STAT6061/STAT5008 – Causal Inference

Part 3-2. Weighting and Matching

An-Shun Tai

[1]*Department of Statistics*
National Cheng Kung University

[2]*Institute of Statistics and Data Science*
National Tsing Hua University

"Design" of nonexperimental studies: Matching

Structuring the nonexperimental study as if it were a randomized experiment, based solely on pre-treatment variables.

- In observational studies, the stratification (and standardization) methods presented in Part 3.1 are designed to emulate the structure of stratified randomized experiments.

- Alternatively, matching methods aim to replicate the structure of paired randomized experiments to improve the balance in covariate distributions.

However, matching and paired randomized experiments differ in two ways:

1. Matching assumes unconfoundedness; paired experiments ensure it by design.
2. Matching is often inexact, leaving some covariate differences between pairs. In contrast, paired experiments use within-pair randomization to ensure equal assignment probabilities and avoid bias.

- Matching acts as a nonparametric imputation technique for estimating causal effects.

Steps in implementing matching methods

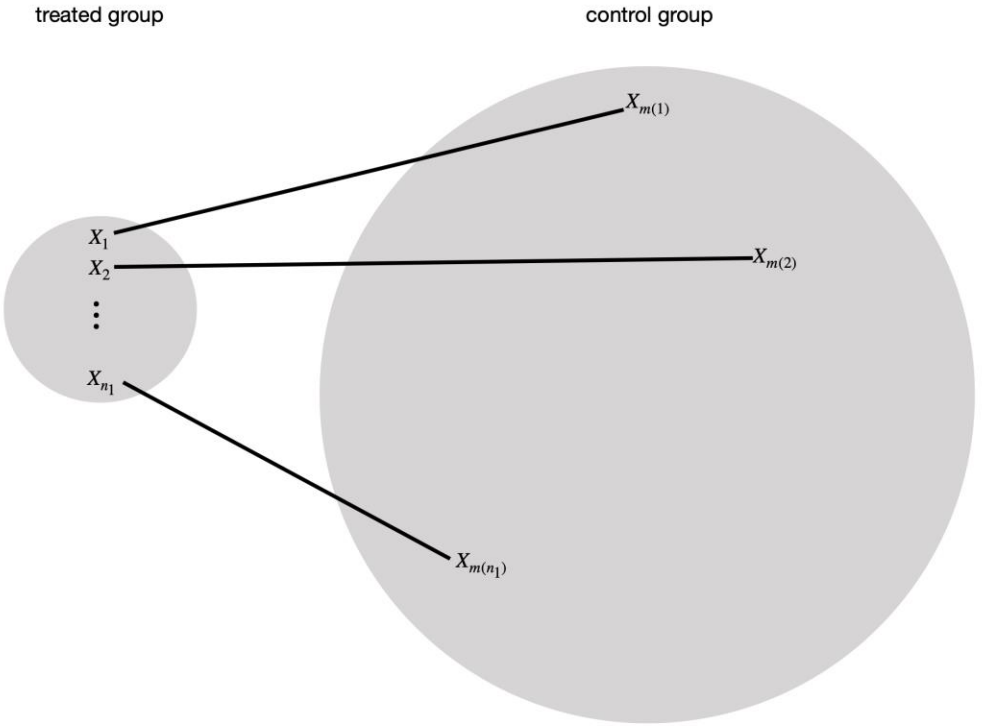
(Stuart, 2010; Ding, 2024)

