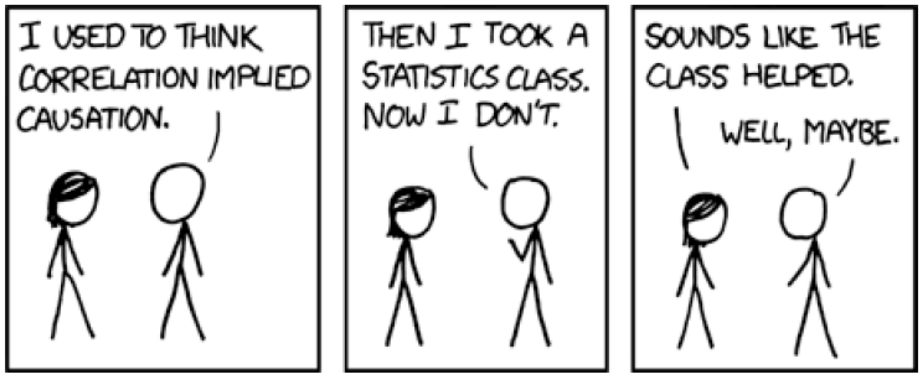
Causal Inference In Observational Studies

Michael Elliott
Professor, Biostatistics
Research Professor, Program in Surveys and Data Science
University of Michigan

mrelliot@umich.edu



Overview

- **Basic Concepts and Notation**
  - Potential Outcomes
  - Treatment/Exposure and the Fundamental Problem of Causal Inference
  - SUTVA
  - Causal Estimands
  - Assignment Mechanisms
- **Casual Inference and Randomized Controlled Trials**
  - Showing the difference in means in an RCT is a causal estimator
- **Propensity Scores**
  - Definition of a propensity score
  - Using propensity scores to obtain ACE
    - Required assumptions
  - Estimating propensity scores
- **Matching, Positivity, and Overlap**
  - Assessing overlap in covariate distributions
  - Using trimming to enhance positivity and improve covariate balance
  - Causal effect estimators using matching

What is causal inference?

- Definition of "cause": "a person or thing that gives rise to an action, phenomenon, or condition."
  - Moon causes tides.
  - The aspirin I took cured my headache.
- A philosophical concept fundamental to scientific inquiry.
- But for many years statistics struggled to put causal inference on a firm mathematical foundation.

Potential Outcomes

- How can we define casual inference with mathematical rigor?
- Donald Rubin and Jamie Robins, among others, undertook this challenge in the 1970's and 1980's.
  - "Rubin Causal Model" (Holland 1986).
- **Potential outcomes**: an outcome $Y(A)$ is defined as the value of an outcome Y at a given level of treatment A: if the treatment is binary (say take treatment $A=1$ versus take placebo $A=0$), we can think of each individual having a pair of potential outcomes $(Y_i(0), Y_i(1))$.
  - We only observe the outcome for the actual treatment given: $Y_i = Y_i(a_i)$.
  - This idea dates back at least to David Hume (1743, *An Enquiry Concerning Human Understanding*): "We may define a cause to be an object, followed by another…where, if the first object had not been, the second had never existed."

Treatment/Exposure

- Often consider treatment to be binary, but could be multivariate (different levels of dose or exposure) or continuous.
  - Many or even an infinite number of potential outcomes for each individual.
- How the treatment is obtained – assignment mechanism – is important for estimating causal effects, but not for defining them.
- Typically must be considered to be manipulable, but there is some controversy here.
  - "The moon causes tides": do we have to be able to displace or destroy the moon to make this statement?
  - Race and gender: do we have to biologically manipulate?
    - Discrimination studies that keep, e.g., CVs identical but manipulate race or gender of applicant: race/ethnicity and gender as social constructs.

Fundamental Problem of Causal Inference (Holland 1986)

Each individual has multiple potential outcomes, but only one can be observed.

- Cannot "replay" to see what the outcome would be under a different exposure.

- Consequently must either live with not fully identified models using Bayesian methods and proper priors to obtain inference from proper posteriors, or make assumptions to obtain fully identified models (possibly with sensitivity analyses).

SUTVA – Stable Unit Value Treatment Assumption

- In general, cannot assume that an assignment to person *i* will not impact the outcome for person *j*.
- Thus in the most general form the potential outcome for an individual in a population of size *N* is $Y_i(A_1,...,A_N)$.
- SUTVA is really two assumptions in one:
  1) Potential outcomes for any unit do not vary with the treatments assigned to other units (also termed no interference): $Y_i(A_1,...,A_N) = Y(A_i)$.
  2) There are no different versions of each treatment level which lead to different outcomes (also termed no hidden variants): $Y_i = Y(a_i)$ for $A_i = a_i$.
     When might these assumptions fail?

Causal Estimands

Under SUTVA, definition of causal estimands are rather straightforward:

- Unit level: change in outcome under treatment vs. control: $Y_i(1) - Y_i(0)$ or more generally $Y_i(A) - Y_i(A')$. (What about $Y_i(1)/Y_i(0)$ or $Y_i(A)/Y_i(A')$?)
- Average causal (treatment) effect (ACE) of treatment:

$$ACE = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right)$$

Causal Estimands

- If have simple random sample from population, can estimate ACE without bias using $\frac{1}{n}\sum_{i=1}^{n}\left(Y_i(1)-Y_i(0)\right)$.
- Can define ACE for subsamples of the population:
  - E.g., females: $\frac{1}{N_f}\sum_{i=1}^{N} I_i(Female)\left(Y_i(1)-Y_i(0)\right)$
  - Effect of treatment among the treated: $\frac{1}{N_t}\sum_{i=1}^{N} I_i(A_i=1)\left(Y_i(1)-Y_i(0)\right)$ where $N_t=\sum_{i=1}^{N} I(A_i=1)$.
  - Again, if simple random sample, replace population quantities with sample quantities to estimate.
    - Weighted estimators could be used for probability surveys
    - Active area of research for non-probability samples.

Assignment Mechanisms

- So far we have discussed estimands in terms of potential outcomes.
- But due to the fundamental problem of causal inference, we cannot compute the quantities – even the sample equivalents – from observed data.
- Hence we need to consider the assignment mechanism of the treatment/exposure.

Assignment Mechanisms

Closely linked with missing data and probability sampling concepts:

Can think of $Y_i(A_i = a_i) = Y_i^{obs}$ and $Y_i(A_i \neq a_i) = Y_i^{mis}$.

Randomized assignment

- Assignment is independent of potential outcomes conditional on observed covariates:
  $P(A \mid X, Y(1), Y(0)) = P(A \mid X)$
  - Equivalent to missing at random, probability sample where probability of selection is known but varies on design variables.
- "Classical experiment" where treatments are assigned with equal probability under the control of the experimenter:
  $P(A \mid X, Y(1), Y(0)) = P(A)$
  - Equivalent to missing completely at random, simple random sample.

There is a close connection between the impact of design on causal inference, missing data, and finite population inference:

Missing Data Mechanism	Sample Design	Causal Inference
Missing Completely At Random	Simple Random Sample	Randomized Experiment
Missing At Random	General Probability Sample	Conditional Randomization
Missing Not At Random	Non-probability Sample	Non-random Assignment

Casual Inference and Randomized Trials

- The development of randomized trials in the 20th century provided the framework for the mathematical and statistical description of causality.
- Ronald Fisher provided the first systematic overview of the design and analysis of randomized experiments in his text *The Design of Experiments* in 1935.
  - This is the basis for randomization inference (including for survey statistics)
- In parallel, Jerzey Neyman laid out a case for the value of randomized experiments in a 1934 paper "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9" published in the Polish language journal *Annals of Agricultural Science* and not translated into English until 1990.
  - Neyman was the first to introduce the concept of the potential outcome (although Rubin formalized and popularized it independently of Neyman)

Unbiased estimator of average causal effect in randomized trials

$$ACE_P = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right)$$

For the moment, assume that we have the entire population at our disposal to conduct a randomized experiment.

- We assign the vector of treatment/controls $(A_1,...A_N)$ at random with probability N_t/N to treatment and N_c/N to control.
- Only A_i is a random variable: we condition on $Y_i(1)$ and $Y_i(0)$. We denote expectation and variance with respect to A as E_A and V_A .

Let $\bar{Y}_1$ and $\bar{Y}_0$ correspond to the observed population means based on the treatment assignment. Then $\bar{Y}_1 = \frac{1}{N_1}\sum_{i:A_i=1} Y_i = \frac{1}{N_1}\sum_{i=1}^{N} A_i Y_i(1)$ where the last equality follows from SUTVA. Similarly $\bar{Y}_0 = \frac{1}{N_0}\sum_{i=1}^{N}(1-A_i)Y_i(0)$, and we have as the estimate of ACE

$$\hat{ACE}_P = \bar{Y}_1 - \bar{Y}_0 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{A_i Y_i(1)}{N_1/N} - \frac{(1-A_i)Y_i(0)}{N_0/N}\right)$$

Under randomization, A_i is independent of the potential outcomes $(Y_i(1), Y_i(0))$, so $\hat{ACE}_P$ is unbiased for the population average causal effect:

$$E_A\left(A\hat{C}E_P\right)=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{E_A(A_i)Y_i(1)}{N_1/N}-\frac{E_A(1-A_i)Y_i(0)}{N_0/N}\right)=$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{(N_1/N)Y_i(1)}{N_1/N}-\frac{(N_0/N)Y_i(0)}{N_0/N}\right)=\frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1)-Y_i(0)\right)=ACE_p$$

(Again, we are conditioning on the potential outcomes since they are fixed – we are making inference only to this population. Only the treatment indicators are random.)

Variance of $\hat{ACE}_P$

Even though we presumably have the entire population available to us, there is still sampling variability because there are $\binom{N}{N_1}$ ways to assign the vector of A_is. We can show

$$V_A\left(\hat{ACE}_P\right) = V_A\left(\bar{Y}_1 - \bar{Y}_0\right) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_{10}^2}{N}$$

where

$S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - \bar{Y}(1)\right)^2$, $S_0^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(0) - \bar{Y}(0)\right)^2$, and

$S_{10}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left((Y_i(1) - Y_i(0)) - (\bar{Y}(1) - \bar{Y}(0)\right)^2$.

