Causal Inference in the Age of Big Data

Jasjeet S. Sekhon

UC Berkeley

March 23, 2017

Causal Inference and Big Data

- Measuring human activity has generated massive datasets with granular population data: e.g.,
    - Browsing, search, and purchase data from online platforms
    - Electronic medical records
    - Individual voter files
    - Individual tax record panels
- Big in size and breadth: wide datasets
- Data can be used for personalization of treatments, creating markets, modeling behavior
- Many inferential issues: e.g., unknown sampling frames, heterogeneity, targeting optimal treatments

Prediction versus Causal Inference

- Causal Inference is like a prediction problem: but predicting something we don't directly observe and possibly cannot estimate well in a given sample

- ML algorithms are good at prediction, but have issues with causal inference:
  - Interventions imply counterfactuals: response schedule versus model prediction
  - Validation requires estimation in the case of causal inference
  - Identification problems not solved by large data
  - Predicting the outcome mistaken for predicting the causal effect
    - targeting based on the lagged outcome

Classical Justifications Versus ML Pipelines

Two different justifications for statistical procedures:

1. (classical) statistical theory:
   it works because we have **relevant** theory that tells us it should
   Hopefully, this is not simply: "Assume that the data are generated by the following model ..." (Brieman 2001)

2. Training/test loop:
   it works because we have validated against ground truth and it works

Classical Justifications Versus ML Pipelines

Two different justifications for statistical procedures:

1. (classical) statistical theory:
   it works because we have **relevant** theory that tells us it should
   Hopefully, this is not simply: "Assume that the data are generated by the following model ..." (Brieman 2001)

2. Training/test loop:
   it works because we have validated against ground truth and it works

On the **normal distribution**:

> *"Everyone believes in it: experimentalists believing that it is a mathematical theorem, mathematicians believing that it is an empirical fact." — Henri Poincaré (quoted by de Finetti 1975)*

Even Classical Justifications Should be Validated

- Question: coverage for the population mean. Is $n = 1000$ enough?
- Sometimes, no. Not for many metrics, even when they are bounded
- For some metrics, asking for 95% CI results in only 60% coverage
- Data is very irregular
  Many zeros, IQR: 0

$$\frac{p100 - p99}{p99 - p50} > 10,000$$

Classical Methods Meet Computing Challenges

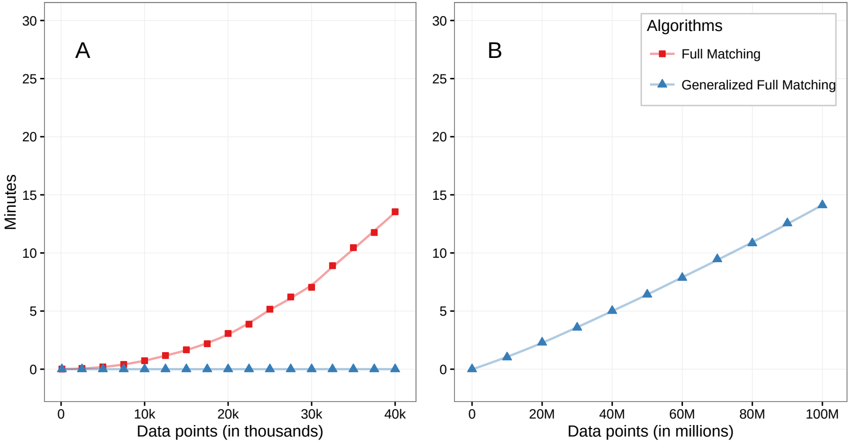
- With big data come small effect sizes: 1e-9
- Some traditional experimental design methods have become computationally infeasible—e.g., blocking, stratification—using the classical methods
- Blocking: create strata and then randomize within strata
- Stratification: create strata after randomization
- Polynomial time solution not quick enough. Linearithmic is survivable.

