Introductory Workshop for Using Propensity Score Weights When Drawing Causal Inferences

Brian Vegetabile, Ph.D.

RAND Corporation

2022 American Causal Inference Conference @ UC Berkeley
May 23rd 2022

Acknowledgements

- This work has been generously supported by NIDA grants 1R01DA034065 and 1R01DA045049
  - https://www.rand.org/statistics/twang.html

- This project and these slides are the product of work by many colleagues

Dan McCaffrey	Beth Ann Griffin
Donna Coffman	Matt Cefalu
Joe Pane	Ricky Sanchez
Craig Martin	Shuangshuang Liu
Marika Suttorp Booth	Lynsay Ayer
Megan Schuler	Geoff Grimm
Mark Godley	Rod Funk
Chuck Stelzner	Katherine Castellano

Goals of this workshop

To increase your understanding of:

- How to define and estimate causal effects
- How to use propensity scores when estimating causal effects
- How to assess the validity of key assumptions

To provide step-by-step instructions for using propensity score weights for binary treatment settings

Approximate Schedule

Time	Subject matter
1:00 to 2:00	Introduction to causal inference, potential outcomes, and propensity scores
2:00 to 2:05	Break
2:05 to 3:10	Propensity Score Weighting in R - binary treatments
3:10 to 3:15	Break
3:15 to 4:00	Additional Topics

Introduction to Causal Inference

Outline

- We'll go over the challenge of confounding in observational studies and how propensity score methods (and weighting in particular) can address it
  - Introduce a motivating case study from the evaluation of a substance use treatment program
- Provide definitions of different causal estimands of interest using the potential outcomes framework
- Provide a high-level overview of mechanics of propensity score estimation
- Present crucial identifying assumptions for causal estimation

Why study causal inference?

Causal effects are of great interest to many fields

- Science is generally interested in isolating **causal effects** and is not typically satisfied with merely observing **correlations**, i.e., the adage that *correlation is not causation*

Causal effects describe the **expected change in outcome** in response to

- A new treatment
- A prevention program
- A new policy
- A new exposure

when compared with some baseline, or referent condition typically taken to be the absence of an exposure or the commonly accepted practice in treatment

Fundamental Challenge of Observational Studies

- Objective: Estimate the effect of a (non-randomized) exposure on an outcome
- A key source of bias in observational studies comes from *selection bias*, i.e., there is a potential for bias when exposure status is correlated with baseline covariates
  - e.g., individuals with more severe substance use tend to enroll in more intensive treatment programs than those with less severe substance use
  - e.g., environmental lead exposure is related to SES
  - e.g., individuals with PTSD are at greater risk for other mental health problems than those without PTSD
- This selection process often means that treated individuals and control individuals look quite different on baseline characteristics, referred to as potential confounders.

Motivating Example

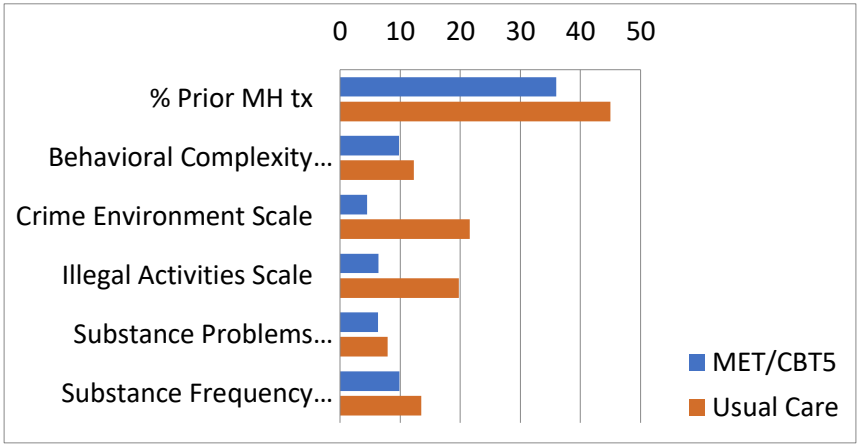
- Observational data on youth enrolled in community-based substance use treatment programs funded by SAMHSA
  - Substance Abuse and Mental Health Services Administration
- As requirement of SAMHSA funding, longitudinal data is collected on youth – substance use behaviors, mental health symptoms, etc.
- Data affords a rich opportunity to assess the relative effectiveness of different treatment modalities in a community setting
- Main challenge: Youth are not randomized, so systematic baseline differences in youth receiving different treatment programs

Motivating Example

- To begin, we will focus on comparing 2 different treatment programs
- **Case study objective:** To estimate the causal effect of MET/CBT5 versus "usual car" on substance use outcomes among adolescent clients (controlling for adolescent characteristics)
  - MET/CBT5: Motivation enhancement therapy plus cognitive behavioral theory in 5 sessions
  - Usual Care: Standard outpatient treatment in community settings

Baseline differences across treatment groups

- Selection bias may occur when individuals receiving different treatment conditions have different characteristics



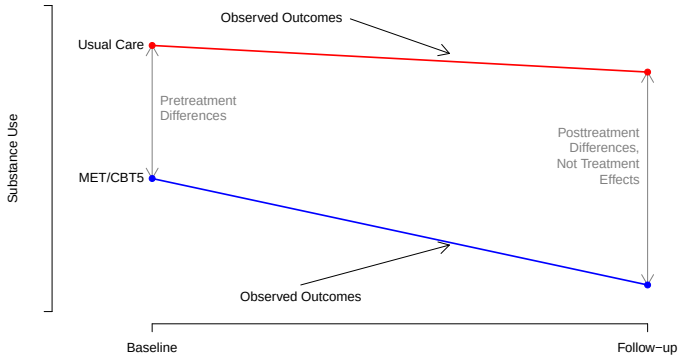
Outcome of interest

We will assess effectiveness of MET/CBT5 vs usual care by comparing substance use frequency at 12-month follow-up

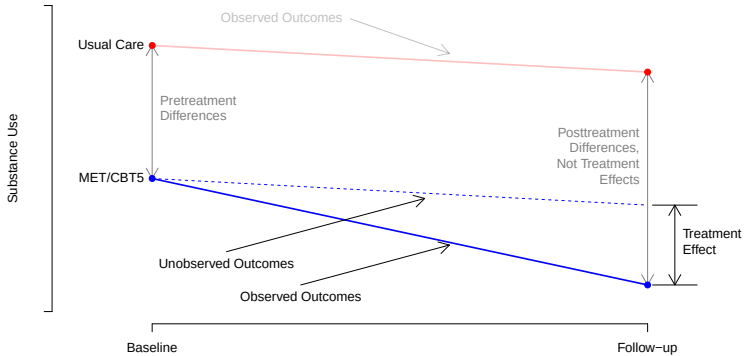
- Mean substance frequency scale (SFS) at 12-months =
  0.11 ($\approx$ 10 days of use per month) for **Usual care**
  0.07 ($\approx$ 6 days of use per month) for **MET/CBT5**
  - Effect size (ES) difference of 0.41

To what extent does higher use at 12-months among youth in usual care youth reflect lower treatment effectiveness vs. youth with more severe substance use were enrolled in usual care?

Observed data: different kids, different treatments



What we would want: same kids, different treatments



Defining the causal estimands of interest using the potential outcomes framework

Notation: Potential Outcomes Framework

- Let A be the exposure of interest ($A = 0$ or $A = 1$)
- Let X represent pre-exposure covariates
- Ideal experiment: Observe each individual under both $A = 0$ and $A = 1$
  - We would want each individual to be their own control
- Reality: Each individual can only be observed under exactly one exposure, not both exposure conditions

Potential Outcomes Framework

- Define: $Y(a)$ as the **potential outcome** when receiving treatment a
- Two potential outcomes for each study participant i:
  - Outcome after receiving treatment $= Y_i(1)$
  - Outcome after receiving comparison treatment $= Y_i(0)$
- Potential outcomes $Y(1)$ and $Y(0)$ will exist for all individuals, regardless of which treatment the individual actually received
  - We will drop subscript i when talking about general effects

Observed Outcomes

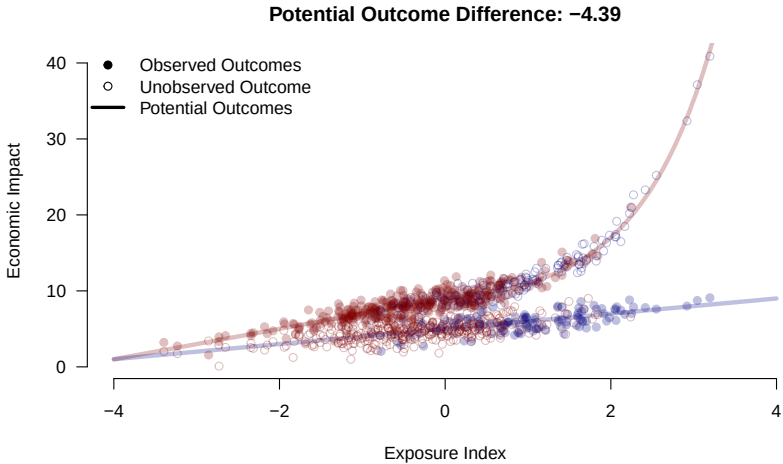
- Unfortunately, we only observe one of these outcomes for each participant
- Define the **observed outcome** as

$$Y \equiv Y^{obs} = \begin{cases} Y(1) & \text{if } A = 1 \\ Y(0) & \text{if } A = 0 \end{cases}$$

- Alternatively, we can write (this is often referred to as a consistency assumption)

$$Y = AY(1) + (1 - A)Y(0) \tag{1}$$

Illustrating the problem



Potential Outcomes Framework

- Ideal estimand of interest is:

$$\tau_i = Y_i(1) - Y_i(0)$$

  a treatment effect for an *individual person*
- As noted, we only observe one outcome for each individual,
- We will focus instead on estimating average treatment effects in some population $\mathcal{P}_X$ represented by covariates X, e.g.,

$$E_{\mathcal{P}_X}[Y(1) - Y(0)]$$

- The focus will be understanding how we estimate this from observed data

Causal Estimands: ATE

Average treatment effect in the population (ATE)

- ATE estimand defined as: $E_X[Y(1) - Y(0)]$
- Aims to answer the question:
  - How effective is the treatment in the population? (If you have a control condition)
  - What is the relative effectiveness of two treatments on average in the population? (If you have 2 treatment conditions)

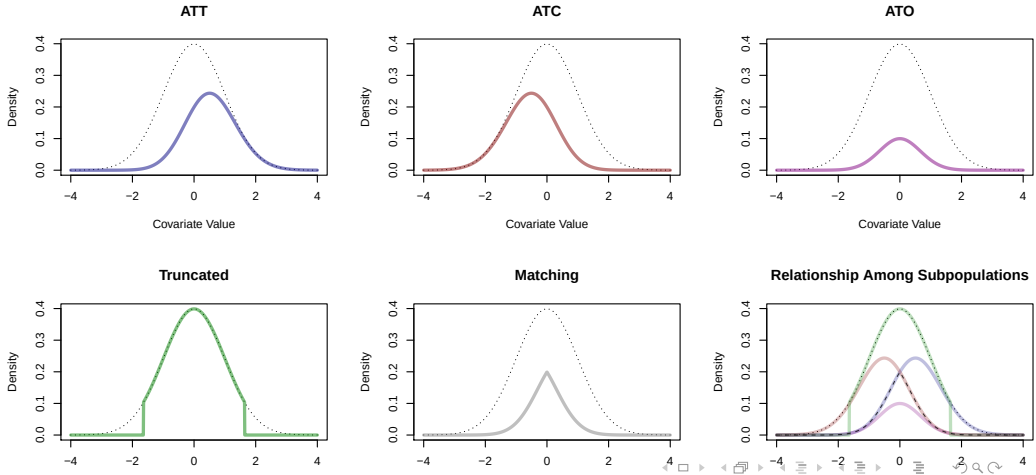
Causal Estimands: ATT

Average treatment effect in the treated population (ATT)

- ATT estimand defined as: $E_{X|A=1}[Y(1) - Y(0)|A = 1]$
- Aims to answer the question:
  - How effective is the treatment in the **population that is most similar to the treated population**

Many other estimands we could define

- Li, Fan, Kari Lock Morgan, and Alan M. Zaslavsky. "Balancing covariates via propensity score weighting." Journal of the American Statistical Association 113, no. 521 (2018): 390-400.



Causal Estimands: Case Study

Average treatment effect in the population (ATE)

- Answers the question:
  - What is the relative effectiveness of MET/CBT5 versus usual care on average in the population?

Average treatment effect in the treated population (ATT)

- Answers the question:
  - What is the relative effectiveness of MET/CBT5 versus usual care on average in the population who had received MET/CBT5?

Potential Outcomes and Assigned Treatment

- In the potential outcome framework, $Y(1)$ and $Y(0)$ exist for all individuals in the population and we are interested in inferences about contrasts of them
- Thus, three sets of distributions to answer our motivating questions will be important:
  1. Distributions for the entire populations (with density functions $p(\cdot)$)

$$p\left(Y(1)\right) \qquad p\left(Y(0)\right) \qquad p\left(Y(1), Y(0)\right)$$

  2. Distributions for the subsample where $A = 1$

$$p\left(Y(1)|A=1\right) \qquad p\left(Y(0)|A=1\right) \qquad p\left(Y(1), Y(0)|A=1\right)$$

  3. Distributions for the subsample where $A = 0$

$$p\left(Y(1)|A=0\right) \qquad p\left(Y(0)|A=0\right) \qquad p\left(Y(1), Y(0)|A=0\right)$$

Potential Outcomes, Treatment Assignment, and Selection

- A key source of bias in observational studies comes from *selection*, i.e., when there are preferences for treatment assignment based on pre-exposure covariates
  - For example, this could be a doctor selecting a treatment level based on the patients medical history
- Selection causes the conditional distributions of individuals with $A = 1$ to look different from individuals with $A = 0$
  - For all other sets of covariates

What we mean by different: defining covariate balance

- A measure of **covariate balance** will be a comparison of the densities

$$p(X|A=1) \text{ and } p(X|A=0)$$

- Define a **balance function** $\mathcal{B}(X, A)$ as any function of the observed data that assessed balance, e.g., the standardized difference in means (or a weighted analog) could be defined as

$$\Delta = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{\frac{1}{2}(s_1^2 + s_0^2)}}$$

- Balance functions will be explored more later

Potential Outcomes, Treatment Assignment, and Selection

- Imbalances also imply that the distributions of the potential outcomes are *different* depending on treatment level, i.e.,

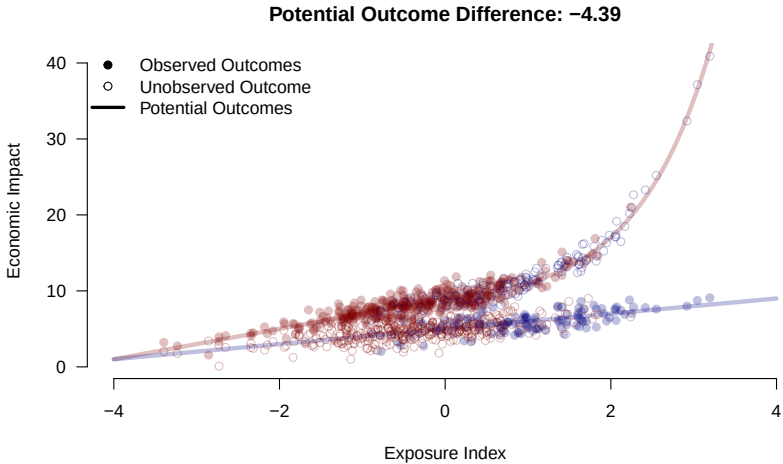
$$p(Y(0), Y(1)|A = 1) \neq p(Y(0), Y(1)|A = 0)$$

- This selection problem will ultimately imply that we cannot take simple mean differences in the treatment and control groups

