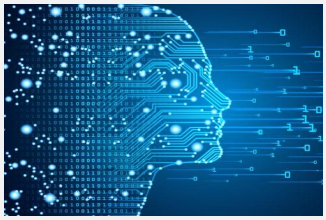


Principles of Causal Inference

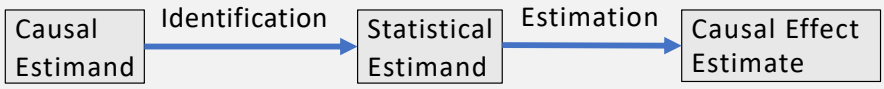
Vasant G. Honavar

Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics,
Public Health Sciences and Neuroscience
Center for Artificial Intelligence Foundations and Scientific Applications
Institute for Computational and Data Sciences
Huck Institutes of the Life Sciences
Clinical and Translational Sciences Institute
Northeast Big Data Hub
Pennsylvania State University
vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

From identification to estimation

- So far, we have focused on identification of causal effects
  - In the Rubin framework under identifiability assumptions
    - Exchangeability, positivity, consistency
  - In the Pearl framework under causal assumptions (structural causal model)
- In either case, we express the causal effect of interest in terms of statistical quantities that can be estimated from observational data
- We now turn to causal effect estimation from data

From identification to estimation



Randomized experiment

Randomized Experiments

- The assignment mechanism is random, known, and controlled by the researcher
- Because the treatments are randomly assigned, the treatment groups should all look similar regarding covariates (observed and unobserved)
- Randomization ensures exchangeability and hence, association implies causation
- In a randomized experiment, potential outcomes are statistically independent of the observed treatment T, given the observed covariates i.e., $\{Y^0, Y^1\} \perp\!\!\!\perp T | \mathbf{X}$
- We have already considered causal effect estimation from randomized experiments

Causal effect estimation from observational data

- An observational study can be viewed as a conditionally randomized experiment if the following conditions hold:
  - Treatments correspond to well-defined interventions that can be imagined in the data
  - The conditional probability of receiving every possible treatment, though not decided by the investigators, depends only on the measured covariates $\mathbf{X}$
  - The probability of receiving every treatment conditional on $\mathbf{X}$ is greater than 0
- These conditions, taken together, are called identifiability assumptions
- We know how to estimate causal effects from conditionally randomized experiments

Causal inference from Observational Data

- In the case of observational studies, we should carefully describe
  - The randomized experiment that we would like to, but cannot, conduct
  - How the observational study emulates that randomized experiment
- In ideal randomized experiments, the data contain sufficient information to identify causal effects
- In contrast, the information in observational data is insufficient to identify causal effects
  - We need causal assumptions (or equivalently, identifiability assumptions)

Estimating causal effects from observational data

- To estimate causal effects from observational data:
  - We specify the randomized control trial that we would like to, but cannot conduct
  - Under "reasonable" assumptions, show how the target trial can be emulated using observational data
    - Identifiability assumptions (Rubin framework)
    - Causal assumptions (Pearl framework)

Study Design versus Analysis

- In a randomized experiment, the design phase (collecting data, balancing covariates, specifying plan) is done before one gets to see treatment outcomes or perform analysis
- In an observational study, you typically get all data together (covariates, treatment, outcomes): there is no clear separation between design and analysis
- Solution?
  - If possible, complete the observational study design before you look at the outcomes
  - Without looking at the outcomes, you may do whatever you need to ensure at least approximate exchangeability

Design Trumps Analysis[1]

- "Design": everything done before access to outcome data
  - Contemplating and collecting data (including covariates)
  - Making causal assumptions
  - Specification of analysis plan that simulates a randomized trial from the observational data as soon as the outcomes are revealed
- Solution?
  - Whenever possible, do all the hard work in the design phase, and the analysis with outcomes will be straightforward

[1]Rubin, D.B., 2008. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, *2*(3), pp.808-840.

Analysis

- In randomized experiments, there is usually a pre-specified protocol for analysis
- In observational studies, people often try many different models and analyses – can introduce subjectivity and bias
- Solution?
  - Specify protocol with outcomes in advance, and do most of your work in the design phase to make analysis easy

Assignment Mechanism

- In a randomized experiment, the assignment mechanism satisfies the identifiability assumptions by design
- In an observational study these, or equivalently an assumed causal graph, are only assumptions, and they are unverifiable from observational data
- Solution?
  - Rely on domain expertise to ensure that the assumptions are plausible
  - Conduct robustness analyses