Blocking/Post-Stratification

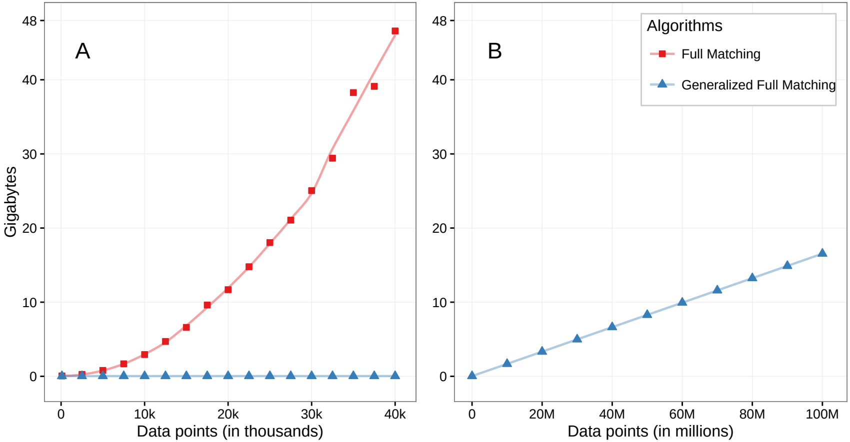
Minimizes the pair-wise Maximum Within-Block Distance: λ
(Higgins, Sävje, Sekhon 2016; Sävje, Higgins, Sekhon 2017)

- Any valid distance metric; triangle inequality
- We prove this is a NP-hard problem
- Ensures good covariate balance by design: approximately optimal: $\leq 4 \times \lambda$
- Works for any number of treatments and any minimum number of observations per block
- It is fast: $O(n \log n)$ expected time
- It is memory efficient: $O(n)$ storage
- Special cases
  1. with one covariate: λ
  2. with two covariates: $\leq 2 \times \lambda$

Time Complexity



Space Complexity



Correct by Design

- Freedman (2008): Can regression adjustments be made to experimental data? Problem: "Since randomization does not justify the models ..."

$$Y = \alpha + \gamma T + \beta X + \epsilon$$

- Analyze behavior under a weaker model. Neyman's non-parametric model: each subject has two potential responses, one if treated and the other if untreated; only one of the two responses is observed; finite sample
- Winston Lin (2013): regression is okay for $p << n$
- Miratrix, Sekhon, and Yu (2013): post-stratification; saturated regression
- Bloniarz, Liu, Zhang, Sekhon, Yu (2015): lasso and $p > n$, but additional sparsity assumptions needed

Regression Adjustment

- Consider estimators of the form

$$\widehat{ATE} = \left[\hat{Y}(t) - (\bar{x}_t - \bar{x})^T \beta^{(t)}\right] - \left[\hat{Y}(c) - (\bar{x}_c - \bar{x})^T \beta^{(c)}\right]$$

- $\beta^{t,c}$: projection coefficients
- We can decompose potential outcome as follows:

$$Y_i(t) = \bar{Y}(t) + (\bar{x}_i - \bar{x})^T \beta^{(t)} + e_i^{(t)}$$

Conditional Average Treatment Effect (CATE)

Individual treatment effect: $D_i := Y_i(t) - Y_i(c)$

Let $\hat{\tau}_i$ be an estimator for D_i

$\tau(x_i)$ is the **CATE** for all units whose covariate vector is equal to x_i:

$$\tau(x_i) := \mathbb{E}\Big[D\Big|X = x_i\Big] = \mathbb{E}\Big[Y(t) - Y(c)\Big|X_i = x_i\Big]$$

Variance of Conditional Average Treatment Effect

Decompose the MSE at x_i:

$$\mathbb{E}\left[(D_i - \hat{\tau}_i)^2 | X_i = x_i\right]$$
$$= \mathbb{E}\left[(D_i - \tau(x_i))^2 | X_i = x_i\right] + \mathbb{E}\left[(\tau(x_i) - \hat{\tau}_i)^2 | X_i = x_i\right]$$

Since we cannot influence the first term, estimating D_i is equivalent to estimating the CATE at x_i.

X-Learner

procedure X–LEARNER(X, Y^{obs}, T)

Estimate μ_1 and μ_0:

$$\hat{\mu}_0 = M_1(Y^0 \sim X^0)$$

$$\hat{\mu}_1 = M_2(Y^1 \sim X^1)$$

Compute pseudo residuals:

$$\tilde{D}_i^1 := Y_i^1 - \hat{\mu}_0(X_i^1)$$

$$\tilde{D}_i^0 := \hat{\mu}_1(X_i^0) - Y_i^0$$

Estimate CATE separately for the treated and control units:

$$\hat{\tau}_1 = M_3(\tilde{D}^1 \sim X^1)$$

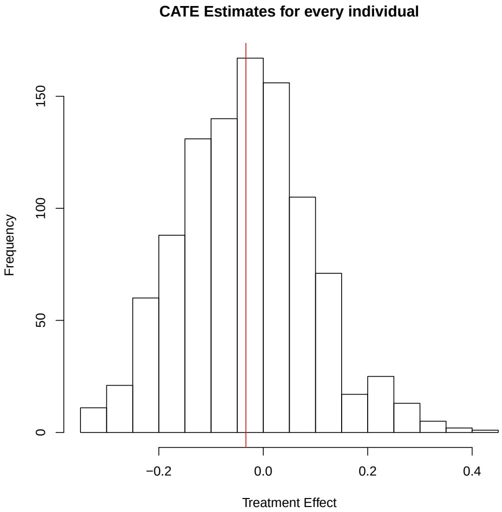
$$\hat{\tau}_0 = M_4(\tilde{D}^0 \sim X^0)$$