1. Defining "closeness": the distance measure used to determine whether an individual is a good match for another.

2. Implementing a matching method, given that measure of closeness.

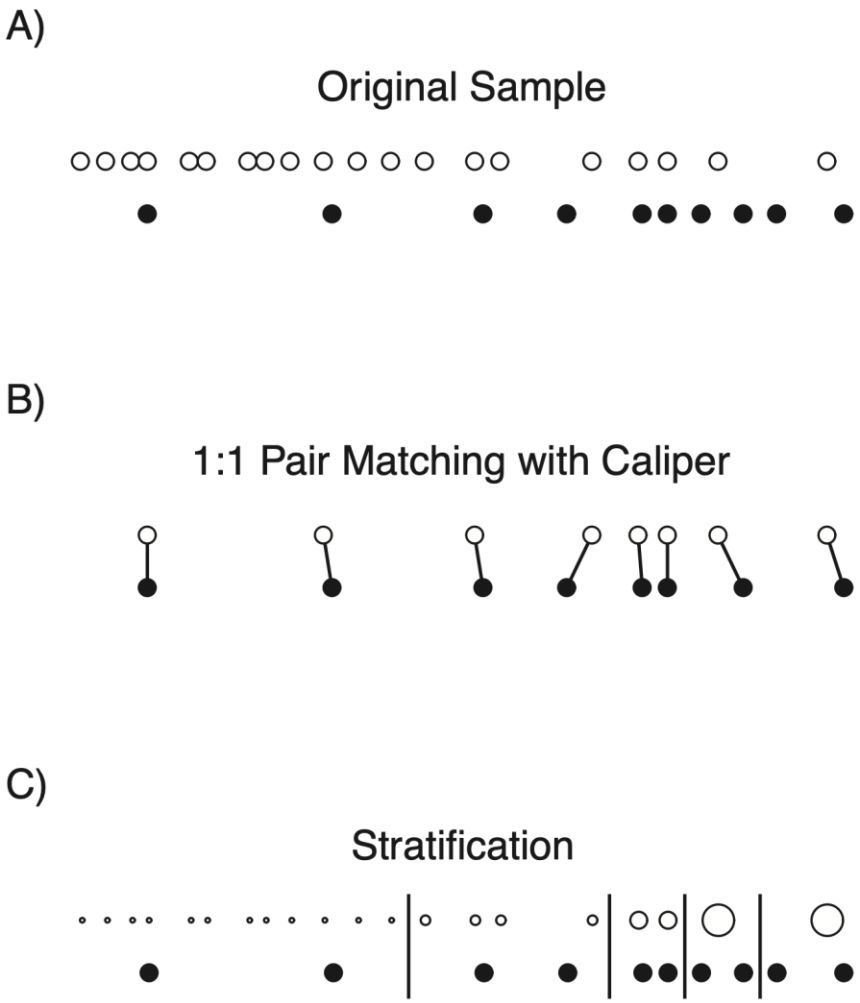
3. Assessing the quality of the resulting matched samples, and perhaps iterating with steps 1 and 2 until well-matched samples result.

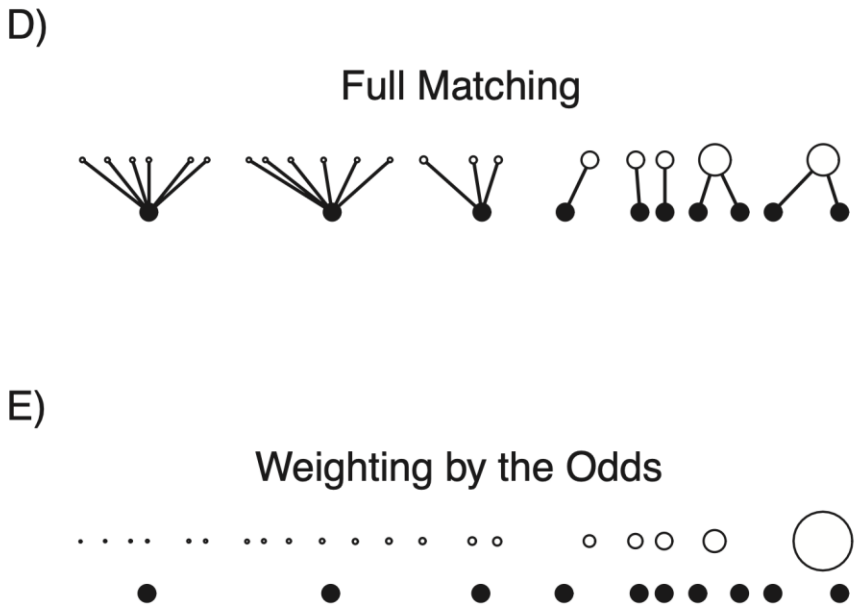
4. Analysis of the outcome and estimation of the treatment effect, given the matching done in step 3.



