

Lab 7

Barum Park

Department of Sociology
New York University

Mar. 8, 2018

Before we start ...

Any questions regarding last class?

!!! WARNING !!!

**!!! PLEASE CONSULT YOUR TEXTBOOKS RATHER THAN
USING THESE SLIDES TO STUDY !!!**

**THE TEXTBOOKS THAT YOU WERE ASSIGNED WENT
THROUGH MANY REVISIONS. SO YOU CAN TRUST THEIR
CONTENT**

These slides, on the other hand, were created by a poor GRADUATE STUDENT from
the top of his head !!

Causal Inference, Basics

Causal inference with observational studies

▶ Only the basics in probability:

1. Morgan, S.L. and Winship, C., 2014. *Counterfactuals and Causal Inference*. Cambridge University Press.

▶ A little bit more math:

1. Gelman, A. and Hill, J., 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge university press. Chapters 9 and 10.
2. Angrist, J.D. and Pischke, J.S., 2008. *Mostly Harmless Econometrics: An empiricist's Companion*. Princeton university press.
3. Rosnbaum, p. R., 2010. *Design of Observational Studies*. Springer, New York, NY.
4. Rosenbaum, P.R., 2002. *Observational studies*. Springer, New York, NY.
5. Athey, S. and Imbens, G. W. 2017. "Econometrics of Randomized Experiements" Ch. 3 in *Handbook of Economic Field Experiments*. Elsevier

Potential outcomes

- ▶ Suppose we have a sample of size n , a treatment variable D , and an outcome of interest y .

- ▶ Let

$$y_i(D_i = 0) = y_i^0 \text{ and } y_i(D_i = 1) = y_i^1$$

be the **potential outcomes**, where

1. y_i^0 is the outcome if individual i **does not receive the treatment**, and
 2. y_i^1 is the outcome if individual i **does receive the treatment**.
- ▶ The (additive) treatment effect for individual i can be thus represented as

$$\tau_i = y_i^1 - y_i^0.$$

The Fundamental Problem of Causal Inference

- ▶ The problem is that y_i^1 and y_i^0 can **never be observed together**: An individual is either treated or not (never both!), so
 1. If i receives the treatment, we observe y_i^1
 2. if she does not, we observe y_i^0
- ▶ It follows that $\tau_i = y_i^1 - y_i^0$ cannot be calculated

(Hypothetical) complete data:

Unit, i	Pre-treatment inputs			Treatment indicator T_i	Potential outcomes		Treatment effect $y_i^1 - y_i^0$
	X_i				y_i^0	y_i^1	
1	2	1	50	0	69	75	6
2	3	1	98	0	111	108	-3
3	2	2	80	1	92	102	10
4	3	1	98	1	112	111	-1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	4	1	104	1	111	114	3

Observed data:

Unit, i	Pre-treatment inputs			Treatment indicator T_i	Potential outcomes		Treatment effect $y_i^1 - y_i^0$
	X_i				y_i^0	y_i^1	
1	2	1	50	0	69	?	?
2	3	1	98	0	111	?	?
3	2	2	80	1	?	102	?
4	3	1	98	1	?	111	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	4	1	104	1	?	114	?

Averages?

- ▶ What if we think about averages? That is

$$\begin{aligned}\tau &= \frac{1}{n} \sum_{i=1}^n (y_i^1 - y_i^0) = \text{Avg}[y_i^1 - y_i^0] \\ &= \text{Avg}[y_i^1] - \text{Avg}[y_i^0] \\ &= \bar{y}^1 - \bar{y}^0\end{aligned}$$

which is the definition of the “average treatment effect.” Can we calculate this quantity?

- ▶ **No.** How can we calculate a function (average) of something that is not observed?

What DO we observe?

- ▶ We observe

$$y_i^1 | D_i = 1 \quad \text{OR(!)} \quad y_i^0 | D_i = 0$$

or, more compactly,

$$y_i = D_i y_i^1 + (1 - D_i) y_i^0.$$

- ▶ Further, we observe

$$\hat{\tau}_{NAIVE} = \text{Avg}(y_i | D_i = 1) - \text{Avg}(y_i | D_i = 0)$$

What assumptions do we need?

- ▶ The crucial question is **what do we have to assume** such that:

$$\tau = E[\hat{\tau}_{NAIVE}]$$

since, **IF(!)** the equality is satisfied, we can **estimate** τ by $\hat{\tau}_{NAIVE}$.