Average the estimates:

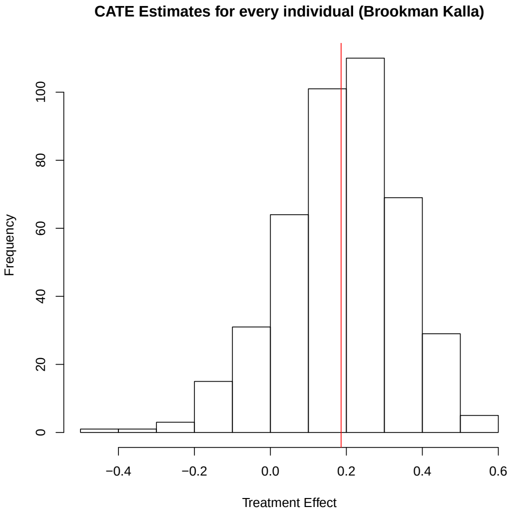
$$\hat{\tau}(x) = \hat{e}(x)\hat{\tau}_0(x) + (1 - \hat{e}(x))\hat{\tau}_1(x)$$

Algorithm 1: $M_k(Y \sim X)$ is here the notation for a regression estimator, which estimates $x \mapsto \mathbb{E}[Y|X = x]$. $\hat{e}(x)$ is an estimator for the propensity score, $e(x) = \mathbb{P}[T = 1|X = x]$. Künzel, Sekhon, Bickel, and Yu (2017)

