

Selection on Observables II: Estimation

Stat 256, Spring 2026

Chad Hazlett

UCLA

`chazlett@ucla.edu`

Where we left off

Conditional ignorability (CI):

$$\{Y_{1i}, Y_{0i}\} \perp\!\!\!\perp D_i \mid X_i$$

Equivalently (backdoor criterion on the DAG): X blocks every backdoor path from D to Y .

Under CI (+ common support), the ATE is identified:

$$\tau_{\text{ATE}} = \mathbb{E}_X[\mathbb{E}[Y \mid D=1, X] - \mathbb{E}[Y \mid D=0, X]] = \mathbb{E}_X[\mu_1(X) - \mu_0(X)].$$

Today: *estimation*. Given a chosen adjustment set X , how do we actually compute $\hat{\tau}$?

One target, many roads

Every estimator we meet today is an approximation of the same object:

$$\hat{\tau} \approx \mathbb{E}_{X \sim F} [\mu_1(X) - \mu_0(X)].$$

where $\mu_d(X) = \mathbb{E}[Y|X, D = d]$

F controls the estimand (e.g. $F = p(X)$ for ATE; $F = p(X | D=1)$ for ATT).

Menu of approximations.

- Subclassification (recap)
- Matching
- Propensity-score methods (matching + IPW)
- Balancing weights (skip π ; balance directly)
- Regression via imputation (Lin; T-learner)
- Augmented / doubly robust (brief)

These differ only in *how* they approximate conditioning on X .

Running example: LaLonde / Dehejia–Wahba

The question. Did the National Supported Work (NSW) job-training program raise 1978 earnings?

The data.

- 185 NSW treated men (experimental).
- 2,490 observational controls from PSID (survey of the general population).
- Covariates X : age, education, race, marriage, 1974/75 earnings, 1974/75 unemployment.

Experimental benchmark. On the *randomised* NSW sample (not shown here), $\hat{\tau} \approx \$1,794$.

Observational challenge. DIM on observational sample = $-\$15,205$. Eek! The PSID controls are nothing like NSW-treated: older, richer, more employed.

Recap: subclassification

Partition X into strata $x_1, x_2, \dots, x_K$ (coarse bins).
Within each stratum, CI $\Rightarrow$ mini-experiment.

$$\hat{\tau}_{\text{sub}} = \sum_{k=1}^K \hat{\Pr}(X = x_k) \cdot [\bar{Y}_1(x_k) - \bar{Y}_0(x_k)]$$

The problem. With continuous or many-dimensional X , strata become empty as K grows — stratification alone runs out of units.

Matching handles this by finding *near*-matches instead of exact ones.

A crude try on LaLonde: stratify on u_{74}, u_{75}

Stratify NSW+PSID by $(u_{74}, u_{75}) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$: four cells by whether the worker was unemployed in 1974, 1975.

Within each cell, compute DIM; re-weight cells by the *treated* share (ATT):

$$\hat{\tau}_{\text{sub}} = \sum_{k=1}^4 \frac{n_{1k}}{n_1} [\bar{Y}_1(k) - \bar{Y}_0(k)] \approx -\$4,286.$$

This actually cuts the bias by a factor of 4, but it's not enough.

Why: Turns out the unemployed (in 1975) people in PSID and NSW are fairly comparable, and within just them you get reasonable estimates, but the employed in PSID and NSW are pretty different.

Matching: a flexible stratification

Matching = fine-grained subclassification: for each unit, find a “stratum of one” made of its closest neighbour(s) in the other arm.

ATT estimator (one-to-one NN):

$$\hat{\tau}_{\text{ATT}} = \frac{1}{n_1} \sum_{i: D_i=1} [Y_i - Y_{j(i)}]$$

where $j(i) = \arg \min_{j: D_j=0} d(X_i, X_j)$.