- ▶ The assumption that have to be met are
 1. $D \perp (y^1, y^0)$ (Ignorability of the treatment assignment)
 2. SUTVA (stable unit treatment value assumption)

Ignorability

- ▶ The statement

$$D \perp y^1, y^0$$

means that the treatment assignment (D) is independent of the potential outcomes (y^1, y^0)

- ▶ In plain English: the probability of being assigned to the treatment (or control) group does not depend on what individuals would do if they were to receive (or not receive) the treatment

► Why is this important (in intuitive terms)?

To say that the treatment assignment is independent of the potential outcomes means that we are sampling **randomly** individuals (irrespective of their potential outcomes) into the treatment and the control group. Hence, the distribution of potential outcomes, y^1 , of individuals in the treatment group (which we observe) will be “representative” of the distribution of the potential outcomes y^1 of individuals in the control group (which we do not observe). It follows that average response of the treatment group will be a “typical” response of the individuals in **both** the treatment and the control group. The same goes for the control group and the distribution of y^0 . To say that the treatment assignment **does** depend on the potential outcomes, on the other hand, means that the distribution of y^1 differs between the treatment and the control group. Thus, the distribution of y^1 in the treatment group (which we observe) is no longer representative of the distribution of y^1 in the control group (which we do not observe). So there is no way to make valid inference regarding the average response of the control group under the counterfactual scenario in which they had received the treatment

- ▶ Recall we want to estimate τ from $\hat{\tau}_{NAIVE}$. This means we want to estimate

$$\left\{ \text{Avg}(y_i^1) - \text{Avg}(y_i^0) \right\} \text{ with } \left\{ \text{Avg}(y_i | D_i = 1) - \text{Avg}(y_i | D_i = 0) \right\}$$

- ▶ Thus, we would like to estimate $\text{Avg}(y_i^1)$ with $\text{Avg}(y_i | D_i = 1)$ (and of course $\text{Avg}(y_i^0)$ with $\text{Avg}(y_i | D_i = 0)$).
- ▶ The first is an average over **all** individuals, while the second is an average over only those **who have actually received the treatment**, so they are different (!)
- ▶ To proceed, let us decompose $\text{Avg}(y_i^1)$ into

$$\text{Avg}(y_i^1) = \underbrace{\pi \text{Avg}(y_i^1 | D_i = 1)}_{\text{observable}} + (1 - \pi) \underbrace{\text{Avg}(y_i^1 | D_i = 0)}_{\text{unobservable}},$$

where $\pi = \text{Prop}[D_i = 1]$.

▶ IF

$$\text{Avg}(y_i^1 | D_i = 1) = \text{Avg}(y_i^1 | D_i = 0)$$

THEN

$$\begin{aligned}\text{Avg}(y_i^1) &= \pi \text{Avg}(y_i^1 | D_i = 1) + (1 - \pi) \text{Avg}(y_i^1 | D_i = 0) \\ &= \pi \text{Avg}(y_i^1 | D_i = 1) + (1 - \pi) \text{Avg}(y_i^1 | D_i = 1) \\ &= \text{Avg}(y_i^1 | D_i = 1)\end{aligned}$$

- ▶ That is, if the equality is satisfied, the average outcome of all individuals had they received the treatment is equal to the average outcome of those who actually received the treatment

- ▶ Of course, in most situations,

$$\text{Avg}(y_i^1 | D_i = 1) \neq \text{Avg}(y_i^1 | D_i = 0)$$

- ▶ However, if $D \perp (y^1, y^0)$, then

$$E[\text{Avg}(y_i^1 | D_i = 1)] = E[\text{Avg}(y_i^1 | D_i = 0)],$$

i.e., the average potential outcomes are, in expectation, equal.

Why (only for the curious) ?

Suppose $D \perp (y^1, y^0)$, and let m be the number of treated individuals, then

$$\begin{aligned} E[\text{Avg}(y_i^1 | D_i = 1)] &= E\left[\frac{1}{m} \sum_{i=1}^n y_i^1 D_i\right] = \frac{1}{m} \sum_{i=1}^n E[y_i^1 D_i] \\ &= \frac{1}{m} \sum_{i=1}^n E[y_i^1] E[D_i] = \frac{1}{n} \sum_{i=1}^n E[y_i^1] \\ &= \frac{1}{n-m} \sum_{i=1}^n E[y_i^1] E[1 - D_i] \\ &= E\left[\frac{1}{n-m} \sum_{i=1}^n y_i^1 (1 - D_i)\right] \\ &= E[\text{Avg}(y_i^1 | D_i = 0)]. \end{aligned}$$

It doesn't matter whether y^1, y^0 are random or fixed; if we treat them as fixed, the only random source is D , so $E[y_i^1]$ is simply y_i^1 .

