Randomization Tests to Assess Covariate Balance When Designing and Analyzing Matched Datasets

Zach Branson

Carnegie Mellon University
Department of Statistics and Data Science

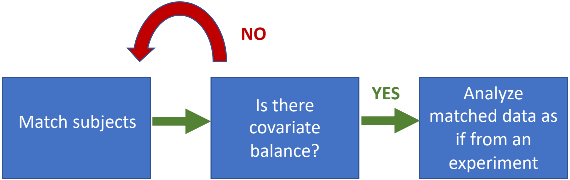
EUROCIM 2021

Experiments, Observational Studies, and Matching

- Experiments → similar treatment/control groups → causal inference
- In observational studies, treatment groups are typically not similar.
  → Biased estimators, sensitivity to model specification.
- **Matching**: Match treated subjects to "similar" control subjects.
  - Pair subjects by propensity score, Mahalanobis distance, etc.
  - Block subjects by coarsened covariates
  - Optimize group-level covariate balance
- Matching → similar treatment/control groups → causal inference
- Common to assume matched datasets ≈ randomized experiments
- But what kind of experimental design are we approximating, if any?
- Completely randomized? Blocked? Something else?
  → The choice has important implications for inference.

Matching and Covariate Balance Assessments

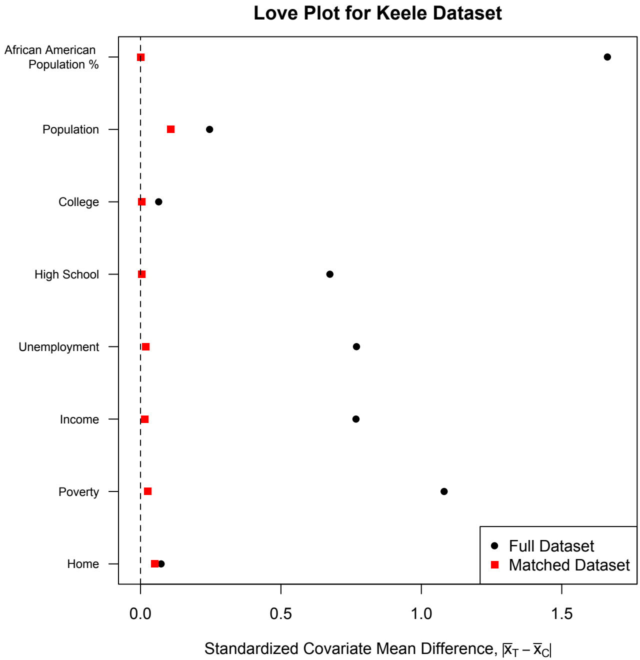
- Matching is only useful if it produces similar treatment/control groups (i.e., covariate balance).
- **Covariate balance assessments** always conducted after matching. (e.g., standardized $|\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C| \leq 0.1$?)

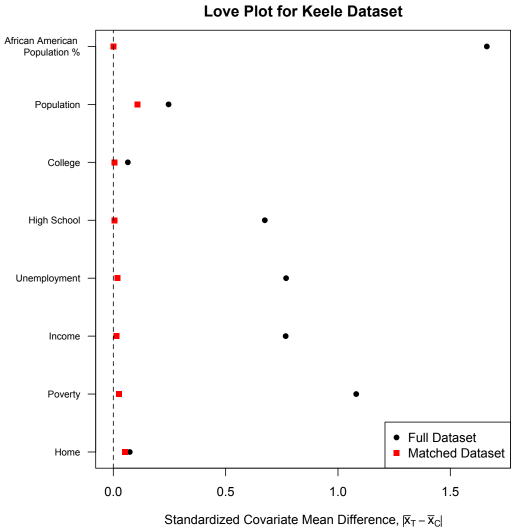


- Covariate balance assessments rely on rules-of-thumb. They do not formally test if an experiment has been approximated.
- Key point of this talk: Provide a valid randomization test to assess if a matched dataset approximates a particular experimental design.
- As an example, let's consider an application.

Application in Political Science

- Keele et al. (2017): Does having at least one African American candidate in Louisiana mayoral elections affect black voter turnout?
- Data: 1,006 elections (356 treatment, 650 control) from 1988-2011.
- Treatment: At least one electoral candidate was African American.
- Outcome: Black voter turnout (measured in percentage points).
- "treatment" cities were quite different from "control" cities.
- Keele et al. (2017) matched 197 pairs of treatment/control elections such that $|\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C| \leq 0.1$ for all covariates.


Love Plot for Keele Dataset
African American Population %
Population
College
High School
Unemployment
Income
Poverty
Home
0.0
0.5
1.0
1.5
Standardized Covariate Mean Difference, $|\bar{x}_T - \bar{x}_C|$
Full Dataset
Matched Dataset



- Used this as justification to analyze the matched data as a **paired experiment**.