- An unbiased estimator of S_1^2 is $s_1^2 = \frac{1}{N_1 - 1}\sum_{i:A_i=1}\left(Y_i - \bar{Y}_1\right)^2$ and similarly an unbiased estimator of S_0^2 is $s_0^2 = \frac{1}{N_0 - 1}\sum_{i:A_i=0}\left(Y_i - \bar{Y}_0\right)^2$.
- However, we cannot estimate S_{10}^2, since that would require simultaneously observing both potential outcomes for an observation (Fundamental Problem of Causal Inference).
- Since $S_{10}^2 \geq 0$, a conservative estimator of the variance of the causal effect is given by $v_A\left(\hat{\tau}_p\right) = \frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}$ (the "standard variance estimator").
  - This estimator is unbiased if there is a constant treatment effect (sometimes termed the "sharp null hypothesis"):
    $$\tau_P = Y_i(1) - Y_i(0) \text{ for all } i.$$

Let's consider a toy example with a "population" of 4 and 2 assigned randomly to each treatment arm: pain measure 90 minutes after taking either an analgesic or a placebo

i	$Y(1)$	$Y(0)$	$Y(1)-Y(0)$
1	2	4	-2
2	3	3	0
3	1	5	-4
4	0	3	-3
$\bar{Y}(A)$	1.5	3.75	-2.25
S_A	1.667	0.917	2.917

There are 6 possible treatment assignments:

i	$Y(1)$	$Y(0)$	A	A	A	A	A	A
1	2	4	1	1	1	0	0	0
2	3	3	1	0	0	1	1	0
3	1	5	0	1	0	1	0	1
4	0	3	0	0	1	0	1	1
$\overline{Y}_1$			2.5	1.5	1	2	1.5	0.5
$\overline{Y}_0$			4	3	4	3.5	4.5	3.5
$\overline{Y}_1 - \overline{Y}_0$			-1.5	-1.5	-3	-1.5	-3	-3

```
> mean(c(-1.5,-1.5,-3,-1.5,-3,-3))
[1] -2.25
> var(c(-1.5,-1.5,-3,-1.5,-3,-3))*5/6
[1] 0.5625
```

$$\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_{10}^2}{N} = \frac{1.667}{2} + \frac{0.917}{2} - \frac{2.917}{4} = 0.5625$$

Superpopulation Inference

Up until now we have treated our sample of size N as our population of interest. But often our object of inference is a super-population, typically an infinitely large population from which we assume our population is drawn at random.

We define these superpopulation quantities as the means and variances of $Y_i(1)$, $Y_i(0)$, and $Y_i(1) - Y_i(0)$:

$$E(Y_i(1)) = \mu_1, \;\; V(Y_i(1)) = \sigma_1^2$$
$$E(Y_i(0)) = \mu_0 \;, V(Y_i(0)) = \sigma_0^2$$
$$E(Y_i(1) - Y_i(0)) = \mu_1 - \mu_0 = \tau, \;\; V(Y_i(1) - Y_i(0)) = \sigma_{10}^2$$

The ACE is given by $E(Y_i(1)) - E(Y_i(0)) = \mu_1 - \mu_0 = \tau$.

We now consider expectation and variance with respect to both the sampling distribution of A (which we have already done) and with respect to the sampling variance of the potential outcomes of Y, using

$$E(\bullet) = E_Y(E_{A|Y}(\bullet))$$

$$V(\bullet) = E_Y(V_{A|Y}(\bullet)) + V_Y(E_{A|Y}(\bullet))$$

$$E(\hat{\tau}_P) = E_Y(E_{A|Y}(\hat{\tau}_P)) = E_Y(\tau_P) =$$

$$E_Y\left(\frac{1}{N}\sum_{i=1}^{N} Y_i(1) - \frac{1}{N}\sum_{i=1}^{N} Y_i(0)\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mu_1 - \frac{1}{N}\sum_{i=1}^{N} \mu_0 = \tau$$

(Note that the last equality assumes that the sample is a simple random sample from the superpopulation, so IID holds.)

$$V(\hat{\tau}_P) = E_Y\left(\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_{10}^2}{N}\right) + V_Y\left(\frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right)\right)$$

$$= \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0} - \frac{\sigma_{10}^2}{N} +$$

$$\frac{1}{N^2}\left(\sum_{i=1}^{N} V\left(Y_i(1) - Y_i(0)\right) + \sum_{i=1}^{N}\sum_{j=1}^{N} I(i \neq j)\underbrace{C\left(\left(Y_i(1) - Y_i(0)\right), \left(Y_j(1) - Y_j(0)\right)\right)}_{=0, \text{ since independent observations}}\right)$$

$$= \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0} - \frac{\sigma_{10}^2}{N} + \frac{1}{N^2}\sum_{i=1}^{N}\sigma_{10}^2 = \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}$$

"Standard" variance estimator holds for superpopulation inference.

Propensity Scores

The development of randomized designs, whether in assignment (clinical trials) or sampling (probability surveys), was a huge leap for scientific investigation in the sciences, particularly in medicine and public health, in the 20th century.

- Randomization allows ACE to be estimated from observed data without additional assumptions since all potential confounders are independent of treatment assignment.

However, there are limits to what can be accomplished, particularly with respect to randomized assignment:

- Ethical limitations (smoking)
- Compliance limitations (side effects)
- Practical limitations (expense, recruitment)

Propensity scores are a generalization of regression modeling that allow for causal inference under a set of assumptions.

Assignment mechanism assumptions

The assignment mechanism is the probability of being assigned a given level of treatment conditional on other covariates:

$$P(A_i \mid \mathbf{A}_{-i}, \mathbf{X}, \mathbf{Y(1)}, \mathbf{Y(0)})$$

There are a several assumptions that need to be made to make progress in using the assignment mechanism to make causal inference.

1) Stable Unit Treatment Value:

$$P(A_i \mid \mathbf{X}, \mathbf{Y(1)}, \mathbf{Y(0)}) = P(A_i \mid \mathbf{X}_i, Y(1)_i, Y(0)_i)$$

We've discussed this previously: the assignment for a given individual is independent of the assignments and outcomes for other individuals.

2) Positivity:

$$0 < P(A_i \mid \mathbf{X}_i, Y(1)_i, Y(0)_i) < 1$$

All units have a non-zero chance of assignment to all possible treatments.

3) Unconfoundedness

$$P(A_i \mid \mathbf{X}_i, Y(1)_i, Y(0)_i) = P(A_i \mid \mathbf{X}_i)$$

Assignment is independent of the potential outcomes conditional on $\mathbf{X}$.

Under these three assumptions,

$$P(A_i \mid \mathbf{A}_{-i}, \mathbf{X}, \mathbf{Y(1)}, \mathbf{Y(0)}) = e(\mathbf{x}_i)^{A_i} \left(1 - e(\mathbf{x}_i)\right)^{1-A_i}$$

where $e(\mathbf{x}_i) = P(A_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i)$ is termed the propensity score.

Note that positivity and unconfoundedness imply

a) $0 < e(\mathbf{x}_i) < 1$ for $i = 1, \ldots, N$

b) $A_i \perp \left(Y(1)_i, Y(0)_i\right) \mid \mathbf{X}_i$

Testing Positivity and Unconfoundedness Assumptions

- The positivity assumption can be tested from the available data by modeling the probability of selection as a function of covariates: substantial numbers of subjects with $\hat{P}(A_i = 1 | \mathbf{X}_i) \approx 0$ or $\hat{P}(A_i = 1 | \mathbf{X}_i) \approx 1$ suggest problems with positivity.
  - Largest concern is that there are values of
    $\hat{P}(A_i = 1 | \mathbf{X}_i) < \min_{i:A_i=1}\left(\hat{P}(A_i = 1 | \mathbf{X}_i)\right)$ for $i : A_i = 0$ and
    $\hat{P}(A_i = 1 | \mathbf{X}_i) > \max_{i:A_i=0}\left(\hat{P}(A_i = 1 | \mathbf{X}_i)\right)$ for $i : A_i = 1$
    - Lack of overlap: some elements in the data do not have a counterpart on the opposite treatment arm with a similar change of assignment.
    - Ignore if small; otherwise drop cases and restrict inference to segment of population that meets positivity assumption.

- Unconfoundedness is <u>not</u> testable:

  Unconfoundedness implies

$$P(Y_i(0) \mid A_i = 1, X_i) = P(Y_i(0) \mid A_i = 0, X_i)$$
$$P(Y_i(1) \mid A_i = 1, X_i) = P(Y_i(1) \mid A_i = 0, X_i)$$

  which in turn implies

$$P(Y_i^{mis} \mid A_i = a, X_i) = P(Y_i^{mis} \mid A_i = 1 - a, X_i)$$

  which is impossible to test (Fundamental Problem of Casual Inference)

Using the Propensity Score to Estimate Causal Effects

- Under the regular assignment mechanism assumptions, we could in principle average treatment effects at each level of X and average across the distribution of X to get an unbiased estimate of the ACE.

- Impractical in many settings because in the absence of actual randomization we will want to use a set of covariates as large as possible to improve the assumption of unconfoundedness.

- Propensity score to the rescue!

Why use propensity scores to obtain causal estimates?

If we have unconfoundedness, it would seem in principle that we could just use (linear) regression to estimate a treatment effect. If $E(Y_i) = \alpha + \tau A_i + \boldsymbol{\beta}^T \mathbf{X}_i$, then

$$\hat{\tau} = \overline{Y}_1 - \overline{Y}_0 - \hat{\boldsymbol{\beta}}^T \left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_0 \right)$$

However, this relies on the regression model being correct.

- What is Y is non-linear in $\mathbf{X}$, or has missing interactions?
- What if Y is not normally distributed?
- What if $\mathbf{X}$ is high dimensional?

Propensity scores are not model-free (require modeling of A on $\mathbf{X}$), but often this relationship may be better understood, and dimension reduction of $\mathbf{X}$ to a scalar propensity score allows for easy assessment of support.

- Also easy to explain to non-statisticians (“compare like to like”)

How do we obtain causal estimates with propensity scores?

