EL COLEGIO DE MÉXICO

CEE 60 AÑOS 1964-2024 CENTRO DE ESTUDIOS ECONÓMICOS

CENTRO DE ESTUDIOS ECONÓMICOS

Maestría en Economía 2024–2026

Microeconometrics for Evaluation

7 Propensity Score Matching

Disclaimer: I AM NOT the original intellectual author of the material presented in these notes. The content is STRONGLY based on a combination of lecture notes (from Aurora Ramirez), textbook references, and personal annotations for learning purposes. Any errors or omissions are entirely my own responsibility.

Jose Daniel Fuentes García
Github : danifuentesga

Índice

Agenda

- Why do we use the **propensity score**?
- How do we build the **propensity score**?
- How is the estimation of the **propensity score** implemented in Stata?
- **Explanation:** The agenda sets up the session: purpose, construction, and software implementation of the propensity score.
- **Intuition:** Think of it like learning a tool: first why it matters, then how to make it, and finally how to use it in practice.

Purpose of the Propensity Score

- When the treatment is not random, **propensity score matching** can be used to compare a treatment group and a control group that are equivalent in observable characteristics.
- **Propensity score matching** also summarizes the covariate information about treatment selection into a single scalar.
- **Explanation:** The propensity score balances treated and control groups on observed covariates, making them comparable. It condenses multiple variables into one score.
- **Intuition:** Imagine condensing many student attributes (age, grades, background) into one "similarity score" so you can match students fairly across groups.

Propensity Score

- Matching on X_i means comparing "close" units based on distance to the nearest neighbor of X_i, but this suffers from the curse of dimensionality.
- With **propensity score matching**, we compare units that, based only on observables, have very similar probabilities of being treated.
- If, conditional on X_i, two units have a similar probability of treatment, we say their **propensity scores** are similar.
- Comparing a treated unit with a control unit that has a similar **propensity score** makes the remaining variation—conditional on the score—random (if the selection-on-observables assumption holds).
- **Tiny Math Example:** Suppose Unit A has $p = 0{,}72$ and Unit B has $p = 0{,}70$. Even if their raw covariates differ, they can be matched because their treatment probabilities are almost identical.
- **Intuition:** Instead of comparing dozens of traits directly, we compress them into one probability and match units with nearly the same score.

Definition and Assumptions

Definition of the Propensity Score

The propensity score is defined as the probability of treatment conditional on covariates:

$$p(X_i) = \Pr(D_i = 1 \mid X_i) = \mathbb{E}(D_i \mid X_i).$$

Identification Assumptions

1. $(Y_{0i}, Y_{1i}) \perp D_i \mid X_i$ (conditional independence)
2. $0 < \Pr(D_i = 1 \mid X_i) < 1$ (common support)

- **Example:** If X_i = age and education, and we observe $p(X_i) = 0{,}65$, this means the unit has a 65 % chance of being treated given those covariates.
- **Intuition:** The score is just a summary probability. The two assumptions say (1) once we control for X_i, treatment is as good as random, and (2) every unit has a positive chance to be in either group.

Conditional Independence (Rosenbaum & Rubin, 1983)

Idea

There exists a set X of observable covariates such that, once we control for X, treatment assignment is independent of potential outcomes:

$$(Y_{0i}, Y_{1i}) \perp D_i \mid X_i.$$

- **Interpretation:** Given X_i, assignment is "as good as random."
- **Implication:** We can build an unbiased counterfactual for treated units using non-experimental controls.
- Also called treatment ignorability, selection on observables, exogeneity, or *unconfoundedness*.
- Not empirically testable.
- **Example:** If two individuals have the same X_i (say age = 30, income = 40k), then $\Pr(D_i = 1 \mid X_i)$ is the same for both, regardless of their unobserved outcomes.
- **Intuition:** Once you match on the right covariates, treatment looks like it was randomly assigned, even if it wasn't in reality.

Common Support

Condition

For each X, there is a positive probability of being treated and not treated:

$$0 < \Pr(D_i = 1 \mid X_i) < 1.$$

- Ensures sufficient overlap to find appropriate "pairs."
- Can be diagnosed with graphs/tables of score overlap.
- **Example:** If for young individuals $\Pr(D = 1|X) = 0{,}9$ and for old individuals $\Pr(D = 1|X) = 0{,}1$, both groups still have a chance of being treated and untreated — overlap exists.
- **Intuition:** To compare fairly, both groups must have some overlap; otherwise, we cannot find valid matches across treatment and control.