- ▶ Thus, if $D \perp (y^1, y^0)$, we have

$$E[\text{Avg}(y_i^1 | D_i = 1)] = \text{Avg}(y_i^1)$$

and, by symmetry,

$$E[\text{Avg}(y_i^0 | D_i = 0)] = \text{Avg}(y_i^0)$$

- ▶ Therefore, we can estimate the unobserved (!) with observables,

$$\begin{aligned} E[\tau_{NAIVE}] &= E[\underbrace{\text{Avg}(y_i^1 | D_i = 1)}_{\text{observable}}] - E[\underbrace{\text{Avg}(y_i^0 | D_i = 1)}_{\text{observable}}] \\ &= \underbrace{\text{Avg}(y_i^1) - \text{Avg}(y_i^0)}_{\text{unobservable}} \\ &= \tau, \end{aligned}$$

i.e., we have an unbiased estimator of the average treatment effect (ATE)!

How to meet the assumption?

- ▶ RANDOMIZE THE TREATMENT ASSIGNMENT. If you randomize the treatment assignment, then, **by the design of the study**, the assignment variable D is independent of the potential outcomes (y^1, y^0) (and we know this even without observing the potential outcomes!)
- ▶ If randomization is not possible, then ... try to find and instrument or natural experiment!
- ▶ If no instrument is available, then ... try to match ...

ATE, ATT, ATC

- ▶ Consider again the decomposition

$$\text{Avg}(y_i^1) = \underbrace{\pi \text{Avg}(y_i^1 | D_i = 1)}_{\text{observable}} + (1 - \pi) \underbrace{\text{Avg}(y_i^1 | D_i = 0)}_{\text{unobservable}}$$

- ▶ Similarly, we have,

$$\text{Avg}(y_i^0) = \underbrace{\pi \text{Avg}(y_i^0 | D_i = 1)}_{\text{unobservable}} + (1 - \pi) \underbrace{\text{Avg}(y_i^0 | D_i = 0)}_{\text{observable}}$$

- ▶ The relationship between ATE, ATT, and ATC is simply

$$\begin{aligned} \text{ATE} - \tau &= \text{Avg}(y_i^1) - \text{Avg}(y_i^0) \\ &= \pi \left(\underbrace{\text{Avg}(y_i^1 | D_i = 1) - \text{Avg}(y_i^0 | D_i = 1)}_{\text{ATT}} \right) \\ &\quad + (1 - \pi) \left(\underbrace{\text{Avg}(y_i^1 | D_i = 0) - \text{Avg}(y_i^0 | D_i = 0)}_{\text{ATC}} \right) \end{aligned}$$

SUTVA

- ▶ Quite everything we have established so far falls apart if SUTVA (Stable Unit Treatment Value Assumption) is not satisfied ... (not even the notation is right ...)
- ▶ So what is SUTVA?

- ▶ Recall that we have defined the potential outcomes of individual i as (y_i^0, y_i^1) .
- ▶ This seemingly innocuous representation of potential outcomes is actually not a “representation” but a **model** of the response process
- ▶ It **assumes** that the response of i depends only on whether i herself receives the treatment or not, **regardless of what treatments are assigned to all other individuals**
- ▶ This assumption is SUTVA

Examples of violations of SUTVA

1. If the value of a college degree (D) depends on how many individuals in the population hold a college degree, SUTVA is violated
2. If randomly assigning some students within a class to participate in extracurricular activities affects the outcomes of not only those who participate (treatment group) but also those who don't (control group), SUTVA is violated

Consequences of violations of SUTVA

- ▶ We cannot write the potential outcomes as (y_i^0, y_i^1) anymore
- ▶ For example, the potential outcome of i receiving the treatment will differ whether j receives the treatment as well or not.
- ▶ In a **two**-individual scenario with interference (i.e., violation of SUTVA), individual i has **four** potential outcomes:

	$D_j = 0$	$D_j = 1$
$D_i = 0$	y_i^{00}	y_i^{01}
$D_i = 1$	y_i^{10}	y_i^{11}

- ▶ In a sample of n individuals out of which m receive the treatment, there are $\binom{n}{m}$ potential outcomes for every individual! (e.g., with $n = 20$ and $m = 10$, there are 184,756 potential outcomes for each of the $i = 1, 2, \dots, n$ individuals)

Possible remedies?