Approach 1: use weighted estimators:

$$\hat{\tau} = \frac{1}{N}\sum_{i:A_i=1}\frac{Y_i^{obs}}{e(x_i)} - \frac{1}{N}\sum_{i:A_i=0}\frac{Y_i^{obs}}{1-e(x_i)}$$

Under unconfoundedness,

$$E\left(Y_i^{obs} \mid A_i = a, X_i = x_i\right) = E\left(Y_i\ (a) \mid X_i = x_i\right), \text{ and}$$

$$E_{Y|X}\left(\frac{1}{N}\sum_{i:A_i=1}\frac{Y_i^{obs}}{e(x_i)}\right) = \frac{1}{N}\sum_{i=1}^{N} E_{Y|X}\left(E_{A|Y,X}\left(\frac{A_i Y_i^{obs}}{e(x_i)}\right)\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N} E_{Y|X}\left(\frac{E_{A|Y,X}(A_i)Y_i^{obs}}{e(x_i)}\right) = \frac{1}{N}\sum_{i=1}^{N} E_{Y|X}\left(\frac{e(x_i)Y_i^{obs}}{e(x_i)}\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N} E\left(Y_i^{obs} \mid X_i = x_i, A_i = 1\right) = \frac{1}{N}\sum_{i=1}^{N} E\left(Y_i\ (1) \mid X_i = x_i\right)$$

Assuming a simple random sample, $\frac{1}{N}\sum_{i=1}^{N} E\left(Y_i\ (1) \mid X_i\right) \to \mu_1$ as $N \to \infty$

Similarly $E\left(\frac{1}{N}\sum_{i:A_i=0} \frac{Y_i^{obs}}{1-e(x_i)}\right) \to \mu_0$ as $N \to \infty$

So $E\left(\hat{\tau}\right) = \mu_1 - \mu_0 = \tau$

(Note this is similar to the Horvitz-Thompson estimator of a mean using survey weights, except instead of the reciprocal of the probability of selection, we use the reciprocal of the probability of assignment.)

Stabilized estimators

Because $e(x_i)$ and $\left(1-e(x_i)\right)$ may be close to 0, $e(x_i)^{-1}$ and $\left(1-e(x_i)\right)^{-1}$ may get very large and lead to unstable estimation of τ.

One way to deal with this is to replace $e(x_i)^{-1}$ with $\frac{Ne(x_i)^{-1}}{\sum_{i:A_i=1} e(x_i)^{-1}}$ and similarly $\left(1-e(x_i)\right)^{-1}$ with $\frac{N(1-e(x_i)^{-1})}{\sum_{i:A_i=0}(1-e(x_i))^{-1}}$,

so that $\hat{\tau} = \frac{\sum_{i:A_i=1} \frac{Y_i^{obs}}{e(x_i)}}{\sum_{i:A_i=1} \frac{1}{e(x_i)}} - \frac{\sum_{i:A_i=0} \frac{Y_i^{obs}}{1-e(x_i)}}{\sum_{i:A_i=0} \frac{1}{1-e(x_i)}}$.

(Equivalent to the weighted Hayek estimator in survey statistics.)

Approach 2: stratified estimators:

Estimate $\hat{\tau}_j$ within classes defined by the ordered propensity score, and find $\hat{\tau} = \sum_{j=1}^{J} \frac{N_j}{N} \hat{\tau}_j$, where N_j is the number of observations in the jth class.

Historically letting J=5 and using the quintiles of the propensity score to form the classes has been used, although more data driven methods are available.

- Somewhat robust to misspecification of the propensity model.
  - May be more biased since large weights are downweighted.
- More stable than fully-weighted estimation.

Variance estimation: stabilized estimator

Can borrow from survey statistics to compute variance treating estimated propensities scores as constant:

$$v(\hat{\tau}) = v\left(\hat{\bar{Y}}(1) - \hat{\bar{Y}}(0)\right) = v\left(\hat{\bar{Y}}(1)\right) + v\left(\hat{\bar{Y}}(0)\right)$$

where

$$v\left(\hat{\bar{Y}}(a)\right) = \frac{\frac{N_a}{N_a - 1}\sum_{i:A_i=a}\left(z_i - \bar{z}(a)\right)^2}{\left(\sum_{i:A_i=a} w_i\right)^2}$$, where $z_i = w_i\left(Y_i^{obs} - \hat{\bar{Y}}(a)\right)$ for

$$w_i = \begin{cases} e(x_i)^{-1} \text{ for } A_i = 1 \\ (1 - e(x_i))^{-1} \text{ for } A_i = 0 \end{cases}$$ and $$\hat{\bar{Y}}(a) = \frac{\sum_{i:A_i=a} w_i Y_i^{obs}}{\sum_{i:A_i=a} w_i},$$

$$\bar{z}(a) = \frac{1}{N_a}\sum_{i:A_i=a} z_i .$$

Variance estimation: stratified estimator:

$v(\hat{\tau}) = \sum_{j=1}^{J} \left(\frac{N_j}{N} \right)^2 v(\hat{\tau}_j)$, where $v(\hat{\tau}_j)$ is determined as on previous slide.

Can use standard survey software to compute variances of weighted means.

- PROC SURVEY procedures in SAS
- "survey" package in R

Propensity Score Construction

Rosenbaum and Rubin (1983) argue that construction of the propensity score should be based only on the covariates **X**; the values of *Y* should be ignored.

- Corresponds to randomization, which balances all covariates regardless of relationship to outcome
  - Avoids "data snooping" to look for PS model that will provide the a priori idea of the effect
  - Counterargument is that only values of *X* associated with *Y* can serve as confounders: should focus on these as "prognostic scores" (Hansen 2008).
  - Will focus on constructing as a function of **X**.

- Typically do not know propensity score and must estimate from available data:
  1) Estimate (based on substantive knowledge about the assignment mechanism to the extent possible)
  2) Check balance in strata based on propensity score
  3) Repeat as necessary until balance is achieved.
- Historically have used logistic regression:

$$\log\left(\frac{P(A_i=1\,|\,\mathbf{X}_i)}{1-P(A_i=1\,|\,\mathbf{X}_i)}\right)=\mathbf{X}_i^T\boldsymbol{\alpha}$$

  - Fit using data and construct $e(x_i)=\dfrac{e^{\mathbf{x}_i^T\hat{\boldsymbol{\alpha}}}}{1+e^{\mathbf{x}_i^T\hat{\boldsymbol{\alpha}}}}$.
  - Interpretation of $\boldsymbol{\alpha}$ is not important, so more recent advanced regression techniques such as random forests, Bayesian Additive Regression Trees, or lasso are increasingly used, particularly if $\mathbf{X}$ is high-dimensional.

Let's first look at an example in simulation.

$$C_i \sim 0.5N(0,1) + 0.5N(1,1)$$
$$A_i \mid C \sim BIN\left(1, e^{(C_i - 0.5)} \Big/ \left(1 + e^{(C_i - 0.5)}\right)\right)$$
$$Y_i \mid A_i, C_i = A_i + 2\exp(C_i) + \varepsilon_{y_i}, \quad \varepsilon_{y_i} \sim N(0,1)$$
$$i = 1,\ldots,5000$$

C is a confounder: continuous, bimodal (low and high modes)
A is the treatment: more likely with larger values of C
Y is the outcome: causal effect is 1, but also associated with C.

```
n<-5000
Z<-rbinom(n,1,.5)
C<-rnorm(n,Z,1)
A<-rbinom(n,1,expit(C-.5))
Y<-2*exp(C)+A+rnorm(n,0,1)
```

Let's consider 4 approaches to estimating treatment effects:

- Unadjusted $(\bar{Y}_1 - \bar{Y}_0)$
- Linear model with C incorrectly specified:
  $Y_i = \beta_0 + \beta_1 A_i + \beta_2 C_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$
- Propensity scores with propensity correctly estimated:
  $$\log\left(\frac{P(A_i = 1)}{1 - P(A_i = 1)}\right) = \alpha_0 + \alpha_1 C_i$$
  $$\frac{\sum_{i:A_i=1} \frac{Y_i}{e(C_i)}}{\sum_{i:A_i=1} \frac{1}{e(C_i)}} - \frac{\sum_{i:A_i=0} \frac{Y_i}{1-e(C_i)}}{\sum_{i:A_i=0} \frac{1}{1-e(C_i)}}, \quad e(C_i) = \frac{\exp(\hat{\alpha}_0 + \hat{\alpha}_1 C_i)}{1 + \exp(\hat{\alpha}_0 + \hat{\alpha}_1 C_i)}$$
- Linear model with C correctly specified:
  $Y_i = \beta_0 + \beta_1 A_i + \beta_2 \exp(C_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$

```
summary(lm(Y~A))

summary(lm(Y~A+C))

p<-glm(A~C,family=binomial)$fitted.values
wt<-A/p+(1-A)/(1-p)
mydesign <- svydesign(ids = ~1, data = data.frame(Y,A), weights = ~wt)
summary(svyglm(Y ~ A, design = mydesign))

expC<-exp(C)
summary(lm(Y~A+expC))
```

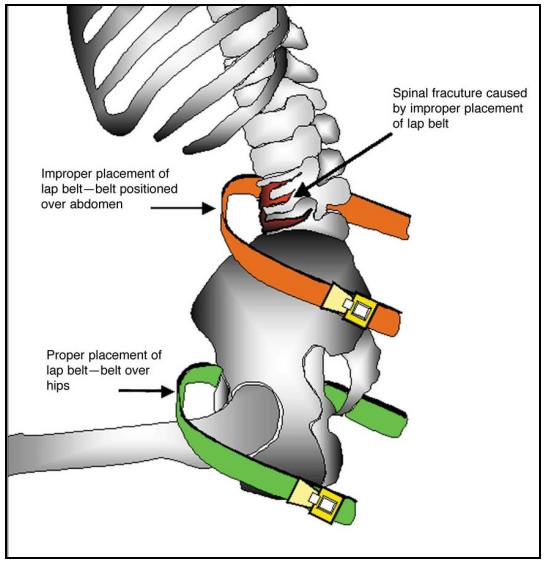
Result from a single simulated dataset:

	$\hat{ACE}$	SE	95% CI
Unadjusted	6.517	0.232	(6.061,6.972)
Misspecified Model	0.594	0.178	(0.244,0.943)
Propensity Score	1.059	0.374	(0.326,1.791)
Correctly Specified Model	0.987	0.030	(0.928,1.045)

- Extreme confounding in unadjusted model.
- Misspecified linear model also substantially biased.
- Propensity score and correctly specified model unbiased, but PS approach far more variable (price to pay for outcome model robustness).

Example: effect of booster seat use on risk of injury to children in passenger vehicle crashes.

- Partners in Child Passenger Safety: surveillance of children in passenger vehicle crashes.
- Probability sample of 8,000 children aged 4-7 seated in booster seats vs. seat belts (2007).