Identification under Strong Ignorability

When both assumptions hold (CIA and common support), treatment assignment is **strongly ignorable** (Rosenbaum & Rubin, 1983).

By definition:

$$\delta(x) = \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = x] \qquad \textbf{Def. of treatment effect} \tag{1}$$

$$\delta(x) = \mathbb{E}[Y_{1i} \mid X_i = x] - \mathbb{E}[Y_{0i} \mid X_i = x] \qquad \textbf{Linearity of expectation} \tag{2}$$

$$\mathbb{E}[Y_{1i} \mid X_i = x] = \mathbb{E}[Y_{1i} \mid D_i = 1, X_i = x] \qquad \textbf{Unconfoundedness: replace with treated} \tag{3}$$

$$\mathbb{E}[Y_{0i} \mid X_i = x] = \mathbb{E}[Y_{0i} \mid D_i = 0, X_i = x] \qquad \textbf{Unconfoundedness: replace with controls} \tag{4}$$

$$\delta(x) = \mathbb{E}[Y_{1i} \mid D_i = 1, X_i = x] - \mathbb{E}[Y_{0i} \mid D_i = 0, X_i = x] \qquad \textbf{Substitute into } \delta(x) \tag{5}$$

$$\delta = \mathbb{E}[\delta(X_i)] \qquad \textbf{Take expectation over } X_i \textbf{ (common support)} \tag{6}$$

$$\delta = \mathbb{E}[Y_i \mid D_i = 1, X_i] - \mathbb{E}[Y_i \mid D_i = 0, X_i] \qquad \textbf{Observed conditional means} \tag{7}$$

$$Y_i = \alpha + \delta D_i + \beta X_i + \varepsilon_i \qquad \textbf{Link with regression model} \tag{8}$$

- **Example:** If at $X = 12$ years schooling, treated mean $= 50$ and control mean $= 45$, then $\delta(12) = 5$. Averaging across X gives the overall effect δ.
- **Intuition:** Within each X group, treatment vs. control looks random. Their outcome difference is causal, and averaging these differences yields the treatment effect.

Propensity Score Theorem (Rosenbaum & Rubin, 1983)

Statement

If $(Y_{1i}, Y_{0i}) \perp D_i \mid X_i$, then:

$$(Y_{1i}, Y_{0i}) \perp D_i \mid p(X_i), \quad p(X_i) = \Pr(D_i = 1 \mid X_i).$$

$$(Y_{1i}, Y_{0i}) \perp D_i \mid X_i \qquad \textbf{Unconf. assumption} \tag{9}$$

$$p(X_i) = \Pr(D_i = 1 \mid X_i) \qquad \textbf{By def. of score} \tag{10}$$

$$\Pr(D_i = 1 \mid X_i) = p(X_i) \qquad \textbf{By def.} \tag{11}$$

$$\Pr(D_i = 1 \mid p(X_i)) = p(X_i) \qquad \textbf{Iterated exp.} \tag{12}$$

$$\mathbb{E}[Y_{1i} \mid D_i, p(X_i)] = \mathbb{E}[Y_{1i} \mid p(X_i)] \quad \textbf{Substitute} \quad (13)$$

$$\mathbb{E}[Y_{0i} \mid D_i, p(X_i)] = \mathbb{E}[Y_{0i} \mid p(X_i)] \quad \textbf{Substitute} \quad (14)$$

$$\Rightarrow (Y_{1i}, Y_{0i}) \perp D_i \mid p(X_i).$$

Dimension Reduction

- Stratifying on X_i suffers from sparse cells in finite samples.
- $p(X_i)$ is scalar; stratifying by score is more feasible.
- **Example:** Suppose X_i = age, income, education. Instead of stratifying on all three, we use $p(X_i) = 0{,}65$. Matching only on this scalar preserves conditional independence.
- **Intuition:** Many variables collapse into one number that balances treatment and control, making matching practical.