What designs can we consider for this matched dataset?

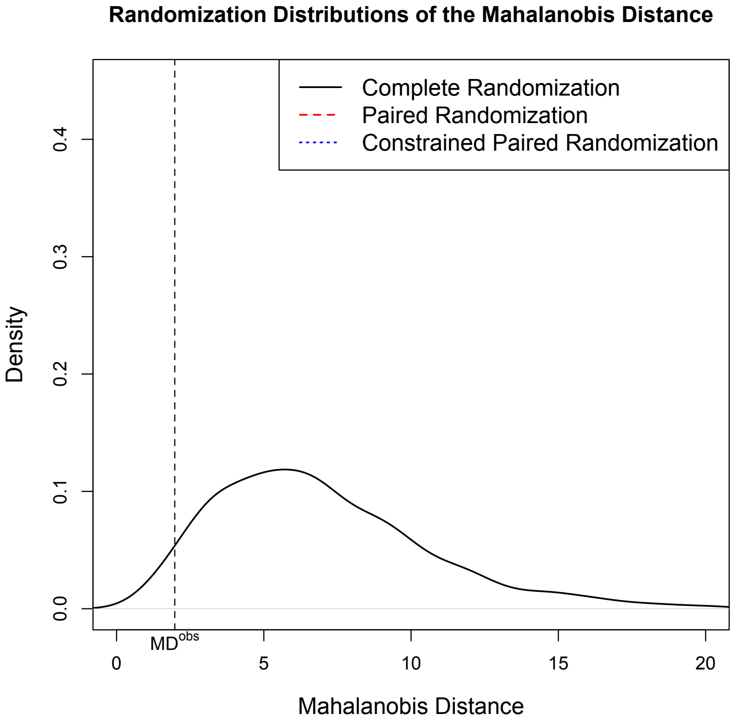
- We have 197 matched pairs such that $|\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C| \leq 0.1$ for all covariates.
- Should we view this dataset as approximating an experiment?
- Three experimental designs we will consider:
    1. **Complete Randomization**: Permutations of treatment.
    2. **Paired Randomization**: Permutations of treatment within pairs.
    3. **Constraind Paired Randomization**: Permutations of treatment within pairs, such that $|\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C| \leq 0.1$.
- We'll present a test for these designs.
- Lets us pinpoint which design—if any—is most appropriate.

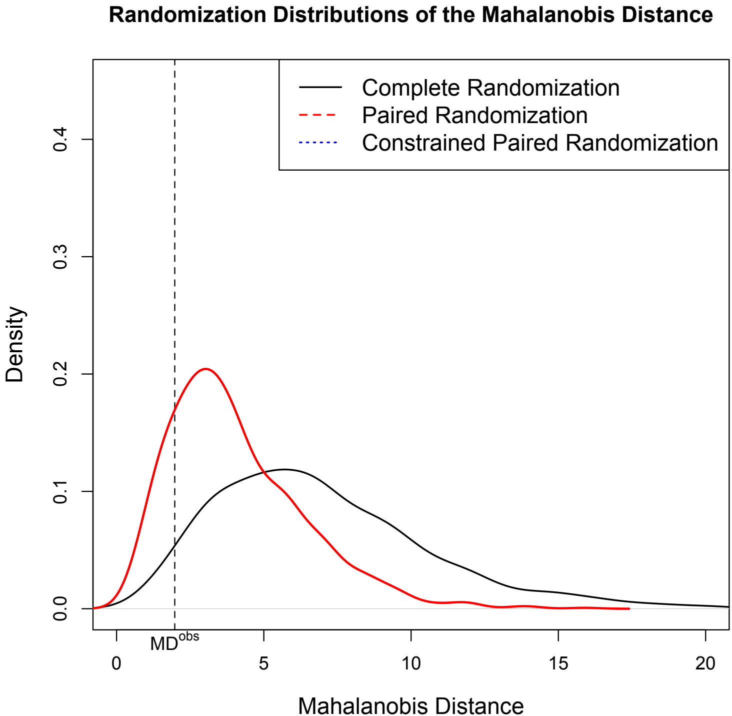
Test for Random Assignment in Matched Data