- Question of interest: do booster seats reduce risk of injury to children in passenger vehicle crash and, if so, by how much?
  - Injury is defined as a concussion or Abbreviated Injury Scale (AIS) > 2 (broken bone, internal injury).
- Booster seats and seatbelts are not randomly assigned:
  - There may be differences in the age of the children, the type of vehicles driven, and the type of drivers between the two groups.

(We will ignore issues of the complex sample design, although relatively straightforward to incorporate given that we will use survey software to analyze.)

Potential confounders of booster seat use and injury:

- Age of child
- Age of driver
- Type of vehicle
- Crash severity
- Exposure to an airbag

	Seat Belt	Booster Seat	p-value
N	5158	2842	---
Mean age (years)	**5.81**	**4.88**	**<.0001**
% passenger car	**50.3**	**44.9**	**<.0001**
% large van	2.2	1.8	.27
% pickup truck	**6.5**	**5.1**	**.0077**
% minivan	**18.8**	**23.3**	**<.0001**
% SUV	**22.1**	**24.9**	**.0054**
% driver<25 years old	**10.7**	**8.8**	**.0053**
% towaway	**60.9**	**53.4**	**<.0001**
% intrusion	**20.6**	**17.7**	**.0016**
% airbag exposure	**4.8**	**0.4**	**<.0001**

Fit logistic regression model with booster seat use (vs. seat belt) as outcome:

```
Call:
glm(formula = boost ~ as.factor(age_yrs) + lvan + truck + mvan +
    suv + youngd + towaway + intru + airbag, family = ("binomial"))

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.67404    0.06356  10.605  < 2e-16 ***
as.factor(age_yrs)5 -0.93963    0.06542 -14.363  < 2e-16 ***
as.factor(age_yrs)6 -1.72812    0.07196 -24.016  < 2e-16 ***
as.factor(age_yrs)7 -2.47917    0.08192 -30.265  < 2e-16 ***
lvan                -0.19346    0.18714  -1.034 0.301228
truck               -0.26540    0.11514  -2.305 0.021169 *
mvan                 0.30624    0.06760   4.530 5.89e-06 ***
suv                  0.22356    0.06551   3.413 0.000643 ***
youngd              -0.40242    0.08931  -4.506 6.61e-06 ***
towaway             -0.20126    0.05500  -3.659 0.000253 ***
intru               -0.10279    0.06965  -1.476 0.139957
airbag              -2.33052    0.31574  -7.381 1.57e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null deviance: 10410.2  on 7999  degrees of freedom
Residual deviance:  8886.8  on 7988  degrees of freedom
```

```
AIC: 8910.8
```

Drop intrusion and large van.

Consider 2-way interactions:

Interaction	$-2(l_{\text{int}} - l_{base})$	df	p-value
age*vehicle	**17.64**	**9**	**.040**
age*driver<25	3.06	3	.38
age*towaway	2.14	3	.54
age*airbag	**11.38**	**3**	**.0098**
vehicle*driver<25	4.21	3	.24
vehicle*towaway	4.63	3	.20
vehicle*airbag	0.43	3	.93
driver<25*towaway	0.01	1	.93
driver<25*airbag	0.41	1	.52
towaway*airbag	0.89	1	.34

Fit model with age*vehicle and age*airbag interaction:

```
Call:
glm(formula = boost ~ as.factor(age_yrs) + truck + mvan + suv +
    youngd + towaway + airbag + as.factor(age_yrs) * truck +
    as.factor(age_yrs) * mvan + as.factor(age_yrs) * suv + as.factor(age_yrs) *
    airbag, family = ("binomial"))

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.72768    0.07382   9.857  < 2e-16 ***
as.factor(age_yrs)5       -0.97937    0.09242 -10.597  < 2e-16 ***
as.factor(age_yrs)6       -1.95718    0.10609 -18.449  < 2e-16 ***
as.factor(age_yrs)7       -2.52269    0.11894 -21.210  < 2e-16 ***
truck                     -0.45942    0.18810  -2.442 0.014587 *
mvan                       0.13256    0.12220   1.085 0.278034
suv                        0.18900    0.12045   1.569 0.116623
youngd                    -0.40677    0.08966  -4.537 5.71e-06 ***
towaway                   -0.22591    0.05241  -4.310 1.63e-05 ***
airbag                    -3.98246    1.01606  -3.920 8.87e-05 ***
as.factor(age_yrs)5:truck  0.21523    0.27721   0.776 0.437490
as.factor(age_yrs)6:truck  0.33462    0.31683   1.056 0.290900
as.factor(age_yrs)7:truck  0.58001    0.35817   1.619 0.105373
as.factor(age_yrs)5:mvan   0.14020    0.17069   0.821 0.411440
as.factor(age_yrs)6:mvan   0.61282    0.18163   3.374 0.000741 ***
as.factor(age_yrs)7:mvan  -0.06000    0.21671  -0.277 0.781871
as.factor(age_yrs)5:suv   -0.04266    0.16523  -0.258 0.796277
as.factor(age_yrs)6:suv    0.18201    0.18341   0.992 0.321030
as.factor(age_yrs)7:suv    0.10068    0.19887   0.506 0.612671
as.factor(age_yrs)5:airbag  1.93014    1.14224   1.690 0.091069 .
as.factor(age_yrs)6:airbag  2.75581    1.10273   2.499 0.012451 *
as.factor(age_yrs)7:airbag -9.57769  160.95225  -0.060 0.952549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null deviance: 10410.2  on 7999  degrees of freedom
Residual deviance:  8860.4  on 7978  degrees of freedom
AIC: 8904.4
```

Refit with explicit dummy variables for age:

```
age5<-as.numeric(age_yrs==5)
age6<-as.numeric(age_yrs==6)
age7<-as.numeric(age_yrs==7)

summary(glm(boost~age5+age6+age7+truck+mvan+suv+youngd+towaway+airbag+
age5*truck+age5*mvan+age5*suv+
age6*truck+age6*mvan+age6*suv+
age7*truck+age7*mvan+age7*suv+
age5*airbag+age6*airbag,family=("binomial")))
```

```
Call:
glm(formula = boost ~ age5 + age6 + age7 + truck + mvan + suv +
    youngd + towaway + airbag + age5 * truck + age5 * mvan +
    age5 * suv + age6 * truck + age6 * mvan + age6 * suv + age7 *
    truck + age7 * mvan + age7 * suv + age5 * airbag + age6 *
    airbag, family = ("binomial"))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.72823    0.07382   9.865  < 2e-16 ***
age5        -0.97990    0.09242 -10.603  < 2e-16 ***
age6        -1.95772    0.10609 -18.454  < 2e-16 ***
age7        -2.52418    0.11891 -21.228  < 2e-16 ***
truck       -0.45949    0.18815  -2.442 0.014599 *
mvan         0.13215    0.12221   1.081 0.279566
suv          0.18865    0.12047   1.566 0.117358
youngd      -0.40685    0.08966  -4.538 5.69e-06 ***
towaway     -0.22593    0.05242  -4.310 1.63e-05 ***
airbag      -4.17296    1.01113  -4.127 3.67e-05 ***
age5:truck   0.21529    0.27724   0.777 0.437428
age5:mvan    0.14060    0.17070   0.824 0.410119
age5:suv    -0.04231    0.16524  -0.256 0.797906
age6:truck   0.33468    0.31686   1.056 0.290863
age6:mvan    0.61323    0.18164   3.376 0.000735 ***
age6:suv     0.18236    0.18343   0.994 0.320125
age7:truck   0.58007    0.35816   1.620 0.105327
age7:mvan   -0.05916    0.21670  -0.273 0.784858
age7:suv     0.10143    0.19886   0.510 0.610011
age5:airbag  2.12064    1.13785   1.864 0.062359 .
age6:airbag  2.94632    1.09817   2.683 0.007298 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


    Null deviance: 10410.2  on 7999  degrees of freedom
Residual deviance:  8860.8  on 7979  degrees of freedom
AIC: 8902.8
```

Unadjusted mean difference in injury rates:

$\overline{Y}_1 = 0.109 \pm .006 \;\; \overline{Y}_0 = 0.145 \pm .005$

$\overline{Y}_1 - \overline{Y}_0 = -.037$, with 95% CI (-.052,-.022)

Now we analyze using weighted estimator: $\hat{\tau} = \overline{y}_1^w - \overline{y}_0^w$, where

$$\overline{y}_a^w = \frac{\sum_{i=1}^n w_{ia} y_i}{\sum_{i=1}^n w_{ia}} \text{ for } w_{ia} = \begin{cases} \left(A_i e(\mathbf{x}_i) + (1 - A_i)(1 - e(\mathbf{x}_i)) \right)^{-1} & \text{if } A_i = a \\ 0 \text{ if } A_i \neq a \end{cases}.$$

$\hat{\tau} = -0.017$ with 95% CI of (-.036,.002)

R code:

```
# Propensity score model
propmod <- glm(boost ~ age5 + age6 + age7 + truck + mvan + suv + youngd + towaway +
               airbag + age5*truck + age5*mvan + age5*suv + age6*truck +
               age6*mvan + age6*suv + age7*truck + age7*mvan + age7*suv +
               age5*airbag + age6*airbag,
               family = "binomial")

# Define estimated propensity score.
propscore <- propmod$fitted.values

# Create a vector of inverse propensity weights.
w <- boost / propscore + (1-boost) / (1-propscore)

# Create a survey design object.
mydesign <- svydesign(ids = ~1, data = data.frame(inj, boost), weights = ~w)

# Fit a weighted regression model.
mymodel <- svyglm(inj ~ boost, design = mydesign)

# Print summary of the weighted model
summary(mymodel)
```

Next, we consider estimating the ACE using stratification in qunitiles based on the propensity score:

$$Y = \beta_0 + \sum_{j=1}^{4} \beta_j I(Q = (j+1)) + \beta_5 I(B = 1) + \sum_{j=6}^{9} \beta_j I(B = 1) I(Q = (j-4))$$

where Y is an indicator for injury, B an indicator for booster seat restraint, and Q an indicator for PS quintile membership.

First, we compute the quintiles of $e(x_i)$: [.001-.150),[.150,.226),[.226-.418),[.418-.623),[.623,.715).

