Weighting Methods

Kosuke Imai

Harvard University

Stat186/Gov2002 Causal Inference

Fall 2019
Motivation

- Matching methods for improving covariate balance
- Potential limitations of matching methods:
  1. inefficient \(\leadsto\) it may throw away data
  2. ineffective \(\leadsto\) it may not be able to balance covariates

- Recall that matching is a special case of weighting:

\[
\hat{r}_{\text{match}} = \frac{1}{n_1} \sum_{i=1}^{n} T_i \left( Y_i - \frac{1}{|M_i|} \sum_{i' \in M_i} Y_{i'} \right)
\]

\[
= \frac{1}{n_1} \sum_{i:T_i=1} Y_i - \frac{1}{n_0} \sum_{i:T_i=0} \left( \frac{n_0}{n_1} \sum_{i':T_{i'}=1} \frac{1}{|M_{i'}|} \right) Y_i
\]

- 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 sampling probability

\[
\hat{E}(Y_i) = \frac{1}{N} \sum_{i=1}^{N} \frac{S_i Y_i}{\Pr(S_i = 1)}
\]

- Weight by the inverse of propensity score:

\[
\begin{align*}
\hat{ATE} &= \frac{1}{n} \sum_{i=1}^{n} \left\{ T_i Y_i \frac{\hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} - (1 - T_i) Y_i \right\} \\
\hat{ATT} &= \frac{1}{n_1} \sum_{i=1}^{n} \left\{ T_i Y_i - \frac{\hat{\pi}(X_i)(1 - T_i) Y_i}{1 - \hat{\pi}(X_i)} \right\} \\
\hat{ATC} &= \frac{1}{n_0} \sum_{i=1}^{n} \left\{ \frac{(1 - \hat{\pi}(X_i)) T_i Y_i}{\hat{\pi}(X_i)} - (1 - T_i) Y_i \right\}
\end{align*}
\]

- Identical propensity score \(\rightsquigarrow\) difference-in-means estimator
Normalized Weights

- Survey sampling when the population size is unknown
- Hajek Estimator:
  \[
  \text{Hajek Estimator: } \hat{E}(Y_i) = \frac{\sum_i S_i Y_i / \Pr(S_i = 1)}{\sum_i S_i / \Pr(S_i = 1)}
  \]

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

\[
\hat{\text{ATE}} = \frac{\sum_{i=1}^n T_i Y_i / \hat{\pi}(X_i)}{\sum_{i=1}^n T_i / \hat{\pi}(X_i)} - \frac{\sum_{i=1}^n (1 - T_i) Y_i / \{1 - \hat{\pi}(X_i)\}}{\sum_{i=1}^n (1 - T_i) / \{1 - \hat{\pi}(X_i)\}}
\]

- Weighted least squares gives automatic normalization:

\[
(\hat{\alpha}_{\text{wls}}, \hat{\beta}_{\text{wls}}) = \arg\min_{\alpha, \beta} \sum_{i=1}^n \frac{T_i(1 - \hat{\pi}(X_i)) + (1 - T_i)\hat{\pi}(X_i)}{\hat{\pi}(X_i)\{1 - \hat{\pi}(X_i)\}}(Y_i - \alpha - \beta T_i)^2
\]
IPW estimator as the method of moments estimator:

1. moment condition from the propensity score model (e.g., score)

\[
\sum_{i=1}^{n} \left\{ \frac{T_i}{\pi_\theta(X_i)} - \frac{1 - T_i}{1 - \pi_\theta(X_i)} \right\} \frac{\partial}{\partial \theta} \pi_\theta(X_i) = 0
\]

2. moment conditions from the weighting estimator

Horvitz/Thompson:

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{T_i Y_i}{\pi_\theta(X_i)} - \mu_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - T_i) Y_i}{1 - \pi_\theta(X_i)} - \mu_0 = 0
\]

Hajek:

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{T_i (Y_i - \mu_1)}{\pi_\theta(X_i)} = \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - T_i) (Y_i - \mu_0)}{1 - \pi_\theta(X_i)} = 0
\]

⇝ 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.