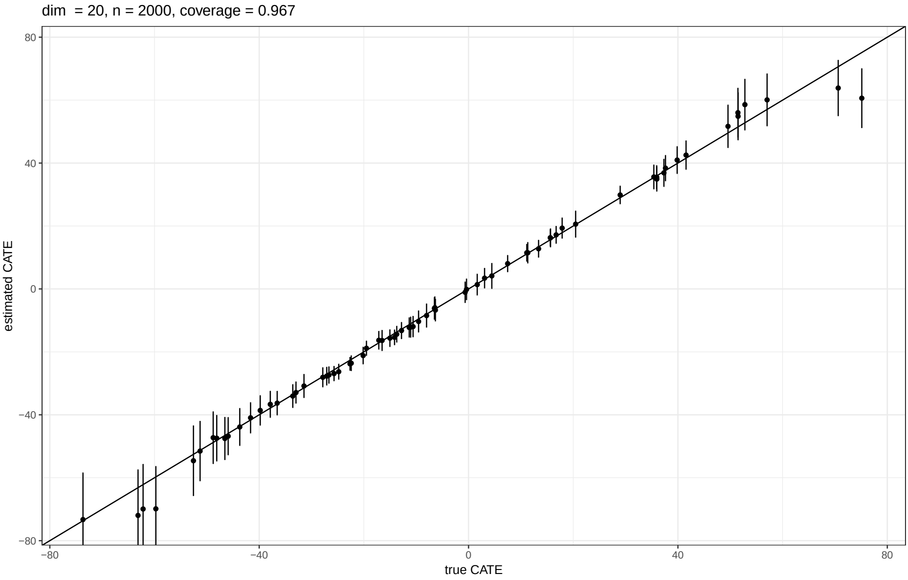
Experiment with No Average Effect



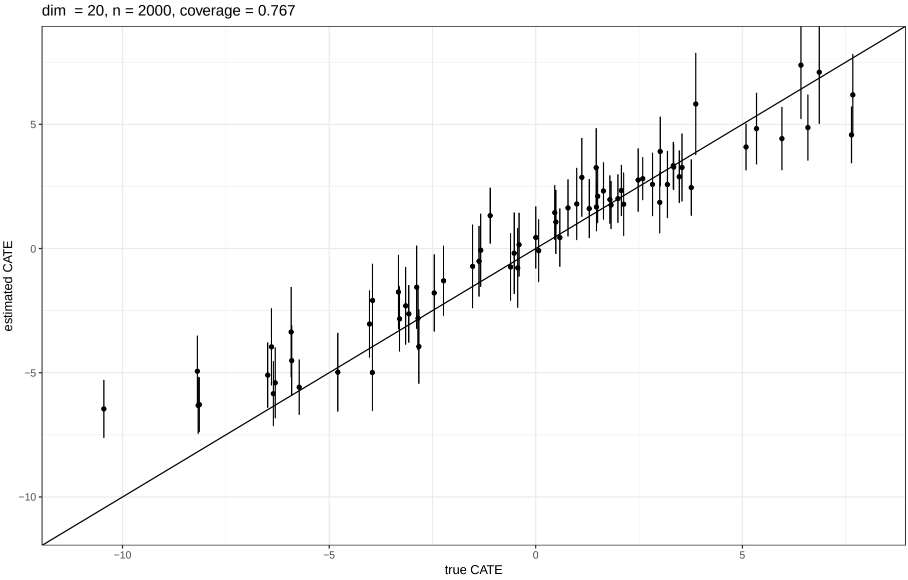
Experiment with Positive Average Effect



Coverage

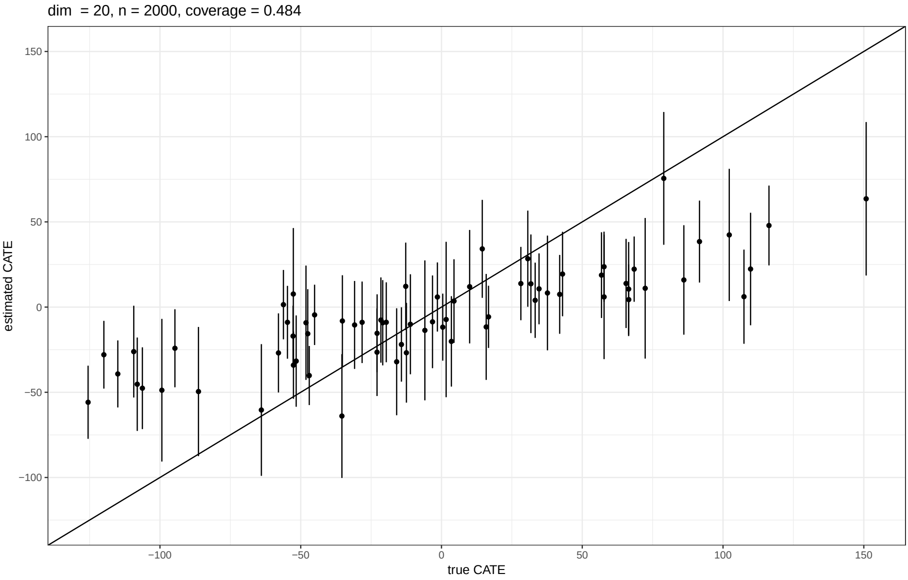

dim = 20, n = 2000, coverage = 0.967
estimated CATE
true CATE
80
40
0
−40
−80
−80
−40
0
40
80

Coverage



dim = 20, n = 2000, coverage = 0.767
estimated CATE
true CATE
5
0
−5
−10
−10
−5
0
5

Coverage



dim = 20, n = 2000, coverage = 0.484
estimated CATE
true CATE
150
100
50
0
−50
−100
−100
−50
0
50
100
150

Conclusion

- Big data doesn't solve the causal identification problem
- Data is now cheap and computing expensive
- Power is a significant concern
- Somethings are easier to validate than others: experiments estimate marginal effects and not general equilibrium
- Lots of observational data: massive push to use it; natural experiments are underutilized
- Validation, validation, and validation

Blocking/Post-Stratification

- Local approximate optimality: Let $\mathbf{b}_{sub} \subseteq \mathbf{b}_{alg}$ be any subset of blocks from a blocking constructed by the algorithm. Define $V_{sub} = \bigcup_{V_x \in \mathbf{b}_{sub}} V_x$ as the set of all vertices contained in the blocks of $\mathbf{b}_{sub}$. Let λ_{sub} denote the maximum edge cost in an optimal blocking of V_{sub}. The subset of blocks is an approximately optimal blocking of V_{sub}:

$$\max_{ij \in E(\mathbf{b}_{sub})} c_{ij} \leq 4\lambda_{sub}.$$

  It is fast:
  - NNG plus $O(d^0 kn)$ time and $O(d^0 kn)$ space
  - K-d trees NN: $O(2^d kn \log n)$ expected time, $O(2^d kn^2)$ worst time, and $O(kn)$ storage
  - Compare with bipartite, network flow methods:
    - e.g., Derigs: $O(n^3 \log n + dn^2)$ worst time and $O(d^0 n^2)$ space