	Q1			Q2			Q3			Q4			Q5		
	SB	BS	p	SB	BS	p	SB	BS	p	SB	BS	p	SB	BS	p
N	1406	170	---	1080	223	---	1298	613	---	811	741	---	563	1095	---
Mean age (years)	**6.74**	**6.85**	**.0054**	6.52	6.49	.418	**5.60**	**5.52**	**.0013**	**4.80**	**4.72**	**.0005**	4.00	4.00	1
% passenger car	73.3	75.2	.58	36.2	36.3	.97	47.7	49.2	.54	34.1	34.3	.96	49.6	46.7	.27
% large van	2.8	2.9	.90	**1.0**	**4.1**	**.030**	2.2	1.8	.52	1.6	0.9	.25	**3.7**	**1.8**	**.034**
% pickup truck	5.0	5.2	.89	11.7	11.7	.99	6.5	7.3	.49	6.9	8.8	.17	0	0	1
% minivan	15.6	14.7	.75	12.9	12.5	.87	20.7	23.5	.18	26.5	28.1	.49	22.0	23.5	.53
% SUV	3.2	1.8	.20	38.1	35.4	.44	**22.8**	**18.1**	**.016**	30.8	27.9	.21	24.7	28.8	.13
% driver<25 years old	16.9	11.8	.056	3.8	3.5	.88	12.6	11.5	.51	**13.2**	**19.6**	**.0007**	0.7	0.6	.87
% towaway	75.6	70.0	.13	61.9	63.7	.63	**55.8**	**61.2**	**.025**	43.3	41.7	.53	**59.9**	**52.3**	**.0033**
% intrusion	22.9	18.8	.21	22.7	17.9	.099	18.7	21.7	.13	**17.2**	**12.7**	**.011**	20.1	18.7	.51
% airbag exposure	**17.7**	**6.5**	**<.0001**	0	0	1	0	0	1	0	0	1	0	0	1

- Balance greatly improved (still not perfect).

Then we fit the model estimating to quintile regression coefficients:

```
fit<-lm(inj~q2+q3+q4+q5+boost+boost*q2+boost*q3+boost*q4+boost*q5)

Call:
lm(formula = inj ~ q2 + q3 + q4 + q5 + boost + boost * q2 + boost *
    q3 + boost * q4 + boost * q5)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.192034   0.008997  21.343  < 2e-16 ***
q2          -0.056849   0.013651  -4.164 3.15e-05 ***
q3          -0.067997   0.012986  -5.236 1.68e-07 ***
q4          -0.078594   0.014876  -5.283 1.30e-07 ***
q5          -0.048162   0.016826  -2.862  0.00422 **
boost       -0.033211   0.027395  -1.212  0.22544
q2:boost     0.028070   0.036964   0.759  0.44763
q3:boost     0.057624   0.031998   1.801  0.07176 .
q4:boost     0.014237   0.032318   0.441  0.65956
q5:boost    -0.026643   0.032506  -0.820  0.41244
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3374 on 7990 degrees of freedom
Multiple R-squared:  0.0102,    Adjusted R-squared:  0.009085
F-statistic: 9.149 on 9 and 7990 DF,  p-value: 6.606e-14
```

Note that

$$\hat{\tau}_1 = \underbrace{\left(\hat{\beta}_0 + \hat{\beta}_5\right)}_{\substack{\text{injury rate for} \\ \text{booster in Q1}}} - \underbrace{\left(\hat{\beta}_0\right)}_{\substack{\text{injury rate for} \\ \text{seat belt in Q1}}} = \hat{\beta}_5$$

$$\hat{\tau}_2 = \underbrace{\left(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_5 + \hat{\beta}_6\right)}_{\substack{\text{injury rate for} \\ \text{booster in Q2}}} - \underbrace{\left(\hat{\beta}_0 + \hat{\beta}_1\right)}_{\substack{\text{injury rate for} \\ \text{seat belt in Q2}}} = \hat{\beta}_5 + \hat{\beta}_6$$

Similarly,

$$\hat{\tau}_3 = \hat{\beta}_5 + \hat{\beta}_7$$

$$\hat{\tau}_4 = \hat{\beta}_5 + \hat{\beta}_8$$

$$\hat{\tau}_5 = \hat{\beta}_5 + \hat{\beta}_9$$

Using an unweighted estimator stratified by PS, we obtain

$$\hat{\tau} = \sum_{j=1}^{5} \frac{N_j}{N} \hat{\tau}_j = \hat{\beta}_5 + \sum_{j=2}^{5} \frac{N_j}{N} \hat{\beta}_{j+4}$$