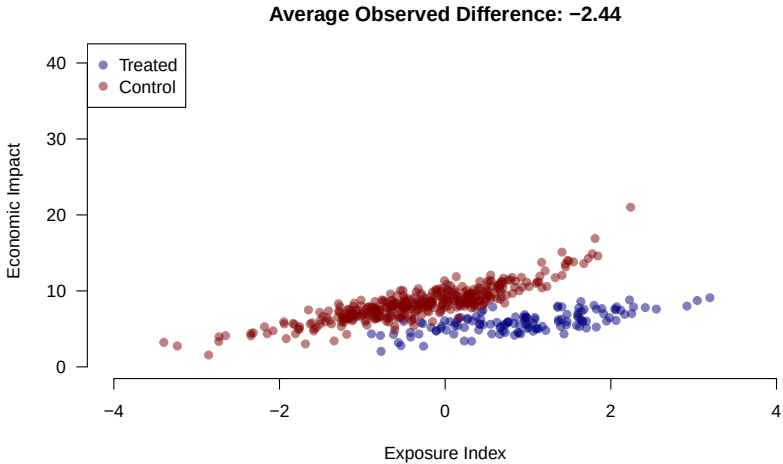
$$\bar{Y}_{A=1} - \bar{Y}_{A=0} = \frac{1}{N_1} \sum_{i:A_i=1} Y_i - \frac{1}{N_0} \sum_{i:A_i=0} Y_i = \frac{\sum_i A_i Y_i}{\sum_i A_i} - \frac{\sum_i (1 - A_i) Y_i}{\sum_i (1 - A_i)}$$

as these could be biased

Illustrating the problem



Illustrating the problem



Challenge of Causal Effect Estimation - The ATE Example

- The average treatment effect is the expected difference in potential outcomes across the entire population or sample
  - $ATE = E[Y(1) - Y(0)]$
- Easy to show that the ATE equals the average of all individuals' potential outcomes under treatment minus the average of all individuals' potential outcomes under control
  - $E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)]$

Challenge of Causal Effect Estimation - The ATE Example

- Those averages include the potential outcomes under treatment for individuals who receive treatment ($A = 1$) and those who do not ($A = 0$)
  - $E[Y(1)] = E[Y(1)|A = 1]\pi + E[Y(1)|A = 0](1 - \pi), \quad \pi = Pr(A = 1)$
  - $E[Y(0)] = E[Y(0)|A = 1]\pi + E[Y(0)|A = 0](1 - \pi)$
- Estimating the treatment effect requires estimates of
  - $E[Y(1)|A = 1]$, $E[Y(1)|A = 0]$, $E[Y(0)|A = 1]$, and $E[Y(0)|A = 0]$
- Two distributions that we do not observe

Challenge of Causal Effect Estimation - The ATE Example

- We can directly estimate $E[Y(1)|A=1]$ and $E[Y(0)|A=0]$
- But if pretreatment differences in the distributions of individuals exist
- Then the average potential outcome under treatment for individuals who receive the treatment will not equal the average potential outcome under treatment for individuals who do not receive treatment
  - $E[Y(1)|A=1] \neq E[Y(1)|A=0]$ and $E[Y(0)|A=1] \neq E[Y(0)|A=0]$
  - Similar results hold for potential outcomes under control
- Need to make additional assumptions to estimate $E[Y(1)|A=0]$ and $E[Y(0)|A=1]$

Solving the Fundamental Problem of Causal Inference

- **Key challenge:** Estimating the means for the unobserved potential outcomes

$$E[Y(1)|A=0] \qquad \text{and} \qquad E[Y(0)|A=1]$$

- **Typical path forward**:
  Find or create groups where the distributions of the expected values of $Y(1)$ are the same for individuals who receive treatment $(A=1)$ and individuals who receive control $(A=0)$
  - Similarly, the distributions or expected values of $Y(0)$ are the same for individuals who receive treatment and individuals who receive control
- Within these groups potential outcomes are unrelated to treatment assignment
  - This is approach and set of assumptions is known as **strong ignorability**

Mathematical Definition of Strong Ignorability Given Covariates

The following conditions were originally outlined in Rosenbaum and Rubin (1983) as **strong ignorability**

- Condition 1: The distribution of potential outcomes, $(Y(0), Y(1))$, is the same for individuals with $A = 1$ and $A = 0$ given a value of X, i.e., formally

$$(Y(0), Y(1)) \perp\!\!\!\perp A \mid X$$

- Effectively, it will provide the following mathematical property

$$E[Y(a)|A = 1, X = x] = E[Y(a)|A = 0, X = x] = E[Y(a)|X = x]$$

Mathematical Definition of Strong Ignorability Given Covariates

The following conditions were originally outlined in Rosenbaum and Rubin (1983) as **strong ignorability**

- Condition 2: There is positive probability of receiving each treatment level, i.e., the **positivity** assumption where,

$$0 < Pr(A = 1|X = x) < 1$$

for any X with "sufficient" density in our population of interest

Mathematical Definition of Strong Ignorability Given Covariates

- The second condition contains an object we'll spend a significant amount of time studying today
- Define the **propensity score** as

$$e(x) = Pr(A = 1|X = x)$$

- The propensity score is also often referred to as an **assignment mechanism**, or the process by which individuals are assigned exposure levels

What do we get with ignorability?

Ignorability will allow us to control for variables by first conditioning on X and then use this to estimate population parameters by integrating over a distribution of interest

$$
\begin{aligned}
& E_X[E[Y|A=1,X=x] - E[Y|A=0,X=x]] \\
& E_X[E[Y(1)|A=1,X=x] - E[Y(0)|A=0,X=x] && \text{(Consistency, or SUTVA)} \\
=& E_X[E[Y(1)|X=x]] - E_X[E[Y(0)|X=x]] && \text{(Ignorability)} \\
=& E_X[E[Y(1) - Y(0)|X=x]] \\
=& E[Y(1) - Y(0)]
\end{aligned}
$$

Positivity ensures we can estimate both $E[Y|A=a,X=x]$ expectations

Methods to Achieve Strong Ignorability Given Covariates

- By experimental study design: (conditionally) randomized experiments
- By assumptions on observations (often untestable):
  - Define groups on the basis of observed covariates and assume strong ignorability holds within these groups
    - Modeling assignment or selection (propensity scores)
  - Regression discontinuity
  - Instrumental variables (IV)

Completely Randomized Experiments

- Collect group of participants and randomly assign to treatment or comparison
  - Researchers control the assignment mechanism **by design**
- Therefore, through design and control, assignment is unrelated to any characteristics of the participant, i.e.,

$$Pr(A = 1|X = x) = Pr(A = a) = c$$

- Provides a random sample of $Y(1)$'s and a random sample of $Y(0)$'s, *that is representative from the population*

Completely Randomized Experiments

- Because of randomization
  - $E(Y(1)|A=1) = E(Y(1)|A=0)$ and $E(Y(0)|A=1) = E(Y(0)|A=0)$
  - $(Y(0), Y(1))$ is independent of treatment assignment A **by design**
- Any realized assignment might result in groups with differences but on average the procedure will provide an unbiased estimate of $E(Y(0))$ and $E(Y(1))$
  - The behavior of differences can be characterized through probability distributions

Experiments vs Observational Studies

- Randomized experiments = gold standard for estimating causal effects
  - Randomization (if it works) makes groups being compared balanced on baseline characteristics
  - Treatment assignment is unrelated to potential outcomes (**strong ignorability**)
- Randomization is not always feasible
- Observational studies provide another way to get at causal effects
  - Treatment assignment is not controlled by the researcher
  - Groups being compared are usually imbalanced
  - Can use causal inference methods to try to replicate what a random study does

Estimation and Propensity Score Fundamentals

Estimation - Simple Nonparametric Estimators

- Recall the estimator below is biased

$$\frac{\sum_i A_i Y_i}{\sum_i A_i} - \frac{\sum_i (1 - A_i) Y_i}{\sum_i (1 - A_i)}$$

- Let $w = (w_1, w_0)$ be a set of weights for the treated and control groups respectively
- Our goal will be to use the propensity score to derive weights so that we can form simple estimators to estimate treatment effects:

$$\hat{\tau}_{\mathcal{P}_X} = \frac{\sum_i w_{1,i} A_i Y_i}{\sum_i w_{1,i} A_i} - \frac{\sum_i w_{0,i} (1 - A_i) Y_i}{\sum_i w_{0,i} (1 - A_i)}$$

- Or other more complex weighting estimators
  - Weighted regression estimators
  - "Doubly robust" estimators

How will the propensity scores help us deal with selection?

- Propensity score methods are a statistical approach to controlling for potential confounding in order to obtain unbiased effect estimates
  - Alternative: regression modeling assumes strong ignorability conditional on the values of the covariates in the model $\rightarrow$ sensitive to functional forms
- The propensity score, $e(x)$, is an individual's probability of receiving the treatment, A, given their pre-treatment characteristics, X

$$e(x) = Pr(A = 1|X = x)$$

  - Used to statistically create **balance** across treatment groups
  - Conditioning on $e(x)$ can balance distributions of X between treatment and comparison groups

Propensity Scores Help us Deal with the Challenge of Selection

- Lasting impact (in my opinion) of RR83 is the following property
- If **strong ignorability** given X holds, then strong ignorability given $e(X)$ also holds
- That is, the propensity score is sufficient to control for observed pretreatment differences between groups
  - Treatment assignment is **strongly ignorable** given the propensity score
    - $(Y(1), Y(0)) \perp\!\!\!\perp A | e(X)$ - e.g, no unobserved confounders
    - $0 < e(x) < 1$ - e.g, overlap between groups
- Provides dimensionality reduction for the problem
  - If X is dimension d, the propensity score is one dimensional

The Propensity Score and Strong Ignorability

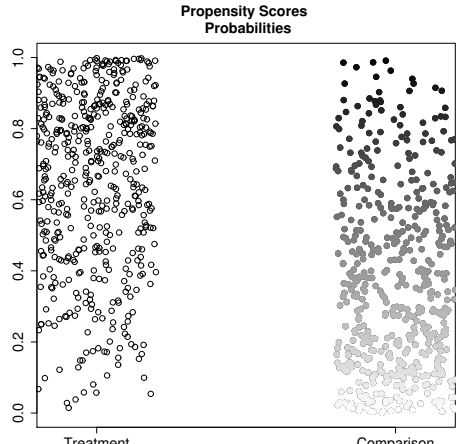
- If distribution of $(Y(0), Y(1))$ is the same for individuals with $A = 1$ and $A = 0$ at each value of X, then it is the same for individuals with $A = 1$ and $A = 0$ at each value of $e(x)$
  - If $(Y(0), Y(1)) \perp\!\!\!\perp A|X$ then $(Y(0), Y(1)) \perp\!\!\!\perp A|e(x)$
- If treatment is strongly ignorable with respect to X (i.e., the distribution of $(Y(0), Y(1))$ is the same for individuals with $A = 1$ and $A = 0$ and at each value and $0 < e(x) < 1$ for all X), then for every value of $e(x)$, the expected value of $Y(1)$ for individuals with $e(x)$ and $A = 1$ is equal to the expected value of $Y(1)$ for individuals with $e(x)$ and $A = 0$ and the same for $Y(0)$

General Framework for a Simple Propensity Score Analysis

1. Select estimand: ATT, ATE, ATO, or others
2. Model and Estimate propensity scores
3. Assess diagnostics to ensure propensity scores are well-estimated
4. Calculate treatment effect by conditioning on the propensity score in some way
   - Stratification
   - Matching
   - **Weighting**

Propensity score weighting intuition

- Propensity score weighting effectively "downweights" some observations that do not provide good matches (in terms of covariates) and "upweights" observations that are good matches



Propensity Score Weights for Different Causal Effects

- Define two weight functions $w_1(x)$ and $w_0(x)$
- Let the analytic weights be defined as

$$w_i = A_i w_1(X_i) + (1 - A_i) w_0(X_i)$$

- These weights allow us to use the estimator (and others)

$$\hat{\tau}_{\mathcal{P}_X} = \frac{\sum_i w_1(X_i) A_i Y_i}{\sum_i w_1(X_i) A_i} - \frac{\sum_i w_0(X_i)(1 - A_i) Y_i}{\sum_i w_0(X_i)(1 - A_i)}$$

  to obtain consistent estimation
- *Note I've been very particular about how I've written the above to demonstrate how all of the data is involved in estimating the effect*

Weights for Appropriate Estimands

Table 1 from *Balancing Covariates via Propensity Score Weighting*

Target Population	Estimand	Weight $w_1(x), w_0(x)$
Combined	ATE	$\left(\frac{1}{e(x)}, \frac{1}{1-e(x)}\right)$
Treated	ATT	$\left(1, \frac{e(x)}{1-e(x)}\right)$
Control	ATC	$\left(\frac{1-e(x)}{e(x)}, 1\right)$
Overlap	ATO	$(1-e(x), e(x))$
Truncated Combined		$\left(\frac{I(\delta < e(x) < 1-\delta)}{e(x)}, \frac{I(\delta < e(x) < 1-\delta)}{1-e(x)}\right)$
Matching		$\left(\frac{\min\{e(x), 1-e(x)\}}{e(x)}, \frac{\min\{e(x), 1-e(x)\}}{1-e(x)}\right)$

A key contribution of Li et al. (2018) was defining the overlap weights

Why this generally works: the simple case

- Assume ignorability and that $Y(a) = f_a(x)$
- Assume the weight functions are estimating the ATE and therefore

$$w_a(x) = \frac{p(a)}{p(a|x)}$$

- It is easy to show that

$$\frac{p(a)}{p(a|x)} = \frac{p(a)}{p(a|x)}\frac{p(x)}{p(x)} = \frac{p(a)p(x)}{p(a,x)} = \frac{p(x)}{p(x|a)}\frac{p(a)}{p(a)} = \frac{p(x)}{p(x|a)}$$

Why this generally works: the simple case

When formed appropriately, the weights can imply that we're estimating with respect to a different density,

$$\begin{aligned} E[w_a(x)Y|A=a] &= \int \frac{p(a)}{p(a|x)} f_a(x) p(x|a) \partial x \\ &= \int \frac{p(x)}{p(x|a)} f_a(x) p(x|a) \partial x \\ &= \int f_a(x) p(x) \partial x \\ &= E[Y(a)] \end{aligned}$$

Why this generally works: the simple case

When formed appropriately, **the weights will also provide balance**,

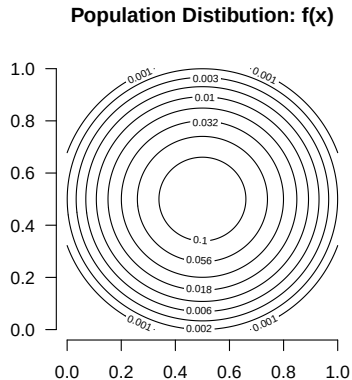
$$
\begin{aligned}
&E[w_a(x)h(X)|A=a] \\
&= \int \frac{p(a)}{p(a|x)} h(x)p(x|a)\partial x \\
&= \int \frac{p(x)}{p(x|a)} h(x)p(x|a)\partial x \\
&= \int h(x)p(x)\partial x \\
&= E[h(X)]
\end{aligned}
\qquad
\begin{aligned}
&E[w_{a'}(x)h(X)|A=a'] \\
&= \int \frac{p(a')}{p(a'|x)} h(x)p(x|a')\partial x \\
&= \int \frac{p(x)}{p(x|a')} h(x)p(x|a')\partial x \\
&= \int h(x)p(x)\partial x \\
&= E[h(X)]
\end{aligned}
$$

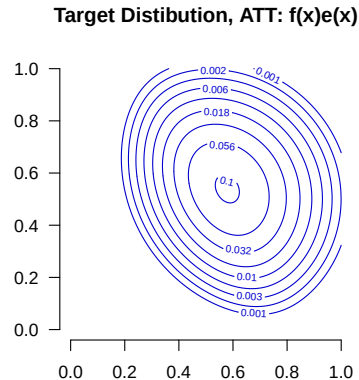
Formalized in:

- Li, Fan, Kari Lock Morgan, and Alan M. Zaslavsky. "Balancing covariates via propensity score weighting." Journal of the American Statistical Association 113, no. 521 (2018): 390-400.