1. Isolate individuals (make sure that they don't interact)
2. Change level of analysis (if SUTVA is violated by students in classrooms, but classrooms are sampled independently from different schools, analyze classroom outcomes)
3. Incorporate all the potential outcomes explicitly into your model (e.g., Aronow & Samii. 2017; Miguel & Kremer. 2004)

Remarks

- ▶ In the literature on causal inference, there is a distinction between the Sample Average Treatment Effect (SATE) and the Population Average Treatment Effect (PATE)
- ▶ This is because when dealing with inferential statistics of causal effects, there are two sources of uncertainty:
 1. The uncertainty by not being able to observe the potential outcomes and
 2. The uncertainty by analyzing a sample and not the population
- ▶ The explanation offered in these slides are for the SATE. However, under ignorability and SUTVA, the estimator

$$\tau_{NAIVE} = \underbrace{\text{Avg}(y_i^1 | D_i = 1)}_{\text{observable}} - \underbrace{E[\text{Avg}(y_i^0 | D_i = 0)]}_{\text{observable}}$$

is still unbiased for the PATE, given that SUTVA holds (although the standard errors will be different)

Propensity Scores

Matching

- ▶ There are, however, many situation in which we cannot assign the treatment randomly. So what would we do?
- ▶ Consider the probability that an individual i receives the treatment, i.e., $\Pr[D_i = 1]$.
- ▶ This probability will often depend on covariates which are observed, unobserved, as well as the potential outcomes (y_i^0, y_i^1) (note that we are assuming SUTVA again!)
- ▶ For convenience, let us denote by x_i the observed covariates, by u_i all unobserved covariates including the potential outcomes, and let $W_i = (x_i, u_i)$ and

$$\pi_i = \Pr[D_i = 1 | W_i].$$

Note: π_i is a “representation” not a “model” of the probability of receiving the treatment. In other words, π_i is equal to the true probability the assignment, since we can always find a u_i that makes the equality true.

- ▶ Now, **assume** that we know the probability π_i for all individuals, $i = 1, 2, \dots, n$, in our sample
- ▶ Then we might find two individuals (i, j) for which $\pi_i = \pi_j$.
- ▶ Suppose further that between i and j only one is treated (and the other receives the control). What is the probability of this event?
- ▶ It is exactly $1/2$.

Why (only for the curious)?

Suppose we have two individuals i and j for which

$$\Pr[D_i = 1|W_i] = \pi_i = \pi_j = \Pr[D_j = 1|W_j],$$

i.e., given the observed and unobserved covariates, and their potential outcomes, the probability that i and j receive the treatment is exactly equal. Recall that for three events, A , B , and C , we have $\Pr(A|B, C) = \Pr(A, B|C) / \Pr(A|C)$. Lastly, notice that the condition that only one of them receives the treatment can be expressed as $D_i + D_j = 1$ as D_i and D_j are either one or zero. So, the probability that i gets the treatment but j does not is

$$\begin{aligned}\Pr[D_i = 1, D_j = 0|W_i, W_j, D_i + D_j = 1] &= \frac{\Pr[D_i = 1, D_j = 0|W_i, W_j]}{\Pr[D_i + D_j = 1|W_i, W_j]} \\ &= \frac{\pi_i(1 - \pi_j)}{\pi_i(1 - \pi_j) + \pi_j(1 - \pi_i)} = \frac{\pi_i(1 - \pi_i)}{\pi_i(1 - \pi_i) + \pi_i(1 - \pi_i)} \\ &= \frac{1}{2}\end{aligned}$$

as desired.

- ▶ In other words, if we can find for each i in the sample a different individual j so that $\pi_i = \pi_j$, then we would have a **paired randomized experiment**! Within pairs, the treatment assignment is ignorable (with each one of the pair having a probability $1/2$ of receiving the treatment)!
- ▶ There are some problems, however:
 1. We don't know the probabilities π_i
 2. Actually, we cannot even observe π_i ; we only observe D_i and x_i .
- ▶ So, again, what should we do?

Exact Matching

- ▶ As it is reasonable to think that whether $D_i = 1$ or 0 depends on the covariates x_i , we might try to find for each individual i in our sample another individual j that has exactly the same covariate profile!
- ▶ That is we try to find two individuals (i, j) , such that $x_i = x_j$, and **pray** that if $x_i = x_j$, then $u_i = u_j$ as well!
- ▶ Since, if this is true, then $\pi_i = \Pr[D_i = 1 | W_i]$ is completely determined by x_i . So, again, we are in the world of a paired randomized experiment!
- ▶ Doing this (except for the prayer) is called **exact matching**.

