Matching and Weighting Methods

Kosuke Imai

Harvard University

Spring 2021

Motivation

- Causal inference $\rightsquigarrow$ inference for counterfactuals
- Comparison between treated and control units
- Consider the Average Treatment Effect for the Treated (ATT):

$$\tau_{\text{ATT}} = \mathbb{E}(Y_i(1) - Y_i(0) \mid T_i = 1)$$

- Regression $\rightsquigarrow$ model-based imputation:

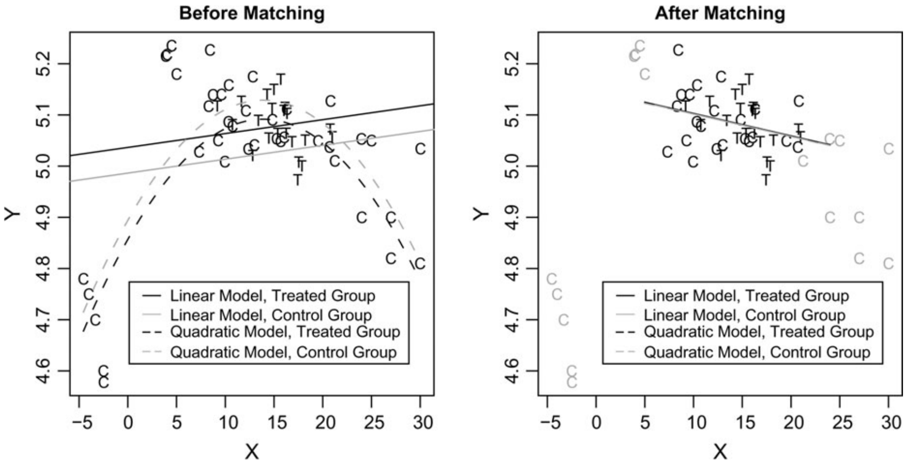
$$\hat{\tau}_{\text{reg}} = \frac{1}{n_1} \sum_{i=1}^{n} T_i \left(Y_i - \hat{\mu}_0(\mathbf{X}_i) \right)$$

- Regression can be model-dependent
- Matching $\rightsquigarrow$ nonparametric imputation:

$$\hat{\tau}_{\text{match}} = \frac{1}{n_1} \sum_{i=1}^{n} T_i \left(Y_i - \frac{1}{|\mathcal{M}_i|} \sum_{i' \in \mathcal{M}_i} Y_{i'} \right)$$

  where $\mathcal{M}_i$ is the "matched set" for treated unit i
- Weighting as a generalization of matching

Matching as Nonparametric Preprocessing for Reducing Model Dependence (Ho, et al. 2007. *Political Anal.*)



Bias in Observational Studies

- Assumptions
  1. Overlap: $0 < \Pr(T_i = 1 \mid \mathbf{X}_i = \mathbf{x}) < 1$ for any $\mathbf{x}$
  2. Ignorability: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid \mathbf{X}_i = \mathbf{x}$ for any $\mathbf{x}$
- Bias decomposition (Heckman et al. 1998. *Econometrica*):

$$
\begin{aligned}
& \mathbb{E}(Y_i(0) \mid T_i = 1) - \mathbb{E}(Y_i \mid T_i = 0) \\
= & \underbrace{\int_{S_1 \setminus S} \mathbb{E}(Y_i(0) \mid T_i = 1, \mathbf{X}_i = \mathbf{x}) dF_{\mathbf{X}_i \mid T_i = 1}(\mathbf{x}) - \int_{S_0 \setminus S} \mathbb{E}(Y_i(0) \mid T_i = 0, \mathbf{X}_i = \mathbf{x}) dF_{\mathbf{X}_i \mid T_i = 0}(\mathbf{x})}_{\text{bias due to lack of common support}} \\
& + \underbrace{\int_S \mathbb{E}(Y_i(0) \mid T_i = 0, \mathbf{X}_i = \mathbf{x}) d\{F_{\mathbf{X}_i \mid T_i = 1}(\mathbf{x}) - F_{\mathbf{X}_i \mid T_i = 0}(\mathbf{x})\}}_{\text{bias due to imbalance of observables within their common support}} \\
& + \underbrace{\int_S \{\mathbb{E}(Y_i(0) \mid T_i = 1, \mathbf{X}_i = \mathbf{x}) - \mathbb{E}(Y_i(0) \mid T_i = 0, \mathbf{X}_i = \mathbf{x})\} dF_{\mathbf{X}_i \mid T_i = 1}(\mathbf{x})}_{\text{bias due to unobservables in common support of observables}}
\end{aligned}
$$