Neyman Model

- $Y_i(t)$, $Y_i(c)$: potential outcomes for unit i under **treatment** and **control**
- T_i: random indicator of treatment for unit i
- Y_i: observed outcome (under no-interference)

$$Y_i = Y_i(t)T_i + Y_i(c)(1 - T_i), \quad i = 1, ..., n$$

- Average of the treated/control

$$\bar{Y}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t), \quad \hat{Y}(t) = \frac{1}{n_t}\sum_{i \in T} Y_i$$

$$\bar{Y}(c) = \frac{1}{n}\sum_{i=1}^{n} Y_i(c), \quad \hat{Y}(c) = \frac{1}{n_c}\sum_{i \in C} Y_i$$

- n, n_t, n_c: number of total, treated, and control units

Simple Estimator

- If assignment is completely randomized, the ATE can be estimated without bias using the simple difference in means:

$$\widehat{ATE}_{\text{unadj}} = \hat{Y}(t) - \hat{Y}(c)$$

- $\hat{\sigma}^2$ is asymptotically conservative estimate of σ^2:

$$\hat{\sigma}^2 = \frac{n}{n_t}\hat{\sigma}_t^2 + \frac{n}{n_t}\hat{\sigma}_t^2$$

Assumptions, Freedman (2008)

- Condition 1: Stability of treatment assignment probability

$$n_A/n \to p_A,\ \ n_B/n \to p_B,\ \text{as } n \to \infty,$$

  for some $p_A, p_B \in (0, 1)$

- Condition 2: The centered moment conditions

$$n^{-1} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^4 \leq L,\ \forall j$$

$$n^{-1} \sum_{i=1}^{n} (e_i^{(a)})^4 \leq L;\ \ n^{-1} \sum_{i=1}^{n} (e_i^{(b)})^4 \leq L$$

- Condition 3: The means $n^{-1} \sum_{i=1}^{n} (e_i^{(a)})^2$, $n^{-1} \sum_{i=1}^{n} (e_i^{(b)})^2$ and $n^{-1} \sum_{i=1}^{n} e_i^{(a)} e_i^{(b)}$ converge to finite limits.

Conservative variance estimate

- Asymptotic variance

$$\sigma^2 = \lim_{n\to\infty} \left[\frac{1-p_A}{p_A}\sigma^2_{e^{(a)}} + \frac{p_A}{1-p_A}\sigma^2_{e^{(b)}} + 2\sigma_{e^{(a)}e^{(b)}} \right]$$

- Let

$$\hat{\sigma}^2_{e^{(a)}} = \frac{1}{n_A - df^{(a)}} \sum_{i\in A} \left\{ a_i - \bar{a}_A - (x_i - \bar{x}_A)^T \hat{\beta}^{(a)} \right\}^2$$

- $\hat{\sigma}^2$ is asymptotically conservative estimate of σ^2:

$$\hat{\sigma}^2 = \frac{n}{n_A}\hat{\sigma}^2_{e^{(a)}} + \frac{n}{n_B}\hat{\sigma}^2_{e^{(b)}}$$

Further assumptions for consistency of Lasso

- Condition 4: Decay and scaling

$$\delta_n = o\left(\frac{1}{s\sqrt{\log p}}\right); \quad (s \log p)/\sqrt{n} = o(1)$$

- Condition 5: Cone invertibility factor

$$\|h_S\|_1 \leq Cs\|\hat{\Sigma}h\|_\infty, \ \forall h \in \mathcal{C} = \{h : \|h_{S^c}\|_1 \leq \xi\|h_S\|_1\}$$

$$\hat{\Sigma} = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

- Condition 6: Tuning parameter

$$\lambda_a \in (\frac{1}{\eta}, M] \times \left(\frac{11\sqrt{L}}{3p'_A}\sqrt{\frac{\log p}{n}} + \delta_n\right)$$

$$\lambda_b \in (\frac{1}{\eta}, M] \times \left(\frac{11\sqrt{L}}{3p'_B}\sqrt{\frac{\log p}{n}} + \delta_n\right)$$