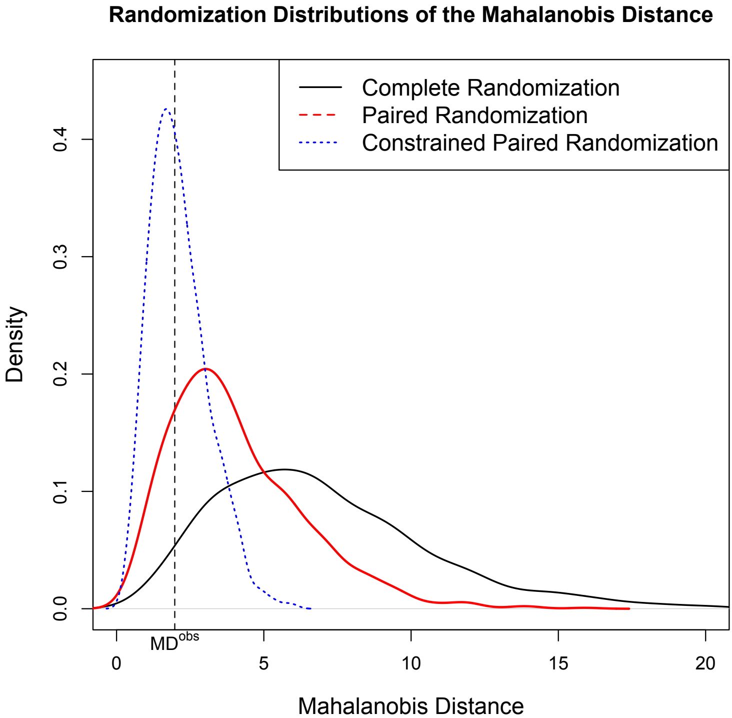
- Assume a matched dataset with N subjects, covariate matrix $\mathbf{X}_{N\times K}$, and binary treatment $\mathbf{W}_{N\times 1}$.
- Here is the test for **Complete Randomization**:
  1. Choose a **test statistic** $\underbrace{B(\mathbf{W},\mathbf{X})}_{\text{covariate balance}}$. We'll use the Mahalanobis distance:

$$B(\mathbf{W},\mathbf{X}) \equiv (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)^T \left[\text{cov}(\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)\right]^{-1} (\bar{\mathbf{x}}_T - \bar{\mathbf{x}}_C)$$

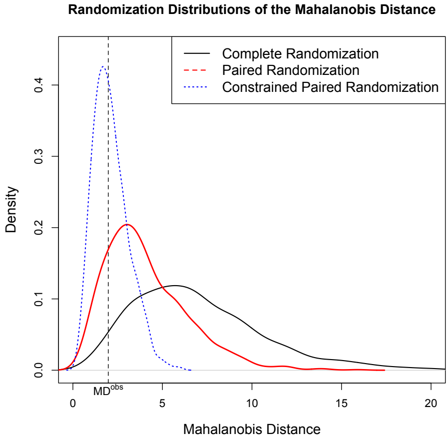
  2. Generate **hypothetical randomizations** $\underbrace{\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(M)}}_{\text{permutations}}$
  3. Compute $\underbrace{B(\mathbf{w}^{(1)},\mathbf{X}),\ldots,B(\mathbf{w}^{(M)},\mathbf{X})}_{\text{randomization distribution of covariate balance}}$
  4. Compare randomization distribution to observed balance.
- If observed balance is very different from randomization distribution, evidence against Complete Randomization.
- For other designs, just change Step 2 accordingly.


Randomization Distributions of the Mahalanobis Distance
Complete Randomization
Paired Randomization
Constrained Paired Randomization
Density
0.0
0.1
0.2
0.3
0.4
MD^obs
0
5
10
15
20
Mahalanobis Distance


Randomization Distributions of the Mahalanobis Distance
Complete Randomization
Paired Randomization
Constrained Paired Randomization
Density
0.0
0.1
0.2
0.3
0.4
MD^obs
0
5
10
15
20
Mahalanobis Distance



Randomization Distributions of the Mahalanobis Distance
Complete Randomization
Paired Randomization
Constrained Paired Randomization
Density
0.0
0.1
0.2
0.3
0.4
MD^obs
0
5
10
15
20
Mahalanobis Distance



- Constrained Paired Randomization appears to be the most plausible.
- Justifies using a CI under this design, which is narrower than under Paired Randomization or Complete Randomization.

Conclusion

- Matching is a popular way to alleviate covariate imbalances. Balance checks are a part of every matching procedure.
- Our work provides a valid, exact test for the hypothesis that matched data approximates a particular experimental design.
  - Doesn't rely on rules-of-thumb that may not be appropriate for a particular dataset.
  - Allows for any experimental design.
  - Can graphically put several designs on the same univariate scale.
- Tests and graphics implemented in `R` package `randChecks`.
  - Can be used to formally assess balance for any binary indicator.
  - For example, balance checks also come up in instrumental variables and regression discontinuity designs.
- Paper in *Observational Studies* (2021)
  - Our test has more power correctly rejecting experimental designs than t-tests and KS tests.
  - Well-designed matched datasets can be analyzed as well-designed experiments, resulting in narrower CIs closer to the nominal level.