Visualizing with an example: the ATT

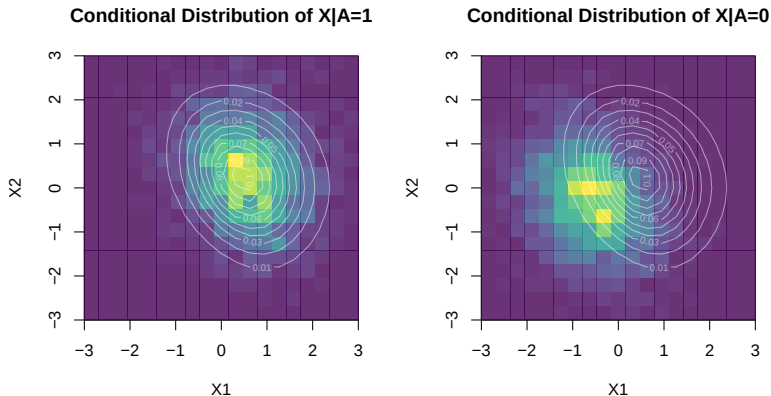
Let $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,1)$, and let $e(x) = \text{logit}^{-1}(2X_1 + X_2)$





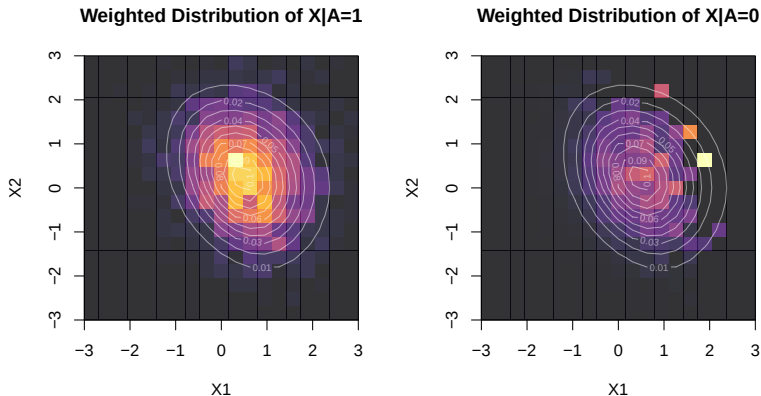
Visualizing with an example: the ATT

Let $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,1)$, and let $e(x) = \text{logit}^{-1}(2X_1 + X_2)$ for a sample of $N = 5000$ and a goal of estimating the ATT



Visualizing with an example: the ATT

Using the appropriate weight function and summing the weights in each bin → weighted distribution is similar to target distribution



Defining Balancing Weights

We can define a set of functions $(w_0(x), w_1(x))$ as **balancing weight functions** if

$$E[w_1(x)h(X)|A = 1] = E[w_0(x)h(X)|A = 0]$$

for any function $h(\cdot)$.

The analytic weights of the form

$$w = Aw_1(X) + (1 - A)w_0(X)$$

are defined to be **balancing weights**

Formalized in:

- Li, Fan, Kari Lock Morgan, and Alan M. Zaslavsky. "Balancing covariates via propensity score weighting." Journal of the American Statistical Association 113, no. 521 (2018): 390-400.

Propensity Score and Balance

- Propensity score weighting can undo the differential sampling of individuals with different values of covariates and make the weighted distributions of covariates the same for treatment and control
- Covariate distributions that are the same between treatment and control groups are “balanced”

Remember: Propensity Scores are Intermediate Outcomes

- Estimate propensity scores as a means to estimating treatment effects
- Goal is not to make the most accurate inference about predictors of treatment but **to make the most accurate inferences about the effects of treatment**

Deciding on Good Balance

- If the propensity scores are adequate, the pretreatment covariates should balance
- Checking the balance is used to determine if the propensity score model is sufficient
  - t-tests for each covariate and higher order terms (not recommended)
  - "Standardized bias" for each covariate and higher order term and other metrics of distance in probability spaces

Standardized Difference in Means

Define the weighted expectation of X in group $A = a$ as

$$\tilde{X}_a = \frac{\sum_i w_{a,i} I(A_i = a) X_i}{\sum_i w_{a,i} I(A_i = a)}$$

and let the standard deviation of X in the target population $\mathcal{P}_X$ be defined as $\sigma_{\mathcal{P}_X}$

The **weighted standardized difference in means** is defined as

$$\Delta = \frac{\tilde{X}_1 - \tilde{X}_0}{\sigma_{\mathcal{P}_X}}$$

Other interesting functions to balance

Define the weighted expectation of $h(X)$ in group $A = a$ as

$$\tilde{h}_a(X) = \frac{\sum_i w_{a,i} I(A_i = a) h(X_i)}{\sum_i w_{a,i} I(A_i = a)}$$

and let the standard deviation of $h(X)$ in the target population $\mathcal{P}_X$ be defined as $\sigma^h_{\mathcal{P}_X}$

The **weighted standardized difference** is defined as

$$\Delta = \frac{\tilde{h}_1(X) - \tilde{h}_0(X)}{\sigma^h_{\mathcal{P}_X}}$$

This can be used to balance other characteristics of the distributions, such as higher moments like the variance of X in the target population

On the difference between balance and significance: problems with using the t-test to assess balance

- With many covariates we should expect some "significant" differences by chance
- The t-statistic is not only a function of bias but also variance

$$t = \frac{\bar{x}_{\text{treat}} - \bar{x}_{\text{control}}}{\sqrt{\frac{s^2_{\text{treat}}}{n^*_{\text{treat}}} + \frac{s^2_{\text{control}}}{n^*_{\text{control}}}}}$$

- n^* is the effective sample size and is a function of the variability in the weights (more on this later)
- When t is small it may indicate that we have eliminated bias, or that we have weighted or so finely stratified so that we have no power

Putting it all together:
Details of A Simple Propensity Score Analysis

Steps of a Common Approach

1. Fit a model for the propensity score
   - Include pretreatment covariates that have a significant bivariate relationship with treatment assignment (possibly identified via prior research and theory or practice) and the outcome of interest
   - One approach would be to model the assignment mechanism using logistic regression:
     - Curate a list of variables from domain experts to identify important variables
     - Use stepwise regression to refine the set of pretreatment covariates

Steps of a Common Approach

1. Fit a model for the propensity score
2. Create weights that match an estimand of interest
   - If the estimand is the ATE for example,

$$w_i = \frac{A_i}{\hat{e}(X_i)} + \frac{1 - A_i}{1 - \hat{e}(X_i)}$$

   - If the estimand is the ATT for example,

$$w_i = A_i + (1 - A_i)\frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)}$$

   - For the ATT, you can see that the the propensity score weights will essentially drop unmatched control subjects for ATT

Steps of a Common Approach

1. Fit a model for the propensity score
2. Create weights that match an estimand of interest
3. Assess balance for each covariate in the data set
   - Calculate weighted and unweighted standardized difference in means
   - Assess balance on higher moments like variance:
     - Huang MY, Vegetabile BG, Burgette LF, Setodji C, Griffin BA. Higher Moments Matter for Optimal Balance Weighting in Causal Estimation. Epidemiology (Cambridge, Mass.). 2022 Apr 12.
   - Good balance is typically when both are small (standardized differnce in means less than 0.1)

Steps of a Common Approach

1. Fit a model for the propensity score
2. Create weights that match an estimand of interest
3. Assess balance for each covariate in the data set
   - If **not** balanced, return to step 1
   - Otherwise, continue

Steps of a Common Approach

1. Fit a model for the propensity score
2. Create weights that match an estimand of interest
3. Assess balance for each covariate in the data set
   - If **not** balanced, return to step 1
   - Otherwise, continue
4. Estimate treatment effect in weighted data set
   - Nonparametric weighted estimator
   - Weighted regression
   - Doubly-robust estimation

Estimation - Simple Nonparametric Estimators

- Let $w = (w_1, w_0)$ be a set of weights for the treated and control groups respectively
- A simple nonparametric weighting estimator is of the form

$$\hat{\tau}_{\mathcal{P}_X} = \frac{\sum_i w_{1,i} A_i Y_i}{\sum_i w_{1,i} A_i} - \frac{\sum_i w_{0,i}(1 - A_i) Y_i}{\sum_i w_{0,i}(1 - A_i)}$$

- **Sensitive to:** specification and estimation of the propensity score

Estimation - Regression Estimation

- Let $w = Aw_1 + (1 - A)w_0$ be a vector of weights for each unit and let $W = \text{diag}(w)$ be a diagonal matrix collecting the weights
- Define $Z = (A\ X)$ or potentially some transformation and interactions of variables
- A regression estimator could be of the form

$$\hat{\beta} = (Z^T W Z)^{-1} Z^T W Y$$

  and estimated effects appropriately defined based on elements of $\hat{\beta}$
- **Sensitive to:** specification and estimation of the propensity score and the outcome model $\rightarrow$ need to specify one right
- Can use the `survey` package in `R` for estimation and standard errors

Doubly-Robust Estimation

- Funk, Michele Jonsson, Daniel Westreich, Chris Wiesen, Til Stürmer, M. Alan Brookhart, and Marie Davidian. "Doubly robust estimation of causal effects." American journal of epidemiology 173, no. 7 (2011): 761-767.
- Glynn, Adam N., and Kevin M. Quinn. "An introduction to the augmented inverse propensity weighted estimator." Political analysis 18, no. 1 (2010): 36-56.
- Kurz, Christoph F. "Augmented Inverse Probability Weighting and the Double Robustness Property." Medical Decision Making (2021): 0272989X211027181.
- Kang, Joseph DY, and Joseph L. Schafer. "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data." Statistical science (2007): 523-539.

Funk et al. (2011)

Overview of Doubly Robust Estimation of Causal Effects

- "Doubly robust estimation combines a form of outcome regression with a model for the exposure (i.e., the propensity score) to estimate the causal effect of an exposure on an outcome."
- Essentially the estimator will be consistent if either outcome regression model or the propensity score model are correct

DR Estimator of Funk et al.

Let A be a binary exposure, Y an observed outcome, X represent pre-exposure covariates and $e(x) = Pr(A = 1|X = x)$ be the propensity score

Assume models for the potential outcomes, i.e., $\hat{Y}(a) = \hat{f}(a, x)$, and a model for the propensity score $\hat{e}(x)$

Define,

$$DR_i^1 = Y_i \frac{A_i}{\hat{e}(X_i)} - \hat{f}(1, X_i) \frac{A_i - \hat{e}(X_i)}{\hat{e}(X_i)}$$

$$DR_i^0 = Y_i \frac{1 - A_i}{1 - \hat{e}(X_i)} + \hat{f}(0, X_i) \frac{A_i - \hat{e}(X_i)}{1 - \hat{e}(X_i)}$$

$$\hat{\tau}_{DR} = n^{-1} \sum_i (DR_i^1 - DR_i^0)$$

Summary of assumptions for estimation

Assumptions

Strong Ignorability: Distribution of the potential outcomes and treatment assignment A are conditionally independent given pre-exposure covariates

$$\mathbf{Y}_a \perp\!\!\!\perp A \mid X$$

- Namely, treatment assignment can be fully explained by observable covariates
- If treatment assignment depends on unobservables, then propensity score methods will not fully adjust for selection bias – effect estimates may still be biased due to unobserved confounders
- Assumption cannot be tested directly with data

Assumptions

Positivity: All individuals have a positive probability of receiving treatment and control:

$$0 < e(x) < 1$$

- Violation: e.g., there are no women in the treatment group
- This cannot be proven/disproved in practice
- Simple summary statistics can help assess plausibility
  - Check for 0/empty cells with binary and categorical covariates in the treatment and control groups
  - Compare the minimums and maximums for continuous covariates between the treatment and control groups

Assumptions

Stable Unit Treatment Value Assumption (SUTVA):

- Single (stable) version of each treatment level: all individuals in the same treatment level receive the same treatment
- No treatment "interference:" An individual's outcome is not affected by other individuals' treatment status (no spill-over, treatment diffusion, etc.)
- Assumption cannot be tested directly with data

Review of Causal Effect Estimation

- An assumption of strong ignorability provides that for individuals with the same values of the covariates, the distributions of both treatment and control potential outcomes are the same for cases assigned to treatment and control
- Strong ignorability given covariates implies strong ignorability given a balancing score, of which the propensity score is one
- Use propensity scores, probability of treatment assignment, to create treatment and control groups with similar covariates so the control group can estimate the counterfactual for the treatment group and vice versa

Weight Estimation in Practice
And an applied example

Propensity score model variable selection

- All variables in the propensity score model must be measured pre-treatment (otherwise possibly on causal path)
- Priority: Covariates associated with treatment and outcome (i.e., confounders)
  - Possibly identified via prior research or theory
- Including Xs that predict only the outcome can help to reduce variance of treatment effect estimates
- Avoid Xs that predict only treatment but not outcome (commonly called *instruments*) - can contribute to bias and inflate variance of treatment effect estimates

Limitations of Common Approach of Estimating Propensity Scores with Logistic Regression

- Which variables to include in PS model?
- Functional form of variables in PS model? (e.g., interactions)
- In logistic regression models, user has to manually implement models in iterative manner which is time consuming and difficult to automate.
- Regression-based significance testing may not indicate good balance across treatment groups

GBM is a flexible alternative approach

- Use a *generalized boosted model* (GBM), a nonparametric machine learning method, to flexibly estimate the propensity scores
- **A departure from normal ML model selection:** Use balance on the covariates to guide model selection
- Fitting algorithm can select among variables to include in model and add interaction and nonlinearity to final function of covariates

GBM

- Combines many piecewise-constant functions of the covariates to estimate the unknown probability function
  - Use of piecewise constant functions was motivated by regression trees
- Use of piecewise constant functions means that the model can include nominal, discrete, and continuous covariates and covariates with missing values without special instructions
- Fit is invariant to monotone transformations of covariates
- Fitting algorithm automatically selects best subset of possible piecewise functions
  - Selection of the piecewise functions automatically selects covariates for inclusion and functional form including interactions

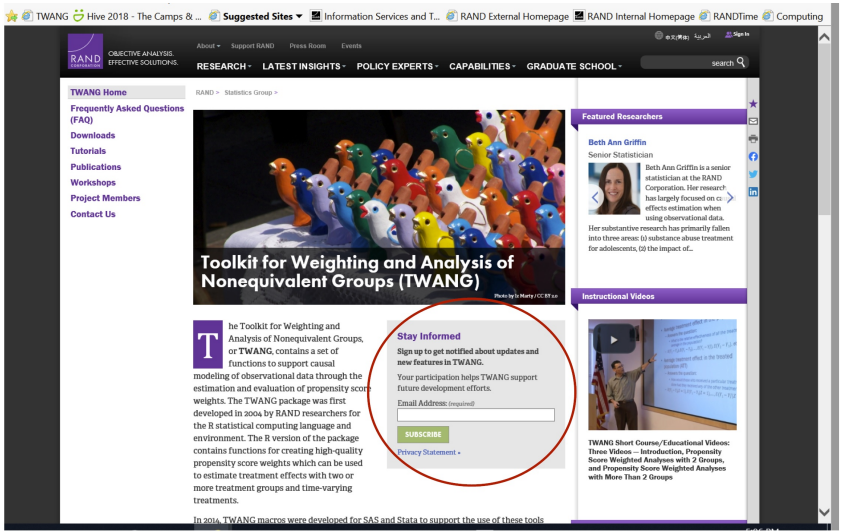
GBM Estimation

- Model fitting occurs through an automated iterative procedure
- Each iteration makes the model more complex by including more variables or making the functional form for included variables more elaborate (i.e., including interactions)
  - Too few iterations and the model misses important relationships between covariates and treatment assignment
  - Too many iterations and the model overfits without finding general patterns about treatment assignment

As an iterative algorithm, GBM needs stopping criteria

- When implementing GBM, must specify "stopping rule" to determine which model iteration will be selected
- Many different stopping rule metrics – selection based on context
- Since our objective is to obtain maximal covariate balance across treatment groups, stopping rules implemented in `twang` are metrics of covariate balance
- `twang` runs many GBM iterations, assesses balance at each iteration, and then picks the iteration that minimizes the balance metric
  - This is automated in the software

TWANG development at RAND



Advantages of TWANG and GBM

- Excellent estimation of $e(x)$
- Balances the Xs with little effort
- The resulting model handles continuous, nominal, ordinal, and missing Xs
- Invariant to 1-to-1 transformations of the Xs
- Model higher interaction terms with more complex regression trees
- Implemented in R in the `twang` library with many tools to make propensity score estimation easy

Example: Phoenix House Academy

- Does inpatient substance abuse treatment have an effect?
- Adolescent probationers in LA County assigned to substance abuse treatment at Phoenix House Academy or other group home facility
- Rich baseline data on covariates clinically selected to measure treatment needs and substance use risk
- Outcomes at 3, 6, and 12 months post intake

