Lab 5: Matching and Weighting

Dias Akhmetbekov

2024-03-01

Motivation

- Causal inference is all about comparing **counterfactuals**, like the ATT:

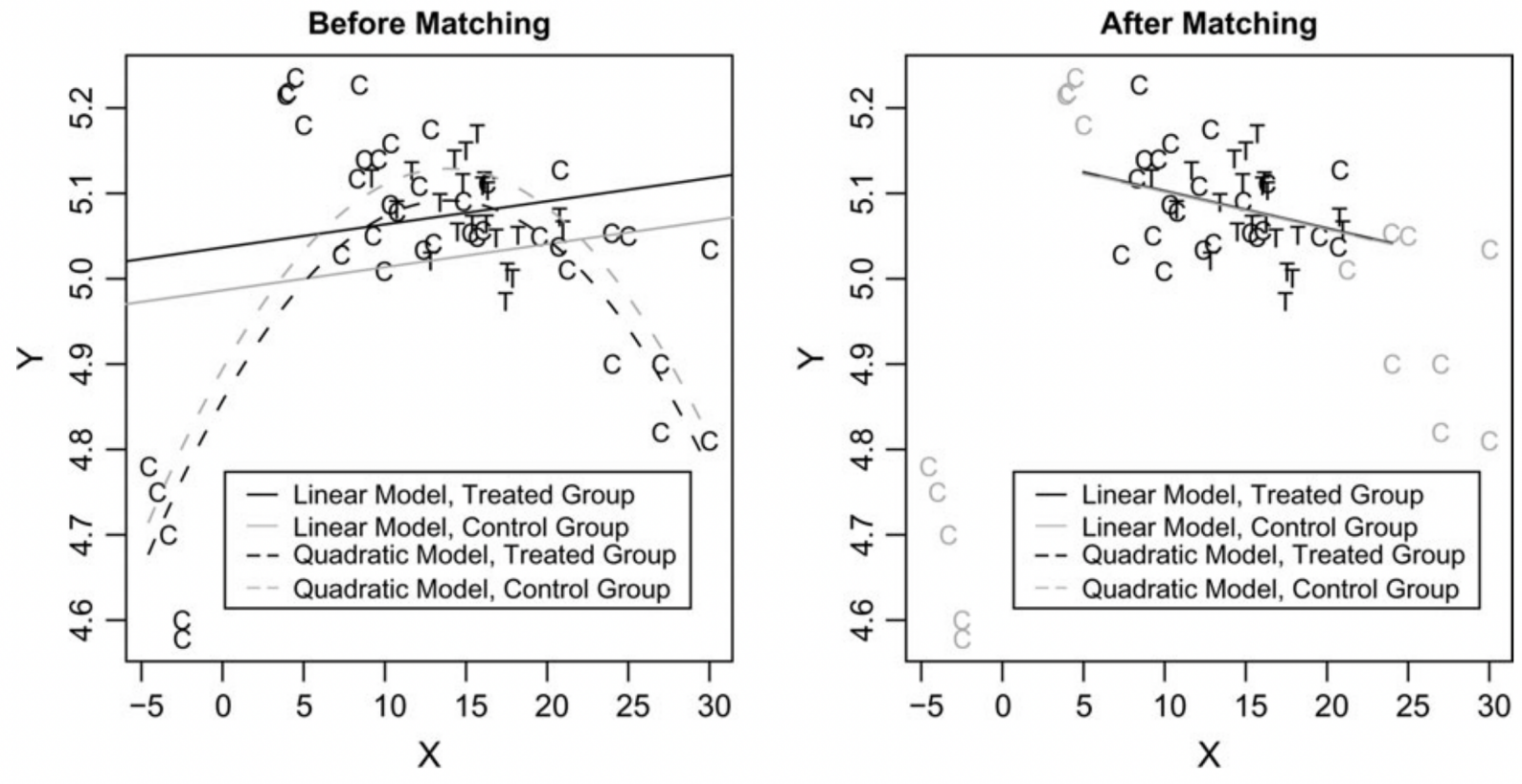
$$\tau_{ATT} = \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]$$

- Recall the **imputation estimators** with regression.

$$\hat{\tau}_{reg} = \frac{1}{n_1} \sum_{i=1}^{n} D_i(Y_i - \hat{\mu}_0(X_i))$$

  - Common solution: use a parametric model for $\hat{\mu}_0(X_i)$
  - For example, could assume it is linear: $\mu_0(x) = x'\beta$
  - Regression, MLE, Bayes, etc.
  - But this model might be wrong $\Longrightarrow$ wrong causal estimates.