Sketch of Proof

Assume $(Y_{1i}, Y_{0i}) \perp D_i \mid X_i$. We want to show:

$$\Pr(D_i = 1 \mid Y_{0i}, Y_{1i}, p(X_i)) = p(X_i).$$

$$\Pr(D_i = 1 \mid Y_{0i}, Y_{1i}, p(X_i)) = \mathbb{E}[D_i \mid Y_{0i}, Y_{1i}, p(X_i)] \quad \textbf{Def. of cond. prob.} \quad (15)$$

$$= \mathbb{E}[\,\mathbb{E}[D_i \mid Y_{0i}, Y_{1i}, X_i] \,|\, Y_{0i}, Y_{1i}, p(X_i)] \quad \textbf{Law of iter. exp.} \quad (16)$$

$$= \mathbb{E}[\,\Pr(D_i = 1 \mid Y_{0i}, Y_{1i}, X_i) \,|\, Y_{0i}, Y_{1i}, p(X_i)] \quad \textbf{Def. of cond. prob.} \quad (17)$$

$$= \mathbb{E}[\,\Pr(D_i = 1 \mid X_i) \,|\, Y_{0i}, Y_{1i}, p(X_i)] \quad \textbf{Ignorability} \quad (18)$$

$$= \mathbb{E}[\,p(X_i) \mid Y_{0i}, Y_{1i}, p(X_i)] \quad \textbf{Def. of score} \quad (19)$$

$$= \int p(x)\, dF(x \mid Y_{0i}, Y_{1i}, p(X_i)) \quad \textbf{Expand cond. exp.} \quad (20)$$

$$= p(X_i) \quad \textbf{Measurability: } p(X_i) \textbf{ known} \quad (21)$$

$$\Rightarrow (Y_{1i}, Y_{0i}) \perp D_i \mid p(X_i). \qquad \blacksquare$$

- **Example:** If $p(X_i) = 0{,}6$, then even conditioning on outcomes (Y_{0i}, Y_{1i}), the probability of treatment is still 0,6. Outcomes add no predictive power once $p(X_i)$ is known.
- **Intuition:** $p(X_i)$ is a sufficient statistic of X_i for treatment assignment. Knowing Y's cannot change the probability, because $p(X_i)$ already summarizes all X-based information.

Balancing Property — Full Proof

Lemma. If $p(X_i) = \Pr(D_i = 1 \mid X_i)$ is the propensity score, then:

$$D_i \perp X_i \mid p(X_i).$$

Goal. Show that:

$$\Pr(D_i = 1 \mid X_i, p(X_i)) = \Pr(D_i = 1 \mid p(X_i)).$$

Left-hand side (LHS):

$$\Pr(D_i = 1 \mid X_i, p(X_i)) = \Pr(D_i = 1 \mid X_i) \qquad \textbf{Knowing } X_i \Rightarrow p(X_i) \qquad (22)$$

$$= p(X_i) \qquad \textbf{Def. of score} \qquad (23)$$

Right-hand side (RHS):

$$\Pr(D_i = 1 \mid p(X_i)) = \mathbb{E}[D_i \mid p(X_i)] \qquad \textbf{Def. of cond. prob.} \tag{24}$$

$$= \mathbb{E}\Big[\, \mathbb{E}[D_i \mid X_i, p(X_i)] \,\Big|\, p(X_i)\Big] \qquad \textbf{Law of iter. exp.} \tag{25}$$

$$= \mathbb{E}\Big[\, \Pr(D_i = 1 \mid X_i, p(X_i)) \,\Big|\, p(X_i)\Big] \qquad \textbf{Replace inner exp.} \tag{26}$$

$$= \mathbb{E}\Big[\, \Pr(D_i = 1 \mid X_i) \,\Big|\, p(X_i)\Big] \qquad p(X_i) \textbf{ encodes } X_i \tag{27}$$

$$= \mathbb{E}[\, p(X_i) \mid p(X_i)] \qquad \textbf{Def. of score} \tag{28}$$

$$= \int p(x)\, dF(x \mid p(X_i)) \qquad \textbf{Expand cond. exp.} \tag{29}$$

$$= p(X_i) \qquad \textbf{Measurability: } p(X_i) \textbf{ known} \tag{30}$$

Conclusion.

$$\Pr(D_i = 1 \mid X_i, p(X_i)) = \Pr(D_i = 1 \mid p(X_i)) = p(X_i),$$

so

$$D_i \perp X_i \mid p(X_i). \qquad \blacksquare$$

- **Example:** Suppose X_i = age, income. If $p(X_i) = 0{,}3$, then $\Pr(D = 1 \mid X_i, p(X_i)) = \Pr(D = 1 \mid p(X_i)) = 0{,}3$. Conditioning on X_i adds nothing once the score is known.
- **Intuition:** The score $p(X)$ is a *sufficient statistic* for treatment. Once you know it, X gives no further information about D.