Types of matching

(Greifer and Stuart, 2021)





Benefits of Matching vs. Stratification

- **Finer covariate control**

Matching allows for unit-level pairing, achieving better covariate balance than broader strata.

- **Reduces model dependence**

Estimation relies less on parametric assumptions due to closer treated-control comparisons.

- **Handles continuous covariates more flexibly**

Matching avoids the need to arbitrarily categorize continuous variables.

- **Improves efficiency in small samples**

Especially when the number of treated units is small, matching can use information more effectively.

- **Enables visual and diagnostic checks**

Balance can be directly assessed before and after matching.

Choice of distance metric: Exact matching

- $d(X_i, X_j)$ represents a distance metric quantifying the dissimilarity between samples i and j in terms of their covariates X_i and X_j.

- $\mathcal{M}_i$ is the "matched set" for treated sample i.

Note: In practice, matching is typically performed by *pairing each treated sample with one or more similar control samples*.

- **Exact matching**

$$d(X_i, X_j) = \begin{cases} 0 & \text{if } X_i = X_j \\ \infty & \text{if } X_i \neq X_j \end{cases}$$

- $\mathcal{M}_i$ is a singleton, meaning each treated unit i is matched to a single control unit, denoted as $\mathcal{M}_i = \{i^*\}$.
- The difference-in-means estimator is unbiased for the sample average treatment effect on the treated (ATT).

$$\hat{\tau}_{\text{match}} = \frac{1}{N_1} \sum_i A_i \left(Y_i - \sum_{i^* \in \mathcal{M}_i} Y_{i^*} \right)$$

- Exact matching scheme is rarely feasible.

Choice of distance metric: Mahalanobis metric matching

➢ **Mahalanobis metric matching**

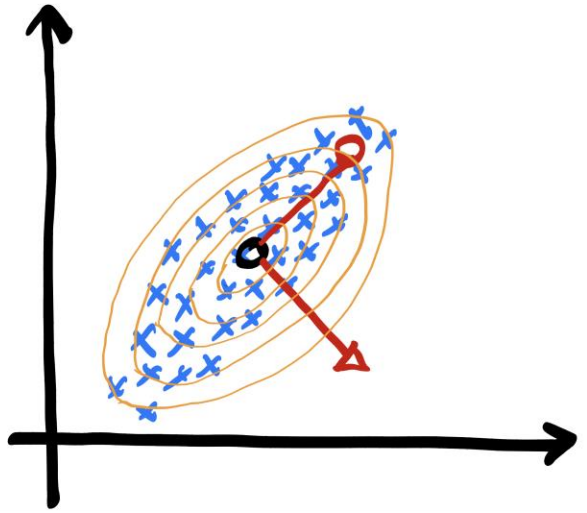
$$d(X_i, X_j) = (X_i - X_j)^T \Sigma^{-1} (X_i - X_j)$$

- Σ is the sample covariance matrix of X, calculated using (1) the pooled treatment and control groups when estimating the ATE, or using (2) only the control group when estimating the ATT.

- Important property: It's not affected by changes in measurement units or affine transformations of the covariates.

Mahalanobis distance

1. Unlike Euclidean distance, Mahalanobis distance standardizes covariates and accounts for the covariance structure of the data.
2. This ensures that variables with larger variances or strong correlations don't disproportionately influence the distance.



Choice of distance metric: Propensity score matching

- Matching directly on **high-dimensional** covariates X (especially using Mahalanobis distance) becomes difficult as the number of covariates grows. Why?