- Matching deals with (1) and (2) but not (3)

Exact and Coarsened Exact Matching

- Exact Matching $\rightsquigarrow$ perfect covariate balance:

$$\widetilde{F}(\mathbf{X}_i \mid T_i = 1) \;=\; \widetilde{F}(\mathbf{X}_i \mid T_i = 0)$$

- No model dependence
- But, exact matching is infeasible when
  - covariate is continuous
  - there are many covariates
- Coarsened Exact Matching (CEM) (Iacus et al. 2011 *Political Anal.*)
  - discretize covariates so that you can match
  - covariates are often discrete
  - discrete categories may have substantive meanings
  - accounts for all interactions among coarsened variables
  - some treated units may have no matched controls (lack of overlap) $\rightsquigarrow$ changes estimand
  - bias-variance tradeoff
  - still infeasible in high dimension

Matching based on Distance Measures

- Common measures used for dimension reduction:
  1. Mahalanobis distance:

$$D(\mathbf{X}_i, \mathbf{X}_j) = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)^\top \widetilde{\Sigma}^{-1} (\mathbf{X}_i - \mathbf{X}_j)}$$

  2. (Estimated) Propensity score:

$$D(\mathbf{X}_i, \mathbf{X}_j) = |\widehat{\pi(\mathbf{X}_i)} - \widehat{\pi(\mathbf{X}_j)}| = |\Pr(\widehat{T_i = 1} \mid \mathbf{X}_i) - \Pr(\widehat{T_j = 1} \mid \mathbf{X}_j)|$$

     or often with the linear predictor of logistic regression

$$D(\mathbf{X}_i, \mathbf{X}_j) = |\text{logit}(\widehat{\pi(\mathbf{X}_i)}) - \text{logit}(\widehat{\pi(\mathbf{X}_j)})|$$

- Classical matching methods (Rubin. 2006. *Matched Sampling for Causal Effects*. Cambridge University Press; Stuart. 2010. *Stat. Sci.*):
  - one-to-one, one-to-many
  - with and without replacement
  - caliper

Propensity Score as a Balancing Score (Rosenbaum and Rubin. 1983. *Biometrika*)

- Probability of receiving the treatment:

$$\pi(\mathbf{X}_i) \quad = \quad \Pr(T_i = 1 \mid \mathbf{X}_i)$$

- Balancing property:

$$T_i \quad \perp\!\!\!\perp \quad \mathbf{X}_i \mid \pi(\mathbf{X}_i)$$

- Exogeneity given the propensity score (under exogeneity given covariates):

$$(Y_i(1), Y_i(0)) \quad \perp\!\!\!\perp \quad T_i \mid \pi(\mathbf{X}_i)$$

- Dimension reduction $\rightsquigarrow$ propensity score matching
- But, true propensity score is unknown: propensity score tautology

Checking Covariate Balance

- Success of matching method depends on the resulting balance
  - Ideally, compare the joint distribution of all covariates
  - In practice, check lower-dimensional summaries (e.g., standardized mean difference, variance ratio, empirical CDF difference)

$$\text{standardized mean difference} = \frac{\overbrace{\frac{1}{n_1}\sum_{i=1}^{n} T_i \left(X_{ij} - \frac{1}{|\mathcal{M}_i|} \sum_{i' \in \mathcal{M}_i} X_{i'j} \right)}^{\text{difference-in-means}}}{\underbrace{\sqrt{\frac{1}{n_1 - 1} \sum_{i=1}^{n} T_i (X_{ij} - \overline{X}_{j1})^2}}_{\text{standard deviation}}}$$