Consequence of Balancing

Claim. Conditional on $p(X_i)$, the distribution of X_i is the same for treated and controls:

$$\Pr(X_i \mid D_i = 1, p(X_i)) = \Pr(X_i \mid D_i = 0, p(X_i)).$$

Proof.
From the balancing property:

$$D_i \perp X_i \mid p(X_i).$$

$$\Pr(X_i \mid D_i, p(X_i)) = \Pr(X_i \mid p(X_i)) \qquad \textbf{Def. of cond. indep.} \qquad (31)$$

$$= \frac{\Pr(X_i, p(X_i))}{\Pr(p(X_i))} \qquad \textbf{Cond. prob. rule} \qquad (32)$$

$$= \frac{\Pr(p(X_i) \mid X_i)\Pr(X_i)}{\Pr(p(X_i))} \qquad \textbf{Bayes' rule} \qquad (33)$$

$$= \frac{1 \cdot \Pr(X_i)}{\Pr(p(X_i))} \qquad p(X_i) \textbf{ determined by } X_i \qquad (34)$$

$$= \frac{\Pr(X_i)}{\Pr(p(X_i))} \qquad \textbf{Simplify} \qquad (35)$$

This expression does not depend on D_i, hence:

$$\Pr(X_i \mid D_i = 1, p(X_i)) = \Pr(X_i \mid D_i = 0, p(X_i)). \quad \blacksquare$$

- **Example:** If $p(X_i) = 0{,}5$, then within that group the distribution of age/income (X_i) looks the same for treated and control units.
- **Intuition:** Conditioning on the score balances the covariates — treatment vs. control groups look statistically identical in X.

Practical Implementation

1. Estimate the **propensity score**.
2. Compute the causal effect by averaging differences in Y between units with similar scores.

Estimation of $p(X_i)$:

$$\Pr(D_i = 1 \mid X_i) = F\{\, h(X_i) \,\} \qquad \textbf{Generic form} \qquad (36)$$

$$F(\cdot) = \text{logit or probit CDF} \qquad \textbf{Choice of link} \qquad (37)$$

$$h(X_i) = \text{linear terms} + \text{higher-order} + \text{interactions} \qquad \textbf{Specification} \qquad (38)$$

- Select $h(\cdot)$ to achieve **balance**, not merely based on statistical significance.

- **Example:** Suppose $h(X_i) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{income}$, with logit link $F(z) = \frac{1}{1+e^{-z}}$. Then $p(X_i)$ gives each individual's treatment probability.
- **Intuition:** Estimate the probability of treatment with a flexible model. The goal is not prediction accuracy, but ensuring treated and control groups look similar after matching on $p(X)$.

Testing the Balancing Property

1. Estimate the score (logit/probit).
2. Sort by score and divide into blocks of similar size (e.g., 5).
3. In each block, test equality of means of X between $D = 1$ and $D = 0$.
4. If the test fails, subdivide the block and re-test.
5. If a covariate fails systematically, enrich $h(\cdot)$ (higher-order terms/interactions) and repeat.

- **Example:** Suppose after estimating $p(X)$ you form 5 blocks. In block 3, $\bar{X}_{D=1} = 10{,}2$ and $\bar{X}_{D=0} = 10{,}1$. A t-test shows no significant difference $\rightarrow$ balance holds in that block.
- **Intuition:** Balance is checked block by block. If covariates are not balanced, you refine the model until treated and control groups look statistically similar within each score stratum.

Inverse Probability Weighting (IPW)

Assume $(Y_{1i}, Y_{0i}) \perp D_i \mid X_i$.
ATE via IPW

$$\delta_{ATE} = \mathbb{E}[Y_{1i} - Y_{0i}] \qquad \textbf{Def. of ATE} \tag{39}$$

$$= \mathbb{E}\left[\frac{Y_i D_i}{p(X_i)} - \frac{Y_i(1-D_i)}{1-p(X_i)}\right] \qquad \textbf{Re-weighted form} \tag{40}$$

Derivation.