Unconfoundedness

- Based on the covariates, is the treatment assignment independent of the outcomes?
- Why do we care? What if assignment did depend on the outcomes, conditional on the covariates?
  - We would have confounding
- In the presence of confounding, potential outcomes could differ between treatment groups *before treatment is even applied,* even if covariate values are identical for the treated and untreated groups!
- Unconfoundedness allows us to compare units with similar covariate values to estimate causal effects
- We have to assume absence of confounding, or turn to a set of causal assumptions to identify the confounders, or rely on methods that can cope with confounders

Unconfoundedness

- The plausibility of unconfoundedness lies in the collection of covariates
- Want to compare "like with like".
- Which covariates do we need data on to ensure that a set of units are comparable?
- Answer:
  - Data on all covariates that matter!
  - How do we know which covariates matter?
  - There is no way to know with certainty unless we have a causal graph

Reality

- Observational studies are rarely truly unconfounded
- We just try to get as close as possible to the truth by
  - collecting the all the relevant covariate data possible
  - making plausible causal assumptions when we can
  - and using the techniques we'll learn…

Positivity?

- Every unit has some chance of being assigned to each of the treatment groups, conditional on covariates
- Solution? If some types of individuals are observed only the treatment group, or only in the control group, eliminate them from analyses
  - Restrict causal inferences to the subset of the data for which the positivity assumption holds
  - Discard individuals in the treated (or control) group that are <u>not</u> similar to any other individual in the other group

Regular Assignment Mechanism

- Regular assignment mechanism satisfies
  - Unconfoundedness
  - (Conditional) Exchangeability $Y^0, Y^1 \perp\!\!\!\perp T \mid \mathbf{X}$
  - Positivity $0 < p(T = 1 \mid \mathbf{X} = \mathbf{x}) < 1$
- The probability that an individual is in the treated group depends only on that individual's covariates

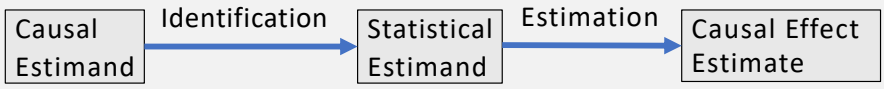
Assignment Mechanism

- In a randomized experiment, the assignment mechanism is known
- In an observational study the assignment mechanism is unknown
- Solution?
  - Estimate the assignment mechanism by modeling it (propensity scores)

Covariate Balance

- In randomized experiments, the randomization creates covariate balance between treatment groups
- In observational studies, treatment groups will be naturally unbalanced regarding covariates
- Solution? compare similar units
- How?
  - Propensity scores
  - Matching
  - Representation learning

From identification to estimation



Causal Estimands

Conditional average treatment effects (CATEs):

$$\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

Always assuming unconfoundedness and positivity

$$\tau \triangleq \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_W\left[\mathbb{E}[Y \mid T = 1, W] - \mathbb{E}[Y \mid T = 0, W]\right]$$

Given W is a sufficient adjustment set

$$\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mathbb{E}_W[\mathbb{E}[Y \mid T = 1, X = x, W] - \mathbb{E}[Y \mid T = 0, X = x, W]]$$

Given $W \cup X$ is a sufficient adjustment set

Estimation Methods

- Conditional outcome modeling and (basic) Machine Learning
- Propensity score and inverse propensity weighting
- Matching
- Non-parametric models - deep learning
- Doubly Robust Machine learning
- Instrumental variables
- Natural experiments

Simple Conditioning

Definition Conditioning calculates treatment effects by identifying groups of individuals with the same covariates, where individuals in one group are treated and in the other group are not.

Intuition Conditioning our analysis of $T \rightarrow Y$ on X breaks the dependence between confounds X and the treatment T

Example Suppose age confounds the causal effect of exercise on cholesterol. By conditioning analysis on age, we can identify the effect of exercise.

Keep in mind How do we know what to condition on?

Grouping becomes harder as dimensionality of X increases

Causal Estimands

Conditional average treatment effects (CATEs):

$$\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

Always assuming unconfoundedness and positivity

$$\tau \triangleq \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_W\left[\mathbb{E}[Y \mid T = 1, W] - \mathbb{E}[Y \mid T = 0, W]\right]$$

Given W is a sufficient adjustment set

$$\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mathbb{E}_W[\mathbb{E}[Y \mid T = 1, X = x, W] - \mathbb{E}[Y \mid T = 0, X = x, W]]$$

