both matching methods.

```
# ATT using logit matching


# ATT using RF matching
```

Interpretation:

[What is the estimated effect of UN intervention on conflict duration?]

2.10 Q5.5. Bias–Variance Tradeoff in Matching Ratios

2.10.1 (a) Conceptual Question

For 1-to-m nearest-neighbor matching without replacement, the ATT estimator is:

$$\hat{\tau}_{\text{ATT}}^{(m)} = \frac{1}{N_T^*} \sum_{i \in \mathcal{T}^*} \left(Y_i - \frac{1}{m} \sum_{j \in \mathcal{J}(i)} Y_j \right)$$

where $\mathcal{J}(i)$ is the set of the m closest control matches for treated unit i.

Tasks:

1. Explain why increasing m tends to:
   - **Decrease variance**
   - **Increase bias**
2. Discuss how this relates to distance in covariate space
3. If overlap is weak, which risk dominates as m grows?

Discussion:

[Your explanation of the bias-variance tradeoff here]

2.10.2 (b) Practical Exercise

Tasks:

1. Re-run matching for 1:1, 2:1, and 3:1 ratios (with replacement and same caliper)
2. Record: number matched, mean distance, ATT estimate
3. Compute cluster-robust standard errors for each design
4. Create results table
5. Plot ATT vs. m with ±1.96 SE error bars

```
# 1:1 matching


# 2:1 matching


# 3:1 matching
```

```
# Create comparison table
```

```
# Plot ATT by matching ratio with error bars
```

Interpretation:

[Do results display expected bias-variance pattern?]

2.10.3 (c) Discussion

Tasks:

- Which design (1:1, 2:1, or 3:1) is most appropriate?
- How does observed pattern relate to Abadie & Imbens (2006)?
- What would happen with infinite data and perfect overlap?

Discussion:

[Your analysis here]

2.11 Q6. Robust Post-Matching Inference (Abadie & Spiess, 2021)

2.11.1 Regression with Cluster-Robust Standard Errors

After matching, fit the regression:

$$Y_i = \alpha + \tau T_i + \varepsilon_i$$

using only matched data.

Let $s(i)$ denote the **subclass (pair id)** of observation i.

Compute **cluster-robust standard errors** for $\hat{\tau}$ by clustering on $s(i)$:

$$\widehat{V}_{\text{CR}}(\hat{\tau}) = (X'X)^{-1} \left(\sum_s X_s' \hat{\varepsilon}_s \hat{\varepsilon}_s' X_s \right) (X'X)^{-1}$$

2.11.2 Tasks:

1. Report $\hat{\tau}$ and its cluster-robust standard error
2. Compare results for logit-matched and RF-matched samples

```
# Regression on logit-matched data with cluster-robust SE
```

```
# Regression on RF-matched data with cluster-robust SE
```

```
# Compare results
```

Interpretation:

[Compare point estimates and standard errors across methods]

2.12 Q7. (Optional) Bootstrap Check

2.12.1 Matched-Pair Bootstrap

Warning: Bootstraps are not theoretically valid for matching estimators, but this serves as a check.

Tasks:

1. Resample matched pairs (subclasses) with replacement
2. Recompute $\hat{\tau}^{(b)}$ for each bootstrap sample $b = 1, \dots, B$
3. Report bootstrap mean, SD, and percentile 95% CI
4. Compare to cluster-robust results

```
# Bootstrap procedure
```

Discussion:

[Do bootstrap and cluster-robust results tell a similar story?]

2.13 Q8. Reflection

2.13.1 Tasks:

1. Why does the propensity score $e(X_i)$ act as a **balancing score**?
2. How does random-forest estimation of $e(X_i)$ change matching results compared to logistic regression?
3. Why is overlap ($0 < e(X_i) < 1$) necessary for identifying the ATT?