We can compute this "by hand", or use the glht function from the R package

```
N<-8000
P<-c(sum(q1==1),sum(q2==1),sum(q3==1),sum(q4==1),sum(q5==1))/N
summary(glht(fit,linfct=t(c(0,0,0,0,0,1,P[2],P[3],P[4],P[5]))))

Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = inj ~ q2 + q3 + q4 + q5 + boost + boost * q2 + boost *
    q3 + boost * q4 + boost * q5)

Linear Hypotheses:
        Estimate Std. Error t value Pr(>|t|)
1 == 0 -0.017634   0.009234   -1.91   0.0562 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So now our ACE is -.018 with 95% CI of (-.036,.000)

Matching, Positivity, and Overlap

- Key to the value of randomization is the fact that the treatment variable is independent of all other variables.
  - $A_i \perp \mathbf{X}_i, \mathbf{U}_i$ for both observed $\mathbf{X}_i$ and unobserved $\mathbf{U}_i$
  - This means that confounding is avoided by design.

- Even if A_i is not randomized, we might be in the happy situation where the distribution of $\mathbf{X}_i$ under treatment and control are approximately equal (e.g., $\bar{\mathbf{X}}_0 \approx \bar{\mathbf{X}}_1$).
  - Covariates are "balanced."
  - Appear to approximate randomized design (but could still be unbalanced based on unobserved covariates $\mathbf{U}_i$).

- More commonly the covariate distribution is unbalanced.
  - As imbalances become extreme, this can lead to lack of overlap.

- Why do we care about positivity/overlap?
  - We don't, if using standard regression; remember $\hat{\tau} = \overline{Y}_1 - \overline{Y}_0 - \hat{\boldsymbol{\beta}}^T \left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_0 \right)$.
    - $\hat{\boldsymbol{\beta}}$ is estimated using all values of X; it doesn't matter if there are values of $\mathbf{X}_1$ that do not appear in $\mathbf{X}_0$ or vice-versa.
    - But this relies on the model being correct – including no interactions with **X** and *A*.
    - If there are interactions included in the model, this requires projection into the unobserved space: there is no data to test whether the model is correctly specified.
  - If we are using propensity score weighting, the weight will approach ∞ if there are values of $\mathbf{X}_1$ that do not appear in $\mathbf{X}_0$ or vice-versa.
  - If we are using matching there will be treated observations for which there is no match in the controls or vice-versa.

Assessing covariate balance with a scalar X

- Normalized difference:

$$\Delta_{ct} = \frac{\mu_1 - \mu_0}{\sqrt{\left(\sigma_1^2 + \sigma_0^2\right)/2}} \ , \quad \mu_1 = E\left(X_i \mid A_i = 1\right) \ , \quad \mu_0 = E\left(X_i \mid A_i = 0\right) \ ,$$

$$\sigma_1^2 = V\left(X_i \mid A_i = 1\right) \ , \ \sigma_0^2 = V\left(X_i \mid A_i = 0\right) \ .$$

- Estimate with sample quantities:

$$\hat{\Delta}_{ct} = \frac{\overline{x}_1 - \overline{x}_0}{\sqrt{\left(s_1^2 + s_0^2\right)/2}}$$

  - Note that this is <u>not</u> the same at the usual t-statistics testing equality of means in the treatment and control samples:

$$t_{ct} = \frac{\overline{x}_1 - \overline{x}_0}{\sqrt{s_1^2/n_1 + s_0^2/n_2}}$$

(There is no sample size, so these measures are independent of sample size.)

- Can interpret $\hat{\Delta}_{ct}$ like a *z*-statistic:
  - $\left|\hat{\Delta}_{ct}\right| \leq 0.25$ excellent to adequate balance.
  - $0.25 < \left|\hat{\Delta}_{ct}\right| \leq 0.5$ adequate to poor balance.
  - $0.5 < \left|\hat{\Delta}_{ct}\right| \leq 1.0$ poor to very poor balance.
  - $1.0 < \left|\hat{\Delta}_{ct}\right| \leq 1.5$ very poor to terrible balance.
  - $\left|\hat{\Delta}_{ct}\right| \geq 1.5$ terrible balance.

Assessing covariate balance with a multivariate X

This idea of normalized differences can be carried forward to multivariate distributions:

$$\Delta_{ct}^{m} = \sqrt{\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\right)^T \left(\frac{\Sigma_1 + \Sigma_0}{2}\right)^{-1} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\right)}$$

where now $\boldsymbol{\mu}_1$ corresponds to the vector of means and Σ_1 for the corresponding variance covariance matrix for $\mathbf{X}_i$ among subjects assigned to treatment, and similarly $\boldsymbol{\mu}_0$ and Σ_0 for subjects assigned to control.

We can estimate with

$$\hat{\Delta}_{ct}^{m} = \sqrt{\left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0\right)^T \left(\frac{\hat{\Sigma}_1 + \hat{\Sigma}_0}{2}\right)^{-1} \left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0\right)}$$

where $\bar{\mathbf{X}}_1 = n_1^{-1} \sum_{i=1}^{n} A_i \mathbf{X}_i$, $\bar{\mathbf{X}}_0 = n_0^{-1} \sum_{i=1}^{n} (1 - A_i) \mathbf{X}_i$,

$\hat{\Sigma}_1 = (n_1 - 1)^{-1} \sum_{i=1}^{n} A_i \left(\mathbf{X}_i - \bar{\mathbf{X}}_1\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}_1\right)^T$,

$\hat{\Sigma}_0 = (n_0 - 1)^{-1} \sum_{i=1}^{n} (1 - A_i) \left(\mathbf{X}_i - \bar{\mathbf{X}}_0\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}_0\right)^T$

- Again, can compare with *z*-statistic for a rough measure of balance.

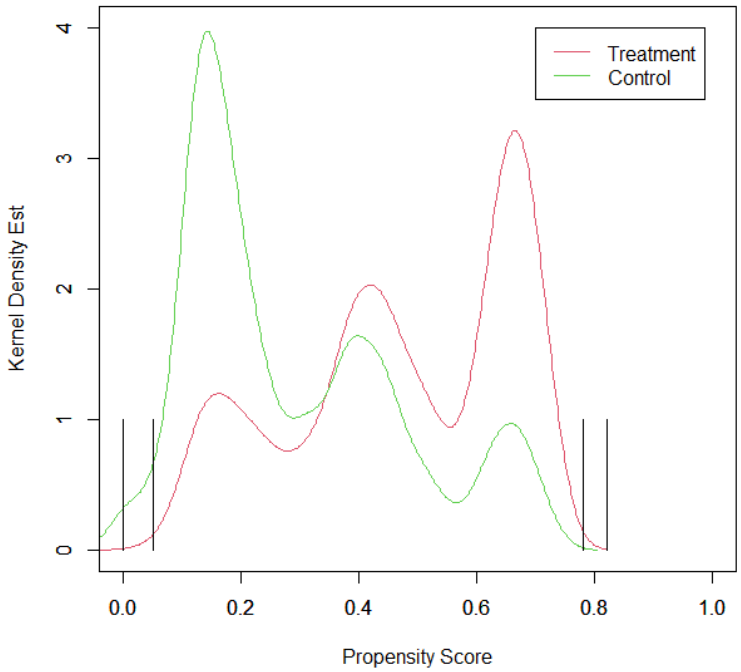
Assessing covariate balance with propensity score

- The propensity score is a way of summarizing all of the information about how X influences treatment assignment.
- Thus a logical way of assessing balance is to work directly with the propensity score.
- Work with logit transformation: $l(x) = \ln\left(\frac{e(x)}{1-e(x)}\right)$
- Consider normalized difference in mean of logit of propensity scores:

$$\hat{\Delta}_{ct}^{l} = \frac{\overline{l}_1 - \overline{l}_0}{\sqrt{\left(s_{l,1}^2 + s_{l,0}^2\right)/2}}$$

  - Again can approximate as z statistic.

Assessing Overlap



- Define each observation as having "sufficient overlap" if there is at least one observation on the opposite treatment arm where the difference in the logit of the propensity scores is less than or equal to some threshold value l^u :

$$\varsigma_i = \begin{cases} 1 \text{ if } \sum_{j, A_j \neq A_i} I\left(\left|l(x_i) - l(x_j)\right| < l^u\right) \geq 1 \\ 0 \text{ otherwise} \end{cases}$$

  - Setting $l^u = 0$ drops all observations for which $\hat{P}(A_i = 1 \mid \mathbf{X}_i) < \min_{i:A_i=1}\left(\hat{P}(A_i = 1 \mid \mathbf{X}_i)\right)$ for $i : A_i = 0$ or $\hat{P}(A_i = 1 \mid \mathbf{X}_i) > \max_{i:A_i=0}\left(\hat{P}(A_i = 1 \mid \mathbf{X}_i)\right)$ for $i : A_i = 1$.
  - Some literature suggests $l^u = 0.1$.

Trimming to improve overlap

- One option to deal with is to restrict analysis to the subset of the sample that has sufficient overlap. So, we if are using propensity weights to adjust for covariate imbalance:

$$\hat{\tau} = \frac{\sum_{i:A_i=1} \varsigma_i w_{1i} Y_i}{\sum_{i:A_i=1} \varsigma_i w_{1i}} - \frac{\sum_{i:A_i=0} \varsigma_i w_{0i} Y_i}{\sum_{i:A_i=0} \varsigma_i w_{0i}}$$

  where $w_{1i} = e(x_i)^{-1}$, $w_{0i} = (1 - e(x_i))^{-1}$.

- Basic concept is that only observations whose covariates allowed them to be assigned to treatment or control can provide information about the effect of treatment without extrapolation.
  - Causal effect estimation is restricted to the subset of the population that lies within these propensity “calipers”.

Car Seat Example Revisited: Balance

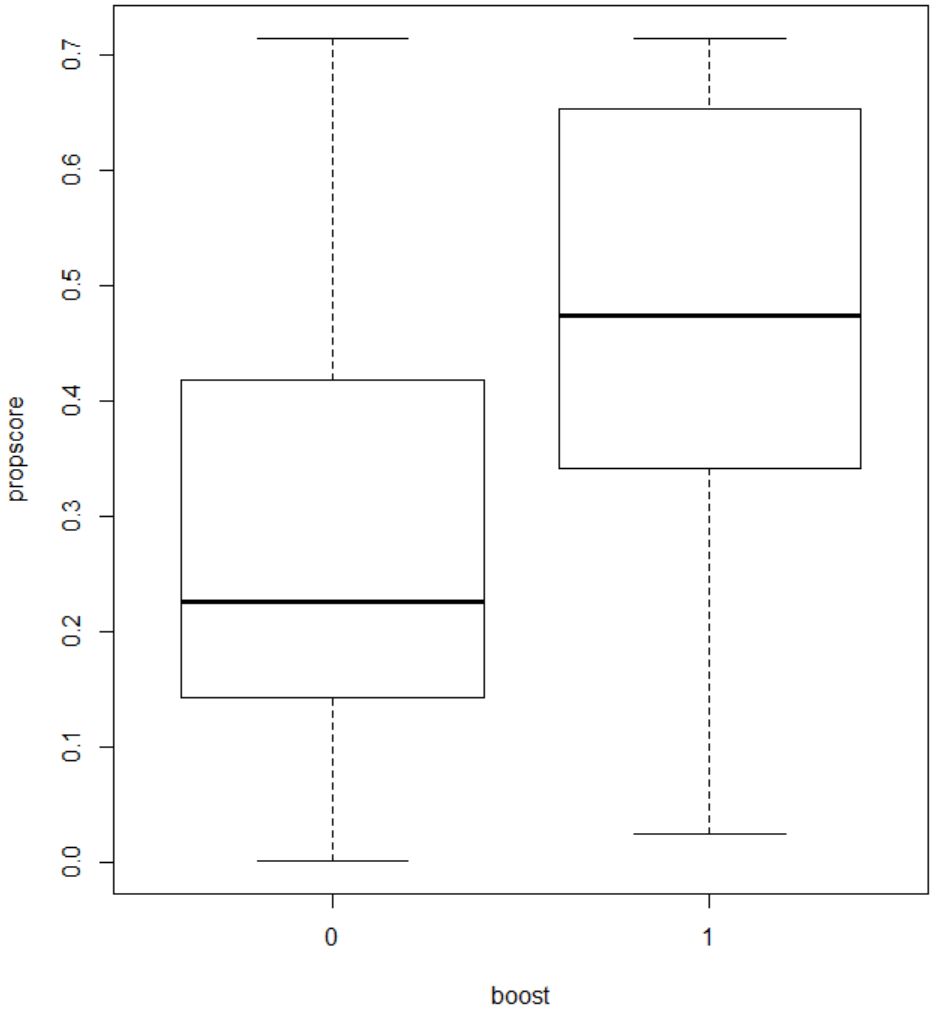
First we consider the normalized difference

	Seat Belt	Booster Seat	p-value	$\hat{\Delta}_{ct}$
N	5158	2842	---	
Mean age (years)	**5.81**	**4.88**	**<.0001**	-0.90
% passenger car	**50.3**	**44.9**	**<.0001**	-0.11
% large van	2.2	1.8	.27	-0.03
% pickup truck	**6.5**	**5.1**	**.0077**	-0.06
% minivan	**18.8**	**23.3**	**<.0001**	0.11
% SUV	**22.1**	**24.9**	**.0054**	0.06
% driver<25 years old	**10.7**	**8.8**	**.0053**	0.06
% towaway	**60.9**	**53.4**	**<.0001**	-0.15
% intrusion	**20.6**	**17.7**	**.0016**	-0.07
% airbag exposure	**4.8**	**0.4**	**<.0001**	-0.28

- Large sample size may overemphasize the problem of imbalance
- Greatest concern is with age and airbag exposure.
- Overall measure of balance $\hat{\Delta}_{ct}^{m} = 0.97$, evidence of poor balance.

R code:

```
tX<-cbind(age_yrs,lvan,truck,mvan,suv,youngd,towaway,intru,airbag)

tX1<-tX[(boost==1),]
tX0<-tX[(boost==0),]

meandiff<-apply(tX1,2,mean)-apply(tX0,2,mean)
var<-(cov(tX1)+cov(tX0))/2
sqrt(t(meandiff)%*%solve(var)%*%meandiff)
```



- Boxplots of predicted probability of being restrained in a car seat.
- Overlap between the booster and seat belt users except at very lowest predicted probability levels.

Now look at the range of $e(\mathbf{x}_i)$ among those in and not in boosters:

Not in boosters: (0.0013,0.7144)

In boosters: (0.0248,0.7144)

Overlapping support: (0.0248,0.7144)

- All children in booster seats had a non-booster seat child who probability of being in a booster seat was similar
- There were 94 non-booster seat children (0.2%) for whom there were no booster seated children with an equivalent low probability of being in a booster seat.

Who are these children?

- Older: mean age 6.6 year
- More likely to be in passenger cars (63%)
- More likely to be young drivers (24%)
- Much more likely to be in towaway crashes (97%)
- All in airbag deployment (100%)

- Restricting the analysis based on overlapping propensity scores implies that the effect of booster seats cannot be assessed for older children in more severe crashes.

- Among those that we can assess, $\hat{\tau} = -0.016$ with 95% CI of (-.035,.003).