Given $W \cup X$ is a sufficient adjustment set

COM Estimators

Key idea

- Just fit a statistical machine learning model $\mu(t, w)$ for $\mathbb{E}(Y|T, W)$
- Approximate $\mathbb{E}_W\ (Y|T) = \mu(t, w)$ with the empirical mean $\frac{1}{n}\sum_i\ \mu(t_i, w_i)$
- Use the model $\mu(t, w)$ to obtain predictions $\hat{\mu}(t_i, w_i)$
- Then COM estimator for ATE is given by:

$$\hat{\tau} = \frac{1}{n}\sum_i\ \hat{\mu}(1, w_i) - \hat{\mu}(0, w_i)$$

- The COM estimator for CATE can be defined in an analogous manner

Conditional Outcome Modeling

$$\tau = \mathbb{E}_W\left[\mathbb{E}[Y \mid T=1, W] - \mathbb{E}[Y \mid T=0, W]\right]$$

$$\tau = \mathbb{E}_W\left[\overbrace{\mu(1, W)}^{\text{model}} - \overbrace{\mu(0, W)}^{\text{model}}\right]$$

Model-assisted estimator:

$$\hat{\tau} = \frac{1}{n}\sum_i \left(\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i)\right)$$

Conditional outcome modeling (COM) of ATE and CATE

ATE COM Estimator: $\hat{\tau} = \frac{1}{n}\sum_i (\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$

CATE Estimand:

$$\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mathbb{E}_W[\mathbb{E}[Y \mid T = 1, X = x, W] - \mathbb{E}[Y \mid T = 0, X = x, W]]$$

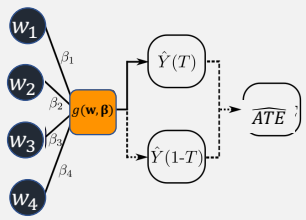
$$\mu(t, w, x) \triangleq \mathbb{E}[Y \mid T = t, W = w, X = x]$$

$$\hat{\tau}_i = \hat{\tau}(x_i) = \hat{\mu}(1, w_i, x_i) - \hat{\mu}(0, w_i, x_i)$$

COM estimators

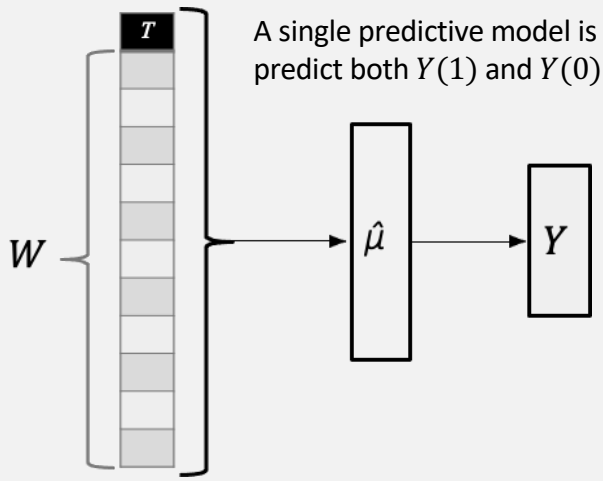
- Fundamentally fit a model (e.g., regression, neural network, random forest, etc.) to estimate μ (as a function of X, W, and T)
- Known by many names
  - G-computation estimators
  - Parametric G-formula
  - Standardization
  - S-learner

ATE Estimation



COM Estimator: S-learner (single learner)

A single predictive model is trained to predict both $Y(1)$ and $Y(0)$



Practical challenges of S-learner

- When the data are high-dimensional, the estimator ignores T, especially when the magnitude of the causal effect is small[1]

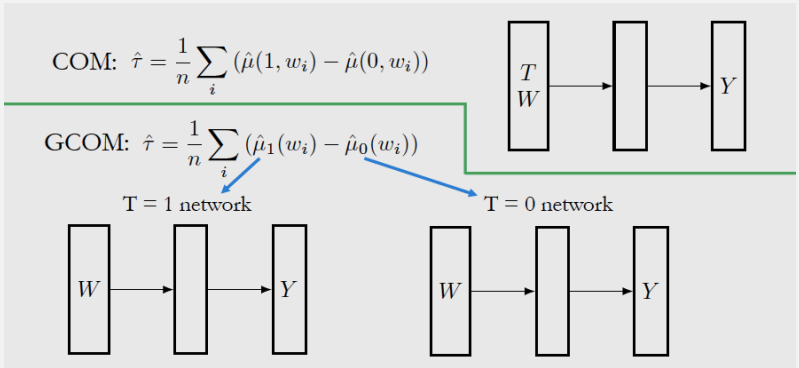
$$\hat{\tau}_i = \hat{\tau}(x_i) = \hat{\mu}(1, w_i, x_i) - \hat{\mu}(0, w_i, x_i)$$

- How can we fix this?

[1] Künzel, S.R., Sekhon, J.S., Bickel, P.J. and Yu, B., 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, *116*(10), pp.4156-4165.