- **Augmented IPW (AIPW) estimator:**

\[
\hat{\tau}_{DR} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{T_i - \hat{\pi}(X_i)}{\hat{\pi}(X_i)} \hat{\mu}_1(X_i) \right\} \\
- \left\{ \frac{(1 - T_i) Y_i}{1 - \hat{\pi}(X_i)} - \frac{T_i - \hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} \hat{\mu}_0(X_i) \right\}
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{\mu}_1(X_i) + \frac{T_i (Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} \right\} \\
- \left\{ \hat{\mu}_0(X_i) + \frac{(1 - T_i) (Y_i - \hat{\mu}_0(X_i))}{1 - \hat{\pi}(X_i)} \right\}
\]

- Consistent if either the propensity score model or the outcome model is correct $\implies$ you get two chances to be correct
- Efficient: smallest asymptotic variance among estimators that are consistent when the propensity score model is correct
The deteriorating performance of propensity score weighting methods when the model is misspecified

Led to improvements of doubly robust estimators \( \Rightarrow \) Cao et al. (2009), Tan (2010), Rotnitzky et al. (2012), Han and Wang (2013) Biometrika. etc.

Setup:
- 4 covariates \( X_i^* \): all are i.i.d. standard normal
- Outcome model: linear model
- Propensity score model: logistic model with linear predictors
- Misspecification induced by measurement error:
  \[
  X_{i1} = \exp(X_{i1}^*/2)
  \]
  \[
  X_{i2} = X_{i2}^*/(1 + \exp(X_{1i}^*) + 10)
  \]
  \[
  X_{i3} = (X_{i1}^* X_{i3}^*/25 + 0.6)^3
  \]
  \[
  X_{i4} = (X_{i1}^* + X_{i4}^* + 20)^2
  \]

Weighting estimators to be evaluated:
1. Horvitz-Thompson
2. Inverse-probability weighting with normalized weights
3. Weighted least squares regression with covariates
4. Doubly-robust least squares regression with covariates
Weighting Estimators Do Fine If the Model is Correct

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Estimator</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>logit</td>
<td>True</td>
</tr>
<tr>
<td>(1) Both models correct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 200)</td>
<td>HT</td>
<td>0.33</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>−0.13</td>
<td>−0.13</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−0.04</td>
<td>−0.04</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>−0.04</td>
<td>−0.04</td>
</tr>
<tr>
<td>(n = 1000)</td>
<td>HT</td>
<td>0.01</td>
<td>−0.18</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>0.01</td>
<td>−0.05</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

(2) Propensity score model correct

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Estimator</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 200)</td>
<td>HT</td>
<td>−0.05</td>
<td>−0.14</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>−0.13</td>
<td>−0.18</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>(n = 1000)</td>
<td>HT</td>
<td>−0.02</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>0.02</td>
<td>−0.03</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>0.04</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Weighting Estimators are Sensitive to Misspecification

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Estimator</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>logit</td>
<td>True</td>
</tr>
<tr>
<td>(3) Outcome model correct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 200$</td>
<td>HT</td>
<td>24.25</td>
<td>−0.18</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>1.70</td>
<td>−0.26</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−2.29</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>−0.08</td>
<td>−0.10</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>HT</td>
<td>41.14</td>
<td>−0.23</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>4.93</td>
<td>−0.02</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−2.94</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>(4) Both models incorrect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 200$</td>
<td>HT</td>
<td>30.32</td>
<td>−0.38</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>1.93</td>
<td>−0.09</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−2.13</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>−7.46</td>
<td>0.37</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>HT</td>
<td>101.47</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>5.16</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>−2.95</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>−48.66</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Covariate Balancing Propensity Score (CBPS)


- How can we improve the estimation of propensity score?
- Estimate the propensity score such that covariates are balanced
- Covariate balancing conditions:

\[
\mathbb{E} \left\{ \frac{T_i}{\pi_\theta(X_i)} - \frac{1 - T_i}{1 - \pi_\theta(X_i)} \right\} f(X_i) = 0
\]

- Usual score condition: \( f(X_i) = \frac{\partial}{\partial \theta} \pi_\theta(X_i) \)
- Balancing intercept \( \leadsto \) normalization of weights, sample boundedness


\[
f(X_i) = \pi_\theta(X_i)\mu_0(X_i) + \{1 - \pi_\theta(X_i)\}\mu_1(X_i)
\]

- double robustness
- smallest asymptotic variance when the propensity score is correct
- Estimation via the (generalized) method of moments
## More Robust Weighting Methods