$$\mathbb{E}\left[\frac{Y_i D_i}{p(X_i)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{Y_i D_i}{p(X_i)} \;\middle|\; X_i\right]\right] \qquad \textbf{Iterated exp.} \quad (41)$$

$$= \mathbb{E}\left[\frac{1}{p(X_i)} \cdot \mathbb{E}[Y_i D_i \mid X_i]\right] \qquad \textbf{Factor } 1/p(X_i) \quad (42)$$

$$= \mathbb{E}\left[\frac{1}{p(X_i)} \cdot \Pr(D_i = 1 \mid X_i)\, \mathbb{E}[Y_i \mid D_i = 1, X_i]\right] \qquad \textbf{Def. of cond. exp.} \quad (43)$$

$$= \mathbb{E}[\mathbb{E}[Y_i \mid D_i = 1, X_i]] \qquad \textbf{Cancel } p(X_i) \quad (44)$$

$$= \mathbb{E}[Y_{1i}] \qquad \textbf{Unconfoundedness} \quad (45)$$

Analogously:

$$\mathbb{E}\left[-\frac{Y_i(1 - D_i)}{1 - p(X_i)}\right] = -\,\mathbb{E}[Y_{0i}]. \quad (46)$$

Thus:

$$\delta_{ATE} = \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}]. \qquad \blacksquare$$

Alternative form.

$$\delta_{ATE} = \mathbb{E}\left[\frac{(D_i - p(X_i))}{p(X_i)(1 - p(X_i))}\, Y_i\right] \qquad \textbf{Algebra rearrangement} \quad (47)$$

TOT via IPW

$$\delta_{TOT} = \mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1] \qquad \textbf{Def. of TOT} \quad (48)$$

$$= \mathbb{E}\left[\frac{(D_i - p(X_i))}{1 - p(X_i)} \cdot \frac{Y_i}{\Pr(D_i = 1)}\right] \qquad \textbf{IPW representation} \quad (49)$$

- **Example:** If $p(X_i) = 0{,}2$, then treated units get weight $1/0{,}2 = 5$, controls get weight $1/0{,}8 = 1{,}25$. This makes treated and controls comparable "as if" randomized.
- **Intuition:** Units that are unlikely to receive the treatment but did are "rare" and thus receive high weight. IPW reconstructs a pseudo-population where treatment is independent of X.

Sample Analogues (IPW)

ATE estimator.
From population form:

$$\delta_{ATE} = \mathbb{E}\left[\frac{(D_i - p(X_i))}{p(X_i)(1-p(X_i))} Y_i\right].$$

$$\hat{\delta}_{ATE} = \frac{1}{N}\sum_{i=1}^{N} \frac{(D_i - \hat{p}(X_i))}{\hat{p}(X_i)(1-\hat{p}(X_i))} Y_i \qquad \textbf{Sample analogue} \tag{50}$$

TOT estimator.
From population form:

$$\delta_{TOT} = \mathbb{E}\left[\frac{(D_i - p(X_i))}{1-p(X_i)} \cdot \frac{Y_i}{\Pr(D_i = 1)}\right].$$

$$\hat{\delta}_{TOT} = \frac{1}{N^T}\sum_{i:D_i=1} \frac{(D_i - \hat{p}(X_i))}{1-\hat{p}(X_i)} Y_i \qquad \textbf{Sample analogue} \tag{51}$$

where $N^T = \sum_i D_i$ is the number of treated units.
Implementation.

1. Estimate $\hat{p}(X_i)$ from a first-stage logit/probit.
2. Plug $\hat{p}(X_i)$ into the formulas above.
3. Compute standard errors with bootstrap over both stages.

- **Example:** Suppose $N = 4$ with $\hat{p}(X) = (0{,}2, 0{,}5, 0{,}7, 0{,}3)$ and outcomes $Y = (10, 8, 9, 7)$. Then each Y_i is weighted by $\frac{D_i - \hat{p}(X_i)}{\hat{p}(X_i)(1-\hat{p}(X_i))}$, giving more influence to rare treatment assignments.
- **Intuition:** Replace expectations with averages. Units with unexpected treatment status get large weights, reconstructing the pseudo-population for causal inference.

Other Matching Methods with the Score

Exact matches on $p(X_i)$ are almost impossible, so algorithms are used:

1. **Nearest-neighbor matching** Match each treated unit with the closest control in terms of $p(X_i)$.

2. **Radius matching** Match treated and control units if their $p(X_i)$ differ by less than a fixed radius r.

3. **Kernel matching** Use weighted averages of controls, with weights decreasing as distance in $p(X_i)$ grows.

4. **Stratification matching** Divide the sample into intervals (strata) of $p(X_i)$ and compare treated vs. control within each stratum.

