ECON 293/MGTECON 634: Machine Learning and Causal Inference

Susan Athey and Stefan Wager
Stanford University

Lecture 1: Randomized Experiments,
Observational Studies, and Matching

6 April 2018

A central goal of machine learning is to understand **what usually happens** in a given situation, e.g.,

- Given today's weather, what's the chance tomorrow's air pollution levels will be dangerously high?

Most economists want to predict **what would happen** if we changed the system, e.g.,

- How does the answer to the above question change if we reduce the number of cars on the road?

Discussion: Eva Ascarza. **Retention Futility.** *J. Marketing Research*, 55(1), 2018.

This class is about the interface of causal inference and machine learning, with both terms understood broadly:

- Our discussion of **causal inference** draws from a long tradition in economics and epidemiology on which questions about **counterfactuals** can be answered using a given type of data, and how these estimands can be **interpreted**.
- We use the term **machine learning** to describe an engineering heavy approach to data analysis. Given a well-defined task in which good performance can be **empirically validated**, we do not shy away from **computationally heavy** tools or **potentially heuristic** approaches (e.g., decision trees, neural networks, non-convex optimization).

[pause for logistics]

Today's lecture:

- The potential outcomes model for causal inference in randomized experiments.
- Observational studies and the propensity score.
- Matching (an engineering approach).

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, \ldots, n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

Following the **potential outcomes** framework (Neyman, 1923; Rubin, 1974), we posit the existence of quantities $Y_i(0)$ and $Y_i(1)$, such that $Y_i = Y_i(W_i)$.

- These correspond to the response we **would have measured** given that the i-th subject received treatment $(W_i = 1)$ or no treatment $(W_i = 0)$.

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, ..., n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

Our first goal is to estimate the **average treatment effect (ATE)**

$$\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right].$$

NB: In reality, we only get to see $Y_i = Y_i(W_i)$.

The potential outcomes framework

The simplest way to **identify** the ATE in the potential outcomes is via a **randomized trial**:

$$\{Y_i(0),\ Y_i(1)\} \perp\!\!\!\perp W_i.$$

In a randomized trial, we can check that:

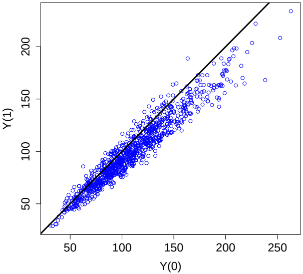
$$\begin{aligned}
\tau &= \mathbb{E}\left[Y_i(1)\right] - \mathbb{E}\left[Y_i(0)\right] \\
&= \mathbb{E}\left[Y_i(1) \,\middle|\, W_i = 1\right] - \mathbb{E}\left[Y_i(0) \,\middle|\, W_i = 0\right] \\
&= \mathbb{E}\left[Y_i \,\middle|\, W_i = 1\right] - \mathbb{E}\left[Y_i \,\middle|\, W_i = 0\right],
\end{aligned}$$

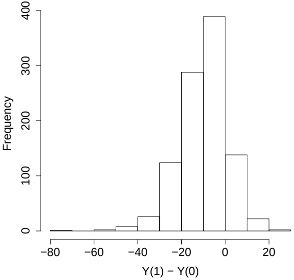
where the last line only has **observable moments**.

Thus, although we never observe $\tau_i = Y_i(1) - Y_i(0)$, we can **consistently estimate** $\tau = \mathbb{E}\left[\tau_i\right]$ in a randomized trial.

Example: The outcome Y_i is daily **air quality index**. The treatment imposes restrictions on driving to reduce traffic.

$Y_i(0)$	$Y_i(1)$	τ_i
154.68	153.49	-1.20
135.67	120.40	-15.27
103.46	117.68	14.23
117.62	95.08	-22.54
161.11	146.73	-14.39
117.89	105.05	-12.84
84.00	75.59	-8.41
73.32	65.68	-7.63
100.07	93.80	-6.28
103.81	82.30	-21.51
...	...	...
111.68	101.47	-10.21





Example: The outcome Y_i is daily **air quality index**. The treatment imposes restrictions on driving to reduce traffic.

$Y_i(0)$	$Y_i(1)$	τ_i
154.68	—	—
135.67	—	—
—	117.68	—
—	95.08	—
—	146.73	—
117.89	—	—
—	75.59	—
—	65.68	—
100.07	—	—
—	82.30	—
...	...	...
110.59	100.52	—

- In practice, we only ever observe a **single** potential outcome.
- However, in a RCT, we can use **averages** over the treated and controls to estimate the ATE.
- We **estimate** $\hat{\tau}$ as $110.59 - 100.52 = 10.07$.