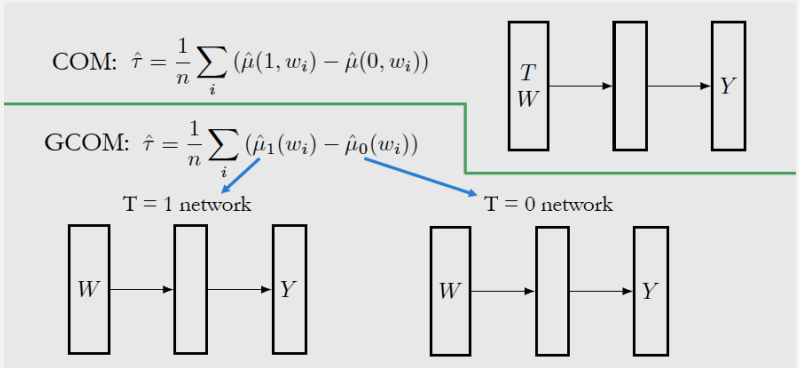
Grouped COM (GCOM) Estimators

- How to ensure that the model does not ignore T?
- Train a separate model for $T = 1$ and $T = 0$!



Now each model is trained on only a subset of the data

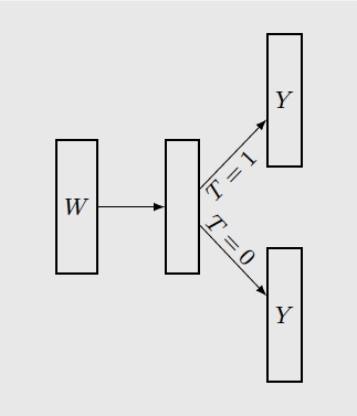
Grouped COM (GCOM) Estimators



- Now each model is trained on only a subset of the data
- Variance of the estimator is higher than that of COM

TARNet Estimator[1]

- The best of COM and GCOM



[1] Shalit, U., Johansson, F.D. and Sontag, D., 2017, Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning* (pp. 3076-3085). PMLR.

X-Learner[1]

1. Estimate $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$ — Assume X is a sufficient adjustment set and is all observed covariates

2a. Impute ITEs

Treatment group: $\hat{\tau}_{1,i} = Y_i(1) - \hat{\mu}_0(x_i)$

Control group: $\hat{\tau}_{0,i} = \hat{\mu}_1(x_i) - Y_i(0)$

2b. Fit a model $\hat{\tau}_1(x)$ to predict $\hat{\tau}_{1,i}$ from x_i in treatment group
Fit a model $\hat{\tau}_0(x)$ to predict $\hat{\tau}_{0,i}$ from x_i in control group

3. $$\hat{\tau}(x) = g(x)\,\hat{\tau}_0(x) + (1 - g(x))\,\hat{\tau}_1(x)$$

where $g(x)$ is some weighing function between 0 and 1. Example: propensity score

[1] Künzel, S.R., Sekhon, J.S., Bickel, P.J. and Yu, B., 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, *116*(10), pp.4156-4165.

Propensity Score

$$e(W) \triangleq P(T = 1 \mid W)$$

Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$.

Even if W is high-dimensional, $e(W)$ is only 1-dimensional!

Propensity Score Theorem

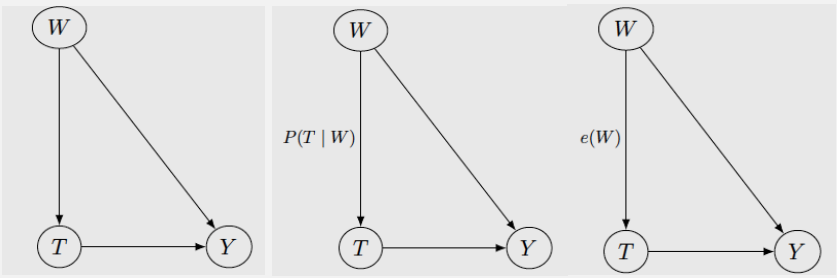
Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$. Equivalently,

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid W \implies (Y(1), Y(0)) \perp\!\!\!\perp T \mid e(W)$$

Propensity Score Theorem

Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$. Equivalently,

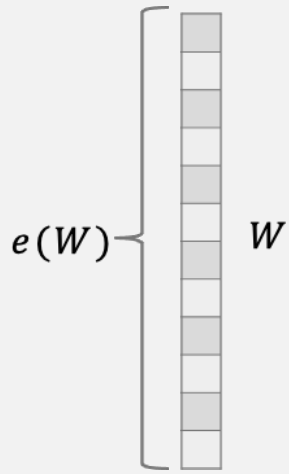
$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid W \implies (Y(1), Y(0)) \perp\!\!\!\perp T \mid e(W)$$