- **Propensity score**

$$e(X) = \Pr(A = 1|X)$$

- The propensity score is typically estimated using logistic regression.
- One-dimensional matching problem.

Propensity score matching: $d(X_i, X_j) = |e(X_i) - e(X_j)|$

Linear propensity score matching: $d(X_i, X_j) = |\text{logit}(e(X_i)) - \text{logit}(e(X_j))|$

- Matching on the linear propensity score can be particularly effective in terms of reducing bias (Rubin, 2001).

Coarsened exact matching (CEM)

(Iacus, King, and Porro, 2011; Iacus, King, and Porro, 2012)

- Propensity score matching relies on model specification and may not ensure covariate balance.

- Core idea

CEM improves causal inference by exactly matching units on coarsened versions of covariates, ensuring pre-processing balance and reducing reliance on modeling assumptions.

- How it works

1. Temporarily coarsen covariates into meaningful bins (e.g., age → decades).
2. Drop unmatched strata (i.e., strata without both treated and control units).
3. Estimate treatment effects on the retained, balanced sample.

- Advantages

- Balance by design: Covariate imbalance is eliminated before analysis.
- Model-free: No need to estimate propensity scores.
- User control: Balance vs. sample size tradeoff is explicit and adjustable.
- Robust to high-dimensional confounding if coarse bins are appropriately defined.

Covariate balance check

- Matching methods are only as effective as the covariate balance they achieve between treatment groups — so how can we ensure good balance?

- View X as a pseudo outcome and assess the difference in the covariate distribution between treatment and control groups.

1. Graphical balance assessment

- Use density plots or histograms to assess the degree of overlap in covariate distributions between groups.

2. Univariate Balance Statistics

- Standardized Mean Difference (SMD), Variance Ratio, Kolmogorov–Smirnov (KS) Statistic

3. Multivariate Balance Statistics

- Mahalanobis Distance, L1 Imbalance Metric, Prognostic Score Balance

Covariate balance check: Univariate Balance Statistics

- The difference-in-means estimator of the covariates: $\hat{\tau}_X = \frac{1}{N_1}\sum_{i=1}^{N} A_i X_i - \frac{1}{N_0}\sum_{i=1}^{N}(1-A_i)X_i$

- **Standardized Mean Difference (SMD)**

$$\frac{\hat{\tau}_X}{\sqrt{(\hat{S}_{X1}^2 + \hat{S}_{X0}^2)/2}}$$

where

$$\hat{S}_{X1}^2 = \frac{1}{N_1 - 1}\sum_{i=1}^{N} A_i (X_i - \bar{X}_{A=1})^2 \,, \hat{S}_{X0}^2 = \frac{1}{N_0 - 1}\sum_{i=1}^{N} (1-A_i)(X_i - \bar{X}_{A=0})^2 \,, \bar{X}_{A=1} = \frac{1}{N_1}\sum_{i=1}^{N} A_i X_i \,, \bar{X}_{A=0} = \frac{1}{N_0}\sum_{i=1}^{N} (1-A_i)X_i$$

- Assesses mean differences for each covariate between treatment groups → Common threshold: SMD < 0.1

- In addition to comparing the differences in location in the two distributions, one may wish to compare measures of dispersion in the two distributions./

- **Log ratio of standard deviations**

$$\ln \hat{S}_{X1}^2 - \ln \hat{S}_{X0}^2$$

Issues in covariate balance when using t-statistics

- Why might t-statistics and p-value be inappropriate in this context?

$$\frac{\hat{\tau}_X}{\sqrt{\hat{S}_{X1}^2/N_1 + \hat{S}_{X0}^2/N_0}}$$

- **Balance is an in-sample property**: It does not depend on a broader population or super-population.

- **Hypothesis tests conflate balance with statistical power**: A change in p-value may reflect a change in power, not actual imbalance. For example, randomly discarding control units may appear to improve balance, but it only reduces power (Imai, King, and Stuart, 2008).