- **Example:** Suppose $p(X) = 0{,}62$ for a treated unit. - Nearest neighbor: pick the control with $p(X) = 0{,}60$. - Radius: accept all controls with $|p(X) - 0{,}62| < 0{,}05$. - Kernel: weight controls by distance, e.g. weight $\exp[-(0{,}62 - 0{,}60)^2]$. - Stratification: put both into the $[0{,}6, 0{,}7]$ block and compare means.
- **Intuition:** Instead of requiring exact matches, use approximate methods (closest neighbor, tolerance bands, weighted averages, or bins) to ensure treated and controls are comparable.

Nearest Neighbor and Radius Matching

Notation:

- T: treated units, C: controls.
- Y_i^T: treated outcome, Y_j^C: control outcome.
- $C(i)$: set of controls matched to treated unit i with score p_i.

Nearest Neighbor:

$$C(i) = \arg\min_{j \in C} \| p_i - p_j \|$$

- Match each treated unit i with the control j whose score p_j is closest.

Radius Matching:

$$C(i) = \{ j \in C : \| p_i - p_j \| < r \}$$

- Match treated unit i with all controls j whose scores fall within a tolerance band of radius r.
- **Example:** Suppose $p_i = 0{,}62$ for a treated unit. - Nearest neighbor: if controls have scores $(0{,}58, 0{,}67, 0{,}40)$, the match is 0,58. - Radius ($r = 0{,}05$): accept all controls with scores in $[0{,}57, 0{,}67]$, so matches are $(0{,}58, 0{,}67)$.
- **Intuition:** Nearest neighbor forces one best match; radius allows multiple acceptable matches within a tolerance window.

TOT Estimator by Matching (NN/Radius)

Estimator:

$$\hat{\delta}^{M}_{TOT} = \frac{1}{N^T} \sum_{i \in T} \left(Y_i^T - \sum_{j \in C(i)} w_{ij} Y_j^C \right),$$

where $M \in \{\text{NN}, \text{Radius}\}$.

Definitions:

- T: set of treated units, C: set of controls.
- $N^T = \#\{i \in T\}$: number of treated units.
- $C(i)$: set of control matches for treated unit i (depends on method M).
- $N_i^C = \#\{j \in C(i)\}$: number of controls matched to unit i.
- $w_{ij} = 1/N_i^C$ if $j \in C(i)$, and 0 otherwise (equal weights).

Algebraic view.

$$\hat{\delta}^{M}_{TOT} = \frac{1}{N^T} \sum_{i \in T} \left(Y_i^T - \frac{1}{N_i^C} \sum_{j \in C(i)} Y_j^C \right) \qquad \textbf{Plug in } w_{ij} \qquad (52)$$

$$= \frac{1}{N^T} \sum_{i \in T} \left(Y_i^T - \bar{Y}^C_{C(i)} \right) \qquad \textbf{Control mean for } i \qquad (53)$$

So the TOT estimator is the average across treated units of the outcome difference between each treated unit and the mean of its matched controls.

- **Example:** Suppose 2 treated units with $Y^T = (10, 12)$. - Matches: $C(1) = (8, 9)$, $C(2) = (11)$. - Control means: $\bar{Y}_{C(1)} = 8{,}5$, $\bar{Y}_{C(2)} = 11$. - Differences: $(10 - 8{,}5) = 1{,}5$, $(12 - 11) = 1$. - Average: $\hat{\delta}_{TOT} = 1{,}25$.
- **Intuition:** For each treated unit, build a synthetic control outcome from its matches. Then average differences across all treated to recover TOT.

Kernel Matching (TOT Estimator)

Estimator:

$$\hat{\delta}^{K}_{TOT} = \frac{1}{N^T} \sum_{i \in T} \left(Y_i^T - \frac{\sum_{j \in C} Y_j^C \, G\left(\frac{p_j - p_i}{h_n}\right)}{\sum_{k \in C} G\left(\frac{p_k - p_i}{h_n}\right)} \right).$$

Definitions:

- N^T: number of treated units.

- p_i: propensity score for treated unit i.
- p_j: propensity score for control unit j.
- $G(\cdot)$: kernel function (e.g. Gaussian, Epanechnikov).
- h_n: bandwidth parameter controlling smoothness.