ATE estimation in randomized trials

We have use the **potential outcomes** framework to justify the classical estimator of an **average treatment effect**:

$$\hat{\tau} = \frac{\sum_{\{i:W_i=1\}} Y_i}{|\{i : W_i = 1\}|} - \frac{\sum_{\{i:W_i=0\}} Y_i}{|\{i : W_i = 0\}|}.$$

This estimator is **unbiased**, **consistent**, **asymptotically Gaussian**, and also very **simple**. But is it the best we can do?

- If one has access to **covariates** X_i and can estimate $\mathbb{E}\left[Y_i \,\middle|\, X_i,\, W_i\right]$ accurately, then one can **improve the precision** of the above estimator.
- Any black-box predictor can be used for this (e.g., a forest, boosted trees, a deep net); the improvement in precision depends on **mean-squared error**.

ATE estimation in randomized trials

The simplest ATE estimator in an RCT is

$$\hat{\tau} = \frac{\sum_{\{i:W_i=1\}} Y_i}{|\{i : W_i = 1\}|} - \frac{\sum_{\{i:W_i=0\}} Y_i}{|\{i : W_i = 0\}|}.$$

How could we possibly improve on this?

- In the **air quality** example, weather has an effect on ozone (hot days have higher levels), independently of treatment.
- If we randomly assign treatment to more hot days and control to more cold days, our estimates we **exaggerate the treatment effect**, and vice-versa.
- In **large samples** these effects cancel out, but in **small samples** they matter. If we could **predict** and **eliminate** the effect of weather, we'd improve accuracy.

The traditional approach to this is via **stratified sampling**; here, we'll discuss an automatic approach that only assumes the existence of a **good predictor**.

ATE estimation in randomized trials

For a set of i.i.d. subjects $i = 1, \ldots, n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

Define the conditional **response surfaces** as

$$\mu_{(w)}(x) = \mathbb{E}\left[Y_i \,\middle|\, X_i = x,\, W_i = w\right].$$

In the potential outcomes model, an **oracle** who knew the $\mu_{(w)}(x)$ could use

$$\bar{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\mu_{(1)}(X_i) - \mu_{(0)}(X_i)\right).$$

Our approach starts by seeking to imitate this oracle.

ATE estimation via prediction

In the potential outcomes model, an **oracle** who knew the $\mu_{(w)}(x)$ could use

$$\bar{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\mu_{(1)}(X_i) - \mu_{(0)}(X_i) \right).$$

A first, naive approach simply sets

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) \right).$$

This is good if $\hat{\mu}_{(w)}(x)$ is obtained via **OLS**. But it breaks down if we use **regularization**.

Example. Suppose that $p \gg n$, but the true model is sparse,

$$\mathbb{E}\left[Y \,\middle|\, X = x, W = w \right] = 2X_1 + 0.1WX_2.$$

A **lasso** might set the coefficient on WX_2 to 0, and estimate $\hat{\tau} = 0$!

ATE estimation via prediction

A better estimator needs to **correct for regularization bias**:

$$
\begin{aligned}
\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)\right) & \qquad \text{(optimistic plug-in)} \\
+ \frac{\sum_{\{i : W_i = 1\}} \left(Y_i - \hat{\mu}_{(1)}(X_i)\right)}{\left|\{i : W_i = 1\}\right|} & \qquad \text{(bias correction for } \hat{\mu}_{(1)}(\cdot)) \\
- \frac{\sum_{\{i : W_i = 0\}} \left(Y_i - \hat{\mu}_{(0)}(X_i)\right)}{\left|\{i : W_i = 0\}\right|} & \qquad \text{(bias correction for } \hat{\mu}_{(0)}(\cdot))
\end{aligned}
$$

Modulo technical details, this is justified **for any** $\hat{\mu}_{(w)}(x)$. If $\hat{\mu}_{(w)}(x)$ can predict Y_i at all, can improve over basic estimator.

If $\hat{\mu}_{(w)}(x)$ is consistent, i.e., $\mathbb{E}\left[(\hat{\mu}_{(W)}(X) - \mu_{(W)}(X))^2\right] \to 0$, then this estimator is **optimal in large samples**.