Discussion:

[Your reflection here]

3 Part II: Synthetic Control - German Reunification Study

3.1 Background

In 1990, West Germany underwent reunification with East Germany. The question: *What was the economic cost (or benefit) of this event on West Germany's GDP per capita?*

Using the synthetic control method, we construct a counterfactual "synthetic West Germany" from a weighted combination of other OECD countries.

Paper: Abadie, Diamond & Hainmueller (2015), *Comparative Politics and the Synthetic Control Method*, AJPS.

Dataset: Available via Harvard Dataverse (doi:10.7910/DVN/24714)

3.2 Load Data

```
# Read German reunification dataset


# Display structure
```

3.3 (a) Conceptual Questions

3.3.1 Tasks:

1. Explain the intuition behind the synthetic control method. What kind of assignment problem does it address?
2. Why is it particularly suitable for the West Germany case?
3. What is the key identification assumption?

Discussion:

[Your conceptual explanation here]

3.4 (b) Mathematical/Optimization Questions

3.4.1 The Optimization Problem

The synthetic control method solves:

$$\min_{w} \sum_{t \leq T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt} \right)^2$$

subject to:

$$w_j \geq 0, \quad \sum_j w_j = 1$$

3.4.2 Tasks:

1. Write and explain each term in the optimization problem
2. What role do v-weights play in predictor balancing?

3. Why is the convex-combination constraint important? What if weights could be negative or sum $\neq 1$?

Mathematical Discussion:

[Your explanation of the optimization problem and constraints]

3.5 (c) Estimation, Balance Before & After

3.5.1 Tasks:

1. Estimate synthetic control for West Germany over pre-treatment period
2. Compute balance table of key predictors (GDP, trade openness, inflation, schooling, investment) showing treated vs. synthetic mean **before treatment**
3. Report non-zero weights w_j
4. Interpret: which donor countries dominate and why?
5. Assess whether pre-treatment fit is acceptable for credible inference

```
# Prepare data for Synth package


# Run synthetic control estimation
```

```
# Create balance table for pre-treatment predictors
```

```
# Report unit weights
```

Interpretation:

[Which countries contribute most to synthetic West Germany? Is pre-treatment balance good?]

3.6 (d) Effect Size & Permutation Test

3.6.1 Tasks:

1. Plot actual vs. synthetic GDP per capita trajectory (pre- and post-treatment)
2. Calculate estimated effect (gap) in first few post-treatment years and average post-treatment gap
3. Perform **permutation (placebo) test** by reassigning treatment to each control country
4. Report where treated unit's gap falls in the distribution (approximate p-value)
5. Interpret: What does this suggest about the economic impact of reunification?

```
# Plot actual vs synthetic West Germany
```

```
# Calculate treatment effect (gap)
```

```
# Permutation test: assign treatment to each control


# Calculate p-value
```

Interpretation:

[What is the estimated effect? Is it statistically significant based on permutation test?]

3.7 (e) Placebo Test on Earlier Years

3.7.1 Tasks:

1. Conduct placebo treatment year **before** actual 1990 treatment (e.g., 1975)
2. Re-estimate synthetic control and plot the gap
3. What does pre-treatment gap behavior tell you about parallel-trajectory assumption?
4. Comment on how convincing you find the main causal estimate

```
# Placebo test with fake treatment year


# Plot placebo gap
```

Interpretation:

[Does the placebo test support the validity of the main estimate?]

4 Conclusion

[Optional: Summarize key findings from both parts]

---

5 References

- Abadie, A., & Spiess, J. (2021). Robust Post-Matching Inference. *Journal of the American Statistical Association.*
- Rosenbaum, P., & Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika.*
- Abadie, A., & Imbens, G. (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica.*
- Gilligan, M., & Sergenti, E. (2008). Do UN Interventions Cause Peace? Using Matching to Improve Causal Inference. *Quarterly Journal of Political Science.*
- Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*, 59(2), 495–510.
- Abadie, A. (2021). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, 59(2), 391–425.