- **Not suitable for use in stopping rules**: As the control sample size increases, t-statistics tend to decrease, which can misleadingly suggest improved covariate balance. However, this does not reflect actual balance and is inappropriate for guiding matching procedures. Why?

Issues in covariate balance when using propensity scores

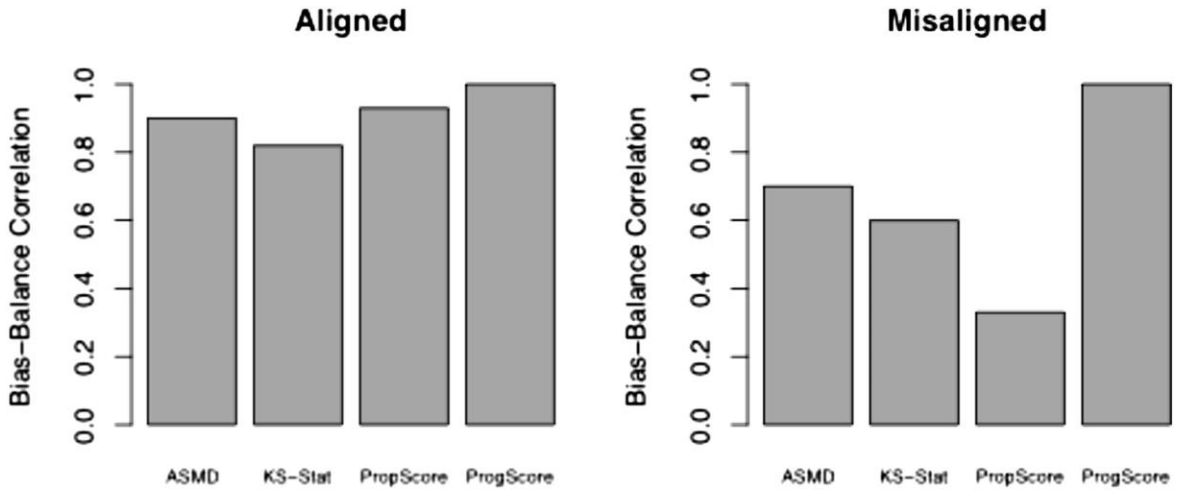
(Stuart, Lee, and Leacy, 2013)

- Although the propensity score is often used to assess balance empirically, simulation studies have shown this approach is conceptually flawed—propensity scores are only meaningful when they actually achieve covariate balance.

Aligned Covariates: When covariates are strongly related to both treatment and outcome, all balance measures tend to show high correlation with bias, making them reliable indicators.

Misaligned Covariates: When covariates influence treatment but not the outcome, propensity score balance shows poor correlation with bias, reducing its effectiveness as a diagnostic tool.

ASMD: Mean Absolute Standardized Mean Difference across covariates.
KS-Stat: Mean Kolmogorov–Smirnov test statistic for distributional differences.
PropScore: Absolute standardized mean difference in the propensity score.
ProgScore: Absolute standardized mean difference in the prognostic score.



Estimation

- In the context of matching, the ATT is often the more natural estimand for assessing causal effects.

- $\mathcal{M}_i = \{j | d(X_i, X_j) \leq \delta\}$ is the "matched set" for treated sample i.
- Matching can be one-to-M, assigning multiple controls to each treated unit.
- The difference-in-means estimator

$$\hat{\tau}_{\text{match}} = \frac{1}{N_1} \sum_i A_i \left(Y_i - \frac{1}{|\mathcal{M}_i|} \sum_{i^* \in \mathcal{M}_i} Y_{i^*} \right)$$

- Comparison with outcome regression (model-based imputation).

$$\frac{1}{N_1} \sum_i A_i (Y_i - \hat{\mu}_0(X_i))$$

Bias of the matching estimator

- Except for exact matching, all other matching methods may still result in covariate imbalance, which can introduce bias.

