Leveraging Observational Outcomes To Improve the Generalization Of Experimental Results

Melody Y. Huang, Naoki Egami, Erin Hartman, Luke Miratrix m.huang@ucla.edu University of California, Los Angeles

Zoom Link (Available 2:30-4:30 PM (ET), T) Meeting PW: 687063

UCLA

Motivation–Towards More Efficient Estimators for Population Effects

Population average treatment effects are often estimated using an inverse propensity weighting estimator (Stuart et al., 2011). These methods rely on the assumption that there exists **conditional ignorability of sampling and treatment heterogeneity**:

$$Y_i(1) - Y_i(0) \perp\!\!\!\perp S_i \mid \mathbf{X}_i$$

Under this assumption, weighted estimators are consistent estimators of PATE. However, in practice, they can be unstable due to the presence of extreme weights, which can lead to a loss in precision.

Population-level observational data usually contain not only pre-treatment covariates, but also outcome data. However, existing methods only utilize the pre-treatment covariate data in the weight estimation process, effectively discarding the population-level information afterwards.

Our method proposes recycling the population-level data in order to incorporate not only the pre-treatment covariate data, but also the population-level outcomes. By leveraging the information that can be mined from the population-level data, we can provide improvements in the generalization of experimental results.

Overview of Method

1. Estimate a selection model to calculate the estimated selection weights, w_i, for all units in the sample.
2. Select a family of auxiliary outcome functions $g(\mathbf{X}_i) : \mathbb{R}^p \to \mathbb{R}$, where $\mathbf{X}_i \in \mathbb{R}^p$. Estimate g, obtaining $\hat{g}(\mathbf{X}_i)$, using the full population data. This function does not need to be correctly specified.
3. Predict $\widehat{Y}_i = \hat{g}(\mathbf{X}_i)$ for each unit in the sample.
4. Residualize the sample outcomes, obtaining $\hat{e}_i = Y_i - \widehat{Y}_i$
5. Estimate the treatment effects using the residualized data and the sampling weights.

Theoretical Properties

Theorem 1 (Consistency of Residualized Weighted Estimator)
Define the residualized weighted estimator as:

$$\hat{\tau}_W^{rec} = \frac{\sum_{i=1}^N S_i T_i \hat{e}_i w_i}{\sum_{i=1}^N S_i T_i w_i} - \frac{\sum_{i=1}^N S_i (1-T_i) \hat{e}_i w_i}{\sum_{i=1}^N S_i (1-T_i) w_i}$$

Assuming that conditional ignorability on selection holds, then $\hat{\tau}_W^{rec}$ will be a consistent estimator for PATE (i.e., $\hat{\tau}_W^{rec} \xrightarrow{p} \tau$).

Implications of Theorem 1:

- The specification of the auxiliary outcome model need not be correct for consistency to hold.

Theorem 2 (Relative Efficiency Gain from Residualizing)
Define the following pseudo-R^2 values:

$$R_1^2 = 1 - \frac{\text{var}(w_i \hat{e}_i(1))}{\text{var}(w_i Y_i(1))} \qquad R_0^2 = 1 - \frac{\text{var}(w_i \hat{e}_i(0))}{\text{var}(w_i Y_i(0))}$$

Let $f = \frac{1}{n_0}\text{var}(w_i Y_i(0)) / \frac{1}{n_1}\text{var}(w_i Y_i(1))$. The relative reduction in variance from residualizing is given by:

$$\begin{aligned} \textit{Relative Reduction} &= \frac{\text{var}(\hat{\tau}_W) - \text{var}(\hat{\tau}_W^{rec})}{\text{var}(\hat{\tau}_W)} \\ &= \frac{f}{1+f} R_0^2 + \frac{1}{1+f} R_1^2 \end{aligned}$$

Implications of Theorem 2:

- Improvement in efficiency is a function of how well the population outcome model fits the sample outcomes, both across the treatment and control groups.
- The relative reduction can be estimated in order to determine if the population outcome model has sufficiently modeled the sample outcomes.

Note: Theorem 2 is calculated with a Horvitz-Thompson style estimator. In practice, a Hajek style estimator is often used instead, in which the estimator is stabilized by the sum of the weights. In that case, the expression in Theorem 2 can be used as an approximation for the relative reduction.