- Frequent use of balance test
  - failure to reject the null $\neq$ covariate balance
  - problematic especially because matching reduces the number of observations

Bias of Matching

- Bias of matching arises because of imbalance:

$$
\begin{aligned}
B(\mathbf{X}_i, \mathcal{X}_{\mathcal{M}_i}) &= \mathbb{E}(Y_i(0) \mid T_i = 1, \mathbf{X}_i) - \mathbb{E}\left\{ \frac{1}{|\mathcal{M}_i|} \sum_{i' \in \mathcal{M}_i} Y_{i'} \;\middle|\; \mathcal{X}_{\mathcal{M}_i} \right\} \\
&= \mu_0(\mathbf{X}_i) - \frac{1}{|\mathcal{M}_i|} \sum_{i' \in \mathcal{M}_i} \mu_0(\mathbf{X}_{i'})
\end{aligned}
$$

  where $\mathcal{X}_{\mathcal{M}_i} = \{\mathbf{X}_{i'}\}_{i' \in \mathcal{M}_i}$ with $\mathcal{M}_i$ denoting the "matched set" for i
- Bias correction (Abadie and Imbens. 2011. *J Bus Econ Stat*):

$$
\begin{aligned}
\widehat{Y_i(0)} &= \frac{1}{|\mathcal{M}_i|} \sum_{i' \in \mathcal{M}_i} Y_{i'} + \mathrm{Bias}\widehat{(\mathbf{X}_i, \mathcal{X}_{\mathcal{M}_i})} \\
&= \frac{1}{|\mathcal{M}_i|} \sum_{i' \in \mathcal{M}_i} \left\{ Y_{i'} + \hat{\beta}^\top (\mathbf{X}_i - \mathbf{X}_{i'}) \right\}
\end{aligned}
$$

  where $\hat{\beta}$ is the estimated coefficient for the regression of $Y_{i'}$ on $\mathbf{X}_{i'}$ using all $i' \in \mathcal{M}_i$

Variance

- All matching estimators can be written as a weighting estimator:

$$
\begin{aligned}
\hat{\tau}_{\text{match}} &= \frac{1}{n_1}\sum_{i=1}^{n} T_i \left(Y_i - \frac{1}{|\mathcal{M}_i|}\sum_{i'\in\mathcal{M}_i} Y_{i'} \right) \\
&= \frac{1}{n_1}\sum_{i:T_i=1} Y_i - \frac{1}{n_0}\sum_{i:T_i=0} \underbrace{\left(\frac{n_0}{n_1}\sum_{i':T_{i'}=1} \frac{\mathbf{1}\{i\in\mathcal{M}_{i'}\}}{|\mathcal{M}_{i'}|} \right)}_{W_i} Y_i
\end{aligned}
$$

- Estimation error for the conditional ATT (CATT):

$$
\begin{aligned}
\hat{\tau}_{\text{match}} - \text{CATT} &= \underbrace{\frac{1}{n_1}\sum_{i:T_i=1} \mu_0(\mathbf{X}_i) - \frac{1}{n_0}\sum_{i:T_i=0} W_i \cdot \mu_0(\mathbf{X}_i)}_{\approx 0 \text{ if matched well and in a large sample}} \\
&\quad + \frac{1}{n_1}\sum_{i:T_i=1} (Y_i(1) - \mu_1(\mathbf{X}_i)) - \frac{1}{n_0}\sum_{i:T_i=0} W_i (Y_i(0) - \mu_0(\mathbf{X}_i))
\end{aligned}
$$