Model Dependence (Ho et al. 2007, Pol.Analysis)



Why matching?

- **Matching** is a nonparametric imputation estimator:

$$\hat{\tau}_m = \frac{1}{n_1} \sum_{i=1}^{n} D_i \left(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_m(i)} Y_j \right)$$

  - $\mathcal{J}(i)$ are the set of M closest control units to i in terms of X_i.
- Matching has strong advantages:
  1. Reduces dependence of estimates on parametric models.
  2. Reduces model-based extrapolation.
  3. Makes counterfactual comparisons more transparent.
- What matching isn't: a solution for selection on unobservables.
  - Matching is an **estimation** technique, not an identification strategy.

Types of matching

- Assumptions:
  - No unmeasured confounders: $D_i \perp\!\!\!\perp (Y_i(0), Y_i(1))|X_i$
  - Overlap/positivity: $0 < P(D_i = 1|X_i = x) < 1$
- Exact Matching
  - Choose matches that have the same value of X_i
  - $\mathcal{J}_m(i)$ is a random set of M control units with $X_j = X_i$
  - Covariate distribution in treated and matched controls exactly the same:

$$\hat{P}(X_i = x|D_i = 1) = \hat{P}(X_i = x|D_i = 0, j \text{ is matched})$$

$$\implies \mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i|D_i = 0, j \text{ is matched}]$$

  - Problem: not feasible with high-dimensional or continuous X_i
- Coarsened Exact Matching (Iacus et al, 2011)
  - Discretize and group covariates into substantively meaningful bins
  - Exact match on these bins accounts for interactions
  - Have to drop treated units in bins with no controls changes estimand.
  - Allows you to control bias/variance tradeoff through coarsening.

Matching in High Dimensions

- High Dimensional X_i and Distance Metrics
  - Even CEM can break down with high dimensional X_i.
  - We can define closeness using lower dimensional **distance metrics**:
  - Reduces dimensionality: maps two vectors to a single number.
- Mahalanobis distance:

$$D(X_i, X_j) = \sqrt{(X_i - X_j)' \hat{\Sigma}^{-1} (X_i - X_j)}$$

  - $\hat{\Sigma}$ is the estimated variance-covariance matrix of the observations:

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'$$

- Estimated propensity score:

$$D(X_i, X_j) = |\hat{\pi}(X_i) - \hat{\pi}(X_j)| = |P(D_i = 1|X_i) - P(D_i = 1|X_j)|$$

Other matching choices

- Matching ratio: how many control units per treated?
  - Lower reduces bias (only use the closest matches)
  - Lower increases variance
- With or without replacement: same control matched to multiple treated?
  - With replacement gives better matches & matching order doesn't matter.
  - Without replacement simplifies variance estimation.
- Caliper: drop poor matches?
  - Only keep matches below a distance threshold, $D(X_i, X_j) \leq c$
  - Reduces imbalance, but if you drop treated units, estimand changes.

Assessing balance

- Goal of matching is to maximize balance: $\hat{F}_1(x) \approx \hat{F}_{0,\delta}(x)$
  - Joint distribution of X_i is similar between treated and matched controls.
  - Difficult to assess balance across many dimensions $\rightarrow$ summaries.
- Options for Assessing Balance:
  - Differences-in-means/medians, standardized.
  - QQ plots/KS statistics for comparing the entire distribution of X_i.
  - L_1: multivariate histogram (for CEM)
  - Choice of metric can change what matching method works best.
- Hypothesis tests for balance are problematic:
  - Dropping units can lower power (increase p-values) without a change in balance.

Bias of inexact matching

- To show the bias on matching, focus on finding a single control match.
- Let $j(i)$ be the matched control for unit i, the bias is:

$$\mathbb{E}[Y_i|D_i = 1, X_i, X_j] - \mathbb{E}[Y_i(0)|D_i = 1, X_i] = (\mu_0(X_i) - \mu_0(X_{j(i)}))$$

  - Bias is 0 if matching is exact since $X_i = X_{j(i)}$
  - Bias grows with **matching discrepancy/imbalance**.
- **Bias correction**: estimate $\hat{\mu}_0(x)$ with regression and estimate bias.

$$\hat{Y}_i(0) = Y_{j(i)} - (\hat{\mu}_0(X_i) - \hat{\mu}_0(X_{j(i)}))$$

  - Imputation of missing potential outcome now matching + regression.
  - Generalizes easily to any number of matches.