Propensity Score and Positivity-Unconfoundedness Tension

- Recall that overlap decreases with the dimensionality of the adjustment set
- The propensity score magically reduces the dimensionality of the adjustment set W to 1!
- Propensity score is unknown but can be learned from data
  $e(\mathbf{w}) = P(T = 1 \mid \mathbf{W} = \mathbf{w})$
- One way to model $e(\mathbf{w})$: logistic regression

$$log\left(\frac{e(\mathbf{w})}{1-e(\mathbf{w})}\right)=\alpha+\boldsymbol{\beta}^{T}\mathbf{w}$$



Propensity Score

In general, we can model the propensity scores using more complex structures, e.g., random forest regression, deep neural networks, etc. trained on the observational data to predict assignment probabilities

- Accurate estimates of propensity score roughly translates to accurate inference of the assignment mechanism
- However, the goal is not to optimize the fit of the model, but doing so while ensuring covariate balance (which is necessary for exchangeability)
- Remember: No peeking at the outcomes!

Why do Propensity Scores work?

Why do propensity scores work?

- Individuals with similar covariates get similar scores, and all individuals mapped to a similar score have similar treatment likelihoods.

What if the estimated propensity score is not accurate? (i.e., can't tell who is treated)

- That's ok. The role of the model is to balance covariates given a score; not to actually identify treated and untreated.

Should we try to predict propensity scores perfectly?

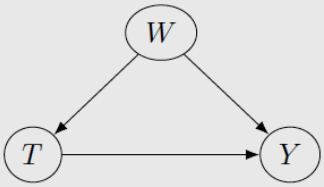
- No! The goal is to use propensity score estimation is to control for confounding and achieve covariate balance
- We must avoid variable selection etc. to optimize propensity score prediction

Balancing

- What if we had 20 covariates, with 4 levels each?
- Over a million million subclasses
- How can we balance across so many covariates?
- One solution: Balance on the propensity score!
- Amazing fact: balancing on just the propensity score balances ALL covariates included in the propensity score model!!!
- We will see why this is the case
- If the above amazing fact is true, we can compare units with similar propensity scores using
  - Stratification
  - Matching
  - Weighting

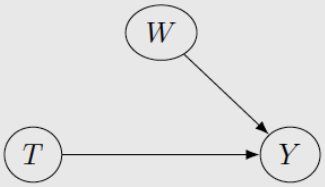
Inverse propensity score weighting: Intuition

Regular population



$P(T \mid W) \neq P(T)$

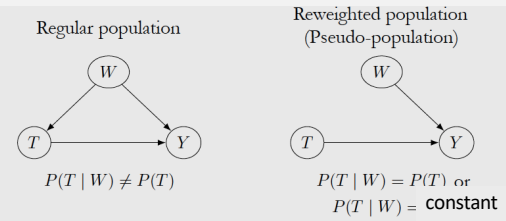
Reweighted population
(Pseudo-population)



$P(T \mid W) = P(T)$ or

$P(T \mid W) =$ constant

Inverse propensity score weighting: Intuition



$$\mathbb{E}[Y(t)] = \mathbb{E}\left[\frac{\mathbb{1}(T=t)Y}{P(t \mid W)}\right]$$

- The effect of W on T is proportional to the propensity score $e(W)$
- We want to neutralize this effect
- One way to do this is to weight samples according to $\frac{1}{e(W)}$
- Note: You don't want propensity scores get close to 0 or 1!
  - Set propensity scores less than ϵ to ϵ (for a small $\epsilon > 0$) and
  - Set propensity scores greater than $(1-\epsilon)$ to $(1-\epsilon)$
  - This can introduce some bias that we have to live with

Inverse propensity weighting: proof

It suffices to show that $\mathbb{E}\left[\frac{YT}{e(W)}\right] = \mathbb{E}[Y(1)]$ and $\mathbb{E}\left[\frac{Y(1-T)}{1-e(W)}\right] = \mathbb{E}[Y(0)]$

$$\mathbb{E}\left[\frac{YT}{e(W)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{YT}{e(W)}|W\right]\right]$$

When $Y = Y(1)$, we have

$$\begin{aligned}\mathbb{E}\left[\frac{YT}{e(W)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{Y(1)T}{e(W)}|W\right]\right] \\ &= \mathbb{E}\left[\frac{\mathbb{E}[Y(1)|W]\mathbb{E}[T|W]}{P(T|W)}\right] \\ &= \mathbb{E}\left[\mathbb{E}[Y(1)|W]\right] \\ &= \mathbb{E}[Y(1)]\end{aligned}$$