Simulation

To empirically test the method, we run a series of simulations. To begin, $(X_1, X_2, X_S, X_\tau) \sim N(0, \Sigma)$, where:

$$\Sigma = \begin{bmatrix} 1 & 0 & 0.45 & 0.5 \\ 0 & 1 & 0 & 0 \\ 0.45 & 0 & 1 & 0.9 \\ 0.5 & 0 & 0.9 & 1 \end{bmatrix}$$

- Treatment Heterogeneity: $\tau_i = X_{\tau i} + \alpha$
- Propensity of being included in experimental sample:

$$P(S_i = 1) \propto \frac{\exp(X_{Si})}{1 + \exp(X_{Si})}$$

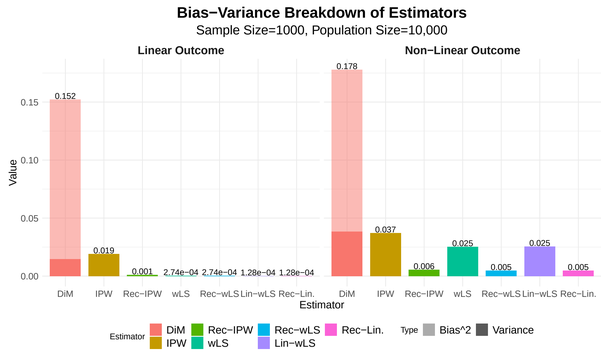
- Outcome model:

$$Y_i(0) = \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma_1 X_{1i}^2 + \gamma_2 \sqrt{|X_{2i}|} + \gamma_3 (X_{1i} \cdot X_{2i})$$

- Potential Outcomes: $Y_i(T_i) = Y_i(0) + \tau_i \cdot T_i + \varepsilon_i$

The population outcome model was estimated using regression with all pair-wise interactions. We ran two scenarios, in which the outcome model comprised of only linear terms (i.e., $\gamma_i = 0$), and one where the outcome model also had non-linear terms (i.e., $\gamma_i \neq 0$).

Three sets of estimators were compared: (1) weighted estimator (IPW), (2) weighted least squares (wLS), and (3) Lin-style weighted least squares (Lin-wLS).



Simulation Results

- Large efficiency gains are observable across all the estimators.
- Residualizing provides benefits when the population outcome model is able to capture interaction terms that are not being accounted for in the covariate adjustments from the wLS and Lin-wLS estimators.
- In the non-linear outcome model case, the population outcome model captures the non-linearities in the outcome model, leading to efficiency gains for not just the IPW estimator but also the wLS and Lin-wLS estimators.

Diverging Population and Sample Data Generating Processes

In the previous simulations, we considered scenarios where the population and sample data generating processes were the same. However, in practice, there may be differences between the population and sample. To see how much divergence can exist between the two processes before residualizing fails to provide efficiency gains, we modify the outcome model to the following:

$$\begin{aligned} Y_i(0) = &\beta_1 X_{1i} + \beta_2 X_{2i} + \gamma_1 X_{1i}^2 + \gamma_2 \sqrt{|X_{2i}|} + \gamma_3 (X_{1i} \cdot X_{2i}) + \\ &\beta_S \cdot (1 - S_i) \cdot (\beta_3 X_{1i} + \gamma_4 X_{1i} \cdot X_{2i}) + \varepsilon_i \end{aligned}$$

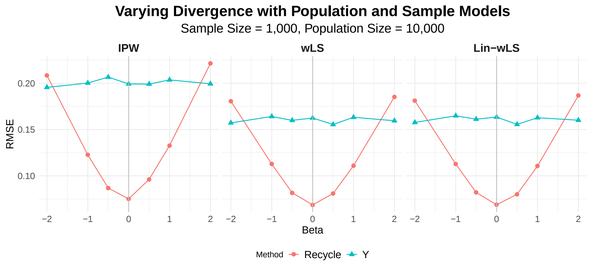
The magnitude of β_S will dictate how different the two processes are.