- **Bias of Matching**

- Let $\chi_{\mathcal{M}_i} = \{X_{i^*} | i^* \in \mathcal{M}_i\}$ denote the set of covariates for all control units matched to treated unit i.
- Bias

$$B(X_i, \chi_{\mathcal{M}_i}) = \mathbb{E}(Y_i(0)|A_i = 1, X_i) - \mathbb{E}\left\{\frac{1}{|\mathcal{M}_i|}\sum_{i^* \in \mathcal{M}_i} Y_{i^*} \middle| \chi_{\mathcal{M}_i}\right\}$$
$$= \mu_0(X_i) - \frac{1}{|\mathcal{M}_i|}\sum_{i^* \in \mathcal{M}_i} \mu_0(X_{i^*})$$

- Bias-corrected matching estimators(Abadie and Imbens, 2011)

Weighting

- Weighting as a generalization of matching

$$\hat{\tau}_{\text{match}} = \frac{1}{N_1}\sum_i A_i\left(Y_i - \frac{1}{|\mathcal{M}_i|}\sum_{i^* \in \mathcal{M}_i} Y_{i^*}\right)$$

$$= \frac{1}{N_1}\sum_{i:A_i=1} Y_i - \frac{1}{N_0}\sum_{i:A_i=0}\left(\frac{N_0}{N_1}\sum_{i':A_{i'}=1}\frac{1(i \in \mathcal{M}_{i'})}{|\mathcal{M}_{i'}|}\right)Y_i$$

$$= \frac{1}{N_1}\sum_{i:A_i=1} Y_i - \frac{1}{N_0}\sum_{i:A_i=0} W_i Y_i$$

Inverse-probability-weighting

- We can infer the average treatment effect by constructing a pseudo-population.

$$\int_x \mathbb{E}(Y|e,x)\Pr(x)\,dx = \int_{x,y} y\Pr(y|e,x)\Pr(x)\,dxdy = \int_{a,x,y} y\frac{I(a=e)}{\Pr(a|x)}\Pr(y,a,x)\,dxdyda = \mathbb{E}\left(\frac{I(A=e)}{\Pr(A|X)}Y\right)$$

- ATE can be estimated by

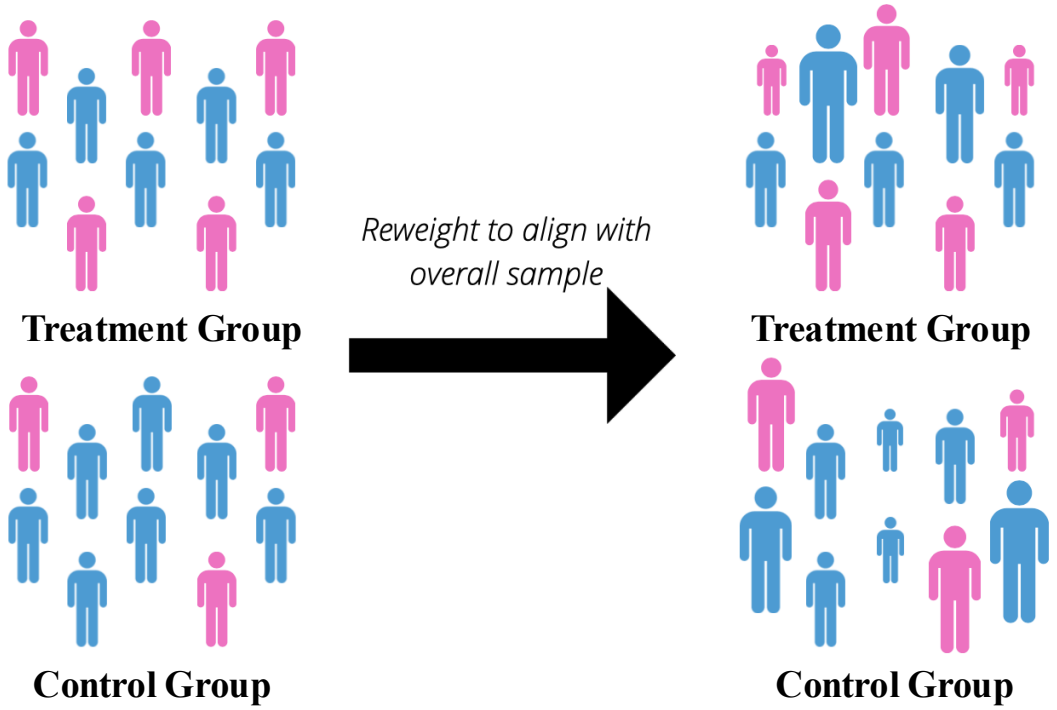
$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{I(A_i=1)}{e(X_i)}Y_i - \frac{I(A_i=0)}{1-e(X_i)}Y_i\right)$$