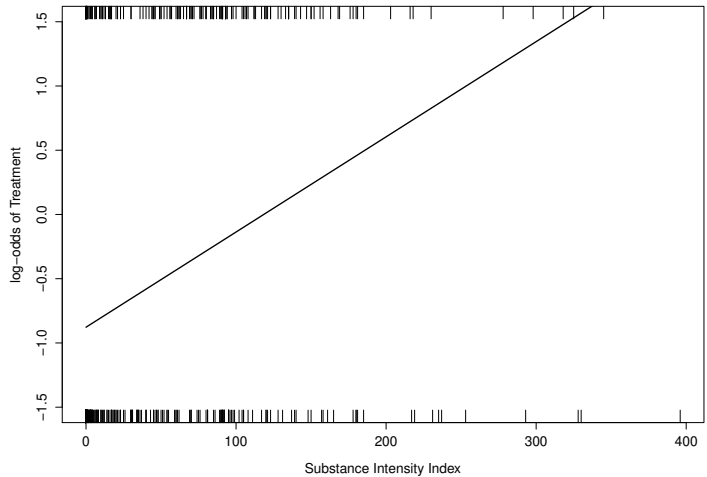
Balance of Pretreatment Features

- Treatment and control groups differ on pretreatment variables

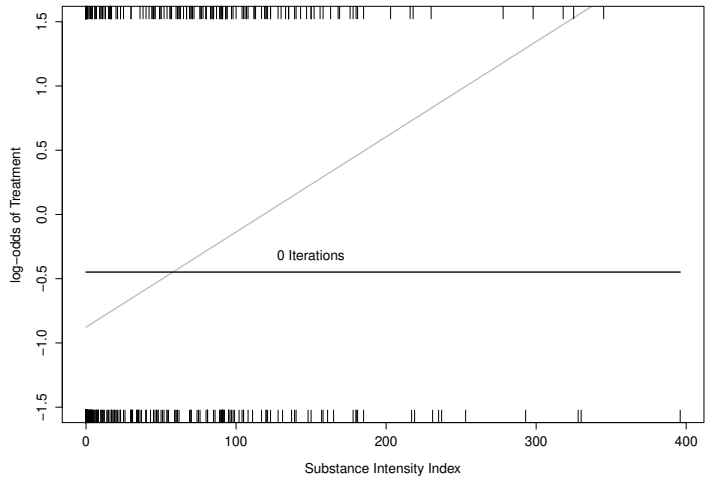
	Treatment	Control	
		unadjusted	*adjusted*
Variable	Mean	Mean	Mean
Treatment motivation	2.52	1.35	?
Environmental risk	30.61	28.94	?
Substance use	7.61	4.59	?
Complex behavior	12.84	12.11	?
Age	15.82	15.31	?
⋮	⋮	⋮	⋮

- The propensity score analysis should match or weight the control subjects so that their covariates are nearly the same as the treatment group's

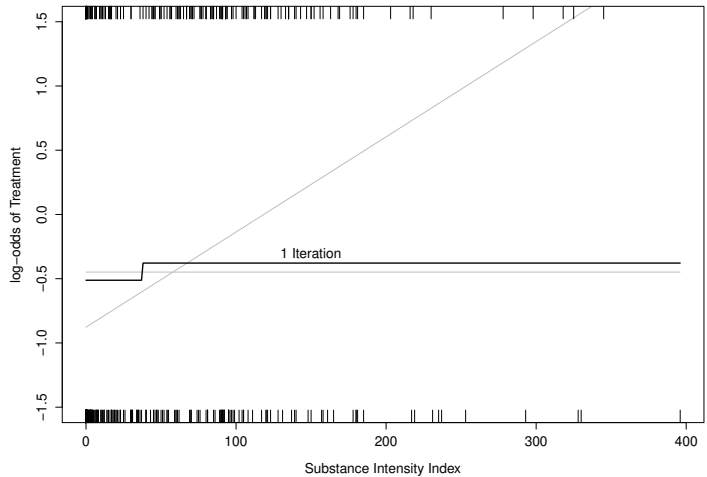
Linear Logistic Regression



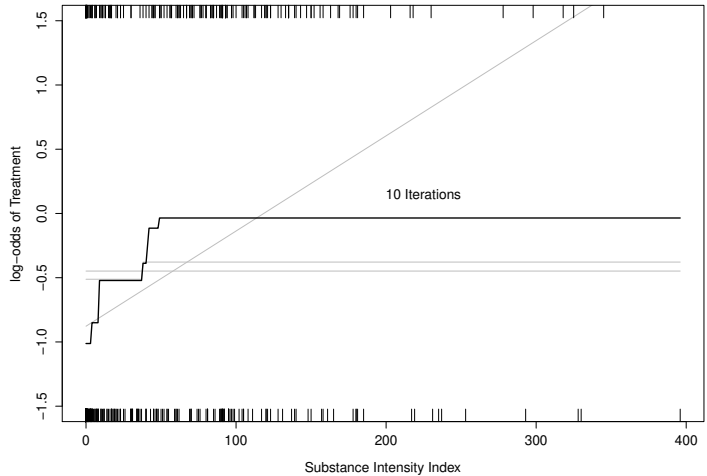
GBM Estimation



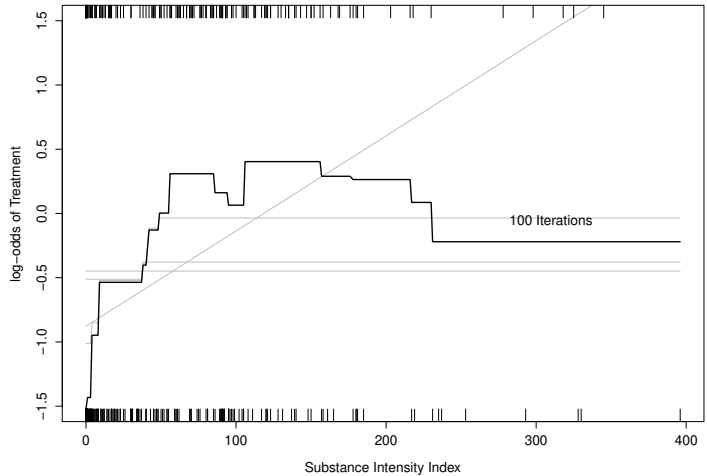
GBM Estimation



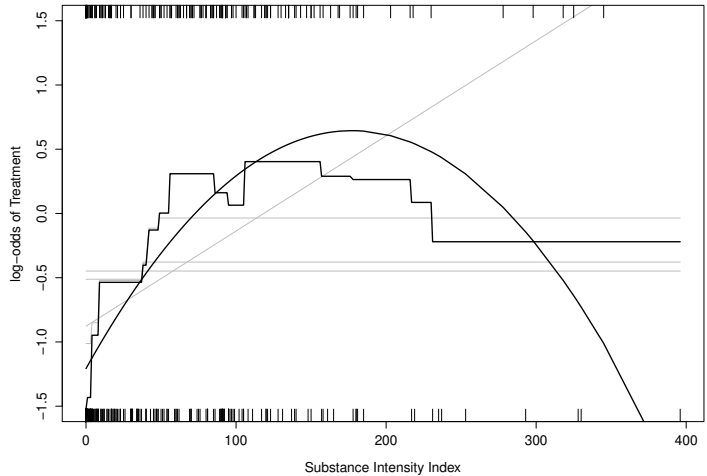
GBM Estimation



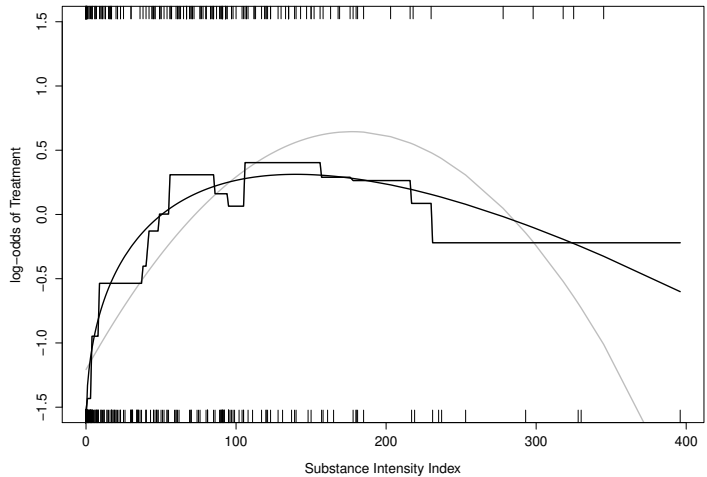
GBM Estimation



GBM Estimation vs. Logistic with a Quadratic in SII



GBM Estimation vs. Logistic with a Quadratic in Square Root SII



Balance of Subject Features, $C = 0$, No Iterations

Variable	treatment mean	weighted control mean	unweighted control mean	Standardized Bias weighted	Standardized Bias unweighted
Treatment motivation	2.52	1.35	1.35	0.89	0.89
Environmental risk	30.61	28.94	28.94	0.17	0.17
Substance use	7.61	4.59	4.59	0.69	0.69
Complex behavior	12.84	12.11	12.11	0.09	0.09
Age	15.82	15.31	15.31	0.56	0.56
⋮	⋮	⋮	⋮		⋮
ESS	175	274	274		
Average |ES|					**0.307**

Balance of Subject Features, 10 Iterations

Variable	treatment mean	weighted control mean	unweighted control mean	effect size weighted	effect size unweighted
Treatment motivation	2.52	1.36	1.35	0.88	0.89
Environmental risk	30.61	28.97	28.94	0.17	0.17
Substance use	7.61	6.94	4.59	0.69	0.69
Complex behavior	12.84	13.00	12.11	0.08	0.09
Age	15.82	15.76	15.31	0.55	0.56
⋮	⋮	⋮	⋮	⋮	⋮
ESS	175	274.0	274		
Average |ES|				**0.315**	**0.316**

Balance of Subject Features, 100 Iterations

Variable	treatment mean	weighted control mean	unweighted control mean	effect size weighted	effect size unweighted
Treatment motivation	2.52	1.32	1.35	0.83	0.89
Environmental risk	30.61	29.24	28.94	0.14	0.17
Substance use	7.61	4.76	4.59	0.65	0.69
Complex behavior	12.84	12.33	12.11	0.06	0.09
Age	15.82	15.33	15.31	0.54	0.56
⋮	⋮	⋮	⋮	⋮	⋮
ESS	175	272.7	274		
Average |ES|				**0.300**	**0.316**

Balance of Subject Features, 1000 Iterations

Variable	treatment mean	weighted control mean	unweighted control mean	effect size weighted	effect size unweighted
Treatment motivation	2.52	1.90	1.35	0.47	0.89
Environmental risk	30.61	30.88	28.94	-0.03	0.17
Substance use	7.61	5.93	4.59	0.39	0.69
Complex behavior	12.84	13.45	12.11	-0.07	0.09
Age	15.82	15.50	15.31	0.36	0.56
⋮	⋮	⋮	⋮	⋮	⋮
ESS	175	211.1	274		
Average |ES|				**0.200**	**.316**

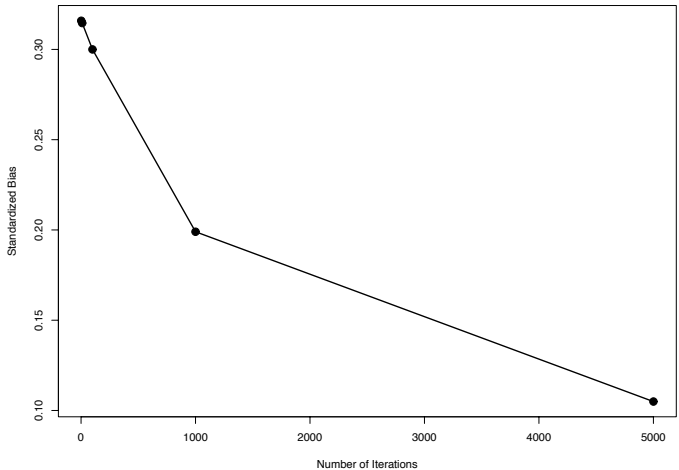
Balance of Subject Features, 5000 Iterations

Variable	treatment mean	weighted control mean	unweighted control mean	effect size weighted	effect size unweighted
Treatment motivation	2.52	2.28	1.35	0.19	0.89
Environmental risk	30.61	31.21	28.94	-0.06	0.17
Substance use	7.61	7.00	4.59	0.14	0.69
Complex behavior	12.84	13.21	12.11	-0.04	0.09
Age	15.82	15.76	15.31	0.07	0.56
⋮	⋮	⋮	⋮	⋮	⋮
ESS	175	104.3	274		
Average |ES|				**0.105**	**0.316**

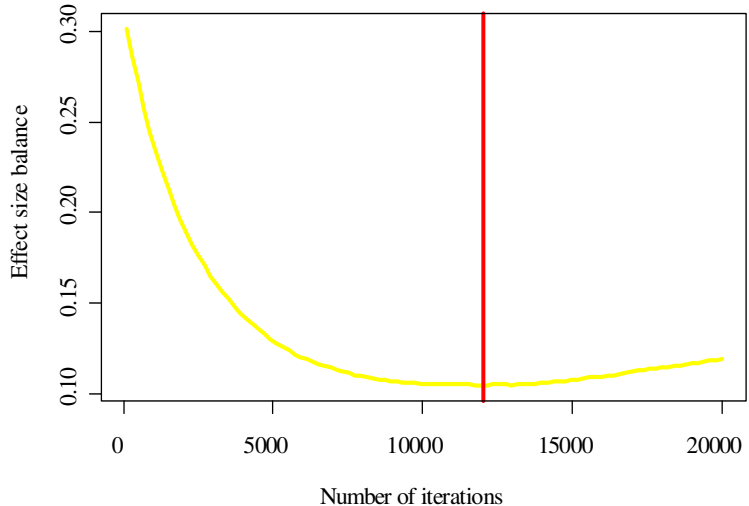
Tuning the GBM

- We must select the number of iterations
  - Within sample prediction error decreases with each additional iteration
  - Too few iterations under fit, miss features of the model, and don't achieve balance
  - Too many iterations overfit, too little variation among weights with group, and don't achieve balance
  - Somewhere in between is the best balance
- Fit with a large number of iterations
- Assess balance at each iteration
- Pick the iteration that minimizes the balance

Increasing Iterations Improves Balance



Model Selection



Assessing Balance

- The propensity score is a balancing score - after weighting, treatment groups should be very similar in terms of covariates
- Multiple metrics for assessing covariate balance
  - Standardized bias (also called "absolute standardized mean difference (ASMD)" or "effect size")
  - Kolmogorov-Smirnov (KS) statistics
- Aggregate across covariates
  - Mean or maximum of the balance metrics for individual covariates

Standardized Bias

- A commonly used measure of balance is standardized bias

$$SB = \frac{\tilde{x}_{\text{treat}} - \tilde{x}_{\text{control}}}{s}$$

- $\tilde{x}_{\text{treat}}$ and $\tilde{x}_{\text{control}}$ are the weighted treatment and control group means
- For ATT, s equals s_{treat}, the unweighted treatment group standard deviation which involves only the treatment subjects so that manipulating the control group never affects the measure's power
- For ATE, s is the pooled within sample unweighted standard deviation
- s does not depend on the weights

Standardized Bias and Balance

- Small values of SB indicate balance
- SB is on the effect size scale: Early rule of thumb was <0.20 indicated good balance, now people tend to aim for less than 0.10
- Can take average or maximum of SBs across all covariates included in the model as single measure of the quality of the propensity score

Kolmogorov-Smirnov Statistic

- Standardized bias only assesses balance of the means
- If outcomes are not linearly related to covariates, then balancing the means might not be sufficient to prevent bias
- Kolmogorov-Smirnov (KS) measures the distance between two distributions and measures balance more generally than standardized bias alone
  - Let $F_T(x)$ be the empirical distribution for the treatment cases

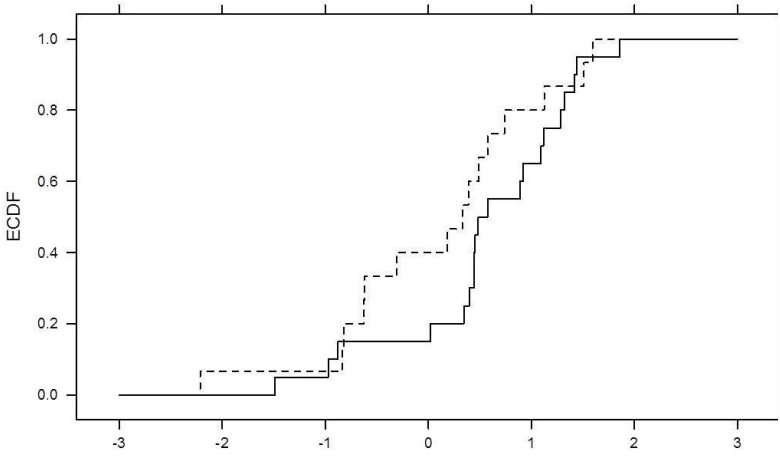
$$F_T(x) = \frac{\sum I(x_i < x)}{N_T}$$