Similar argument holds for $\mathbb{E}[Y(0)]$

Inverse propensity score weighting: ATE

$$\mathbb{E}[Y(t)] = \mathbb{E}\left[\frac{\mathbb{1}(T=t)Y}{P(t \mid W)}\right]$$

$$\tau \triangleq \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}\left[\frac{\mathbb{1}(T=1)Y}{e(W)}\right] - \mathbb{E}\left[\frac{\mathbb{1}(T=0)Y}{1-e(W)}\right]$$

$$\hat{\tau} = \frac{1}{n_1}\sum_{i:t_i=1}\frac{y_i}{\hat{e}(w_i)} - \frac{1}{n_0}\sum_{i:t_i=0}\frac{y_i}{1-\hat{e}(w_i)}$$

Inverse propensity score weighting: CATE

- The same idea as for ATE
- except restrict the estimate to data samples where $x_i = x$

$$\hat{\tau}(x) = \frac{1}{n_x} \sum_{i:x_i=x} \left(\frac{\mathbb{1}(t_i = 1) y_i}{\hat{e}(w_i)} - \frac{\mathbb{1}(t_i = 0) y_i}{1 - \hat{e}(w_i)} \right)$$

Combining COM and propensity scores

$$\hat{\tau} = \frac{1}{n} \sum_i [\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i)]$$

$$\hat{\tau} = \frac{1}{n} \sum_i [\hat{\mu}(1, \hat{e}(w_i)) - \hat{\mu}(0, 1 - \hat{e}(w_i))]$$

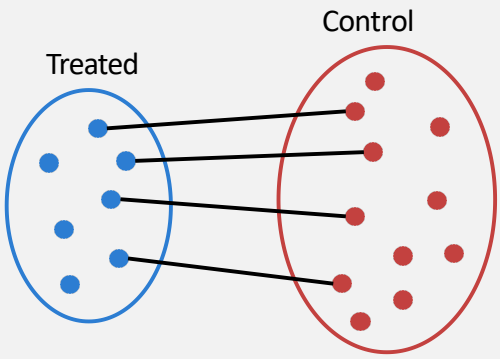
Basic intuition:

- Instead of matching on W, match on (predicted) $e(W)$

Generalizing propensity score-based sample reweighting methods

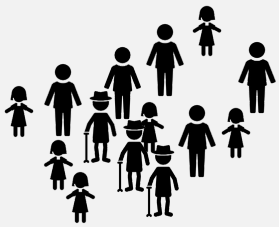
- **Balancing Score**: Balancing score $b(W)$ is a general weighting score, which is the function of covariates W satisfying: $T \perp\!\!\!\perp W \mid b(W)$.
  - Neutralizes the dependence of T on W
  - We will see examples of this later
- Propensity Score is a special case of balancing score

Matching

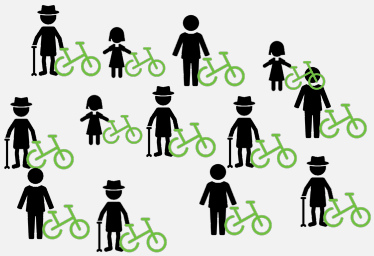


- Matching can be in
  - Original space
  - Learned low-dimensional space (modern representation learning methods)
- Different criteria for "close enough"

Matching and Stratification

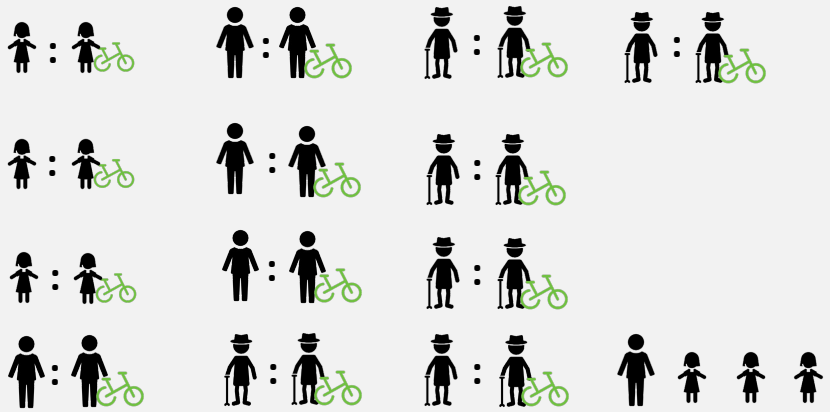


Avg Cholesterol = 200



Avg Cholesterol = 202

Matching and Stratification



Exact Match

Simple:

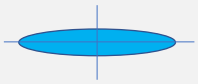
$$Distance(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 0, & \mathbf{x}_i = \mathbf{x}_j \\ \infty, & \mathbf{x}_i \neq \mathbf{x}_j \end{cases}$$

- Use this in low-dimensional settings when overlap is abundant
- But in most cases, there will be too few exact matches …

Mahalanobis Distance

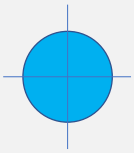
Mahalanobis distance accounts for unit differences by normalizing each dimension by the standard deviation.

$$Mahalanobis(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T S^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$$



And S is the covariance matrix.