Variance

- Matching with Replacement
  - Can either use clustered standard errors (SEs) or cluster bootstrap.
  - Valid for post-matching regression (Abadie and Spiess, 2021).
- Matching without Replacement
  - More complicated due to the same control unit matched to multiple treated.
  - $K_m(i)$ is the number of times a unit is used as a match.
- Assuming units are well-matched so bias can be ignored:

$$V(\hat{\tau}_m) = \frac{1}{n_1}\left(\mathbb{E}\left[(\tau(X_i) - \tau_{ATT})^2 | D_i = 1\right] + V(\hat{\tau}_m | X, D)\right)$$

  - Abadie and Imbens (2006) provide matching-based variance estimators.

Why weighting?

- Downsides of Matching
  - Inefficient: it may throw away data.
  - Ineffective: crude tool so it may not be able to balance covariates.
- Matching is actually a special case of a weighting estimator:

$$\hat{\tau}_m = \frac{1}{n_1} \sum_{i=1}^{n} D_i \left(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_m(i)} Y_j \right)$$

$$= \frac{1}{n_1} \sum_{i:D_i=1} Y_i - \frac{1}{n_0} \sum_{i:D_i=0} \left(\frac{n_0 K_m(i)}{n_1 M} \right) Y_i$$

  - $K_m(i)$ is the number of times i is used as a match.
  - Weighting estimators choose the weights directly to reduce imbalance.

Estimation

- Recall Two Results
  - **Horvitz-Thompson Estimator**

$$\hat{\tau}_{HT} = n^{-1} \sum_i \pi_i^{-1} Y_i D_i - (1 - \pi_i)^{-1} Y_i (1 - D_i)$$

  - Unbiased estimator of τ_{ATE}, here $\pi_i = Pr(D_i = 1|X_i)$
- **Conditional Strong Ignorability**
  - D_i is strongly ignorable conditional on a vector X_i if:
    1. $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i|X_i$
    2. $\exists \epsilon > 0$ s.t. $\epsilon < Pr(D_i = 1|X_i) < 1 - \epsilon$
- Key: $\pi(X_i) = Pr(D_i = 1|X_i)$ is important.
  - This is the **propensity score**.

Why does the PS matter?

- Note our strong ignorability condition, $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i | X_i$ conditions on X_i, which can be quite high dimensional.
- Key result from Rosenbaum-Rubin: if the above holds, then so does $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i | \pi(X_i)$.
  - The intuition comes from the fact that conditional on $\pi(X_i)$, the distribution of X is the same for the treated and untreated, and thus X_i and D_i are independent.
- Crucially, solves a high-dimensional problem - now we just need to condition on a single scalar value (π_i).

How to match? (Aronow and Miller, 2019)

i	$Y_i(0)$	$Y_i(1)$	D_i	X_{i1}	X_{i2}	$\pi(\mathbf{X}_i)$
1	-	2	1	1	7	0.33
2	5	-	0	0	7	0.14
3	-	3	1	10	3	0.73
4	-	10	1	3	1	0.35
5	-	2	1	5	2	0.78
6	0	-	0	7	0	0.70

How to match? (Aronow and Miller, 2019)

i	$Y_i(0)$	$Y_i(1)$	D_i	X_{i1}	X_{i2}	$\pi(\mathbf{X}_i)$
1	5	2	1	1	7	0.33
2	5	2	0	0	7	0.14
3	0	3	1	10	3	0.73
4	5	10	1	3	1	0.35
5	0	2	1	5	2	0.78
6	0	3	0	7	0	0.70

How to match?

- Ideally, match on exactly the propensity score
  - Exact matches are often impossible; methods seek approximate matches. This approximation can introduce bias.
- Consider a scenario where we must choose matches based on the propensity score.
  - Closest p-value matching may be used, but it has issues:
  - It may not always select the closest unit.
  - It can create challenges for inference, especially when $\pi(X)$ is unknown.
- We need to construct $E(Y_i(1)|\pi(X))$ and $E(Y_i(0)|\pi(X))$ for each observation.
  - How do we choose now? Closest p-value can be misleading.
  - We picked unit $i = 2$ for unit 1, but why not unit 4 which is very close?
  - More generally, this approach has challenges for inference, especially with $\pi(X)$ unknown (see Abadie and Imbens (2008))

What to do instead of matching?