  - Let $F_C(x)$ be the weighted empirical distribution for the control cases

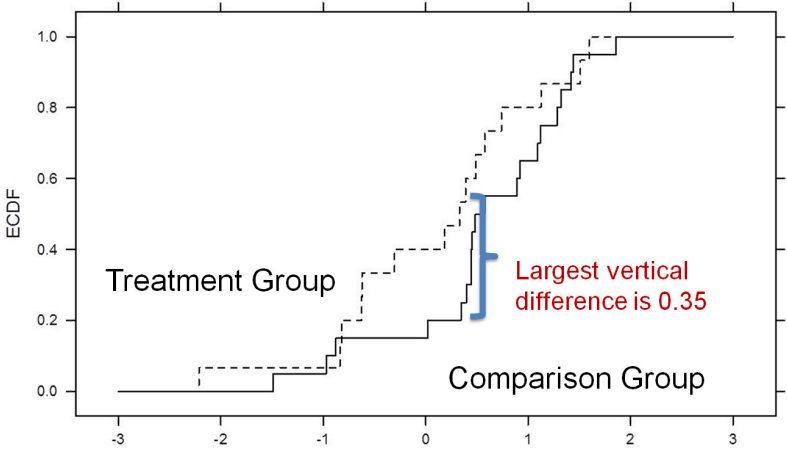
$$F_C(x) = \frac{\sum w_i I(x_i \leq x)}{\sum w_i}$$

  - For ATT, $KS = max|F_T(x) - F_C(x)|$, for ATE use weighted empirical distribution for treatment and control
- Can take average or maximum KS across all covariates included in the model as single measure of the quality of the propensity score

KS Example



KS Example


ECDF
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
Treatment Group
Largest vertical
difference is 0.35
Comparison Group

Issues with KS

- There are no clear standards for what is a large or small value
- Size of KS depends on sample size
- Very useful for comparing alternative models, but less obvious if has met an objective standard for "good" balance
- No standard tests for weighted KS
- We developed approximate tests for individual covariates and the maximum across all variables
- Tests depend on the sample size

Modeling with Multiple Measures of Balance

- The TWANG software considers both the KS and standardized bias
- It contains a sample of each measure across multiple covariates
- It is possible to choose the maximum or the mean (or some other summary statistic) for each balance measure
- One preference is to tune GBM using multiple measures and choose the solution that appears most robust across the alternatives
- If standardized bias is much greater than .10 for any variables or somewhat greater for many variables, often it is worth considering reformulating the problem – becomes an art again

Alternative Stopping Rule Given Similar but not the Same Weights

		Standardized Bias		KS		Number of
Stop Method	ESS	maximum	mean	maximum	mean	Iterations
Mean KS Statistic	88.04	0.264	0.093	0.136	0.058	720
Mean Standardized Bias	88.39	0.261	0.093	0.134	0.058	730
Maximum KS Statistic	95.01	0.269	0.099	0.132	0.060	572
Maximum Standardized Bias	88.49	0.260	0.094	0.135	0.058	742

Estimating the Treatment Effect

- After balancing the only *observed* difference between the two groups is the treatment assignment
- Average Treatment Effect in Select Population

$$\tau = \frac{\sum_i w_1(X_i) A_i Y_i}{\sum_i w_1(X_i) A_i} - \frac{\sum_i w_0(X_i)(1 - A_i) Y_i}{\sum_i w_0(X_i)(1 - A_i)}$$

Effective Sample Size

- Propensity score weighting effectively "downweights" some observations that do not provide good matches (in terms of covariates) and "upweights" observations that are good matches
- Concern in PS weighting is having too few matches – i.e., so many individuals are "downweighted" that the "effective sample size" of weighted data is much smaller than original data
  - ESS will always be smaller, but matter of degree

Variable Weights can add to Variance in Estimated Treatment Effect

$$\operatorname{var}\left(\frac{\sum_{i\in T} w_i e_i}{\sum_{i\in T} w_i} - \frac{\sum_{i\in C} w_i e_i}{\sum_{i\in C} w_i}\right) \approx \operatorname{var}(e_i)\frac{\sum_{i\in T} w_i^2}{\left(\sum_{i\in T} w_i\right)^2} - \operatorname{var}(e_i)\frac{\sum_{i\in C} w_i^2}{\left(\sum_{i\in C} w_i\right)^2}$$

- $\frac{\sum_{i\in T} w_i^2}{\left(\sum_{i\in T} w_i\right)^2} \geq \frac{1}{N_T}$, equality holds only when weights are constant
  - Same for control
- Since without weighting variance would equal $\operatorname{var}(e_i)/N_T$ or $\operatorname{var}(e_i)/N_C$, variable weights can increase the variance of the estimated treatment effect

Effective Sample Size

ESS interpretation:

- ESS is the number of independent observations from a simple random sample that would yield the same precision as the weighted observations used in estimating the treatment effect

$$ESS = \frac{\left(\sum_{i \in C} w_i\right)^2}{\sum_{i \in C} w_i^2}$$

  similar formula holds for treatment for ATE
- Since weighted means are less precise than simple averages, ESS is always less than the observed sample size in the group
- In ATT context, effective sample size (ESS) is number of observations in the control group that "match" with the treatment group
- If ESS is too small, data may not be robust enough to address causal question of interest

Causal Effects for Binary Treatments: Step-by-Step in R

Overview of the `twang` package

TWANG: Toolkit for the Weighting and Analysis of Nonequivalent Groups

- Used to estimate propensity score weights for observational data
- Has robust, built-in diagnostics to assess how well propensity score weighting has balanced groups on baseline covariates
- `twang` uses generalized boosted modeling (GBM), a nonparametric machine learning algorithm, to estimate propensity scores
- After obtaining weights from `twang`, user then estimates treatment effects

Revisiting the Motivating Example

Context: Comparing effectiveness of different types of adolescent substance use treatment for youth receiving SAMHSA-funded community-based treatment
Aim: Estimate the causal effect of MET/CBT5 treatment versus usual care treatment

MET/CBT5

- Longitudinal, observational
- 37 sites from EAT study
- $N = 2459$
- Timeframe: 2003/04 - 2007

Usual care

- Longitudinal, observational
- 4 sites from ATM study
- $N = 444$
- Timeframe: 1998-1999

All youth take GAIN survey (assesses substance use, mental health, etc.) at baseline, 6 months, and 12 months

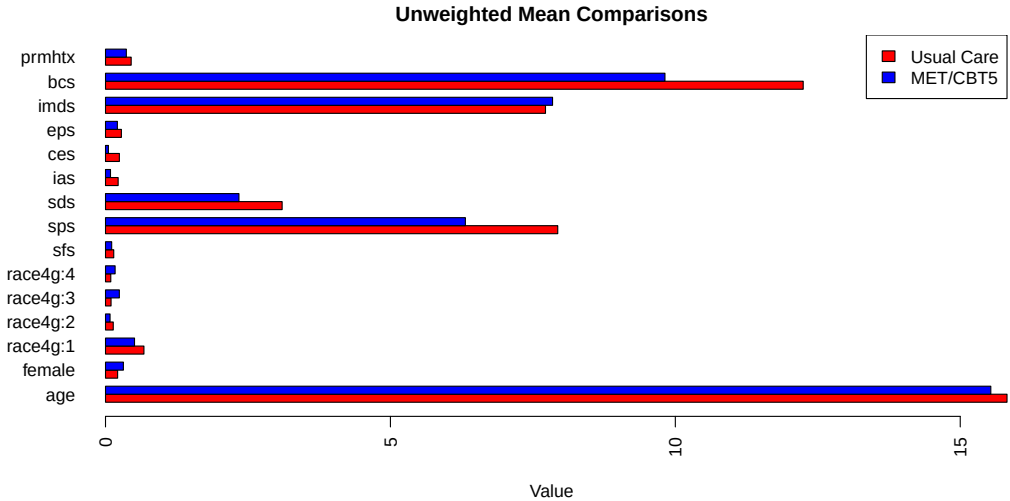
Preliminary analysis - Describing the data

Baseline covariates

Variable Name	Description
tx	Usual care treatment indicator
age	Age of the participant
female	Indicator of female gender
race4g	Race category variable
sfs	Substance Frequency Scale
sps	Substance Problem Scale
sds	Substance Dependence Scale
ias	Illegal Activities Scale
ces	Crime Environment Scale
eps	Emotional Problem Scale
imds	Internal Mental Distress Scale
bcs	Behavior Complexity Scale
prmhtx	Prior mental health treatment

Selection Exists

Regardless of the estimand, we should be cautious because these groups look different at baseline on these covariates



Key steps in a weighted propensity score analysis

The goals can be more plainly be expressed as the steps in a propensity score analysis

1. Choose the primary estimand of interest (ATE or ATT)
2. Estimate propensity score weights
3. Evaluate the quality of the weights
4. Estimate the treatment effect

Step 1 - Choosing the estimand: ATE or ATT

Today we will focus on estimating an ATT (effect on the treated) type estimand

- We want to draw inferences about the effect of MET/CBT5 treatment (relative to the comparison treatment, here usual care), for youth who actually received usual care
- Because the usual care treatment sites want to know if MET/CBT5 works for youth they typically treat
- This is actually ATC (Average Treatment on the Control)
- Can be accomplished by creating a treatment indicator so that "tx" $= 1$ if youth in usual care

Step 2 - Estimate propensity score weights

- Today, we will focus on syntax for estimating PS weights in the R `twang` package, using `ps` command
- Additionally, macros for STATA and SAS are available (run R version of `twang` through STATA and SAS interface)
- Also, can run `twang` through a menu-driven interactive application, built using R Shiny

Setting up twang in R

- Download R:
  https://cran.r-project.org
- Within R, download twang package:

```
install.packages(twang)
```

- Call twang prior to use

```
library(twang)
```

Command to estimate propensity score weights in R

See twang vignette for more details on all arguments

```
ps(formula = formula(data),
   data,
   n.trees = 10000,
   interaction.depth = 3,
   shrinkage = 0.01,
   bag.fraction = 1.0,
   perm.test.iters=0,
   print.level = 2,
   iterlim = 1000,
   verbose = TRUE,
   estimand = "ATE",
   stop.method = c("ks.mean", "es.mean"),
   sampw = NULL,
   multinom = FALSE, ...)
```

Command to estimate propensity score weights in R

Back to our motivating example

```
## fit gbm and extract propensity score weights
my.ps_bin <- ps(tx ~ age + female + race4g + sfs + sps +
                  sds + ias + ces + eps + imds + bcs + prmhtx,
                data=AOD,
                estimand="ATT",
                n.trees=10000,
                shrinkage = 0.001,
                stop.method="es.max")
```

Variable	Description
data	Points to the data object
estimand	Tells ps the estimand is the ATT
n.trees	Specifies the numbers of trees for gbm
shrinkage	Specifies specifies smoothness of the model; smaller is smoother
stop.method	Specifies the balance criteria that should be used to select weights - this says minimize the maximum effect size - i.e., standardized bias

Step 3 - Evaluate the quality of the weights

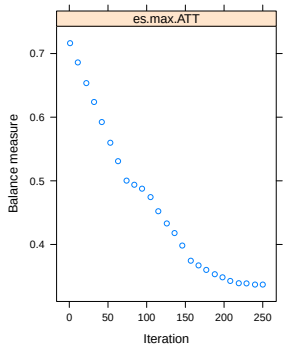
There are three key issues that should be checked after running the ps function:

1. *Convergence* - Did the algorithm run long enough to find a minimum for the selected stopping criteria?
2. *Balance* - how comparable are the two weighted groups to each other? In this example does the weighted MET/CBT5 group look like the usual care group?
3. *Overlap* - To what degree are there individuals in the comparison group that have the same PS values as those in the treatment group? Assessed using boxplots of the propensity scores in the two groups

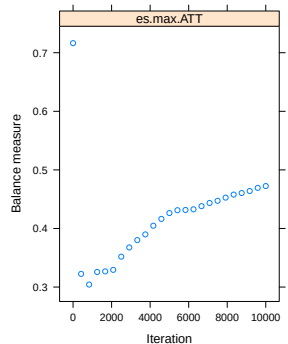
Assessing Convergence

```
plot(my.ps_bin)
```

Bad Convergence



Good Convergence



More on Bad Convergence

This warning message can be helpful in diagnosing convergence

```
Warning message:
In ps(tx ~ age + female + race4g + sfs + sps + sds + ias + ces +  :
  Optimal number of iterations is close to the specified n.trees. n.trees
     is likely set too small and better balance might be obtainable by
    setting n.trees to be larger.
```

Assessing Balance

The `summary()` function provides an overall balance summary.

```
summary(my.ps_bin)
```

```
      n.treat n.ctrl ess.treat ess.ctrl  max.es mean.es max.ks mean.ks
    iter
unw         442   2408     442     2408    0.720   0.323   0.483   0.153
      NA
es.max.ATT  442   2408     442   498.69    0.295   0.089   0.149   0.054
    710
```

Checking Covariate Balance

- Balance table

```
bal.table(my.ps_bin)
```

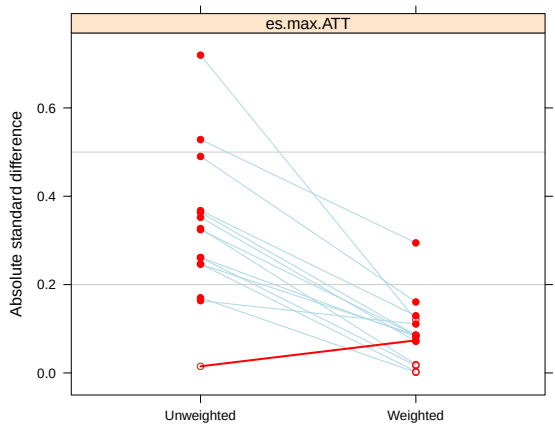
- Balancing Diagnostic Plots

```
plot(my.ps_bin, plots = 2)
```

Descriptive Argument	Numeric Argument	Description
optimize	1	Balance measure as a function of GBM iterations
boxplot	2	Boxplot of treatment/control propensity scores
es	3	Standardized effect size of pretreatment variables
t	4	t-test p-values for weighted pretreatment variables
ks	5	KS p-values for weighted pretreatment variables
histogram	6	Histogram of weights for treatment/control

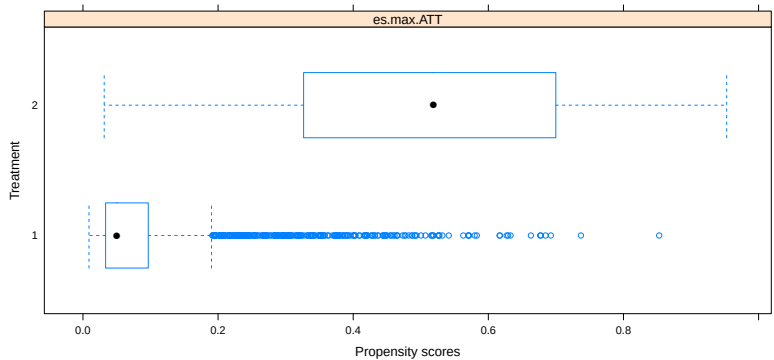
Standardized Effect Size

```
plot(my.ps_bin, plots = 3)
```



Boxplot - Useful for assessing overlap

```
plot(my.ps_bin, plots = 2)
```



Assessing Overlap - An additional check

Compare the conditional distributions, checking for zero cells

Distribution of Race

Category	MET/CBT5	Usual care
Race = 1	1226	298
Race = 2	192	60
Race = 3	584	43
Race = 4	406	41

Example code:

```
t(table(AOD$tx, AOD$race4g))
```

Distribution of SFS

Cutpoints	MET/CBT5	Usual care
[0, 0.00694]	501	76
(0.00694, 0.0347]	496	70
(0.0347, 0.106]	489	79
(0.106, 0.219]	478	94
(0.219, 0.757]	444	123

Example Code:

```
t(table(AOD$tx,
        cut(AOD$sfs,
            quantile(AOD$sfs,
                     seq(0,1,0.2)),
            include.lowest = T)))
```