Details: Wager et al. **High-Dim. Regression Adjust. in RCTs.** *PNAS*, 113(45), 2016.

ATE estimation via prediction

Example: We have $n = 1000$, $p = 400$, and $\mathbb{P}\left[W = 1\right] = 0.4$, with

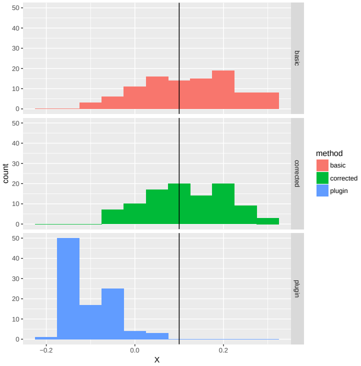
$$\mathbb{E}\left[Y \,\middle|\, X = x, W = w\right] = 2X_1 + 0.1WX_2, \quad X_{ij} \overset{\text{iid}}{\sim} U\left([0,\,2]\right).$$

Predictions made via a **cross-validated lasso** (no intercept).

Consider 3 estimators, with **mean-square errors** for τ:

- **basic**: 0.105.
- **bias-corrected**: 0.092.
- **plug-in**: 0.210.

Distribution of estimates:



Today's lecture:

- The potential outcomes model for causal inference in randomized experiments.
- Observational studies and the propensity score.
- Matching (an engineering approach).

Beyond randomized trials

The simplest way to move beyond randomized controlled trials is to let randomization probabilities depend on **covariate information**.

- We are interested in giving teenagers **cash incentives** to discourage them from **smoking**.
- A random subset of $\sim 5\%$ of teenagers in **Palo Alto, CA**, and a random subset of $\sim 20\%$ of teenagers in **Geneva, Switzerland** are eligible for the study.

Palo Alto	Non-S.	Smoker
Treat.	152	5
Control	2362	122

Geneva	Non-S.	Smoker
Treat.	581	350
Control	2278	1979

This is **not a randomized controlled study**, because Genevans are both more likely to smoke whether or not they get treated, and more likely to get treated.

Beyond randomized trials

The Palo Alto experiment and Geneva experiment are both individually randomized controlled studies—and looking at the numbers clearly shows that the treatment helps prevent smoking.

Palo Alto	Non-S.	Smoker
Treat.	152	6
Control	2362	122

Geneva	Non-S.	Smoker
Treat.	581	395
Control	2278	1979

Looking at aggregate data is misleading, and makes it look like the treatment hurts.

Palo Alto + Geneva	Non-Smoker	Smoker
Treatment	733	401
Control	4640	2101

This phenomenon is an example of Simpson's "paradox".

Beyond randomized trials

Formally, we have covariates $X_i \in \{\text{Palo Alto, Geneva}\}$, and know that the treatment assignment was random conditionally on X_i:

$$\left\{Y_i^{(0)},\, Y_i^{(1)}\right\} \perp\!\!\!\perp W_i \mid X_i.$$

We then estimate the overall average treatment effect as:

$$\hat{\tau} = \sum_{x \in \mathcal{X}} \frac{\left|\{X_i = x\}\right|}{n}\, \hat{\tau}(x),$$

$$\hat{\tau}(x) = \frac{\sum_{\{i : X_i = x,\, W_i = 1\}} Y_i}{\left|\{i : X_i = x,\, W_i = 1\}\right|} - \frac{\sum_{\{i : X_i = x,\, W_i = 0\}} Y_i}{\left|\{i : X_i = x,\, W_i = 0\}\right|}.$$

Covariates and unconfoundedness

For a set of i.i.d. subjects $i = 1, \ldots, n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

We assume that the treatment is **unconfounded** (aka selection on observables) (Rosenbaum & Rubin, 1983):

$$\left\{ Y_i^{(0)}, Y_i^{(1)} \right\} \perp\!\!\!\perp W_i \mid X_i.$$

We seek the ATE $\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$. If the X_i is discrete, we can **stratify**: estimate an ATE for each x separately, and aggregate. But what if X is continuous and/or high-dimensional?

The propensity score

The confounding effects of X_i can be summarized via the **propensity score**,

$$e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right].$$

The key fact about the propensity score is that

$$\tau = \mathbb{E}\left[\frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)}\right].$$

The same idea underlies **importance weighting**, **Horvitz-Thompson sampling**, etc.