Choices:

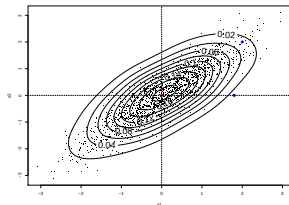
- **Distance metric** $d(\cdot, \cdot)$ — *how do we measure “close”?*
- **Matching routine** — *how do we assemble the matched sets?*

Distance metrics

Euclidean: $d(X_i, X_j) = \|X_i - X_j\|_2$. Scale-sensitive.

Standardised Euclidean: divide each covariate by its SD. Fixes scale, ignores correlation.

Mahalanobis: $d(X_i, X_j) = \sqrt{(X_i - X_j)' \hat{\Sigma}_X^{-1} (X_i - X_j)}$. Scale-invariant + corrects for correlation.



Others: coarsened exact matching, genetic matching, ...

Matching routines

1-to-1 vs 1-to-M. Match each treated to one control, or to M nearest (then average). Larger M : lower variance, higher bias.

With vs without replacement.

- *With*: a control may match many treated. Lower bias, higher variance.
- *Without*: each control used once, can force bad matches when n_0 small.

Greedy vs optimal.

- *Greedy*: step through treated units, grab nearest available control.
- *Optimal*: minimise *total* within-pair distance across the whole match.

Full / complete matching (Hansen 2004): every treated matched to ≥ 1 control *and* vice versa; variable-size sets. Uses all data.

Limits of matching (I): discrepancy

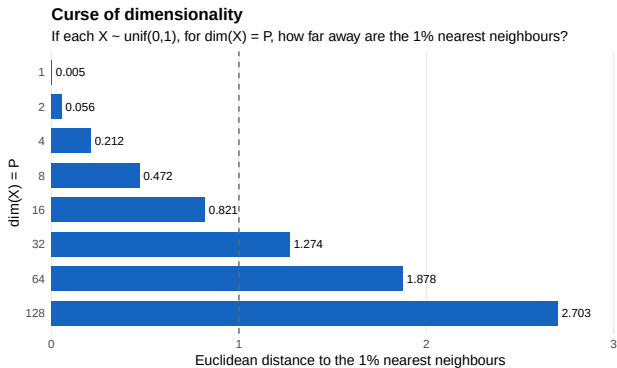
Matched pairs are never exactly equal in X (except under exact match). Discrepancy $\|X_i - X_{j(i)}\|$ leaves residual bias on each pair.

Abadie & Imbens (2006): matching bias does not shrink at $\sqrt{n}$ once $\dim(X) \geq 2$ with continuous X . Matching alone is not $\sqrt{n}$ -consistent.

Bias-adjusted matching: fit $\hat{\mu}_0(X)$ among controls, replace $Y_{j(i)}$ with $Y_{j(i)} - [\hat{\mu}_0(X_{j(i)}) - \hat{\mu}_0(X_i)]$. A “little regression” fills the gap.

Limits of matching (II): curse of dimensionality

As $\dim(X)$ grows, everyone is far away.



Broader point: Matching is appealingly nonparametric, but you pay a price in bias, especially as P grows.

Propensity score

Define $\pi(X) := \Pr(D = 1 \mid X)$.

Rosenbaum & Rubin (1983). If $\{Y_1, Y_0\} \perp\!\!\!\perp D \mid X$, then $\{Y_1, Y_0\} \perp\!\!\!\perp D \mid \pi(X)$.

Why it matters: a scalar summary of X is enough to achieve CI.

Two families of use:

- **PS matching / stratification**: match or stratify on $\hat{\pi}(X)$.
- **IPW**: reweight by $1/\hat{\pi}(X)$ (next section).