Algebraic view.

$$\hat{\delta}^K_{TOT} = \frac{1}{N^T} \sum_{i \in T} \left(Y_i^T - \sum_{j \in C} w_{ij} Y_j^C \right) \qquad \textbf{Rewrite with weights} \qquad (54)$$

$$w_{ij} = \frac{G\left(\frac{p_j - p_i}{h_n}\right)}{\sum_{k \in C} G\left(\frac{p_k - p_i}{h_n}\right)} \qquad \textbf{Normalized kernel weights} \qquad (55)$$

So each treated unit i is compared to a weighted average of controls, where weights are higher for controls with p_j closer to p_i.

- **Example:** Suppose treated $p_i = 0{,}60$, controls $p_j = (0{,}58, 0{,}80)$. - With Gaussian kernel $G(z) = e^{-z^2}$ and $h_n = 0{,}1$: $G((0{,}58 - 0{,}60)/0{,}1) = e^{-0{,}2^2} = 0{,}9608$, $G((0{,}80 - 0{,}60)/0{,}1) = e^{-2^2} = e^{-4} \approx 0{,}0183$. - Normalized weights: $w_{i1} = 0{,}981$, $w_{i2} = 0{,}019$. - The matched control outcome is essentially a weighted average dominated by $p_j = 0{,}58$.
- **Intuition:** Instead of one "nearest" match, all controls contribute, but closer scores get much higher weight. Bandwidth h_n controls how local the comparison is.

Stratification Matching

Idea: Based on the same blocks used to test balance of the score.
Within block q:

$$\hat{\delta}^S_q = \frac{1}{N^T_q} \sum_{i \in T(q)} Y_i^T - \frac{1}{N^C_q} \sum_{i \in C(q)} Y_i^C,$$

where:

- $T(q)$: set of treated in block q,
- $C(q)$: set of controls in block q,
- $N^T_q = \#\{i \in T(q)\}$,
- $N^C_q = \#\{i \in C(q)\}$.

Aggregation across blocks:

$$\hat{\delta}^S_{TOT} = \sum_{q=1}^{Q} \hat{\delta}^S_q \cdot \frac{\sum_{i \in T(q)} D_i}{\sum_i D_i}.$$

This weights each block's effect by the share of treated units in that block.
Algebraic view.

$$\hat{\delta}^S_{TOT} = \sum_{q=1}^{Q} \left(\frac{1}{N^T_q} \sum_{i \in T(q)} Y^T_i - \frac{1}{N^C_q} \sum_{i \in C(q)} Y^C_i \right) \cdot \frac{N^T_q}{N^T} \qquad \textbf{Substitute weights} \qquad (56)$$

$$= \frac{1}{N^T} \sum_{q=1}^{Q} \left(\sum_{i \in T(q)} Y^T_i - \frac{N^T_q}{N^C_q} \sum_{i \in C(q)} Y^C_i \right) \qquad \textbf{Rearrange sums} \qquad (57)$$

So stratification matching is a weighted average of block-level treated vs. control differences.

- **Example:** Suppose two blocks: - Block 1: $N^T_1 = 2$, $N^C_1 = 3$, $\bar{Y}_T = 12$, $\bar{Y}_C = 10$. $\hat{\delta}_1 = 2$. - Block 2: $N^T_2 = 1$, $N^C_2 = 2$, $\bar{Y}_T = 15$, $\bar{Y}_C = 14$. $\hat{\delta}_2 = 1$. - Weights: Block 1 has 2/3 of treated, Block 2 has 1/3. - Aggregate: $\hat{\delta}_{TOT} = 2 \cdot (2/3) + 1 \cdot (1/3) = 1{,}67$.
- **Intuition:** Estimate treatment effects block by block, then average them giving more weight to blocks with more treated units.

Diagnosing Common Support

- Summarize the score for $D = 1$ and $D = 0$, and count how many units fall outside the overlap region.
- Example: if treated scores lie in $[0{,}10, 0{,}85]$, then trim controls with scores below 0,10 or above 0,85.
- Histograms or density plots by group ($D = 1$ vs. $D = 0$) help visualize overlap.

Formal condition:

$$\min p(X_i \mid D = 1) \quad \leq \quad p(X_i) \quad \leq \quad \max p(X_i \mid D = 1),$$

ensuring controls exist in the same score range as treated units.

- **Tiny Math Example:** Suppose treated scores range from 0,20 to 0,75. A control with $p(X) = 0{,}90$ is trimmed (outside support), while one with $p(X) = 0{,}40$ is kept (inside overlap).

- **Intuition:** Common support means both groups must be comparable. Units without overlap cannot be matched and should be dropped.