The propensity score

Inverse-propensity weighting is unbiased because:

$$
\begin{aligned}
\tau &= \mathbb{E}\left[Y_i(1) - Y_i(0)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i(1) \,\middle|\, X_i\right] - \mathbb{E}\left[Y_i(0) \,\middle|\, X_i\right]\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left[W_i \,\middle|\, X_i\right]\mathbb{E}\left[Y_i(1) \,\middle|\, X_i\right]}{e(X_i)} - \frac{\mathbb{E}\left[1 - W_i \,\middle|\, X_i\right]\mathbb{E}\left[Y_i(0) \,\middle|\, X_i\right]}{1 - e(X_i)}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left[W_i Y_i(1) \,\middle|\, X_i\right]}{e(X_i)} - \frac{\mathbb{E}\left[(1 - W_i)\, Y_i(0) \,\middle|\, X_i\right]}{1 - e(X_i)}\right] \\
&= \mathbb{E}\left[\frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)}\right].
\end{aligned}
$$

The 5-th equality depends on consistency of the **potential outcomes**, and the 4-th equality relies on **unconfoundedness**,

$$\left\{Y_i^{(0)},\, Y_i^{(1)}\right\} \perp\!\!\!\perp W_i \mid X_i.$$

Inverse-propensity weighting

We know that the **average treatment effect** is

$$\tau = \mathbb{E}\left[\frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)}\right].$$

A natural idea is to **estimate** $\hat{e}(\cdot)$ via some machine learning method (e.g., an L_1-penalized logistic regression in high dimensions), and then use

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{e}(X_i)}\right).$$

This strategy has several pitfalls, however:

- Getting properly **calibrated** $\hat{e}(\cdot)$ estimates is hard.
- **Regularization bias** is still a problem.

We will discuss how to improve this estimator in Lecture 3.

Propensity stratification

We know that the **average treatment effect** is

$$\tau = \mathbb{E}\left[\frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)}\right].$$

Another simple way to use this fact is via **propensity stratification**. Pick a number of strata M, and for each $k = 1, \ldots, M$, define $\mathcal{S}_k = \{x : (k-1)/M \leq \hat{e}(x) < k/M\}$ and

$$\hat{\tau}_k = \frac{\sum_{\{i : W_i = 1,\, X_i \in \mathcal{S}_k\}} Y_i}{|\{i : W_i = 1,\, X_i \in \mathcal{S}_k\}|} - \frac{\sum_{\{i : W_i = 0,\, X_i \in \mathcal{S}_k\}} Y_i}{|\{i : W_i = 0,\, X_i \in \mathcal{S}_k\}|},$$

and **aggregate** these estimates as $\hat{\tau} = \frac{1}{n} \sum_{k=1}^{M} |\{i : X_i \in \mathcal{S}_k\}| \, \hat{\tau}_k$.

We are **matching** samples with comparable propensities to each other. Empirically, this is less sensitive to **miscalibration** of $\hat{e}(\cdot)$.

Matching

Matching is a simple and **widely used** approach to treatment effect estimation. The remainder of this lecture will give a brief overview of matching, and discuss how we can take an **engineering approach** to make matching better.

The upcoming presentation relies on the package `designmatch` for `R` by José Zubizarreta, Cinar Kilcioglu and Juan P. Vielma.

Matching

The basic idea in matching is simple. For $k = 1, ..., K$, make **non-overlapping pairs** $(i_{k0},\, i_{k1})$ such that:

$$W_{i_{k0}} = 0, \quad W_{i_{k1}} = 1, \quad X_{i_{k0}} \approx X_{i_{k1}}.$$

We then estimate the **treatment effect** by comparing outcomes:

$$\hat{\tau} = \frac{1}{K} \sum_{k=1}^{K} \left(Y_{i_{k1}} - Y_{i_{k0}} \right),$$

and estimate **standard errors** as $\hat{\sigma}^2 = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(Y_{i_{k1}} - Y_{i_{k0}} \right)^2$.

Estimate a causal quantity given **unconfoundedness**,

$$\left\{ Y_i^{(0)},\, Y_i^{(1)} \right\} \perp\!\!\!\perp W_i \;\big|\; X_i,$$

but the target estimand depends on **where** the matches are.

Matching

We want to study whether **green-certified** commercial buildings command higher rents than non-certified buildings.