```
nw<-w*(propscore>=.0248)

mydesign<-svydesign(ids=c(1:N),
variables=data.frame(inj,boost),weights=nw)
mymodel<-svyglm(inj~boost,design=mydesign)
summary(mymodel)
```

- All in all, reasonably good overlap for booster seat analysis.
- Differences dropping “extreme” cases modest (although statistical significance changes).

Causal Estimation using Matched Pairs

- Conceptually the idea is to find two observations with identical covariates but different treatments.
  - Stratified propensity score analysis is a step in this direction.
- Averaging this over many observations gives a causal effect if there are no unmeasured confounders.
  - Again, if treatment assignment is not randomized, there may be unmeasured confounders: unobserved factors that affect both treatment assignment and outcome.
  - But is an attempt to recreate randomized trial conditions as best as can be done in the absence of randomization.
- Easier to explain to non-statisticians than regression modeling or propensity scores.
  - Fits with intuition about how to conduct experiments.

- Matching does not rely on a model for propensity of assignment.
  - But in anything other than very low dimensional setting must decide on an algorithm.
- Can be inefficient if there are relatively few cases that match
  - Could view this as automatically “trimming.”
  - Can expand definition of match at risk of introducing confounding.

One-to-one matching (matching without replacement)

- To start, we will focus on matching n_1 treatment cases with $n_0 \geq n_1$ controls.
- Let $m_i^c \in \{1,...,n_0\}$ be the index of the control case such that $\mathbf{x}_i = \mathbf{x}_{m_i^c}$ for $i = 1,...,n_1$. This estimates the counterfactual under control: $\hat{Y}_i(0) = Y_{m_i^c}, i = 1,...,n_1$.
  - Only practical in settings with low-dimensional categorical covariates.
  - Even if dimension is low enough to allow exact matching, may require a much larger number of controls than cases since each time a control case is matched it is removed from the set available for future matches.

- ACE (among the treated) is given by

$$\hat{\tau}_t^{match} = \frac{1}{n_1} \sum_{i:A_i=1} \left(Y_i - Y_{m_i^c} \right)$$

with variance estimator

$$v\left(\hat{\tau}_t^{match} \right) = \frac{1}{n_1(n_1 - 1)} \sum_{i:A_i=1} \left(Y_i - Y_{m_i^c} - \hat{\tau}^{match} \right)^2$$

- P-values computed using z-statistic and CIs constructed using normal approximation.
- Derivation of the variance assumes that $Y_i(0) = Y_{m_i^c}$ and thus that the matched data correspond to a pairwise randomized experiment.

- In most practical settings, there will not be enough, if any, exact matches among covariates.
  - We are trying to ensure $A_i \perp (Y_i(1), Y_i(0)) | \mathbf{X}_i$, which is more plausible as the dimension of ***X*** increases.
- So we take a step back and define matching based on minimizing a generic distance function $\|\bullet\|$:

$$m_i^c = \min_j \left\| \mathbf{X}_i - \mathbf{X}_j \right\|, \; j = 1, ..., n_0$$

- In practice we have the problem that a single control might be optimal for multiple treated observations: that is, $m_i^c = m_{i'}^c$ for $i \neq i'$.
- Deal with this through optimal allocation:

$$\arg\min_{\left\{m_1^c, ... m_{N_1}^C\right\}} \sum_{i=1}^{n_1} \left\| \mathbf{X}_i - \mathbf{X}_{m_i^c} \right\|$$

  - Impossible to solve directly until sample is extremely small ($n_1 n_0$ possible combinations to consider).

- Use "greedy" or "nearest available matching" algorithm:

$$m_1^c = \min_j \left\| \mathbf{X}_i - \mathbf{X}_j \right\|$$

$$m_2^c = \min_{j: j \neq m_1^c} \left\| \mathbf{X}_i - \mathbf{X}_j \right\|$$

$$m_3^c = \min_{j: j \neq m_1^c, m_2^c} \left\| \mathbf{X}_i - \mathbf{X}_j \right\|$$

$$\vdots$$

$$m_{n_1}^c = \min_{j: j \neq m_1^c, m_2^c, \ldots, m_{n_1-1}^c} \left\| \mathbf{X}_i - \mathbf{X}_j \right\|$$

- Because potential controls are used up in order of the treated observations, the ordering of the treated cases plays a role.
  - Suggest ordering by propensity score (largest to smallest), since high propensities to treat will be rarest in the control group, so need to "save" those for matching treated cases.

- Statistical inference conducted using same methods as for exact matching.

- One issue is if there are multiple controls that match (are equidistant from) a treated observation (can happen in exact matching as well):
  - Pick one at random for the match
    - Picking one at random leaves more controls for future matching.
  - Use the mean of all (matching) observations
    - Using mean reduces variance, conservative inference

- How do we determine distance?
- Two common approaches are Mahalanobis distance and Euclidian distance

Mahalanobis:

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

$$V_M = \frac{1}{2}\left(\frac{1}{n_0} \sum_{i:A_i=0} \left(\mathbf{X}_i - \bar{\mathbf{X}}_0\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}_0\right)^T + \frac{1}{n_1} \sum_{i:A_i=1} \left(\mathbf{X}_i - \bar{\mathbf{X}}_1\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}_1\right)^T \right)$$

Euclidian:

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

$$V_E = diag\left(V_M\right)$$

Another option is to match using propensity scores (usually on the logit scale):

$$d_p(\mathbf{x}_i, \mathbf{x}_j) = \left(l(\mathbf{x}_i) - l(\mathbf{x}_j)\right)^2 = \left(\log\left(\frac{e(\mathbf{x}_i)}{1-e(\mathbf{x}_i)}\right) - \log\left(\frac{e(\mathbf{x}_j)}{1-e(\mathbf{x}_j)}\right)\right)^2$$

Matching with Replacement

- Rather than just using a single control to match to a single treatment, if there are number of controls available with similar distance metrics, these can all be used in the analysis.
- Let M_i^C represent the set of controls that are matched to a given treatment observation (each control can be matched to more than one observation). This could be a fixed number M or vary by treated subject M_i (notation assumes the latter).
  - $M_i \equiv M$ controls with smallest values of $d(\mathbf{x}_i, \mathbf{x}_j)$.
  - All M_i observations within a certain distance measure c: $M_i^c = \left\{ j : d(\mathbf{x}_i, \mathbf{x}_j) \leq c \right\}$.
- Estimate the counterfactual for the control as

$$\hat{Y}_i(0) = \frac{1}{M_i} \sum_{j \in M_i^c} Y_j, i = 1, \ldots, n_1$$

Matched causal effect estimates for the whole population

- Recall that, in this discussion of matching estimators, we have been obtaining causal effect estimates among the treated ("treatment effect among the treated")
  - In non-randomized studies the treated by differ systematically from the non-treated, including with respect to the treatment effect itself.
- We can play this same game in the other direction: obtaining estimates of the treatment effect among the controls by matching to treated cases.
- Thus in the example above we have

$$\hat{\tau}_c^{match,M} = \frac{1}{n_0} \sum_{j:A_i=0} \left(\frac{1}{M_j} \sum_{i \in M_j^c} Y_i - Y_j \right)$$

$$v\left(\hat{\tau}_c^{match,M} \right) \approx \frac{1}{n_0} \left(s_0^2 + \frac{s_1^2}{\bar{M}} \right), \quad \bar{M} = \frac{1}{n_0} \sum_{i=1}^{n_0} M_j$$

The can be combined to give an overall ACE:

$$\hat{\tau}^{match,M} = \left(\frac{n_0}{n_0 + n_1} \right) \hat{\tau}_c^{match,M} + \left(\frac{n_1}{n_0 + n_1} \right) \hat{\tau}_t^{match,M}$$

Variance can be estimated as

$$v\left(\hat{\tau}^{match,M} \right) = \left(\frac{n_0}{n_0 + n_1} \right)^2 v\left(\hat{\tau}_c^{match,M} \right) + \left(\frac{n_1}{n_0 + n_1} \right)^2 v\left(\hat{\tau}_t^{match,M} \right)$$

but this ignores the potential substantial correlation between $\hat{\tau}_c^{match,M}$ and $\hat{\tau}_t^{match,M}$.

Maybe better to use bootstrapping.

1. Within each set of treatment and controls, resample with replacement.
2. For the bth resampling, compute $\hat{\tau}_c^{match,M(b)}$, $\hat{\tau}_t^{match,M(b)}$, and $\hat{\tau}^{match,M(b)} = \left(\frac{n_0}{n_0+n_1}\right)\hat{\tau}_c^{match,M(b)} + \left(\frac{n_1}{n_0+n_1}\right)\hat{\tau}_t^{match,M(b)}$. (Note that resampling within each treatment arm stratum conditions on the sample size within each stratum; does not change across replications.)
3. Repeat 1. and 2. $b = 1,...,B$ times.

The variance estimator is given by the empirical variance of the bootstrap estimates:

$$v\left(\hat{\tau}^{match,M}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\tau}^{match,M(b)} - \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}^{match,M(b)}\right)^2$$

- Could also use empirical percentiles (e.g., 2.5 and 97.5).

Example: return to booster seat for matching estimators

- $N_0 = 5158,\ N_1 = 2842$: we have enough "donors" from the control group to do a 1-1 match.
- Match using Malhalanobis distance and greedy algorithm.
- $\hat{\tau}_t = -0.010$, (-0.026,0.006)
- Is this exactly comparable with the previous estimator of booster seat effectiveness we obtained using the propensity score strata/weights? How is it different?

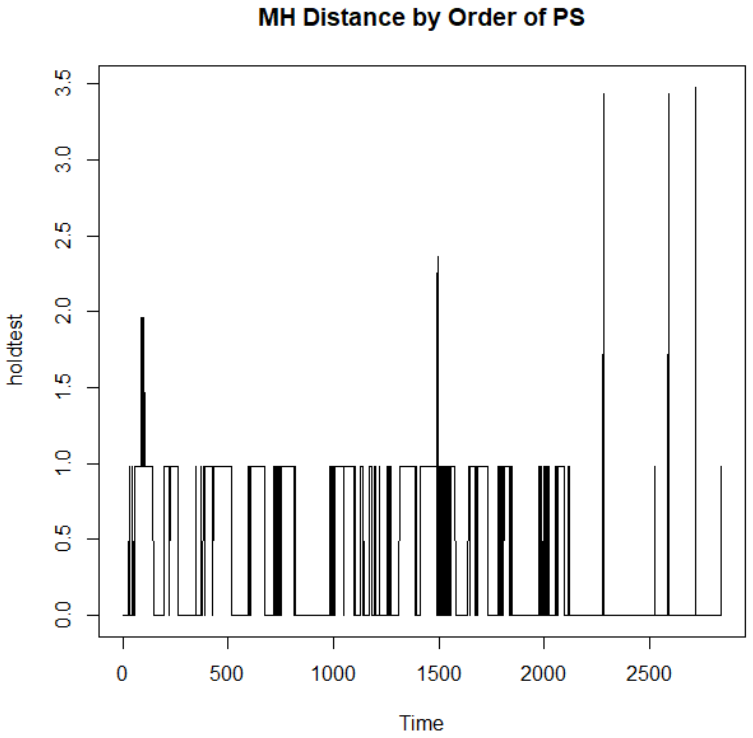
A bit of "hand coding" here as I couldn't find full R packages to match the notes examples:

```
###Function to generate Malhalanobis distance###
malhal<-function(x0,x1){
return(
sqrt(
(x0-x1)%*%Siginv%*%(x0-x1)
)
)
}

###Generate covariate matrix###
tX<-cbind(age_yrs,lvan,truck,mvan,suv,youngd,towaway,intru,airbag)
```