Costs:

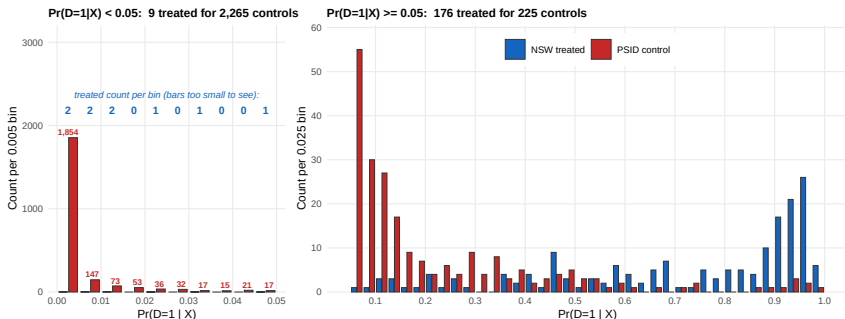
- You have to specify a model for $\pi(X)$, if wrong, $\hat{\pi}$ inconsistent.
- You don't get “exact balance” on X in any sense.

PS matching:

First idea: match treated i to control j minimizing $|\hat{\pi}(X_i) - \hat{\pi}(X_j)|$.

Let's see what $\hat{\pi}(X)$ looks like in LaLonde (logit model)

P-score overlap: NSW treated vs PSID controls



Here the smart play is: only ask about ATT.

If two-sided poor overlap, you wouldn't be able to do anything.

IPW: reweighting to the target distribution

Matching suffers discrepancies and throws away observations— just weight!

- Treated arm has covariate density $p(X \mid D=1)$; we want $p(X)$.
- Multiply each treated unit by $p(X)/p(X \mid D=1)$ to match $p(X)$.
- But $p(X \mid D=1)$ is poorly estimated, so rewrite
 $p(X)/p(X \mid D=1) = p(D=1)/p(D=1 \mid X) = p(D=1)/\pi(X)$.
- Likewise for controls, weight by $p(D=0)/(1 - \pi(X))$.

These are the (**stabilised**) IPW weights:

$$w_i = \frac{D \Pr(D=1)}{\hat{\pi}(X_i)} + \frac{(1-D) \Pr(D=0)}{1 - \hat{\pi}(X_i)}$$

Use them to weight the DIM for the ATE:

$$\hat{\tau}_{\text{IPW}} = \frac{\sum_i w_i D_i Y_i}{\sum_i w_i D_i} - \frac{\sum_i w_i (1 - D_i) Y_i}{\sum_i w_i (1 - D_i)}.$$

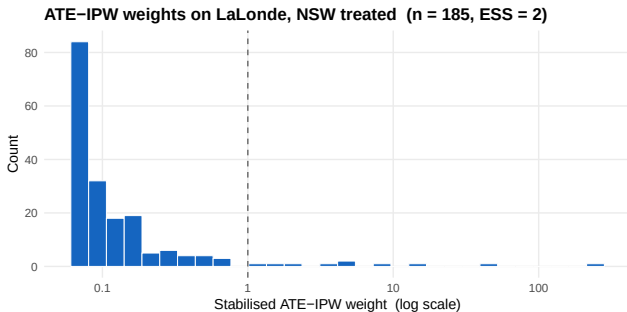
ATT version: treated $w_i = 1$ (already at target); controls

$$w_i = \frac{\hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} \cdot \frac{\Pr(D=0)}{\Pr(D=1)} \text{ (Bayes for } p(X \mid D=1)/p(X \mid D=0)).$$

Worry about the weights

Weights blow up when $\hat{\pi} \rightarrow 0$ (treated) or $\rightarrow 1$ (controls).

The few treated units that look PSID-like (low $\hat{\pi}$) get astronomical weights:



- One treated unit gets a weight over 240, carrying 70% of the treated weight;
- For treated, $ESS = (\sum w_i)^2 / \sum w_i^2 \approx 2!$ (from 185).
- For controls, many get downweighted but $ESS \approx 711$ (from 2490). ATT fine.

Balancing weights: solve directly

If you want $p(X \mid D = 1) \approx p(X \mid D = 0)$, why not just choose weights $\{w_i\}$ s.t. weighted control distribution matches treated (ATT) or target (ATE) on (chosen moments of) X ?