- Data on $\mathbf{n_1 = 694}$ green buildings and $\mathbf{n_0 = 7,411}$ non-green buildings.
- Measure several **covariates**, including localtion, age, amenities, number of stories, quality, etc.
- Assume **unconfoundedness** given these covariates.

The answer is a treatment effect. But what is the **question**?

Estimands for matching

We want to study whether **green-certified** commercial buildings command higher rents than non-certified buildings. The outcome Y_i is log-rent. Potential **questions** we could ask include:

- Estimate the **average treatment effect** (ATE) $\mathbb{E}\left[Y_i(1) - Y_i(0)\right]$.
- Estimate the **average treatment effect on the treated** (ATT) $\mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, W_i = 1\right]$.
- Assume that $\tau = \tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right]$ is **constant**. Estimate τ.
- Estimate a **representative treatment effect** for a specific sample of interest.

Estimands for matching

The **average treatment effect on the treated** (ATT) $\mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, W_i = 1\right]$ often has simple interpretation.

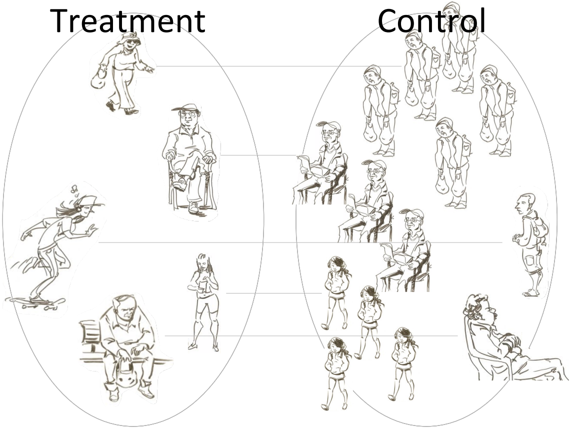


Image credit: J. Zubizarreta.

Estimands for matching

Assuming that $\tau = \tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right]$ is **constant** is often helpful in practice. In case of heterogeneity, the **actual estimand** is

$$\tau_\alpha = \mathbb{E}\left[\alpha(X)\tau(X)\right], \quad \text{for some} \;\; \mathbb{E}\left[\alpha(X)\right] = 1,$$

where the **weighting function** $\alpha(x)$ favors regions with many candidate matches.

Discussion: F. Li et al. **Balancing covariates via p. score weighting.** *JASA*, 2017.

Estimators for matching

Suppose we have n observations, the first $n_1 \ll n$ of which are **treated**. For any pair X_i and X_j, define a **distance** $\Delta(X_i, X_j)$.

ATT matching finds the best possible control match for each treated unit, such that $K = n_1$, $i_{k1} = k$, and

$$\{i_{k0}\}_{k=1}^{K} \text{ minimizes } \sum_{k=1}^{K} \Delta\left(X_{i_{k0}}, X_{i_{k1}}\right).$$

FREE matching lets K float, and solves

$$\underset{K, \{i_{k0}, i_{k1}\}_{k=1}^{K}}{\text{minimize}} \quad \sum_{k=1}^{K} \Delta\left(X_{i_{k0}}, X_{i_{k1}}\right) - \lambda K \quad \text{with} \quad W_{i_{k0}} = 0, \; W_{i_{k1}} = 1.$$

Can be solved as **min-cost flow** problems (Rosenbaum, 1989).

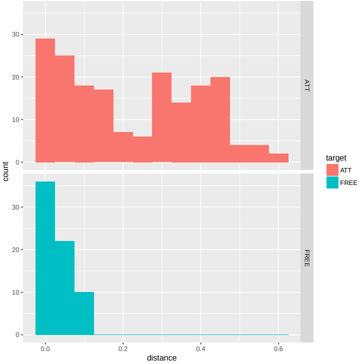
Estimators for matching

Application: **Lalonde** data, with 185 treated units / 260 controls. Covariates include age, education, race, ethnicity, marital status, prior income.

ATT matching creates 185 pairs, with mean Δ of 0.22.

FREE matching creates 68 pairs, with mean Δ of 0.03.

Distribution of distances:



Assessing matching

ATT matching creates 185 pairs, with mean Δ of 0.22. Is the quality of this match acceptable?