- ATT can be estimated by

$$\frac{1}{N_1}\sum_{i=1}^{N}\left(A_iY_i - \frac{e(X_i)I(A_i=0)}{1-e(X_i)}Y_i\right)$$

Pseudo-population

- In observational studies, treatment groups may differ systematically in covariates, making direct comparisons biased.
- Inverse probability weighting reweights individuals by the inverse of their probability of receiving the treatment (i.e., Propensity score) they actually received (based on covariates).
- This reweighting creates a pseudo-population where treatment assignment is independent of measured covariates—mimicking randomized experiment.
- In the pseudo-population, the distribution of covariates is balanced across treatment groups, enabling unbiased estimation of ATE.



Modified from https://causallycurious.com/posts/ip-weighting/ip_weighting

Remark 1: Overadjustment?

*"Matching methods are not designed to **compete with** modeling adjustments such as linear regression, and, in fact, the two methods have been shown to work best in combination (Stuart, 2010)"*

- **Does Matching + Regression = Overadjustment?**
- Matching handles design-stage confounding control.
- Regression addresses residual imbalance—systematic differences remaining after matching—and enhances estimation efficiency.

⇒Together, they help reduce bias and variance

Rubin, D. B., & Thomas, N. (2000). **Combining propensity score matching with additional adjustments for prognostic covariates.** *Journal of the American Statistical Association*, *95*(450), 573-585.

Table 4. Average Conditional Bias

	High-weight special variables			*Low-weight special variables*		
	Approximate linear	*Nonlinear additive*	*Nonlinear multiplicative*	*Approximate linear*	*Nonlinear additive*	*Nonlinear multiplicative*
1–1 matching						
Method 1	.04	.11	.09	.07	.09	.08
Method 2	.04	.05	.05	.08	.09	.09
Method 3	.23	.27	.29	.55	.53	.63
Method 4	.07	.05	.08	.08	.07	.08
Unmatched	.60	.55	.57	.47	.44	.44
Method 1 + reg	.02	.12	.09	.03	.08	.06
Method 2 + reg	.01	.03	.03	.03	.04	.04
Method 3 + reg	.03	.18	.10	.06	.13	.09
Method 4 + reg	.01	.04	.04	.03	.04	.03
Unmatched + reg	.02	.13	.09	.03	.09	.06
1–5 matching						
Method 1	.05	.15	.10	.05	.11	.08
Method 2	.04	.05	.05	.05	.05	.05
Method 3	.15	.19	.18	.35	.35	.38
Method 4	.14	.15	.13	.11	.13	.10
Unmatched	.58	.52	.55	.45	.43	.39
Method 1 + reg	.01	.12	.08	.02	.09	.05
Method 2 + reg	.01	.04	.03	.02	.04	.03
Method 3 + reg	.02	.10	.06	.04	.08	.05
Method 4 + reg	.01	.09	.05	.03	.07	.04
Unmatched + reg	.02	.13	.09	.03	.09	.06

NOTE: Square root of the average conditional squared bias. The treated population standard deviation is 1, because the matching leaves the treated group fixed. Simulation error is beyond the reported decimal value.