We estimate the expected relative reduction from residualizing:

Estimated Relative Reduction

β_S	Estimated Value
-5.00	-7.50
-2.00	-0.47
-1.00	0.53
-0.75	0.67
0.75	0.66
1.00	0.51
2.00	-0.50
5.00	-7.56

When $|\beta_S| < 2$, we expect residualizing first will result in efficiency gains for the weighted estimator.



Key Takeaways

- When $|\beta_S|$ takes on very large values, the residualized estimators fail to provide much improvements in efficiency, and in some extreme cases, see a loss in efficiency.
- When $|\beta_S|$ is relatively small, though the data generating processes are different, residualizing still provides efficiency gains.
- The empirical performance of the estimators aligns with what we expected from estimating the relative reduction.
- In practice, practitioners can check for whether or not the population outcome model is explaining enough of the variation in the sample outcomes to provide precision gains by using Theorem 2.

Empirical Application: Get-Out-the-Vote

We now apply our method to a set of Get-Out-the-Vote experiments, originally conducted by Green, Gerber, and Nickerson, 2003.

The goal of the original study was to assess whether in-person canvassing increases voter turnout. Experiments were deployed across six different sites–Bridgeport, Columbus, Detroit, Minneapolis, St. Paul, and Raleigh–prior to the 2001 local elections.

Set-Up

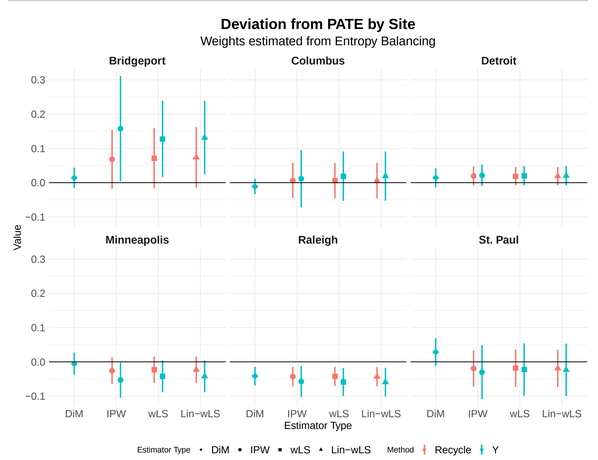
- For each of the six sites, we define PATE to be the estimated intent-to-treat (ITT) effect from the other 5 sites, with the goal of extrapolating the estimated ITT from one site to the others
- **Weighting Method:** entropy balancing on the available covariates (age, family size, and past voting history from primary and general elections) (Hainmueller, 2012).
- **Population Outcome Model:** Super Learner ensemble model across the population outcomes with available covariates.

Proportion of Explained Variance across Cities

City	Population	Within Sample		
		No Weights	Trimmed	Original Weights
Bridgeport	0.56	0.19	0.63	0.85
Raleigh	0.46	0.44	0.55	0.58
Minneapolis	0.61	0.37	0.39	0.39
Detroit	0.62	0.28	0.11	0.11
Columbus	0.58	0.36	0.77	0.86
St. Paul	0.61	0.38	0.61	0.54

From the proportion of explained variances, we expect that certain sites, such as Bridgeport and Columbus, should see greater improvements in efficiency than sites such as Detroit, in which the population outcome model does not do as well in explaining the sample outcomes.

Results



Discussion. In sites where the proportion of explained variance is high, we see a larger precision gains from residualizing first (i.e., Bridgeport, Columbus).

References

Buchanan, Ashley L et al. (2018). "Generalizing evidence from randomized trials using inverse probability of sampling weights". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181.4, pp. 1193–1209.

Cole, Stephen R and Elizabeth A Stuart (2010). "Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial". In: *American journal of epidemiology* 172.1, pp. 107–115.

Green, Donald P, Alan S Gerber, and David W Nickerson (2003). "Getting out the vote in local elections: Results from six door-to-door canvassing experiments". In: *Journal of Politics* 65.4, pp. 1083–1096.

Hainmueller, Jens (2012). "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies". In: *Political analysis*, pp. 25–46.

Laan, Mark J Van der, Eric C Polley, and Alan E Hubbard (2007). "Super learner". In: *Statistical applications in genetics and molecular biology* 6.1.

Stuart, Elizabeth A et al. (2011). "The use of propensity scores to assess the generalizability of results from randomized trials". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174.2, pp. 369–386.