Mean-balance constraint (in ATT case):

$$\sum_{i: D_i=1} \phi(X_i) = \sum_{i: D_i=0} w_i \phi(X_i), \quad \text{s.t.} \quad \sum w_i = n_1, \quad w_i \geq 0 \quad \forall i$$

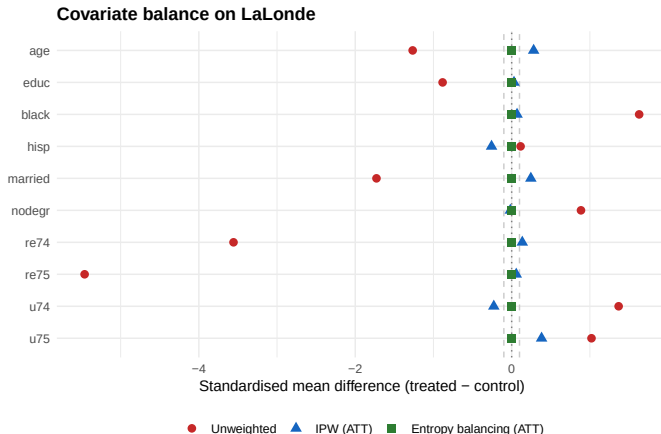
for some map $\phi(X)$ (e.g. $\phi(X) = X$ for mean balance; $\phi(X) = K(X, \cdot)$ for kernel balance).

Constraint alone is under-determined. Add a regulariser on w such as:

- Maximum entropy: $\min \sum_i w_i \log w_i \rightarrow$ entropy balancing (Hainmueller 2012)
- L_2 penalty: $\min \sum_i w_i^2 \rightarrow$ stable balancing weights (Zubizarreta 2015)
- Empirical likelihood, $\sum_i \log(w_i)$

Balance on LaLonde: unweighted vs IPW vs ebal

Standardised mean differences across covariates: unweighted is far off; IPW shrinks the gap but *doesn't* hit zero (and can still miss); entropy balancing zeroes the means by construction.



CBPS (Imai & Ratkovic 2014): a pscore tuned for balance

What if we still want a pscore (interpretable, parametric, easy to inspect) but tune it so the resulting IPW gives exact finite-sample balance? Stack the moment conditions per unit:

$$\phi(X_i, D_i; \beta) = \begin{pmatrix} [D_i - \pi(X_i; \beta)] X_i \\ \left[\frac{D_i}{\pi(X_i; \beta)} - \frac{1 - D_i}{1 - \pi(X_i; \beta)} \right] X_i \end{pmatrix}.$$

Top: standard logistic MLE score (what logistic regression solves). Bottom: IPW balance moment — forces IPW-reweighted means of X to match across arms.

Choose β to set the moments to zero:

- **M-estimator** (just-identified): keep k of the $2k$ moments — e.g. balance only — and solve $\sum_i \phi^{(j)}(X_i; \beta) = 0$.
- **GMM** (over-identified, both rows): $\hat{\beta} = \arg \min_{\beta} \left[\frac{1}{n} \sum_i \phi \right]' W \left[\frac{1}{n} \sum_i \phi \right]$.

Beyond means; prognostic balancing

What does mean balance buy? Choosing $\phi(X) = X$ balances X 's means across arms — which translates into balance on *linear functions* of X .

- If $\mu_0(X)$ is linear in X , that's exactly what you need: $\mathbb{E}[\mu_0(X)]$ matches across arms and $\hat{\tau}$ is unbiased.
- If $\mu_0(X)$ has nonlinearities or interactions, mean-balanced X does not imply balanced $\mu_0(X)$. Residual bias remains.

Prognostic balancing (for ATT): what we really want is $\mathbb{E}[\mu_0(X)]$ equal across arms. Balance directly on $\hat{\mu}_0(X)$, the prognostic score.

- Fit $\hat{\mu}_0(X)$ on controls; project onto all units.
- Solve for control weights so $\sum_{i:D=0} w_i \hat{\mu}_0(X_i) = \sum_{i:D=1} \hat{\mu}_0(X_i)$.
- $\hat{\mu}_0$ can be OLS, ML, anything. Flexibility propagates into the weights.