Asymptotic Normality

Theorem

Assume conditions 1 - 6 hold. Then

$$\sqrt{n}\left(\widehat{ATE}_{\text{Lasso}} - ATE\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right),$$

$$\sigma^2 = \lim_{n\to\infty}\left[\frac{1-p_A}{p_A}\sigma^2_{e^{(a)}} + \frac{p_A}{1-p_A}\sigma^2_{e^{(b)}} + 2\sigma_{e^{(a)}e^{(b)}}\right]$$

which is no greater than the asymptotic variance of the $\sqrt{n}\left(\widehat{ATE}_{\text{unadj}} - ATE\right)$. *The difference is* $\frac{1}{p_A(1-p_A)}\Delta$.

$$\Delta = -\lim_{n\to\infty}\|X\beta_E\|_2^2 \le 0, \quad \beta_E = (1-p_A)\beta^{(a)} + p_A\beta^{(b)}$$

Two quantities needed for high-dim case

- Sparsity measures (number of nonzero coefficients)

$$s = |\{j : \beta_j^{(a)} \neq 0 \text{ or } \beta_j^{(b)} \neq 0\}|$$

- Maximum covariance

$$\delta_n = \max_{\omega=a,b} \left\{ \max_j \left| \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j) \left(e_i^{(\omega)} - \bar{e}^{(\omega)} \right) \right| \right\}$$

Combining Population Data with Experimental Data

- Can using observational population data help one estimate experimental treatment effects?
- Various approaches: design based, shrinkage, test/validation

Simulation Setup

$$X \sim \mathcal{N}(0, I)$$
$$W \sim Bern(e(X))$$
$$Y = W\mu_t(X) + (1 - W)\mu_c(X) + \varepsilon$$

with $\varepsilon \overset{i.i.d.}{\sim} N(0, 1)$.

- First setup;

$$\mu_t(x) = 3x_1 + 5x_2 + 30x_3$$
$$\mu_c(x) = 3x_1 + 5x_2$$
$$e(x) = .1$$

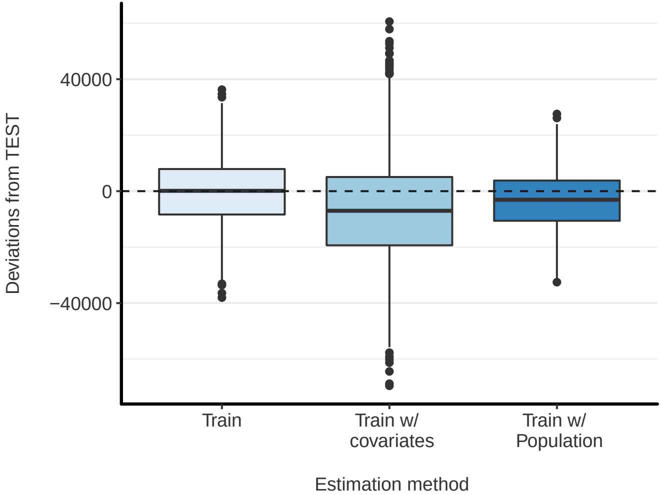
- Second setup:

$$\mu_t(x) = 3x_1 + 3x_2 + 4x_3$$
$$\mu_c(x) = 3x_1 + 5x_2$$
$$e(x) = .1$$

- Third setup:

$$\mu_t(x) = x^T \beta_t \qquad \text{with} \qquad \beta_t \sim \text{Unif}[(1, 30)^{\text{dim}}]$$
$$\mu_c(x) = x^T \beta_c \qquad \text{with} \qquad \beta_c \sim \text{Unif}[(1, 30)^{\text{dim}}]$$
$$e(x) = .5$$

Using Large Population Data



Combing RCTs and Observational Data

$$\left(\lambda \cdot \beta_{RCT}^{training} + (1-\lambda) \cdot \beta_{NRS} - \beta_{RCT}^{test}\right)^2$$

$$\Rightarrow \lambda = \underset{\lambda \in \{\lambda:\ 0 \leq \lambda \leq 1\}}{\text{argmin}} \left\{ \left(\lambda \cdot \beta_{RCT}^{training} + (1-\lambda) \cdot \beta_{NRS} - \beta_{RCT}^{test}\right)^2 \right\}$$

Where, $0 \leq \lambda \leq 1$

The optimal weight to assign $\beta_{RCT}^{training}$ is,

$$\hat{\lambda} = \min\left(1, \max\left(0, \frac{\beta_{RCT}^{test} - \beta_{NRS}}{\beta_{RCT}^{training} - \beta_{NRS}}\right)\right)$$