```
###Generate propensity scores and order data###
propmod<-
glm(boost~age5+age6+age7+truck+mvan+suv+youngd+towaway+airbag+age5*truck+age5*mvan+age5*suv+age
6*truck+age6*mvan+age6*suv+age7*truck+age7*mvan+age7*suv+age5*airbag+age6*airbag,
family=("binomial"))

propscore<-propmod$fitted.values

X<-tX[order(propscore,decreasing=TRUE),]
Y<-inj[order(propscore,decreasing=TRUE)]
BOOSTER<-boost[order(propscore,decreasing=TRUE)]

X1<-X[(BOOSTER==1),]
X0<-X[(BOOSTER==0),]
Y1<-Y[(BOOSTER==1)]
Y0<-Y[(BOOSTER==0)]

###Estimate means and covariance for MH distance computation###
mu1<-apply(X1,2,mean)
mu0<-apply(X0,2,mean)
Sig1<-cov(X1)
Sig0<-cov(X0)
Sig<-.5*Sig1+.5*Sig0
Siginv<-solve(Sig)
```

```
###Compute MH distance and determine match###
N1<-sum(boost)
NO<-N-N1

holdj<-rep(0,N1)
holdtest<-rep(0,N1)
XO<-cbind(c(1:NO),XO) ###Attach IDs for matching###
nXO<-XO

for(i in 1:N1){
test<-rep(0,(NO-(i-1))) ##Holds MH distance for ith case#
for(j in 1:(NO-(i-1))){
test[j]<-malhal(nXO[j,2:length(nXO[1,])],X1[i,])
}
holdtest[i]<-min(test) #Save min MH distance#
holdj[i]<-nXO[order(test)[1],1] #ID of control match#
j<-order(test)[1]
nXO<-nXO[-j,] #Drop matched observation from controls#
print(i)
}
```

How well did this matching work?



Variable	Difference Before Matching	Difference After Matching
Age	-0.92	-0.35
% Towaway	-7.54	-0.14
% Airbag	-4.44	0

To obtain the full ACE, we use multiple matching with a caliper match of MH distance less than 1 and combine $\hat{\tau}_t = -0.014$ and $\hat{\tau}_c = -0.006$:

$$\hat{\tau} = \left(\frac{5158}{8000}\right)\hat{\tau}_c + \left(\frac{2842}{8000}\right)\hat{\tau}_t = -.009$$

$$v(\hat{\tau}) = \left(\frac{5158}{8000}\right)^2 v(\hat{\tau}_c) + \left(\frac{2842}{8000}\right)^2 v(\hat{\tau}_t) = 1.464 \times 10^{-5}$$

95% CI=(-0.017,-0.002)

(3 treatment and 119 control cases could not be matched)

```
###Compute tau_t###
diff1<-rep(0,N1) ###Hold Y(1)-mean of matched Y(0)###
M1<-rep(0,N1)
for(i in 1:N1){
test<-rep(0,N0)
for(j in 1:N0){
test[j]<-malhal(X0[j,2:length(X0[1,])],X1[i,])
}
M1[i]<-sum(test<=1) ###Number of matched controls###
###Difference of outcome under treatment and###
###mean of matched control###
diff1[i]<-Y1[i]-sum(Y0[test<=1])/sum(test<=1)
print(i)
}

tau1<-mean(diff1[M1>0]) ###Drop non-matched cases###
s12<-var(Y1)
s02<-var(Y0)
vtau1<-(s12+s02/mean(M1[M1>0]))/sum(M1>0) ###variance###
```

```
###Compute tau_c same as tau_t###
diff0<-rep(0,N0)
M0<-rep(0,N0)
for(i in 1:N0){
test<-rep(0,N1)
for(j in 1:N1){
test[j]<-malhal(X0[i,2:length(X0[1,])],X1[j,])
}
M0[i]<-sum(test<=1)
diff0[i]<-sum(Y1[test<=1])/sum(test<=1)-Y0[i]
print(i)
}

tau0<-mean(diff0[M0>0])
s12<-var(Y1)
s02<-var(Y0)
vtau0<-(s02+s12/mean(M0[M0>0]))/sum(M0>0)

###Combine tau_t and tau_c###
tau<-(N0/(N0+N1))*tau0+(N1/(N0+N1))*tau1
vtau<-((N0/(N0+N1))^2)*vtau0+((N1/(N0+N1))^2)*vtau1
c((tau-1.96*sqrt(vtau)),(tau+1.96*sqrt(vtau)))
```

Variance using boostrapping:
$v(\hat{\tau}) = 8.13 \times 10^{-5}$

95% CI=(-0.027,0.008)

Boostrapping accounts for the correlation in the matching leading to wider intervals.

```
BOOT<-200
bootau<-rep(0,BOOT)
for(boot in 1:BOOT){

resamp1<-sample(c(1:N1),replace=TRUE)
resamp0<-sample(c(1:N0),replace=TRUE)

tX0<-X0[resamp0,]
tX1<-X1[resamp1,]
tY0<-Y0[resamp0]
tY1<-Y1[resamp1]

diff1<-rep(0,N1)
M1<-rep(0,N1)
for(i in 1:N1){
test<-rep(0,N0)
for(j in 1:N0){
test[j]<-malhal(tX0[j,2:length(tX0[1,])],tX1[i,])
}
M1[i]<-sum(test<=1)
diff1[i]<-tY1[i]-sum(tY0[test<=1])/sum(test<=1)
#print(i)
}

tau1<-mean(diff1[M1>0])
```

```
diff0<-rep(0,N0)
M0<-rep(0,N0)
for(i in 1:N0){
test<-rep(0,N1)
for(j in 1:N1){
test[j]<-malhal(tX0[i,2:length(tX0[1,])],tX1[j,])
}
M0[i]<-sum(test<=1)
diff0[i]<-sum(tY1[test<=1])/sum(test<=1)-tY0[i]
#print(i)
}

tau0<-mean(diff0[M0>0])

bootau[boot]<-(N0/(N0+N1))*tau0+(N1/(N0+N1))*tau1

print(boot)
}
var(bootau)
```

Now consider matching on propensity scores.

First, consider the quality of the 1-1 match of treated to controls:

Variable	Difference Before Matching	Difference After Matching Mahalanobis	Difference After Matching Propensity Score
Age	-0.92	-0.35	0.09
% Towaway	-7.54	-0.14	5.70
% Airbag	-4.44	0	0.39

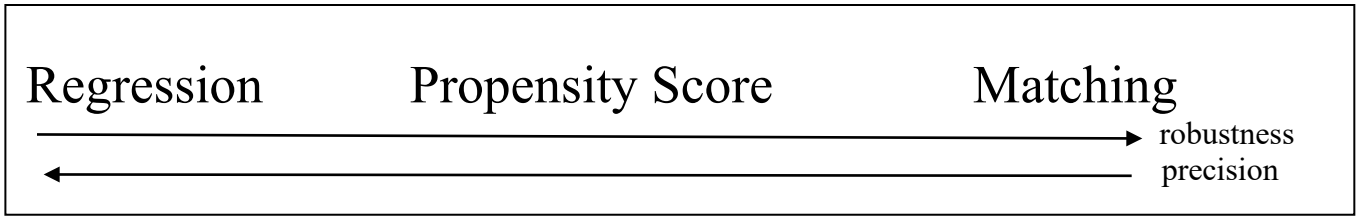
Using multiple matching with a caliper match of d_p distance less than 0.05 yields $\hat{\tau}_t = -0.026$ and $\hat{\tau}_c = -0.010$:

$$\hat{\tau} = \left(\frac{5158}{8000}\right)\hat{\tau}_c + \left(\frac{2842}{8000}\right)\hat{\tau}_t = -0.016$$

(0 treatment and 95 control cases could not be matched)
95% CI by bootstrap: (-0.025,0.009)

Final Thoughts

- Randomized controlled trials are the "gold standard" for causal inference, as they control for both observed and unobserved confounders.
- Regression, propensity scores, and matching are all attempts to approximate what would have happened had we been able to actually run a randomized controlled trial with the treatment/exposure assignment.



- Regression assumes a fully parametric form for the outcome model and confounders.
- Propensity score in model free for the outcome but assumes a model for the treatment assignment.
- Matching is fully model free (but does rely on the availability and choice of covariates to match).

All approaches assume key assumptions of SUTVA, positivity, and conditional unconfoundedness (all confounders are observed)

- Positivity can be assessed from data
- SUTVA not assessed from data but can usually be deduced from study design
- Conditional unconfoundedness cannot be assessed from data or (usually) from study design

Topics I have not addressed:

- Instrumental variables
  - If there are measures that impact treatment/exposure but are independent of potential outcomes, two-stage regression (using predicted value of the treatment/exposure as a function of the IV) can estimate the unconfounded effect of the treatment.
    - Changes in laws, natural disasters, lotteries, etc.
- Sensitivity analyses for unconfoundedness
  - Assume existence of unobserved confounder *U* and set correlations with treatment/exposure and outcome to assess impact of unobserved confounding.

- Doubly-robust estimators
  - Assume both a regression model and a propensity score model: as long as one is correct, the resulting estimator will be consistent for the ACE.
    - Borrows from model-assisted estimators in surveys: if the mean model is correct, the bias adjustment from the propensity-weighted residuals will have mean 0, but if the mean model is misspecified, the bias adjustment from the propensity-weighted residuals will be in the equal and opposite direction.
- Mediation
  - Consider causal pathways where the treatment effect is decomposed into a "direct effect" that is independent of a mediator and an "indirect effect" that requires the treatment to change the mediator and the mediator to change the outcome.
    - Diabetes as a mediator of obesity on survivals: obesity can cause diabetes, which impacts survival, but can also impact through other paths (CVD, etc.)
  - Multiple mediators, multiple pathways

I have provided code that allows you to hopefully better understand the underlying statistical issues.

But you might explore the “propensity” package in R for using propensity scores and the “matching” package in R for using matching methods.

Thanks for your attention!

References:

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95, 481-488.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945-960.

Imbens G. and Rubin D.R. (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press. Chapters 1, 2, 3, 6, 8, 12, 13, 14, 15, 16, 18

Little, R. J., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, 21, 121-145.

Lunceford, J.K., Davidian, M. (2004). Stratification and weighting via the propensity score in the estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23, 2937-2960.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516-524.

Rubin, D. (1990). [On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.] Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies. *Statistical Science, 5*, 472-480.