Additional Model Diagnostics

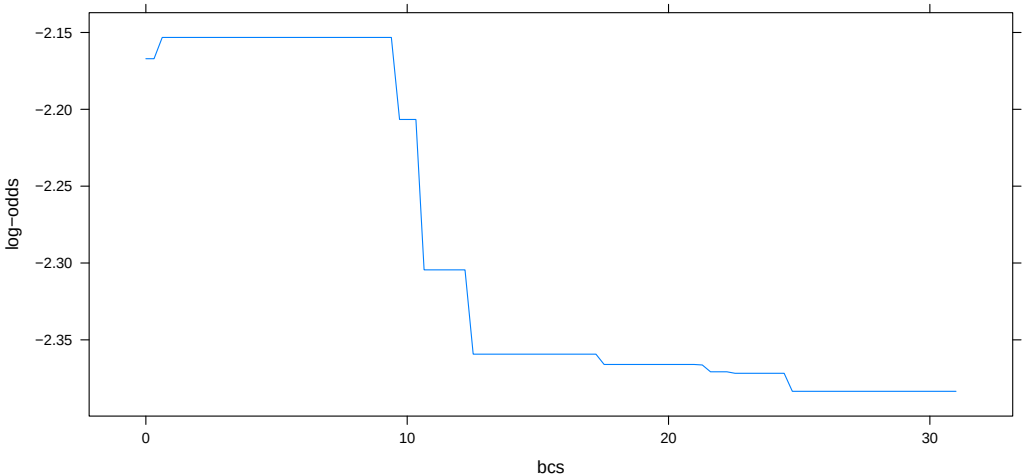
- Marginal relationship between a covariate and treatment

```
plot(my.ps_bin$gbm.obj, i.var="bcs",
     n.trees=my.ps_bin$desc$es.max.ATT$n.trees,
     ylab = 'log-odds')
```

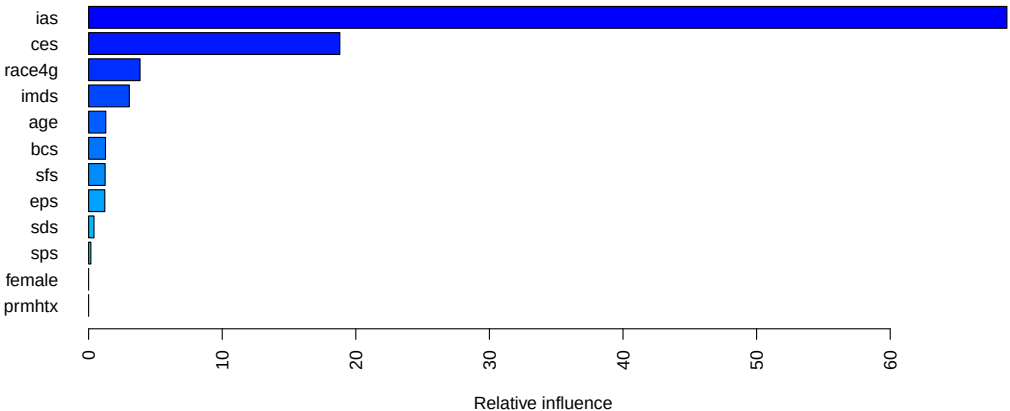
- Relative influence of each covariate

```
summary(my.ps_bin$gbm.obj,
        n.trees = my.ps_bin$desc$es.max.ATT$n.trees,
        las = 2)
```

Marginal Relationships



Relative Influence Plots



What comes next...

Note: We have not even begun to talk about the outcome yet

- Steps 1, 2, and 3 do not involve any outcomes
- We first focus on dealing with selection/pre-treatment group differences
- Then, if we do a good job, we will move to outcome analyses

Step 4 - Estimate treatment effects

Estimate the difference in means between the two groups of interest weighted by the inverse propensity score weights

- These analyses will require us to adjust the standard errors for estimation of the propensity score weights
- Analogous to fitting regression models to survey data with survey weights

We will use the `survey` package in R to estimate treatment effects

Step 4 - Estimate treatment effects

```
# use survey package - install if not already installed
# install.packages("survey")
library(survey)

# extract the propensity score weights
AOD$w <- get.weights(my.ps_bin, estimand = "ATT",
                     stop.method = "es.max")

# use svyglm to incorporate weights
design.ps <- svydesign(ids = ~1, weights = ~w, data = AOD)
glm1 <- svyglm(sfs8p12 ~ tx, design = design.ps)
summary(glm1)
```

Step 4 - Estimate treatment effects

```
> summary(glm1)

Call:
svyglm(formula = sfs8p12 ~ tx, design = design.ps)

Survey design:
svydesign(ids = ~1, weights = ~w, data = AOD)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.094218   0.006259  15.053   <2e-16 ***
tx          0.020147   0.009161   2.199   0.0279 *
```

Results show that, for youth like those in usual care group, usual care treatment is associated higher substance use as 12 months than if they had received MET/CBT5

Additional regression adjustment

Additional regression adjustment for pre-treatment covariates is sometimes necessary and is often referred to as a "doubly-robust" model

- Combines fitting a propensity score weighted regression model with the inclusion of additional pre-treatment control covariates
- As long as one model is correct (either the multivariate outcome model or the propensity score model), consistent treatment effect estimates are obtained
- Typically one focuses on adding in covariates with lingering imbalances

Doubly-robust model

```
glm2 <- svyglm(sfs8p12 ~ tx + ias + race4g + ces + eps, design = design.
    ps)
summary(glm2)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.074936   0.009264   8.089 8.83e-16 ***
tx           0.020755   0.009759   2.127  0.03353 *
ias          0.085368   0.027582   3.095  0.00199 **
race4g2     -0.022325   0.013851  -1.612  0.10713
race4g3      0.029889   0.016879   1.771  0.07670 .
race4g4     -0.035147   0.011999  -2.929  0.00342 **
ces         -0.016002   0.016876  -0.948  0.34308
eps          0.025190   0.026039   0.967  0.33344
```

Results are consistent even after additional regression adjustment

Conclusions

- Use of propensity score weighting greatly improved balance on observed pre-treatment covariates
  - Magnitude of covariate imbalance across treatment groups decreased from 0.05 effect size difference to 0.02 after weighting
- PS weighting helped us produce less biased (i.e., more robust) estimates of the effectiveness of usual care relative to MET/CBT5
  - Recall, unadjusted comparison found usual care youth had SFS scores 0.04 higher at 12 months
  - PS-weighted comparison finds that usual care youth had SFS scores that were 0.02 higher at 12 months, a statistically significant difference
- We conclude, after adjusting for differential baseline characteristics across groups, that usual care youth would be expected to have lower substance use outcomes at 12 months if instead they had received MET/CBT5

Brief Discussion on Variable Selection for Causal Modeling

Review of Variable Selection for Causal Inference

- The goal: select a subset of observed covariates to include in the propensity score model
- The issue: traditional variable selection methods are designed for prediction, not effect estimation
- The solution: ???

Review of Variable Selection for Causal Inference

- While propensity score methods have gained widespread use, there still remains confusion and a lack of guidance on how best to carry out variable selection
- The tools we have discussed today seek to optimize pretreatment covariate balance using GBM
  - Variable selection is embedded within the algorithm
- What are the alternatives?

Controlling for the Widest Set Possible

- Rubin (2009) argues that controlling for the widest set of pretreatment characteristics protects against unobserved confounding
- What happens if the number of pretreatment covariates is large?
- VanderWeele and Shpitser (2011) show that controlling for all pretreatment covariates may lead to biased treatment effect estimates

Instruments

- Covariates that predict treatment assignment but are not causally related to the outcome are often called "instruments"
- They are not confounders
- Controlling for instruments as confounders can exacerbate bias due to omitted variables
- We should avoid including instruments in the covariates used to control for confounding

Outcome Only Predictors

- Covariates that are causally related to the outcome but not treatment are outcome only predictors
- They are not confounders
- Controlling for these outcome only predictors in the same way that we control confounders can reduce the estimation error of treatment effects
- We can include outcome only predictors among the controlled covariates

Pre-Treatment Variables Only

- We should avoid controlling for variables measured after treatment

The Use of Expert Knowledge

- Robins (2001) and Hernan et al. (2002) argue that full or partial knowledge of the underlying causal structure is required to select confounders
- In other words, we must use substantive knowledge to select the covariates to include in the analysis
- What happens if the number of covariates is large?

VanderWeele and Shpitser (2011)

- Argue that bias is removed by adjusting for all covariates that cause treatment or outcome
- A quote: An investigator simply needs to ask, "Is the covariate a cause of the treatment?" and "Is the covariate a cause of the outcome?"
  - If the answer is yes to either, include in the analysis
- What happens if this set of covariates is large?
- Suggest iteratively discarding variables unassociated with the outcome

Data Driven Approaches for Variable Selection

- Previous methods may be sufficient for removing bias but are likely to be inefficient for estimating treatment effects
- Recent data driven approaches for variable selection improve efficiency without sacrificing unbiasedness
- General theme: confounders are related both treatment and outcome

Data Driven Approaches for Variable Selection

- Three possibilities:
  - Variable selection based only on the propensity score model
  - Variable selection based only on the outcome model
  - Combination of the two

Alternatives to TWANG

Review

Goal is to estimate the effect of an intervention A on an outcome Y

- X denotes variables that are potentially related to the exposure assignment and outcome (potentially confounders) and occur temporally before exposure (pre-exposure covariates)
- A is an indicator of assigned exposure $\mathcal{A} = \{0, 1\}$
- Y measures an output of interest
- Focused on (general) estimators of the form;

$$\hat{\tau} = \frac{\sum_i w_1(X_i) A_i Y_i}{\sum_i w_1(X_i) A_i} - \frac{\sum_i w_0(X_i)(1 - A_i) Y_i}{\sum_i w_0(X_i)(1 - A_i)}$$

  where the weights are a function of X through the propensity score $e(X)$
- Require strong assumptions to have unbiased estimation

Review

- Focused on (general) estimators of the form;

$$\hat{\tau} = \frac{\sum_i w_1(X_i) A_i Y_i}{\sum_i w_1(X_i) A_i} - \frac{\sum_i w_0(X_i)(1 - A_i) Y_i}{\sum_i w_0(X_i)(1 - A_i)}$$

where the weights are a function of X through the propensity score $e(X)$

ATE Estimation

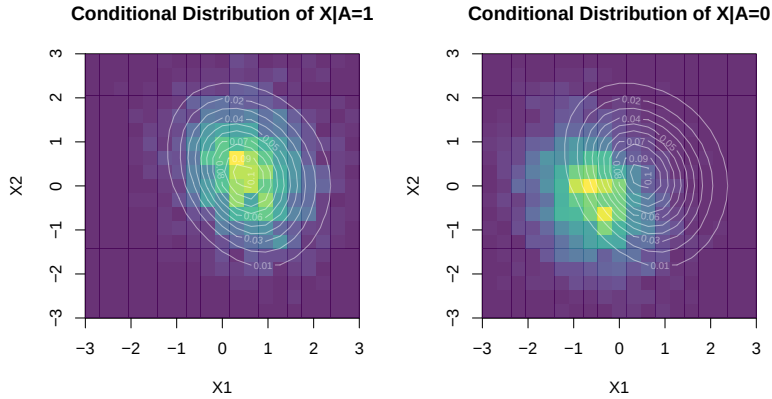
$$w_1(X_i) \propto \frac{1}{\hat{e}(x)} \qquad w_0(X_i) \propto \frac{1}{1 - \hat{e}(x)}$$

ATT Estimation

$$w_1(X_i) \propto 1 \qquad w_0(X_i) \propto \frac{\hat{e}(x)}{1 - \hat{e}(x)}$$

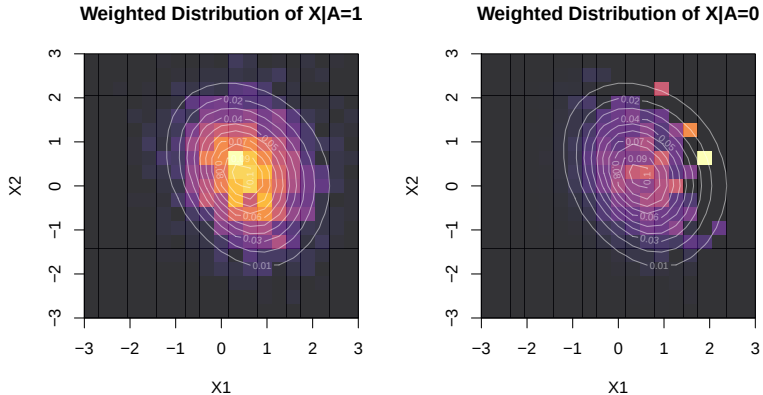
Continuing with the ATT - Hypothetical Observed Distributions

Let $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,1)$, and let $e(x) = \text{logit}^{-1}(2X_1 + X_2)$ for a sample of $N = 5000$ and a goal of estimating the ATT



Continuing with the ATT - Weighted Distributions

Using the appropriate weight function and summing the weights in each bin → weighted distribution is similar to target distribution



Propensity Score Estimation

- We have discussed two methods for propensity score estimation to create weights
  - GBM & TWANG
  - Logistic regression
- Many other methods exist
  - Covariate Balancing Propensity Scores (CBPS, Imai and Ratkovic)
  - Entropy Balancing, exponential tilting, minimum discriminant information adjustment (Hainmueller, Graham et al., Haberman)
  - Stable Balancing Weights (Zubizarreta)
  - Super learning (Polley and van der Laan)
  - High Dimensional Propensity Scores (hd-PS, Schneeweiss and Rassan)
  - Optimally balanced Gaussian Process propensity scores (Vegetabile)

Achieving Balance in the Parameter Estimation Criteria

- For GBM and logistic regression we use balance to guide model selection
  - Art of logistic regression for propensity score estimation is picking and choosing terms to get good balance and improving balance guides choice – similar to using AIC or BIC in other contexts
  - For GBM we use balance to pick the number of iterations that controls the variables included and functional form of the models – replaces cross-validation
- Parameter estimation criteria does not involve balance
  - For logistic regression, we find the MLE for the parameters of the model at each step of modeling process
  - For GBM, the algorithm select the tree models that maximize the likelihood for each iteration
- **What if it did?**

Achieving Balance in the Parameter Estimation Criteria

- For logistic regression want to estimate the parameters for the model

$$e(X;\beta) \equiv Pr(A = 1|X = x, \beta) = 1/(1 + \exp(-\beta' X))$$

- Goal is typically to find the value of β to maximize the log likelihood $\ell(\beta; X)$ or find β that solves $\psi(X;\beta) = 0$ (i.e., a defined estimating equation)
- Suppose we added k additional optimization criteria such that the β should also satisfy

$$\sum_{i=1}^{n} \frac{1}{e(X_i;\beta)} g_k(X_i) A_i = \sum_{i=1}^{n} \frac{1}{1 - e(X_i;\beta)} g_k(X_i)(1 - A_i)$$

- The functions $g(X)$ are typically the moments and cross-moments of variables
- Using balance to determine the values of β – coefficient values will differ from traditional logistic regression fit

Covariate Balancing Propensity Score

- Covariate Balancing Propensity Scores (CBPS) follows this logic within the estimating equation framework
- For logistic regression, maximizing the likelihood solves the estimating equation $\psi(X;\beta) = 0$
- CBPS includes balance in the parameter estimation criteria when estimating the parameters of the logistic regession model
  - Extends estimating equation from $\psi(X;\beta) = 0$ to
    $$\begin{pmatrix} \psi(X;\beta) \\ \sum_{i=1}^{n} \frac{1}{e(X;\beta)} X_i A_i - \sum_{i=1}^{n} \frac{1}{1-e(X;\beta)} X_i (1 - A_i) \end{pmatrix} = 0$$
- Developed by Imai and Ratkovic (2014) who also developed the R package CBPS to implement the method
- Recent extensions exist, such as those for continuous treatments

Selecting Weights to Get Balance for Specific Target Distributions

- Given a target distribution $\mathbb{Q}$, the goal is finding weights w such that weighted summary statistics in each group match the moments of this target distribution, i.e.,

$$\sum_i w_i A_i g_k(X_i) = \sum_i w_i (1 - A_i) g_k(X_i) = E_{\mathbb{Q}}[g_k(X)]$$

- Typical moment functions g_k to balance: $g_1 = X$, $g_2 = X^2, \ldots$
- For example, to estimate the ATT the target distribution $\mathbb{Q}$ is the treated group and we could set a set of targets

$$E_{X|A=1}[X] = \frac{1}{N_1} \sum_{A_i=1} X_i \qquad E_{X|A=1}[X^2] = \frac{1}{N_1} \sum_{A_i=1} X_i^2$$

  and find weights for the control group that give exact balance,i.e., $\sum_{i:A_i=0} w_i(1 - A_i)X_i = \frac{1}{N_1} \sum_{A_i=1} X_i$
- Weights will exists as long as groups are not too distinct, multiple possible solutions