Remark 2: Are propensity scores suitable for matching?

(King and Nielsen, 2019)

➢ **Propensity score paradox (King and Nielsen, 2019)**

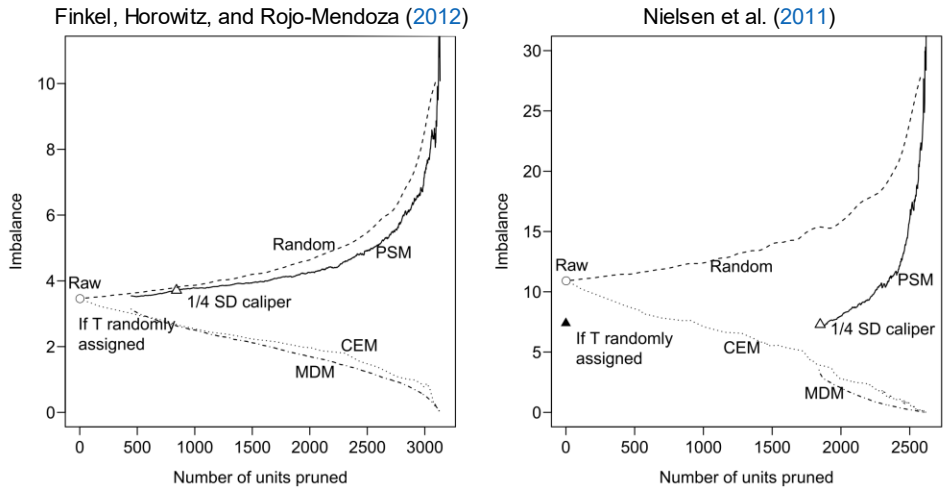
It refers to the counterintuitive situation where using propensity score matching (PSM) can actually increase imbalance or bias in some settings—despite being designed to reduce them.

➢ Why?

- Unlike Mahalanobis matching, which directly matches on observed covariates, PSM matches based on the estimated probability of treatment assignment.
- Close propensity scores do not guarantee similarity in covariates, so residual imbalance can still remain after matching.

➢ Damage caused in real data

Study	Context	Sample Size	Covariates
Finkel et al. (2012)	Civic education in Kenya	3,141 (1,347 treated)	Demographic, socioeconomic, leadership
Nielsen et al. (2011)	Aid shocks and conflict onset	2,627 (393 treated)	Democracy, wealth, population, prior conflict, ethnic & religious fractionalization



➢ Echoes Remark 1's conclusion: **Matching + Regression**

Remark 3: True or estimated propensity score?

- A well-known result in the literature (Rosenbaum, 1987) shows that using the estimated propensity score often leads to better efficiency and covariate balance than using the true propensity score.

- Why?

1. Estimated propensity scores adjust for random covariate imbalances in the sample, whereas the true propensity score does not.

2. While the true propensity score ensures unbiasedness, using an estimated score can reduce variance.

This phenomenon is akin to the benefits seen in regression adjustments within randomized experiments, where adjusting for covariates can lead to more precise estimates.

References

Abadie, A., & Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, *29*(1), 1-11.

Ding, P. (2024). *A First Course in Causal Inference.*

Greifer, N., & Stuart, E. A. (2021). Matching methods for confounder adjustment: an addition to the epidemiologist's toolbox. *Epidemiologic reviews*, *43*(1), 118-129.

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political analysis*, *27*(4), 435-454.

Iacus, S. M., King, G., & Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, *106*(493), 345-361.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, *20*(1), 1-24.

References

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American statistical Association*, *82*(398), 387-394.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*, 169-188.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *25*(1), 1.

Stuart, E. A., Lee, B. K., & Leacy, F. P. (2013). Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of clinical epidemiology*, *66*(8), S84-S90.