Efficient; handy especially when poor overlap on (non-prognostic) X . Work in progress.

G-computation / imputation

Instead of weighting, *model* $\mu_1(x)$ and $\mu_0(x)$. Then impute each unit's missing potential outcome:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n [\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)].$$

Fit $\hat{\mu}_1$ on treated only, $\hat{\mu}_0$ on controls only (**T-learner**).

For ATE, average over all X_i ; for ATT, over treated X_i ; for ATC, over controls.

Functional form of $\hat{\mu}_d$ is your choice: OLS, random forest, kernel regression,

...

Lin's estimator (interacted OLS)

Lin (2013). Fit the fully interacted model

$$Y = \alpha + \tau D + \beta' X + \gamma'(X \cdot D) + \varepsilon.$$

Equivalent to fitting two separate OLS models ($\hat{\mu}_1$ on treated, $\hat{\mu}_0$ on controls) and reading off $\hat{\mu}_d(x) = \hat{\alpha}_d + \hat{\beta}'_d x$.

Two things to want from this regression.

- The coefficient on D should be the *average* treatment effect, not “effect at $X = 0$.”
- The coefficient should equal the g-computation imputation estimate $\frac{1}{n} \sum_i [\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]$.

We get both if we center X first.

Centering X : why and what it buys

Let $\tilde{X}_i = X_i - \bar{X}$ (sample-mean-centred). The interacted model becomes

$$Y = \alpha + \tau D + \beta' \tilde{X} + \gamma'(\tilde{X} \cdot D) + \varepsilon.$$

Unit-level treatment effect at $X = x$:

$$\partial Y / \partial D = \tau + \gamma' \tilde{X} = \tau + \gamma'(x - \bar{X}).$$

(i) Average effect.

$$\frac{1}{n} \sum_{i=1}^n \partial Y_i / \partial D = \tau + \gamma' \underbrace{\frac{1}{n} \sum_i (X_i - \bar{X})}_{=0} = \tau.$$

And the (linear) imputation estimator does just that:

Fitted means:

$$\hat{\mu}_1(\mathbf{x}) = \hat{\alpha} + \hat{\tau} + \hat{\beta}'(\mathbf{x} - \bar{\mathbf{X}}) + \hat{\gamma}'(\mathbf{x} - \bar{\mathbf{X}})$$

$$\hat{\mu}_0(\mathbf{x}) = \hat{\alpha} + \hat{\beta}'(\mathbf{x} - \bar{\mathbf{X}})$$

Imputation (g-comp) estimate:

$$\frac{1}{n} \sum_{i=1}^n [\hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)] = \hat{\tau} + \hat{\gamma}' \underbrace{\frac{1}{n} \sum_i (\mathbf{X}_i - \bar{\mathbf{X}})}_{=0} = \hat{\tau}.$$

So with centered $\mathbf{X}$, three things collapse to one number:

- the coefficient $\hat{\tau}$ on D ,
- the average unit-level slope,
- the g-computation / T-learner imputation estimate.

All roads lead to the same place

Under CI, every method is estimating $\mathbb{E}_{X \sim F}[\mu_1(X) - \mu_0(X)]$.