Turning this into a General Optimization Problem

To constrain the number of solutions for w, typically the problem is transformed into a constrained optimization to find weights to match target distribution $\mathbb{Q}$

$$\min_{\boldsymbol{w}} \; Loss(\boldsymbol{w}) \quad \text{subject to}$$

$$1) \sum_i w_i I(A_i = a) g_k(X_i) = E_{\mathbb{Q}}[g_k(X)] \qquad \text{for specified } a, \text{ and each } g_k$$

$$2) \sum_i w_i = 1$$

$$3) \; w_i > 0$$

Entropy Balancing for ATT Estimation as in Hainmueller (2012)

For entropy balancing[1]: Given a set of "base weights"[2] $\boldsymbol{q}$ select weights $\boldsymbol{w}$ to

$$\min_{\boldsymbol{w}} \sum_{i:A_i=0} w_i \log\left(\frac{w_i}{q_i}\right) \quad \text{subject to}$$

$$1) \sum_{i:A_i=0} w_i g_k(X_i) = \hat{m}_k(X) \qquad \text{and each moment } m_k$$

$$2) \sum_{i:A_i=0} w_i = 1$$

$$3) \; w_i > 0$$

typically each $q_i = 1/N_0$ and this optimizes the Kullback discriminant information or Kullback-Leibler distance comparing weights to equal weighting

[1] or minimum discriminant information adjustment (MDIA)

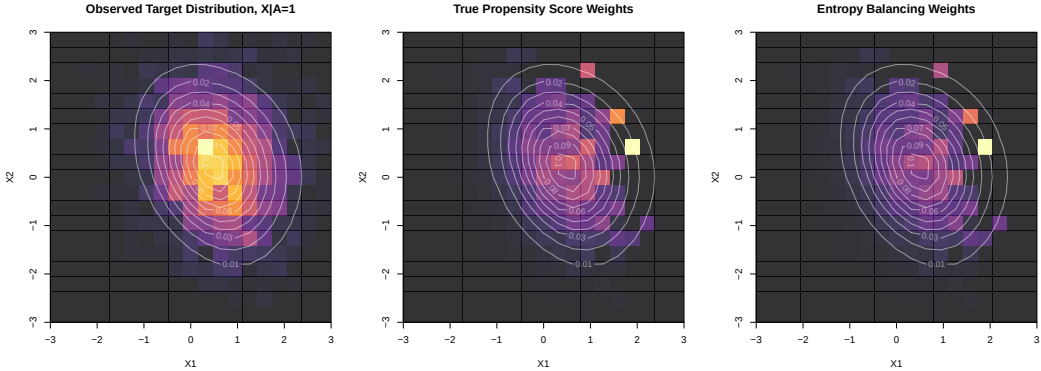
[2] $\boldsymbol{q}$ may be survey weights or other information

Form of Entropy Balancing Weights

- Entropy balancing is computationally efficient because it admits a closed form solution for the weights $\longrightarrow$ dramatically reduces parameter space
  - Can be tricky to implement computationally
- Entropy balancing weights are of the form $\exp(\alpha + \gamma' X)$
- Same form as ATT weights $(e(X)/[1 - e(X)])$ with logistic regression propensity scores using X
  - But the coefficients will differ

Continuing the Simple Example - Entropy Balancing

Let $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,1)$, and let $e(x) = \text{logit}^{-1}(2X_1 + X_2)$ for a sample of $N = 5000$ and a goal of estimating the ATT



Alternative Criteria for Exact Balance Weights

- Can replace the Kullback-Leibler distance with other distance measure
- Zubizarreta (2015) minimized variance of the weights
- Survey sampling calls these Generalized Regression weighting and consider several alternative distances (see Deville and Särndal, 1992 for examples)
- Entropy balancing has nice property that weights are positive and admits a reduced parameter space

Minimal Approximately Balancing Weights

- Similar to previous optimization problems, but relaxes the equality constraint in the optimization
- $|\sum_{(i|A=0)} w_i X_{ki} - \tau_0| < \delta_k \text{ for } k = 1, \ldots, K$
- Select δ_k to control variance of the weights and MSE of the treatment effect estimator

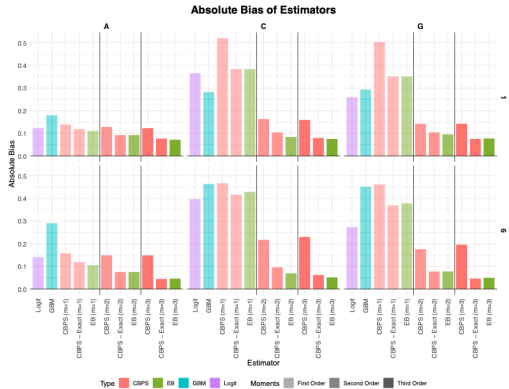
Selecting X and the balancing functions

- Entropy balancing weights (and many optimal balancing algorithms) give exact balance to linear function of covariates used in balancing $\beta' X$
- If $E[Y_0 \mid X]$ is not linear in X then there can be remaining bias
- We can use functions of the covariates to create "Xs" for balancing
  - For example, include covariates and their squares and cross-products

Higher moments matter with nonlinearities

Recent paper:

- Huang, Melody Y., Brian G. Vegetabile, Lane F. Burgette, Claude Setodji, and Beth Ann Griffin. "Higher Moments Matter for Optimal Balance Weighting in Causal Estimation." Epidemiology (Cambridge, Mass.) (2022).



Optimal Balance Vs Direct Propensity Score Estimation

If we can obtain exact balance why bother with propensity scores?

- Exact balance only on selected Xs and only for defined functions
- Not clear how well other (not selected) functions of covariates will balance
- With good propensity score model all functions of covariates should balance (at least in expectation) if strong ignorability holds
- We have model building schemes so propensity score models will tend to be good (for large samples)
- We don't have modeling building schemes for picking the covariates and functions of the covariates to balance exactly
- Exact balance may come at the price of greater variability in the weights (smaller effective sample sizes) – that may not be useful if we pick the wrong covariates or functions of covariates to exactly balance

Using Exact Balance

- Pick functions of the covariates and obtain exact balance
- Obtain propensity scores and the apply exact balance to selected covariates on top of propensity score weighting
- `ebal` package in R and `ebalance` package in Stata will conduct entropy balancing
- Forthcoming package that includes more bells and whistles: https://github.com/bvegetabile/entbal/
- Additionally, many solutions available in the 'WeightIt' package

SuperLearner Ensemble (using WeightIt)

- SuperLearning (Polley et al.) combines multiple propensity score estimation method
  - Combines predictions to get propensity score and coverts those to weights
- Greifer developed the WeightIt package to implement SuperLearner, among other methods, with ease
  - The "super" method in WeightIt estimates propensity scores using the SuperLearner algorithm (Polley et al.)
  - Users have many options for methods to combine including: gbm, glm, glmnet, randomForest, xgboost among others

Example: CBPS with the AOD Data

- Compare the ACRA and MET/CBT-5 conditions from AOD data
- Test the relative effects of two treatment among youth like those that receive ACRA
- Treatment on the treated with ACRA as "the treated" and MET/CBT-5 as the "control"
- For this demonstration we use casewise deletion to remove records with incomplete covariate data
- Conduct the analysis in R

Prepare the Data

```
library(CBPS)
aod <- read.csv("AOD.csv")
atmeat <- subset(aod,
                 subset=(trtvar %in% c("ATM", "EAT")))
nrow(atmeat)
atmeat <- na.omit(atmeat)
nrow(atmeat)
atmeat$race4g <- as.factor(atmeat$race4g)
```

Compare to Logistic Regression: Fit the Model

- CBPS is a modification of standard logistic regression model for propensity scores, so we will compare the results of logistic regression to CBPS
- Use the covariates discussed earlier

```
plog <- glm(atm ~ age + female + race4g + sfs + sps +
                  sds + ias + ces + eps + imds + bcs +
                  prmhtx,
                  family=binomial,
                  data=atmeat)

atmeat$ps1 <- ifelse(atmeat$atm==1,
                                  1, exp(predict(plog)))
```

Compare to Logistic Regression: Use dx.wts to Check Balance

```
b1 <- dx.wts(atmeat$ps1,
             data=atmeat,
             vars=c("age", "female", "race4g", "sfs",
                    "sps", "sds", "ias", "ces", "eps",
                    "imds", "bcs", "prmhtx"),
             treat.var="atm",
             estimand="ATT",
             x.as.weights=TRUE,
             sampw=NULL,
             perm.test.iters=0)
```

Run CBPS:Fit the Model and Get Weights

```
pcbps <- CBPS(atm ~ age + female + race4g + sfs + sps
                    + sds + ias + ces + eps + imds +
                    bcs + prmhtx,
                    data=atmeat,
                    ATT=TRUE)
atmeat$ps2 <- pcbps$weights
```

Run CBPS: Use dx.wts to Check Balance

```
b2 <- dx.wts(atmeat$ps2,
             data=atmeat,
             vars=c("age", "female", "race4g", "sfs",
                    "sps", "sds", "ias", "ces", "eps",
                    "imds", "bcs", "prmhtx"),
             treat.var="atm",
             estimand="ATT",
             x.as.weights=TRUE,
             sampw=NULL,
             perm.test.iters=0)
```

Run GBM: Fit the Model and Get Weights

```
pgbm <- ps(atm ~ age + female + race4g + sfs +
                 sps + sds + ias + ces + eps +
                 imds + bcs + prmhtx,
                 data=atmeat,
                 estimand="ATT",
                 n.trees=10000,
                 stop.method="es.max")

atmeat$ps3 <- unlist(pgbm$w)
```

Compare Weights

- Compare the summary tables to check overall balance and effective sample sizes

```
b1$summary
b2$summary
summary(pgbm)
```

Compare Weights

- Compare the summary tables to check overall balance and effective sample sizes

```
> rbind(b1$summary[2,],b2$summary[2,],summary(pgbm)[2,-8])

  type n.treat n.ctrl ess.treat  ess.ctrl     max.es    mean.es
2          442   2408       442   84.3727  0.6248657   0.1277555
2          442   2408       442  254.6165  0.0096218 0.005278614
es.max.ATT 442   2408       442  498.6925  0.2944589   0.0892948

             max.ks    mean.ks    iter
2          0.2120477 0.0782432    NA
2          0.2579716 0.05458636   NA
es.max.ATT 0.1491926 0.05359044  710
```

- Substance use frequency and illegal activities remain imbalanced with logistic weights
- The environment scale remains imbalanced with GBM (only ES > 0.20)

Compare ATT Estimates

```
d1 <- svydesign(id=~1, weights=~ps1, data=atmeat)
d2 <- svydesign(id=~1, weights=~ps2, data=atmeat)
d3 <- svydesign(id=~1, weights=~ps3, data=atmeat)

f1 <- svyglm(sfs8p12 ~ atm, design=d1)
f2 <- svyglm(sfs8p12 ~ atm, design=d2)
f3 <- svyglm(sfs8p12 ~ atm, design=d3)

res <- list(logit=summary(f1)$coef,
            cbps=summary(f2)$coef,
            gbm=summary(f3)$coef)

print(res)
```

Compare ATT Estimates

```
> print(res)
$logit
               Estimate  Std. Error    t value     Pr(>|t|)
(Intercept) 0.10139089  0.01884309    5.3807991   8.017795e-08
atm         0.01297383  0.01999528    0.6488445   5.164912e-01

$cbps
               Estimate  Std. Error    t value     Pr(>|t|)
(Intercept) 0.08779229  0.009401344   9.338270    1.912081e-20
atm         0.02657243  0.011538391   2.302958    2.135292e-02

$gbm
               Estimate  Std. Error    t value     Pr(>|t|)
(Intercept) 0.09421772  0.006259164   15.052766   2.487814e-49
atm         0.02014700  0.009161131   2.199183    2.794505e-02
```

Get Exact Balance Using Entropy Balancing or MDIA Weights

- The package `ebal` and the function `ebalance` find the entropy balancing or MDIA weights
- `ebalance` requires a data matrix of covariates including transforming class variables, I use `model.matrix` to generate this matrix

```
library(ebal)

tmp <- model.matrix(atm ~ age + female + race4g + sfs +
                          sps + sds + ias + ces + eps +
                          imds + bcs + prmhtx,
                          data=atmeat)

## drop the intercept ##
tmp <- tmp[,-1]
```

Get Exact Balance Using Entropy Balancing or MDIA Weights: Run Fit and Get Weights

```
pbal <- ebalance(Treatment=atmeat$atm, X=tmp)

atmeat$ps4 <- 1
atmeat$ps4[atmeat$atm==0] <- pbal$w
```

- Generates weights only for the control cases

Compare Weights

- Compare the summary tables to check overall balance and effective sample sizes

```
rbind(b1$summary[2,],b2$summary[2,],summary(pgbm)[2,-8],b4$summary[2])
```

	type	n.treat	n.ctrl	ess.treat	ess.ctrl	max.es	mean.es
1	logit	442	2408	442	84.3727	0.6248657	0.1277555
2	CBPS	442	2408	442	254.6165	0.0096218	0.00527861
3	GBM	442	2408	442	498.6925	0.2944589	0.0892948
4	EB	442	2408	442	257.4996	2.21122e-05	4.91149e-06

	max.ks	mean.ks	iter
1	0.2120477	0.0782432	NA
2	0.2579716	0.05458636	NA
3	0.1491926	0.05359044	710
4	0.2613977	0.05403091	NA

- ebalance yields ES of effectly zero but the KS is not zero, only balances the means
- Better mean balance for CBPS and EB comes at the cost of smaller ESS

Compare ATT Estimates

```
$logit
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 0.10139089 0.01884309 5.3807991 8.017795e-08
atm         0.01297383 0.01999528 0.6488445 5.164912e-01

$cbps
              Estimate  Std. Error  t value     Pr(>|t|)
(Intercept) 0.08779229 0.009401344 9.338270 1.912081e-20
atm         0.02657243 0.011538391 2.302958 2.135292e-02

$gbm
              Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 0.09421772 0.006259164 15.052766 2.487814e-49
atm         0.02014700 0.009161131  2.199183 2.794505e-02

$ebal
              Estimate  Std. Error  t value     Pr(>|t|)
(Intercept) 0.08734734 0.009234442 9.458866 6.291983e-21
atm         0.02701738 0.011402811 2.369361 1.788516e-02
```

Compare Coefficients

	logit	CBPS	entropy
(Intercept)	-3.24975893	-2.13708597	-2.189729100
age	0.03101638	-0.03416526	-0.029536458
female	-0.17825380	-0.30058025	-0.322173568
race4g2	0.33919320	0.32815272	0.348741470
race4g3	-1.14100021	-1.05909266	-1.061779956
race4g4	-0.93096936	-0.96710013	-0.985629576
sfs	-0.05263512	-0.26199278	-0.258247705
sps	0.03433795	0.07466073	0.070599036
sds	0.04362259	-0.01342978	-0.008163907
ias	6.23940874	4.98807125	4.920726580
ces	2.71148703	2.77443999	2.801470226
eps	3.49192440	3.37508938	3.261896874
imds	-0.09699256	-0.08302634	-0.080216288
bcs	-0.02381459	-0.02660521	-0.023874070
prmhtx	-0.02625957	-0.16914946	-0.199614380

Software - `entbal`

Installing the package from https://github.com/bvegetabile/entbal/

```
# install.packages("devtools")
devtools::install_github("bvegetabile/entbal")
```

Submission to `CRAN` in progress

To load the library

```
library(entbal)
```

The `entbal` function

```
entbal(
  formula,
  data = NULL,
  eb_pars = entbal::ebpars_default_binary(),
  suppress_warnings = FALSE
)
```

`formula`	R style formula representing exposure given pre-exposure covariates e.g., `TA` $\sim$ `X1 + X2 + X3`
`data`	DataFrame that contains the treatment variable `TA` and the covariates `X1`, `X2`, `X3`. If a DataFrame is not provided the variables will be searched for in the global environment
`eb_pars`	List of parameters required for the entropy balancing optimization.
`suppress_warnings`	A logical to print warnings from a function that checks the input variables to `eb_pars`. Defaults to `FALSE`.