The table below shows that matched controls and treated units are **systematically different** on average.

```
          Mis Min       Max  Mean T  Mean C Std Dif
age         0  16     55.00   25.82   26.86   -0.11
education   0   1     18.00   10.35   10.39   -0.02
black       0   0      1.00    0.84    0.46    1.00
hispanic    0   0      1.00    0.06    0.07   -0.04
married     0   0      1.00    0.19    0.30   -0.25
nodegree    0   0      1.00    0.71    0.65    0.11
re74        0   0  35040.07 2095.57 2621.03   -0.09
re75        0   0  25142.24 1532.06 1629.90   -0.03
```

Assessing matching

FREE matching creates 68 pairs, with mean Δ of 0.03. The moments are also now better balanced.

Are our only options to use a very small matched set with **good balance**, or a big matched set with poor balance?

```
          Mis Min      Max   Mean T   Mean C Std Dif
age         0  16    53.00    24.00    24.50   -0.05
education   0   2    17.00    10.40    10.43   -0.01
black       0   0     1.00     0.78     0.78    0.00
hispanic    0   0     1.00     0.04     0.04    0.00
married     0   0     1.00     0.21     0.21    0.00
nodegree    0   0     1.00     0.63     0.63    0.00
re74        0   0 20279.95  2244.63  2089.50    0.03
re75        0   0 17976.15  1699.08  1378.23    0.10
```

Balance-constrained matching

Balance-constrained matching selects a maximal imbalance t, and solves

$$\begin{aligned} \underset{K,\{i_{k0},i_{k1}\}_{k=1}^{K}}{\text{minimize}} \quad & \sum_{k=1}^{K} \Delta\left(X_{i_{k0}},\, X_{i_{k1}}\right) - \lambda K \quad \text{with} \quad W_{i_{k0}} = 0,\ W_{i_{k1}} = 1 \\ \text{subject to:} \quad & \left\| \frac{1}{k} \sum_{k=1}^{K} \left(X_{i_{k1}} - X_{i_{k0}}\right) \right\|_{\infty} \leq t. \end{aligned}$$

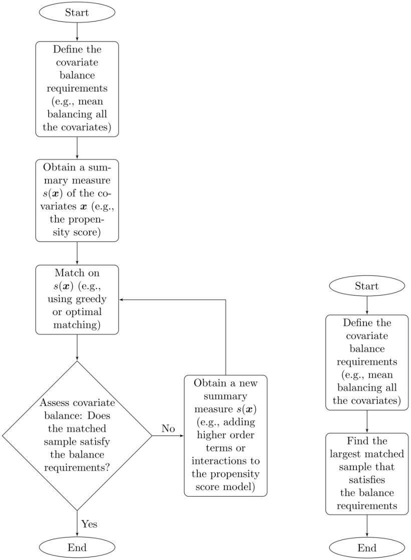
This problem is now a **mixed-integer program**, but can still often be solved using commercial software (e.g., `designmatch` uses `gurobi`, `CPLEX`, etc.)

NB: Standardizing the features is recommended. Now, λ is typically selected to be large.

Balance-constrained matching

Balance-constrained matching creates 122 pairs, with mean Δ of 0.11 and max imbalance t of 0.13. In contrast, ATT gets (185, 0.22, 1) and FREE gets (68, 0.03, 0.1).

```
          Mis Min      Max  Mean T  Mean C Std Dif
age         0  16    55.00   24.97   25.87   -0.10
education   0   1    17.00   10.27   10.22    0.02
black       0   0     1.00    0.76    0.71    0.13
hispanic    0   0     1.00    0.09    0.08    0.03
married     0   0     1.00    0.24    0.27   -0.07
nodegree    0   0     1.00    0.68    0.65    0.07
re74        0   0 35040.07 2511.12 2674.19   -0.03
re75        0   0 25142.24 1820.34 1696.87    0.04
```



Start

Define the covariate balance requirements (e.g., mean balancing all the covariates)

Obtain a summary measure $s(\boldsymbol{x})$ of the covariates $\boldsymbol{x}$ (e.g., the propensity score)

Match on $s(\boldsymbol{x})$ (e.g., using greedy or optimal matching)

Assess covariate balance: Does the matched sample satisfy the balance requirements?

No

Obtain a new summary measure $s(\boldsymbol{x})$ (e.g., adding higher order terms or interactions to the propensity score model)

Yes

End

Start

Define the covariate balance requirements (e.g., mean balancing all the covariates)

Find the largest matched sample that satisfies the balance requirements

End

Source: Kilcioglu & Zubizarreta. *AOAS*, 10(4), 2017.