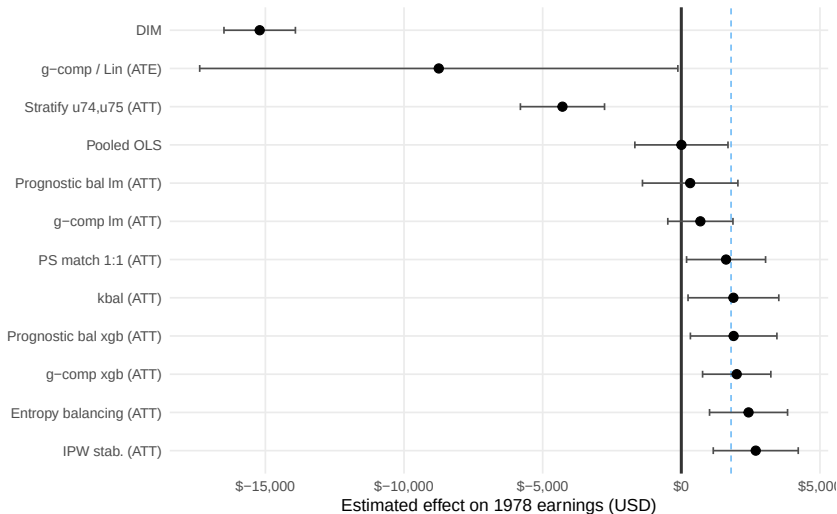
Method	How it approximates conditioning on X
Subclassification	coarse strata of X
Matching	nearest-neighbour “strata of one” in X
PS matching	match on scalar $\pi(X)$
IPW	reweight by $1/\pi(X)$
CBPS	estimate π s.t. balance constraint holds
Balancing weights	solve for weights balancing X (or $\phi(X)$)
Kernel balancing (kbal)	balance on a basis expansion of X
Prognostic balancing	balance on $\hat{\mu}_0(X)$
Regression imputation	model $\mu_d(X)$ (Lin); impute

One angle: Choose the appropriate set of these for your problem, and implement them all. If they disagree, diagnose.

LaLonde: tough one due to overlap

LaLonde / Dehejia-Wahba: observational estimates

NSW treated vs PSID controls. Experimental benchmark ~ \$1794 (light blue dashed).



Lessons from LaLonde

ATE on LaLonde is a trap. Two faces of the same issue:

- g-comp / Lin (ATE) lands at $-\$8,746$: $\hat{\mu}_1(X)$ must be extrapolated to 2,490 PSID-like X 's where no treated unit lives. Huge variance ensues.
- IPW (ATE) fails because the rare *treated* units that look PSID-like get huge weights — effective treated N collapses from 185 to ≈ 2 .

Retreat to ATT. The treated do have controls around them, so ATT is fine.

- But if you need to learn $\mu_0(X)$, learn it in a way that can specialise local structure since many controls are far from treated.

The ATT cluster near \$1,794. PS match, IPW, entropy balancing, kbal, g-comp xgb, and prognostic balancing with xgb all land close to the benchmark. When covariate shift is severe, reweighting or locally-adaptive imputation works well. **General rule: choose the estimand and the family of estimators for your situation.**

Augmented / doubly robust: a signpost

Methods that use **both** $\hat{\pi}$ and $\hat{\mu}_d$:

- **AIPW**: IPW estimator plus a regression adjustment (equivalently, g-comp plus IPW-weighted residuals). Consistent if *either* $\hat{\pi}$ or $\hat{\mu}_d$ is right.
- **TMLE**: iteratively tilt $\hat{\mu}_d$ using $\hat{\pi}$ to solve an efficient score equation.
- **DML** (Chernozhukov et al. 2018): put flexible ML $\hat{\pi}, \hat{\mu}_d$ together with *cross-fitting*. Delivers $\sqrt{n}$ -valid inference despite ML bias.

These are candidates for the Advanced Topic Presentation: “efficiency and flexibility in covariate adjustment.”

If nobody picks it, I might carve out a lecture.

Takeaways

One target. Under CI, everything estimates $\mathbb{E}_{X \sim F}[\mu_1(X) - \mu_0(X)]$. Choose F for the estimand; choose a method for how to approximate conditioning on X .

Method menu. Subclass / matching / PS-matching / IPW / balancing weights / regression imputation (Lin). Pick based on what you're willing to model.

Lin: interacted OLS on *centered* X gives a D -coefficient equal to the g-computation estimate. Center at the right mean for your estimand.

Extrapolation is not reweighting. On LaLonde, OLS-ATE extrapolates $\hat{\mu}_1$ far outside the treated support and lands at $-\$8,746$; the ATT version is $+\$688$; weighting methods cluster near the truth. *Pick estimand and method with an eye on support.*

What's still missing. Unobserved confounding. When CI fails, every method on this menu is biased.