- Assume matching is done well and the sample is relatively large
- Conditional variance:

$$\begin{aligned}
& \mathbb{V}(\hat{\tau}_{\text{match}} \mid \mathbf{X}, \mathbf{T}) \\
\approx \; & \frac{1}{n_1^2} \sum_{i:T_i=1}^{n} \mathbb{V}(Y_i(1) \mid \mathbf{X}, \mathbf{T}) + \frac{1}{n_0^2} \sum_{i:T_i=0}^{n} W_i^2 \cdot \mathbb{V}(Y_i(0) \mid \mathbf{X}, \mathbf{T}) \\
= \; & \sum_{i=1}^{n} \left\{ \frac{T_i}{n_1} + (1 - T_i) \frac{W_i}{n_0} \right\}^2 \mathbb{V}(Y_i \mid \mathbf{X}, \mathbf{T})
\end{aligned}$$

1. estimate $\mathbb{V}(Y_i \mid \mathbf{X}, \mathbf{T})$ via matching (Imbens and Rubin, Chapter 19))
2. heteroskedasticity-robust standard errors using regression

- Bootstrap (Abadie and Spiess, in-press, *J. Am. Stat. Assoc*)
  - sample matches, not units
  - cluster standard errors are valid under misspecification
  - does not work for matching with replacement

Motivation

- Matching methods for improving covariate balance
- Potential limitations of matching methods:
  1. inefficient $\rightsquigarrow$ it may throw away data
  2. ineffective $\rightsquigarrow$ it may not be able to balance covariates
- Recall that matching is a special case of weighting:

$$
\begin{aligned}
\hat{\tau}_{\text{match}} &= \frac{1}{n_1}\sum_{i=1}^{n} T_i \left(Y_i - \frac{1}{|\mathcal{M}_i|}\sum_{i' \in \mathcal{M}_i} Y_{i'} \right) \\
&= \frac{1}{n_1}\sum_{i:T_i=1} Y_i - \frac{1}{n_0}\sum_{i:T_i=0} \underbrace{\left(\frac{n_0}{n_1}\sum_{i':T_{i'}=1} \frac{\mathbf{1}\{i \in \mathcal{M}_{i'}\}}{|\mathcal{M}_{i'}|} \right)}_{W_i} Y_i
\end{aligned}
$$

- Idea: weight each observation in the control group such that it looks like the treatment group (i.e., good covariate balance)

Inverse Probability-of-Treatment Weighting (IPW)

- Weighting for surveys: down-weight over-sampled respondents
- Sampling weights inversely proportional to samplig probability
- Horvitz-Thompson estimator (1952. *J. Am. Stat. Assoc.*):

$$\widehat{\mathbb{E}(Y_i)} = \frac{1}{N}\sum_{i=1}^{N}\frac{S_i Y_i}{\Pr(S_i = 1)}$$

- Weight by the inverse of propensity score:

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i Y_i}{\hat{\pi}(\mathbf{X}_i)} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(\mathbf{X}_i)}\right\}$$

$$\widehat{\text{ATT}} = \frac{1}{n_1}\sum_{i=1}^{n}\left\{T_i Y_i - \frac{\hat{\pi}(\mathbf{X}_i)(1-T_i)Y_i}{1-\hat{\pi}(\mathbf{X}_i)}\right\}$$

$$\widehat{\text{ATC}} = \frac{1}{n_0}\sum_{i=1}^{n}\left\{\frac{(1-\hat{\pi}(\mathbf{X}_i))T_i Y_i}{\hat{\pi}(\mathbf{X}_i)} - (1-T_i)Y_i\right\}$$

- Identical propensity score $\rightsquigarrow$ difference-in-means estimator

Normalized Weights

- Survey sampling when the population size is unknown
- Hajek Estimator:

$$\widehat{\mathbb{E}(Y_i)} = \frac{\sum_i S_i Y_i / \Pr(S_i = 1)}{\sum_i S_i / \Pr(S_i = 1)}$$