What about the propensity score?

Recall: we know that the **average treatment effect** is

$$\tau = \mathbb{E}\left[\frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)}\right].$$

Another simple way to use this fact is via **propensity stratification**. Pick a number of strata M, and for each $m = 1, \ldots, M$, define $\mathcal{S}_m = \{x : (m-1)/M \leq \hat{e}(x) < m/M\}$ and

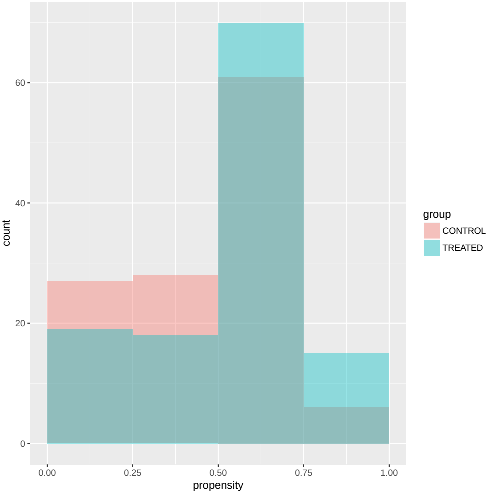
$$\hat{\tau}_m = \frac{\sum_{\{i : W_i = 1,\, X_i \in \mathcal{S}_m\}} Y_i}{|\{i : W_i = 1,\, X_i \in \mathcal{S}_m\}|} - \frac{\sum_{\{i : W_i = 0,\, X_i \in \mathcal{S}_m\}} Y_i}{|\{i : W_i = 0,\, X_i \in \mathcal{S}_m\}|},$$

and **aggregate** these estimates as $\hat{\tau} = \frac{1}{n} \sum_{m=1}^{M} |\{i : X_i \in \mathcal{S}_m\}|\, \hat{\tau}_m$.

This **propensity-stratified** estimator is a popular choice without modern optimization tools. In our new setup, good matching should still **balance** propensity strata.

What about the propensity score?

Our **balance-constrained** matches do a decent, but not perfect job at evening out propensity strata.



Balance-constrained matching with propensity score

Thanks to our optimization-based approach, we can add **effective propensity stratification** as a constraint to **balance-constrained matching**. Given a set of propensity strata $\mathcal{S}_m$ and propensity estimates $\hat{e}(x)$, we solve

$$\underset{K, \{i_{k0}, i_{k1}\}_{k=1}^{K}}{\text{minimize}} \quad \sum_{k=1}^{K} \Delta\left(X_{i_{k0}}, X_{i_{k1}}\right) - \lambda K \quad \text{with} \quad W_{i_{k0}} = 0, \; W_{i_{k1}} = 1$$

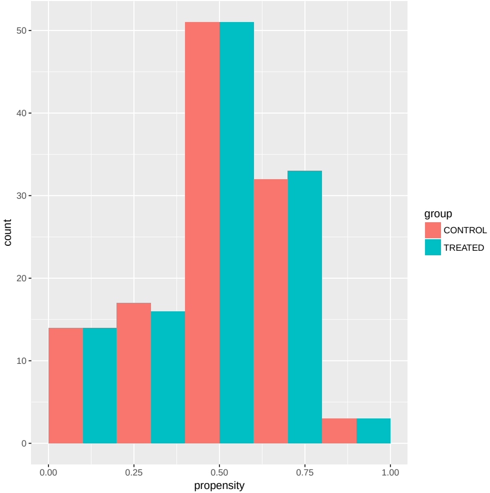
$$\text{subject to:} \quad \left\| \frac{1}{k} \sum_{k=1}^{K} \left(X_{i_{k1}} - X_{i_{k0}}\right) \right\|_{\infty} \leq t$$

$$\text{and} \quad \sum_{k=1}^{K} 1\left(\{X_{k0} \in \mathcal{S}_m\}\right) = \sum_{k=1}^{K} 1\left(\{X_{k1} \in \mathcal{S}_m\}\right) \text{ for all } m.$$

This problem is still a **mixed-integer program** that can be solved with `gurobi`, etc.