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Estimator</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>GLM</td>
<td>CBPS1</td>
</tr>
<tr>
<td>(3) Outcome model correct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 200</td>
<td>HT</td>
<td>24.25</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>1.70</td>
<td>-1.37</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>-2.29</td>
<td>-2.37</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>-0.08</td>
<td>-0.10</td>
</tr>
<tr>
<td>n = 1000</td>
<td>HT</td>
<td>41.14</td>
<td>-2.02</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>4.93</td>
<td>-1.39</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>-2.94</td>
<td>-2.99</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>(4) Both models incorrect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 200</td>
<td>HT</td>
<td>30.32</td>
<td>1.27</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>1.93</td>
<td>-1.26</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>-2.13</td>
<td>-2.20</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>-7.46</td>
<td>-2.59</td>
</tr>
<tr>
<td>n = 1000</td>
<td>HT</td>
<td>101.47</td>
<td>-2.05</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>5.16</td>
<td>-1.44</td>
</tr>
<tr>
<td></td>
<td>WLS</td>
<td>-2.95</td>
<td>-3.01</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>-48.66</td>
<td>-3.59</td>
</tr>
</tbody>
</table>
Calibration Methods

- Avoid modeling the propensity score \( \sim \) just balance covariates
  - no modeling assumption about treatment assignment
  - but implicit assumptions about the outcome model
  - in theory, propensity score balances the entire distributions
  - validation and interpretation are more difficult

- Entropy balancing (Hainmueller. 2012. *Political Anal.*)

\[
\{ w_1^*, w_2^*, \ldots, w_{n_0} \} = \arg\min_w \sum_{i: T_i = 0} w_i \log(\frac{w_i}{q_i})
\]

subject to

\[
w_i \geq 0, \quad \sum_{i: T_i = 0} w_i = 1, \quad \sum_{i: T_i = 0} w_i f(X_i) = \frac{1}{n_1} \sum_{i: T_i = 1} f(X_i)
\]

- convex optimization problem
- exact balance in moments
- possibly extreme weights
Fig. 3 Covariate balance in the LaLonde data. Left panel shows plot of covariate-by-covariate standardized bias in the unadjusted data and after the various preprocessing methods. The standardized bias measures the difference in means between the treatment and control group (scaled by the standard deviation). Zero bias indicates identical means, dots to the right (left) of zero indicate a higher mean among the treatment (control) group. The right panel shows the p-value for a covariate-by-covariate t-test for the differences in means after the unadjusted data and after the various preprocessing methods.

In order to investigate the reduction in model dependency, we follow Ho et al. (2007) and examine the sensitivity of the effect estimates in both the unadjusted and the preprocessed data across a wide range of possible specifications of the outcome model. In particular, we fit one million regressions of the outcome on the treatment variable and a subset of covariates that we randomly draw from the set of all possible subsets of the 52 covariates. We fit each regression twice, once with the unadjusted data (unweighted) and once with the preprocessed data (regressions are weighted by the entropy balancing weights). Figure 4 provides the densities of the estimates. The results are extremely model dependent in the unadjusted data with effect sizes ranging from $-8500$ to over $4000$. In the preprocessed data, however, all regressions yield the exact same estimate that is expected because the weights orthogonalize the treatment indicator with respect to all 52 covariate combinations that are included in the reweighting. This suggests that model dependency is reduced after entropy balancing.

19 Notice that there are over 4.5 quadrillion possible subsets of the 52 covariates (\(\sum_{i=1}^{52} \binom{52}{i}\)) so we cannot run all possible regressions.

- minimize the variance of weights while keeping a pre-specified level of covariate balance

\[
\text{minimize } \|w - \bar{w}\|_2^2
\]

subject to

\[
\begin{align*}
    w_i & \geq 0, \\
    \sum_{i:T_i=0} w_i & = 1, \\
    \left| \sum_{i:T_i=0} w_i X_{ij} - \frac{1}{n_1} \sum_{i:T_i=1} X_{ij} \right| & \leq \delta_j
\end{align*}
\]

- quadratic convex programming problem
- choice of \(\delta_j\)

Recent developments on weighting methods:

Summary

- Weighting methods as a generalization of matching methods
  - more efficient and possibly more effective
  - be careful about extreme weights

- Propensity score weighting
- Doubly robust estimation: combine weighting and regression
- Robust estimation of propensity score for balancing covariates
- Calibration methods

Recommended readings:
- Imbens and Rubin. Chapter 17 (Section 8)