- Weights are normalized but no longer unbiased
- Normalization of weights may be important when propensity score is estimated

$$\widehat{\text{ATE}} = \frac{\sum_{i=1}^n T_i Y_i / \hat{\pi}(\mathbf{X}_i)}{\sum_{i=1}^n T_i / \hat{\pi}(\mathbf{X}_i)} - \frac{\sum_{i=1}^n (1 - T_i) Y_i / \{1 - \hat{\pi}(\mathbf{X}_i)\}}{\sum_{i=1}^n (1 - T_i) / \{1 - \hat{\pi}(\mathbf{X}_i)\}}$$

- Weighted least squares gives automatic normalization:

$$(\hat{\alpha}_{\text{wls}}, \hat{\beta}_{\text{wls}}) = \underset{\alpha,\beta}{\text{argmin}} \sum_{i=1}^n \left\{ \frac{T_i}{\hat{\pi}(\mathbf{X}_i)} + \frac{1 - T_i}{1 - \hat{\pi}(\mathbf{X}_i)} \right\} (Y_i - \alpha - \beta T_i)^2$$

Variance

- IPW estimator as the method of moments estimator:
  1. moment condition from the propensity score model (e.g., score)

$$\sum_{i=1}^{n} \left\{ \frac{T_i}{\pi_\theta(\mathbf{X}_i)} - \frac{1 - T_i}{1 - \pi_\theta(\mathbf{X}_i)} \right\} \frac{\partial}{\partial \theta} \pi_\theta(\mathbf{X}_i) = 0$$

  2. moment conditions from the weighting estimator

$$\text{Horvitz/Thompson: } \frac{1}{n} \sum_{i=1}^{n} \frac{T_i Y_i}{\pi_\theta(\mathbf{X}_i)} - \mu_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - T_i) Y_i}{1 - \pi_\theta(\mathbf{X}_i)} - \mu_0 = 0$$

$$\text{Hajek: } \frac{1}{n} \sum_{i=1}^{n} \frac{T_i (Y_i - \mu_1)}{\pi_\theta(\mathbf{X}_i)} = \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - T_i)(Y_i - \mu_0)}{1 - \pi_\theta(\mathbf{X}_i)} = 0$$

  $\rightsquigarrow$ large sample variances are readily available

- If the propensity score model is correctly specified, these variances are smaller than those with the true propensity score

Doubly Robust Estimator (Robins et al. 1994. *J. Am. Stat. Assoc.*)

- Augmented IPW (AIPW) estimator:

$$\begin{aligned}\hat{\tau}_{\text{DR}} &= \frac{1}{n}\sum_{i=1}^{n}\left[\left\{\frac{T_i Y_i}{\hat{\pi}(\mathbf{X}_i)} - \frac{T_i - \hat{\pi}(\mathbf{X}_i)}{\hat{\pi}(\mathbf{X}_i)}\hat{\mu}_1(\mathbf{X}_i)\right\}\right. \\ &\qquad \left. - \left\{\frac{(1-T_i)Y_i}{1-\hat{\pi}(\mathbf{X}_i)} - \frac{T_i - \hat{\pi}(\mathbf{X}_i)}{1-\hat{\pi}(\mathbf{X}_i)}\hat{\mu}_0(\mathbf{X}_i)\right\}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\left\{\hat{\mu}_1(\mathbf{X}_i) + \frac{T_i(Y_i - \hat{\mu}_1(\mathbf{X}_i))}{\hat{\pi}(\mathbf{X}_i)}\right\}\right. \\ &\qquad \left. - \left\{\hat{\mu}_0(\mathbf{X}_i) + \frac{(1-T_i)(Y_i - \hat{\mu}_0(\mathbf{X}_i))}{1-\hat{\pi}(\mathbf{X}_i)}\right\}\right]\end{aligned}$$

- Consistent if either the propensity score model or the outcome model is correct $\rightsquigarrow$ you get two chances to be correct
- Efficient: smallest asymptotic variance among estimators that are consistent when the propensity score model is correct