- Matching addresses the problem by focusing on X, given ignorability is with respect to X.
  - However, we should not lose sight of the estimand - always focus on the estimand!
- **Key result**: The population version of the Horvitz-Thompson estimator can be seen as an inverse probability weighting (IPW) estimator:

$$E(\tau_i) = E\left(\frac{Y_i D_i}{\pi(X)} - \frac{Y_i(1-D_i)}{1-\pi(X)}\right)$$

  - This is an amazing result!
  - Under discrete X, this simplifies to what we would logically do anyway.

A more stable IPW

- The IPW approach works well, but in small samples can be high variance if you get big $\pi(X)$ values.
  - We can slightly improve on it using the stabilized IPW estimator:

$$\hat{\tau}_{SIPW} = \frac{\frac{1}{n}\sum_i \frac{Y_i D_i}{\hat{\pi}(X_i)}}{\frac{1}{n}\sum_i \frac{D_i}{\hat{\pi}(X_i)}} - \frac{\frac{1}{n}\sum_i \frac{Y_i(1-D_i)}{1-\hat{\pi}(X_i)}}{\frac{1}{n}\sum_i \frac{(1-D_i)}{\hat{\pi}(X_i)}}$$

- This estimator benefits by adjusting for unusually high or low values of $\pi(X)$

True vs. Estimated p-scores

- True propensity scores are only known sometimes (e.g., randomized experiments). In most non-experimental settings, the p-score is unknown and must be estimated
- When estimating, we have two cases:
  - If X is discrete, we know that $\hat{\pi}(X)$ can be an exact approximation (why?)
  - If X is not discrete (or high-dimensional), how should we approximate it?
- We need to estimate $\pi(X)$ in a way that is flexible and will converge to the truth in the limit – e.g., semi-parametric estimation of π.
  - Note a linear model of π will inherently be wrong b/c probabilities are bounded between 0 and 1
  - Practical implication: logit estimation of $\pi(X)$ is reasonable, allowing for flexible specification of X
  - As dimension of X grows, ML / lasso style models grow in value

True vs. Estimated p-scores

- Important result: even if you know the true function $\pi(X)$, better to use the estimated function than the truth (Imbens, Hirano and Ridder (2002))
  - Intuition: the deviations from the "true" propensity score $\hat{\pi}(X) - \pi(X)$ are informative for the estimation of the treatment effects (a la extra moment restrictions in GMM)
- Clear tension – as dimension of controls increases, the noisiness in π grows as well

True vs. Estimated p-scores

```
set.seed(123)
ht.est <- function(y, d, w) {
  n <- length(y)
  (1/n) * sum((y * d * w) - (y * (1-d) * w))
}
n <- 200
x <- rbinom(n, size = 1, prob = 0.5)
dprobs <- 0.5*x + 0.4*(1-x)
d <- rbinom(n, size = 1, prob = dprobs)
y <- 5 * d - 10 * x + rnorm(n, sd = 5)
true.w <- ifelse(d == 1, 1/dprobs, 1/(1-dprobs))
pprobs <- predict(glm(d ~ x))
est.w <- ifelse(d == 1, 1/pprobs, 1/(1 - pprobs))
ht.est(y, d, est.w)
```

```
## [1] 5.029735
```

```
ht.est(y, d, true.w)
```

```
## [1] 5.740815
```