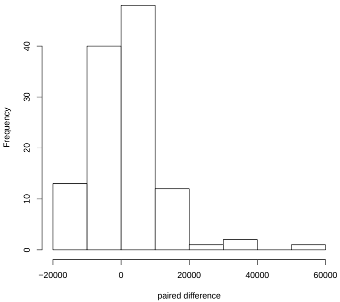
Balance-constrained matching with propensity score

Given the propensity strata constraints, we get 117 pairs with average Δ of 0.12 and worst-case imbalance of 0.11. We learn $\hat{e}(\cdot)$ via a random forest.



So what is the treatment effect?

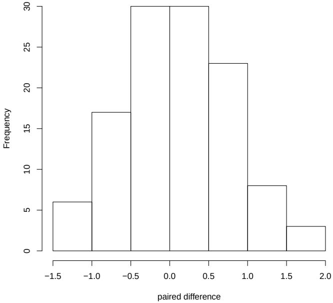
Outcome is post-intervention income. Considering the 117 paired difference, we get a 95% confidence interval of $\tau \in 1705 \pm 1978$. The histogram of the differences is:



Note that we chose matches **before looking at the outcomes**. Further questions: What is the interpretation of $\hat{\tau}$? What about **bias** from imperfect matches? Which matching strategy is **MSE**-optimal?

So what is the treatment effect?

What is we use a **log-stabilized outcome** to avoid outliers, $Y = \log(1 + \text{income}/\text{mean}(\text{income}))$? We get a 95% confidence interval of $\tau \in 0.126 \pm 0.124$, and histogram



Note that we chose matches **before looking at the outcomes**. Further questions: What is the interpretation of $\hat{\tau}$? What about **bias** from imperfect matches? Which matching strategy is **MSE**-optimal?

Representative matching

So far, we have built match sets that are **large**, **balanced** and respect **propensity strata**. But this is potentially at the cost of some interpretability relative to simpler ATT matching.

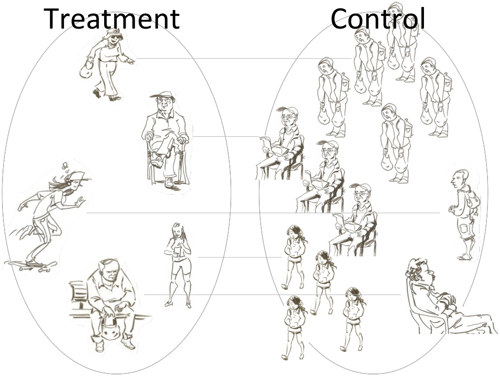


Image credit: J. Zubizarreta.

Representative matching

The solution to a problem in optimization is more optimization...
Idea: Make matched pairs **representative** of treated sample:

$$\underset{K,\,\{i_{k0},\,i_{k1}\}_{k=1}^{K}}{\text{minimize}} \quad \sum_{k=1}^{K} \Delta\left(X_{i_{k0}},\, X_{i_{k1}}\right) - \lambda K \ \ \text{with} \ \ W_{i_{k0}} = 0,\ \ W_{i_{k1}} = 1$$

$$\text{subject to:} \quad \left\| \frac{1}{k} \sum_{k=1}^{K} \left(X_{i_{k1}} - X_{i_{k0}}\right) \right\|_{\infty} \leq t$$

$$\text{and} \ \ \sum_{k=1}^{K} 1\left(\left\{\hat{e}\left(X_{k0}\right) \in \mathcal{S}_m\right\}\right) = \sum_{k=1}^{K} 1\left(\left\{\hat{e}\left(X_{k1}\right) \in \mathcal{S}_m\right\}\right) \ \text{for all } m$$

$$\text{and} \ \ \left\| \frac{1}{k} \sum_{k=1}^{K} X_{i_{kw}} - \frac{1}{n_1} \sum_{\{i : W_i = 1\}} X_i \right\|_{\infty} \leq t',\ \ w \in \{0,\, 1\}\,.$$

The last constraint makes the average of the $X_{i_{kw}}$ in the treated pairs roughly match the features of the mean treated unit.

Matching designs

This last procedure creates matches that:

- Tune the **number of matches** to avoid very poor distances.
- Enforce approximate **aggregate balance** to control bias.
- Enforce exact balance on **propensity strata** for robustness.
- Chooses matched pairs to be **representative**.

All these ideas can be generalized. For example:

- We could enforce exact balance on **important categorical variables** (e.g., state dummy, or demographic category).
- We could make matches representative of **specific subpopulations** (e.g., estimate ATE on black or white participants).

Eventually, questions get high-dimensional / non-parametric / complicated, and **adaptive modeling strategies** are needed.