The parameters: `eb_pars`

```
> ebp <- entbal::ebpars_default_binary()
$exp_type
[1] "binary"
$n_moments
[1] 3
$max_iters
[1] 1000
$estimand
[1] "ATE"
$verbose
[1] TRUE
$optim_method
[1] "l-bfgs-b"
$bal_tol
[1] 1e-08
$opt_constraints
[1] -1000  1000
```

Notes on the parameters

- Typical to balance three moments
- To adjust parameters from default lists

```
ebp$estimand <- 'ATT'
ebp$verbose <- FALSE
```

These defaults are often helpful for quickly iterating with the software and reducing frustrations by ensuring that the list of `eb_pars` has the required variables.

```
# Returns list of default parameters for binary exposure weights
ebpars_default_binary()
# Returns list of default parameters for multi-exposure weights
ebpars_default_multi()
# Returns list of default parameters for continuous exposure weights
ebpars_default_cont()
```

Example Analysis - LaLonde Data

Load data sets

```
# Experimental Evaluation Data
data("dw99nsw")
# Observation Controls from Current Population Survey
data("dw99cps1")
```

Variable Name	Type	Description/Options
`TA`	Exposure	Indicator of Treated Unit: $1 \equiv$ exposed.
`age`	Covariate	Age of individual - *Continuous*
`education`	Covariate	Highest Grade of Education - *Continuous*
`black`	Covariate	Indicator of being "Black" - *Binary*
`hispanic`	Covariate	Indicator of being "Hispanic" - *Binary*
`married`	Covariate	Indicator of being "Married" - *Binary*
`nodegree`	Covariate	Indicator of having "No Degree" - *Binary*
`RE74`	Covariate	Retrospective Earings: 1974 - *Continuous*
`RE75`	Covariate	Retrospective Earings: 1975 - *Continuous*
`RE78`	Outcome	Earings: 1978 - *Continuous*

Example Analysis - Experimental Evaluation Data

```
> summary(lm(RE78 ~ TA, data = dw99nsw))

Residuals:
   Min     1Q Median     3Q    Max
 -6349  -4555  -1829   2917  53959

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4554.8      408.0  11.162  < 2e-16 ***
TA            1794.3      632.9   2.835  0.00479 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6580 on 443 degrees of freedom
Multiple R-squared:  0.01782, Adjusted R-squared:  0.01561
F-statistic: 8.039 on 1 and 443 DF,  p-value: 0.004788
```

Example Analysis - Observational Data

```
> summary(lm(RE78 ~ TA, data = dw99cps1))

Residuals:
   Min     1Q Median     3Q    Max
-14847  -8992   1496  10718  53959

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14846.66      76.14  194.98   <2e-16 ***
TA          -8497.52     712.02  -11.93   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9629 on 16175 degrees of freedom
Multiple R-squared:  0.008729,  Adjusted R-squared:  0.008667
F-statistic: 142.4 on 1 and 16175 DF,  p-value: < 2.2e-16
```

Entropy Balancing

Data set-up

```
outcome <- dw99cps1[,ncol(dw99cps1)]
dat <- dw99cps1[,-ncol(dw99cps1)]
ebpars <- ebpars_default_binary(estimand = 'ATT')
```

Entropy Balancing

```
eb_lalonde <- entbal(TA ~ .,
                     eb_pars = ebpars,
                     data = dat)
#> iter   10 value -3.573880
#> iter   20 value -3.608738
#> iter   30 value -3.612042
#> iter   40 value -3.612062
#> final  value -3.612062
#> converged
#> Warning message:
#> In .check_pars(eb_pars, n_classes, n_obs, uniq_ta) :
#>   Reference exposure not chosen: eb_pars$which_z set to 1
```

Entropy Balancing Results

```
summary(eb_lalonde)
```

```
> res <- summary(eb_lalonde)
Reference levels for headers:
--------------------------------------------------------------------------------
Exposure 0: 0
Exposure 1: 1
--------------------------------------------------------------------------------
Unweighted Balance Statistics:
--------------------------------------------------------------------------------
           MeanGroup1 SEGroup1 MeanGroup0 SEGroup0 StdDiffMeans LogRatioSE MaxKS
age             25.82     7.14      33.23    11.04        -0.80      -0.44  0.34
education       10.35     2.01      12.03     2.87        -0.68      -0.36  0.41
black            0.84     0.36       0.07     0.26         2.43       0.13  0.77
hispanic         0.06     0.24       0.07     0.26        -0.05       0.06  0.01
married          0.19     0.39       0.71     0.45        -1.23       0.15  0.52
nodegree         0.71     0.45       0.30     0.46         0.91       0.21  0.41
RE74          2095.57  4873.40   14016.80  9569.50        -1.57      -0.67  0.60
RE75          1532.06  3210.54   13650.80  9270.11        -1.75      -1.06  0.65
--------------------------------------------------------------------------------
```

Entropy Balancing Results

```
summary(eb_lalonde)
```

```
--------------------------------------------------------------------------------
Weighted Balance Statistics:
--------------------------------------------------------------------------------
           MeanGroup1 SEGroup1 MeanGroup0 SEGroup0 StdDiffMeans LogRatioSE MaxKS
age             25.82     7.14      25.82     7.14            0          0  0.07
education       10.35     2.01      10.35     2.01            0          0  0.04
black            0.84     0.36       0.84     0.36            0          0  0.00
hispanic         0.06     0.24       0.06     0.24            0          0  0.00
married          0.19     0.39       0.19     0.39            0          0  0.00
nodegree         0.71     0.45       0.71     0.45            0          0  0.00
RE74          2095.57  4873.40    2095.60  4873.43            0          0  0.16
RE75          1532.06  3210.54    1532.14  3211.08            0          0  0.10
--------------------------------------------------------------------------------
TA: 1,   Original N = 185
       Weighted ESS = 185
TA: 0,   Original N = 15992
       Weighted ESS = 116.45
--------------------------------------------------------------------------------
```

Estimating the Outcome - `survey` package

```
library(survey)
sdesign <- svydesign(ids = ~1,
                     weights = eb_lalonde$wts,
                     data = dw99cps1)
sglm <- svyglm(RE78 ~ TA, sdesign)
summary(sglm)
```

Final Model Results

```
> summary(sglm)

Call:
svyglm(formula = RE78 ~ TA, design = sdesign)

Survey design:
svydesign(ids = ~1, weights = eb_lalonde$wts, data = dw99cps1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4957.1      434.9  11.397   <2e-16 ***
TA            1392.1      722.5   1.927    0.054 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 61441400)

Number of Fisher Scoring iterations: 2
```

Overview of Sensitivity Analysis

No Unobserved Covariates

- Causal effects estimates are unbiased only if there are no unobserved confounders
- We assume treatment assignment is ignorable given the observed pretreatment variables used in the modeling

Checking the Assumption

- We cannot test for ignorability
- Instead we conduct sensitivity analysis
- Try to assess if estimated treatment effect could plausibly be explained by an omitted variable

Exploring Possible Omitted Variables

- What are the two characteristics of an omitted variable required for creating bias in estimated treatment effects?
  1. It must be related to treatment assignment, that is its distribution differs across treatment and control groups and the more it differs, the greater the potential impact
  2. It must be related to (Y_0, Y_1) and the stronger relationship, the greater the potential for impact

Two Approaches

- Find how strong the relationship between the omitted variable and treatment or the outcome must be to have a consequential impact on inferences about treatment
  - Rosenbaum (2002), Ridgeway (2006)
- Estimated adjusted treatment effects that allow for uncertainty due to potential omitted variables
  - Greenland (2001), VanderWeele and Arah (2011)

Impact on Omitted Variable (1)

- Concern is that there exists an omitted variable which differs between the treatment and control groups even after weighting by the observed covariates X
- Let U denote the omitted variable
- If distribution of U differs in the treatment and control groups, even after weighting for X, then we have

$$Pr(A = 1 \mid X) \neq Pr(A = 1 \mid X, U)$$

  - This could be the definition of an omitted variable

Impact of an Omitted Variable (2)

- To make things concrete we will focus on ATT
- For a fixed U

$$\left(\frac{Pr(A=1 \mid X)}{1-Pr(A=1 \mid X)}\right) c = \frac{Pr(A=1 \mid X,U)}{1-Pr(A=1 \mid X,U)}$$

for some c $\neq 1$

- The further c is from 1 (in either direction) the stronger the effect of U on treatment or the more separated the distributions of U for treatment and control
  - $\frac{f(U|X,A=1)}{f(U|X,A=0)} = c$
- Doesn't matter if U is single omitted variable or a group of omitted variables

Simulating the Impact of the Omitted Variable

- Let $w_i = Pr(A = 1 \mid X_i)/(1 - Pr(A = 1 \mid X_i)$ = our ATT weight without U
- Let c_i be the distortion to the weight due to the omitted variable for observation i
- Then, $w_i^* = c_i \times w_i$ = the correct weight for observation i
- We can approximate the quantity of interest as

$$E[Y_0 \mid A = 1] = \sum_{i \in A_i = 0} w_i^* Y_i$$

How do we Simulate the U's

- What are the 2 characteristics of an omitted variable required for creating bias in estimated treatment effects?
  1. It must be related to treatment assignment, i.e., its distribution differs across treatment and control groups and the more it differs, the greater the potential for impact
  2. It must be related to (Y_0, Y_1) and the stronger the relationship, the greater the potential for impact

Simulating the U's

- We want to simulate $U_i's$ to reflect
  1. Different levels for the strength of the relationship between U and A
     - We capture this using effect size differences

$$ES = \frac{(\overline{U}_{A=1} - \overline{U}_{A=0})}{sd(U)}$$

  2. Different levels for the strength of the relationship between U and Y_0 (we can restrict to Y_0 here since we are estimating ATT)
     - We capture this using the correlation between U and Y (denoted by $\rho_{U,Y}$)

Now Map the c's back to User-Friendly Quantiles

- It can be shown that:

$$c = \pi \times (1 + exp\{-ES \times U\})$$

where $\pi = \Pr(A{=}1)$

- So - once we generate U's based on a fixed ES and a fixed $\rho_{U,Y}$ we can compute the needed c's
- Then we can compute updated weights and TE estimates

Our General Approach

- State relationship between confounder and treatment status as an effect size
- State relationship between confounder and outcome as correlation
- Simulate unobserved confounder with exactly that ES and correlation within sample
- Translate U into needed c_i's and calculate updated propensity score weights (w_i^*)
- Re-run outcome models & capture adjusted p-value and treatment effect size
- Produce user-friendly graphics showing key findings

Using OVtool - 1

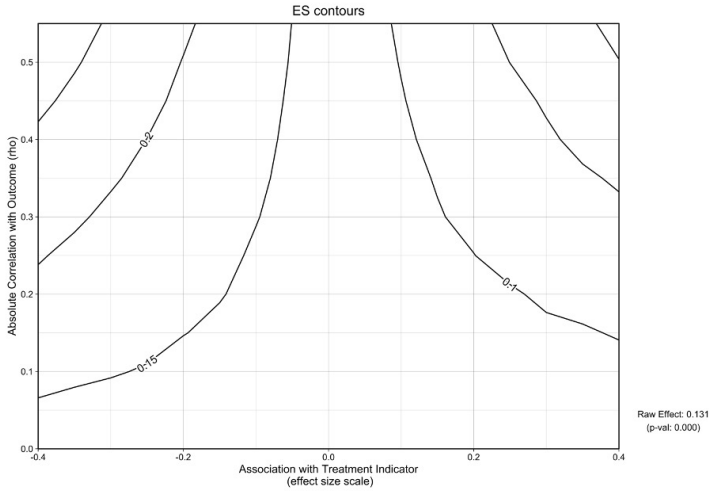
```
library(OVtool)
resultsSFS <- outcome_model(ps_object = NULL,
                                    stop.method = NULL,
                                    data = SUD,
                                    weights = "w_twang",
                                    treatment = "treat",
                                    outcome = "sfs8p12",
                                    model_covariates = c("eps7p_0", "sfs8p_0
    ",
                                                         "sati_0", "ada_0",
                                                         "recov_0", "tss_0",
                                                         "dss9_0"),
                                    estimand = "ATE")

summary(resultsSFS$mod_results)
```

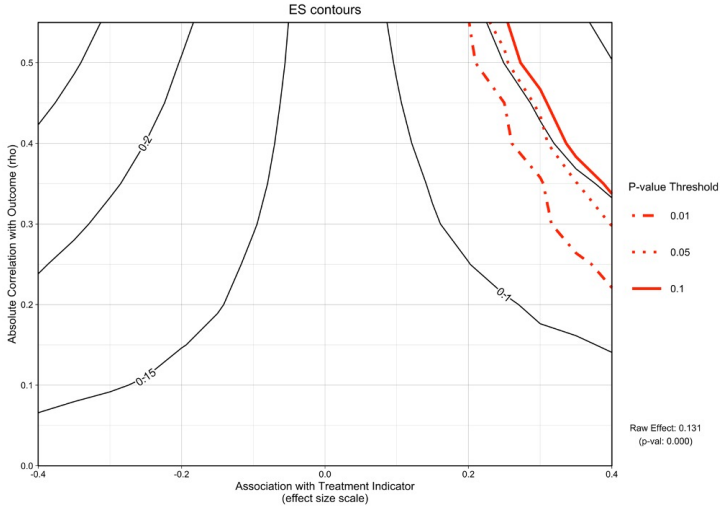
Using OVtool - 2

```
ovtool_results <- ov_sim(model_results = resultsSFS,
                         plot_covariates=c("eps7p_0", "sfs8p_0",
                                           "sati_0", "ada_0",
                                           "recov_0", "tss_0",
                                           "dss9_0","........"),
                         es_grid = NULL,
                         rho_grid = seq(0, 0.40, by = 0.05),
                         n_reps = 50,
                         progress = TRUE,
                         add = FALSE,
                         sim_archive = NULL)
```

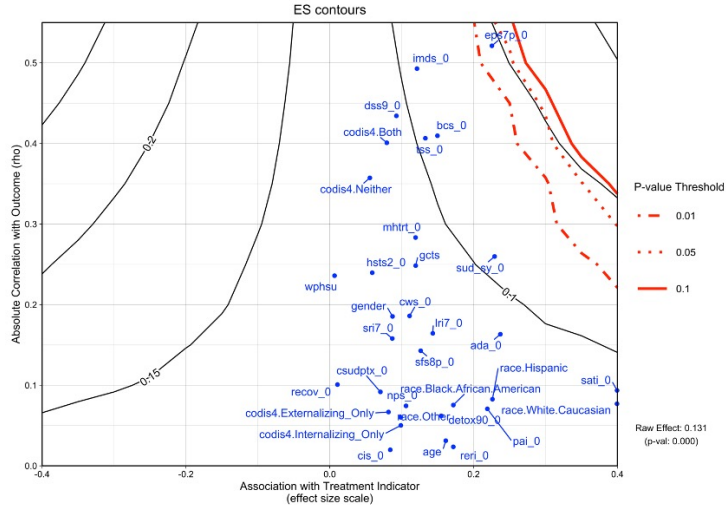
Graphic Comparing Two Evidence-Based Treatments: ACRA vs MET/CBT5



Graphic Comparing Two Evidence-Based Treatments: ACRA vs MET/CBT5



Graphic Comparing Two Evidence-Based Treatments: ACRA vs MET/CBT5



Useful References

- Liu, W., et al. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science, 14*(6), 570-580.
- Carnegie, N.B., et al. (2016). Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness, 9*(3), 395-420.
- Schneeweiss, S. (2006). Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiology Drug Safety, 15*(5), 291-303.
- Blackwell, M. (2014). A selection bias approach to sensitivity analysis for causal effects. *Political Analysis, 22*(2), 169-182.
- VanderWeele, T.J. and P. Ding (2017). Sensitivity analysis in observational research: Introducing the E-value. *Annals of Internal Medicine, 167*(4), 268-274.
- Griffin, B. A., et al. (2020). Expanding outcomes when considering the relative effectiveness of two evidence-based outpatient treatment programs for adolescents Journal of Substance Abuse Treatment, 118: 108075. PMCID: PMC519172

Other Useful Packages

- tipr - Tipping Point Analysis - D'Agostino-McGowan dissertation
  - Tells you tipping points at which results will lose significance
  - Only get simple stats; 1 value for one specific scenario
- treatSens - Bohme Carnegie, Harada and Hill 2006 JREE
  - Much closer to what we are trying for
  - Simulate U and then produce graphics like we have discussed, including use of observed covariates
  - BART can be used to estimate PS weights and/or outcome model
  - No user-friendly tutorial which would've been appreciated given complexity; the example code helped but ran into a bug which didn't have bandwidth to debug