Other distance measures may be used.



- Appropriate for low-dimensional settings when overlap is abundant
- But in most cases, there will be too few exact matches …

Matching

- Identify pairs of treated and untreated individuals who are very similar or even identical to each other
  - Very similar ::= $Distance(\mathbf{x}_i, \mathbf{x}_j) < \epsilon$
- Paired individuals provide the counterfactual estimate for each other

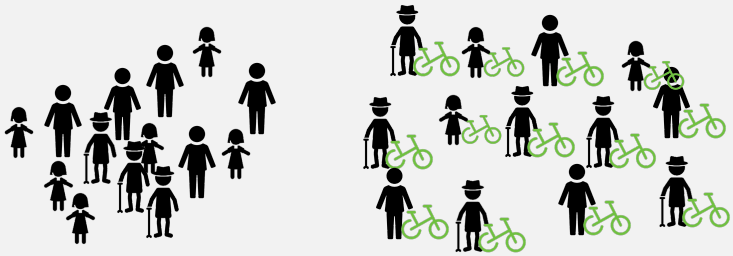
Additional Considerations in Matching

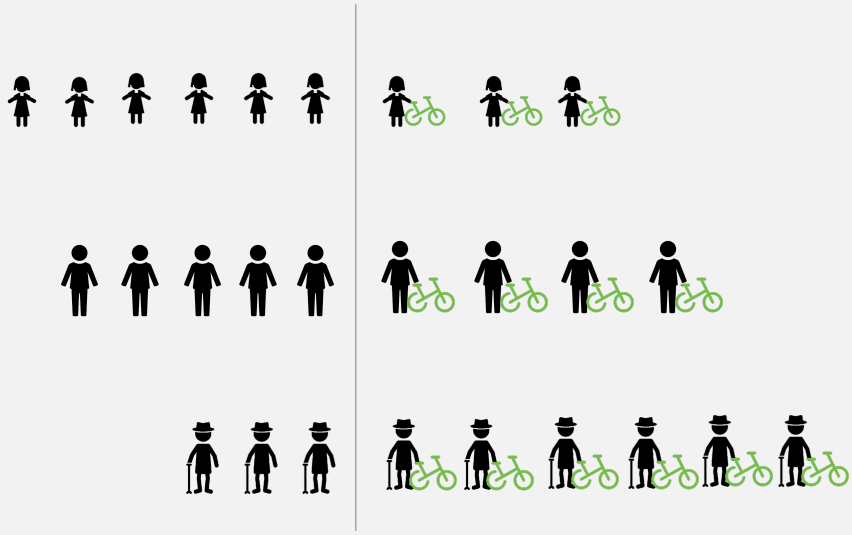
- When matching, should we allow replacement?
  - bias / variance trade-off
- When matching, what if nearest neighbor is far away?
  - Use a "caliper" threshold to limit acceptable distance
- What if not all treated individuals are matched to untreated?
  - This will bias results. Consider redefining original cohort to exclude treated individuals who won't have matches in the untreated population.
- In the simplest case, treatment is binary
  - Advanced variants allow multi-valued, and other treatment regimens

Matching in low-dimensional representation space

- Learn a low-dimensional information preserving mapping of data using representation learning
  - e.g., deep autoencoder networks
- Allow matching methods to be extended to data with complex structure – images, graphs, etc.
- Perform matching in the representation space
- Matching in representation space far more reliable than matching in the original covariate space

From Matching to Stratification





From Matching to Stratification

- 1: 1 matching generalizes to many:many matching.
- Stratification identifies paired *subpopulations* whose covariate distributions are similar.
- There can still be bias, if strata are too large.

How to stratify with propensity score

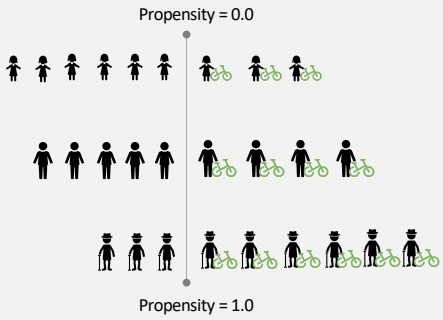
- Train a machine learning model to predict treatment status
  - **Supervised learning:** We are trying to predict a known label (treatment status) based on observed covariates.
  - Conventionally, use a logistic regression model, but any ML model can be used
  - But score must be well-calibrated. i.e., $(100p)\%$ of individuals with score of p are observed to be treated
- Distance is the difference between propensity scores

$$Distance(\mathbf{x}_i, \mathbf{x}_j) = |\hat{e}(\mathbf{x}_i) - \hat{e}(\mathbf{x}_j)|$$

Propensity Score Stratification

We can use propensity score to stratify populations

1. Calculate propensity scores per individual as in matching.
2. But instead of matching, stratify based on score.
3. Calculate average treatment effect as weighted average of outcome differences per strata.
4. Weight by number of treated in the population for ATE on treated.