True vs. Estimated p-scores

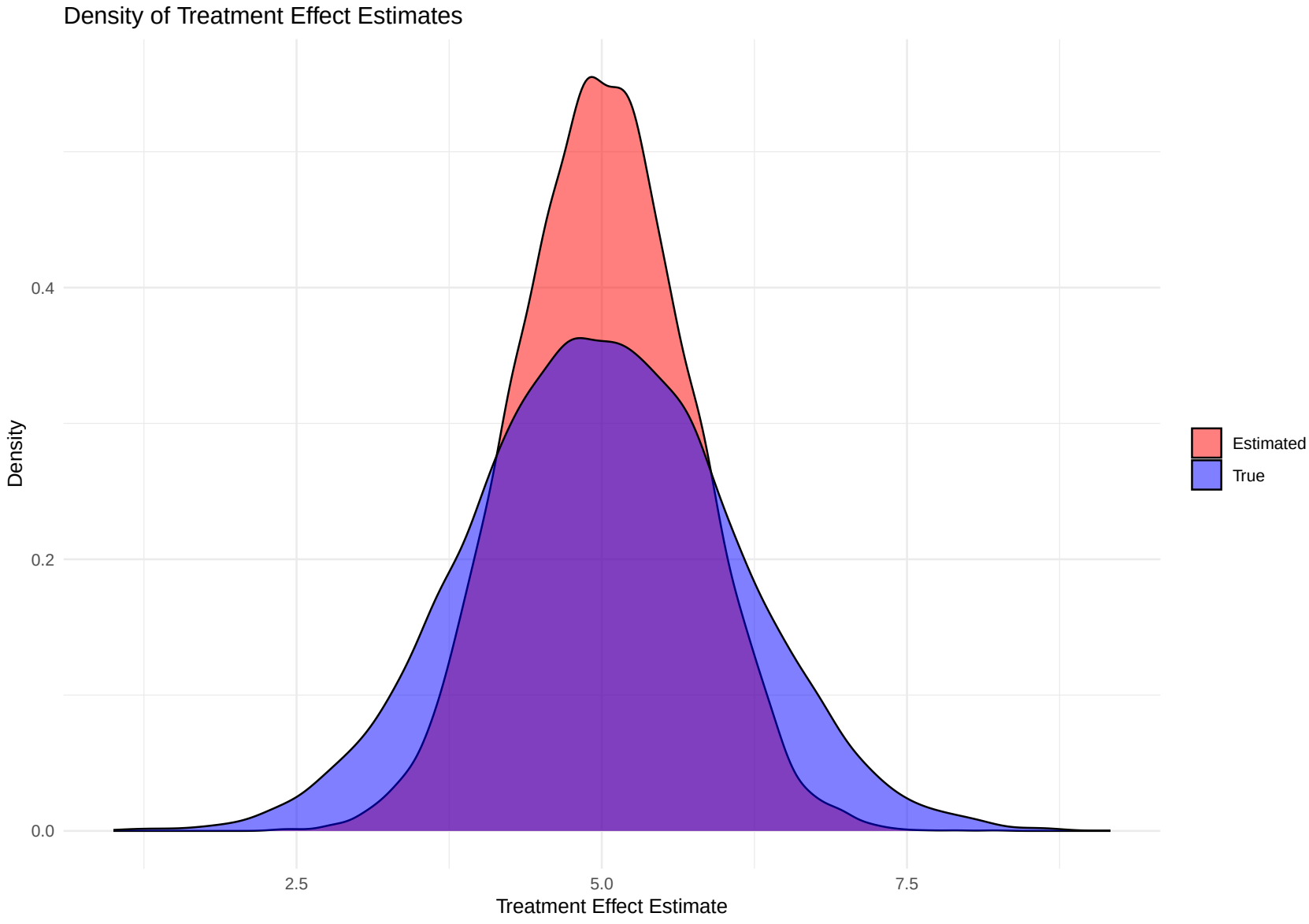
```
sims <- 10000
true.holder <- rep(NA, sims)
est.holder <- rep(NA, sims)
for (i in 1:sims) {
  x <- rbinom(n, size = 1, prob = 0.5)
  dprobs <- 0.5*x + 0.4*(1-x)
  d <- rbinom(n, size = 1, prob = dprobs)
  y <- 5 * d - 10 * x + rnorm(n, sd = 5)
  true.w <- ifelse(d == 1, 1/dprobs, 1/(1-dprobs))
  pprobs <- predict(glm(d ~ x))
  est.w <- ifelse(d == 1, 1/pprobs, 1/(1 - pprobs))
  est.holder[i] <- ht.est(y, d, est.w)
  true.holder[i] <- ht.est(y, d, true.w)
}
var(est.holder)
```

```
## [1] 0.5062535
```

```
var(true.holder)
```

```
## [1] 1.147964
```

True vs. Estimated p-scores



So??

- Why is the estimated propensity score more efficient than the true PS?
- Removing chance variations using $\hat{\pi}(X_i)$ adjusts for any small imbalances that arise because of a finite sample.
- True PS only adjusts for the expected differences between samples.
- Only true if propensity score model is correctly specified!!