Propensity Score Stratification

$$ATT = \sum_{s \in strata} \frac{1}{N_{s,T=1}} \left(\bar{Y}_{s,T=1} - \bar{Y}_{s,T=0}\right)$$

where,

$\bar{Y}_{s,T}$ is the average outcome at strata s and treatment status T

And $N_{s,T=1}$ is the number of treated individuals in strata s

Stratification – Practical Considerations

- How many strata do we pick?
  - Scale will depend on data. Want each stratum to have enough data.
  - Conventional, small-data literature (e.g., ~100 data points) picked 5.
  - With 10k to 1M or more data points, we can pick 100 to 1000 strata.
  - Set strata boundaries to split observed population evenly
  - Aside: why not always pick a small number of strata?
    - Bias-variance trade-off…
- What if there aren't enough treated or untreated individuals in some stratum to make a meaningful comparison?
  - This often happens near propensity score 0.0 and near 1.0
  - Drop ("Clip") these strata from analysis.
  - This essentially redefines the cohort

Stratification

Definition Stratification calculates treatment effects by identifying groups of individuals with similar distributions of covariates, where individuals in one group are treated and in the other group are not.

Intuition The difference in average outcome of paired *groups* tells us the effect of the treatment on that subpopulation. Observed confounds are balanced, due to covariate similarity across paired groups.

Example In our cartoon example, we stratified based on propensity score into 3 strata. ATE is the weighted sum of differences in avg outcomes in each strata.

Keep in mind Make sure there are enough comparable individuals in each strata

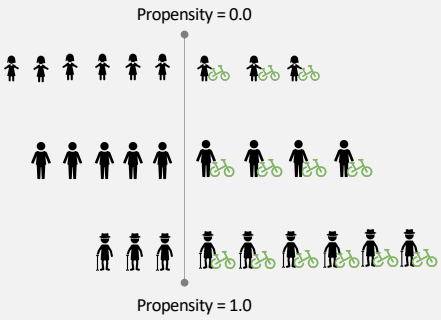
Weighting: An alternative to conditioning

What if we assign weights to observations to simulate randomized experiment?

- Stratification weights strata results by number of treated
- Weighting by treated population ~ weighting by propensity score.

<u>Generalized weighting</u>:

- Calculate effect by weighted sum over all individual outcomes
- Many weighting methods to generate a balanced dataset



Weighting

$$ATE = \frac{1}{N_{T=1}} \sum_{i \in treated} b_i y_i - \frac{1}{N_{T=0}} \sum_{j \in untreated} b_j y_j$$

Inverse Probability of Treatment Weighting (IPTW)

$$ATE = \frac{1}{n} \sum_{i=1}^{n} \frac{t_i y_i}{\hat{e}(w_i)} - \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - t_i) y_i}{1 - \hat{e}(w_i)}$$

Weighting: Caveats and Practical notes

- High variance when e close to 0 or 1
  A single value can derail the estimate.
- Many heuristics for clipping weights; stabilizing weights; etc.
- Assumes propensity score model is correctly specified (i.e., that e is correctly estimated for all individuals)
- Variants of weighting: calculate average treatment effect on treated

Weighting

Definition Weighting calculates average treatment effect as the difference between the weighted sum of the treated and untreated populations

Intuition Weights on each individual act to balance the distribution of covariates in the treated and untreated groups. (i.e., break the dependence between treatment status and covariates)

Keep in mind High variance when propensity scores are very high or very low
Many variants of weighting schemes

Regression (or supervised machine learning)

As we have seen earlier,

- In regression analysis, we build a model of Y as a function of covariates $\mathbf{X}$ and T, and interpret coefficients of X and T causally:

$$E(Y|\mathbf{X}, T) = \alpha_1 X_1 + \alpha_2 X_2 + \cdots \alpha_n X_n + \alpha_T T$$

Model is fit with standard methods (e.g., MLE)

The bigger α_T is, the stronger the causal relationship of T to Y

Regression: Caveats

Causal interpretation of regressions requires many assumptions

Threats to validity include:

- Modeling assumptions : e.g., what if we use a linear model and causal relationship is non-linear
- Multicollinearity: if covariates are correlated, we can't get accurate coefficients
- Omitted variables: Omission of confounders can invalidate findings