- ▶ There is however, again, a problem.
- ▶ First, there is no guarantee that $x_i = x_j$ implies $u_i = u_j$. This, however, cannot be solved unless we have an experiment ... (so, for the time being let's ignore this)
- ▶ The practical problem is that if we have, for example, 20 covariates (each of which is either zero or one), then there are over one million different covariate profiles!
- ▶ Thus, there will be a lot of individuals for which no match can be found even when all covariates are binary. If some of the covariates are continuous, then the number of possible profiles would be infinite ...
- ▶ So, again, what should we do?
- ▶ It is for this reason that the idea of **propensity scores** becomes important.

Propensity Scores

- ▶ The propensity score for individual i , $e(x_i)$, is simply

$$e(x_i) = \Pr[D_i = 1|x_i],$$

i.e., it is the probability of receiving the treatment given the **observed** covariates.

- ▶ Note that $e(x_i)$ is, in general, not equal to π_i . The former is defined in terms of observed variables D_i and x_i , while the latter is defined in terms of observed **and** unobserved variables.
- ▶ The propensity score will be equal to π_i **if** the treatment assignment depends **ONLY** on the observed covariates, i.e.,

$$\text{if } \pi_i = \Pr[D_i = 1|x_i, u_i] = \Pr[D_i = 1|x_i] \text{ then } \pi_i = e(x_i)$$

by definition.

Why caring about propensity scores?

- ▶ **Balancing property:** Treated and control units with the same propensity scores will have the same **distribution** of **observed** covariates x_i .
- ▶ This means the following: if you have two individuals, i and j , with the same propensity score, $e(x_i) = e(x_j)$, then their covariate-profile might differ, $x_i \neq x_j$, but within this pair, the values of the covariates (x_i, x_j) will be unrelated to the treatment assignments (D_i, D_j) . Further, if you look over **many** pairs matched in this way, the **distribution** of the covariates of the treatment and the control group will be equal.

Why (only for the curious)?

We want to show that $\Pr[x_i | D_i = 1, e(x_i)] = \Pr[x_i | D_i = 0, e(x_i)]$. Recall that $E[E[A|B]] = E[A]$ and $E[E[A|B, C]|C] = E[A|C]$. Now, as $e(x_i)$ is a function of x_i , fixing x_i will fix $e(x_i)$. Thus,

$$\Pr[D_i = 1 | x_i, e(x_i)] = \Pr[D_i = 1 | x_i] = e(x_i).$$

Using this result, we obtain

$$\begin{aligned}\Pr[D_i = 1 | e(x_i)] &= E[D_i | e(x_i)] = E[E[D_i | x_i, e(x_i)] | e(x_i)] \\ &= E[\Pr[D_i = 1 | x_i, e(x_i)] | e(x_i)] = E[e(x_i) | e(x_i)] \\ &= e(x_i).\end{aligned}$$

Hence,

$$\Pr[D_i = 1 | x_i, e(x_i)] = \Pr[D_i = 1 | e(x_i)]$$

which is the definition of D_i and x_i being conditionally independent given $e(x_i)$. Thus, we have

$$x_i \perp D_i | e(x_i) \text{ or, equivalently, } \Pr[x_i | D_i = 1, e(x_i)] = \Pr[x_i | D_i = 0, e(x_i)]$$

implying that the distribution of x_i for the group with $D = 1$ and the group with $D = 0$ is equal.

Remark on Balance

- ▶ Using the “right” model to estimate the propensity score, $e(x_i)$, will lead to estimates which induce good balancing of the covariates.
- ▶ However, randomization will balance the covariates as well
- ▶ Furthermore, by design, randomization will not only balance observed covariates x_i , but also all unobserved variables u_i and is thus a much more powerful procedure
- ▶ We use propensity scores and check balance precisely because we want to mimic the scenario of a randomized experiment.
- ▶ When using propensity scores, we have to “hope” that the u_i ’s are balanced as well; if you randomize, we “know” that they’ll be balanced. This is a big difference.

Causal inference using propensity scores

- ▶ Now, IF(!)

$$\pi_i = \Pr[D_i|x_i, u_i] = \Pr[D_i|x_i]$$

then we have

$$\pi_i = e(x_i).$$

- ▶ Recall if $\pi_i = \pi_j$ for two individuals $i \neq j$, then, given that only one of them receives the treatment, the probability of i receiving it is $1/2$. But **if** the above condition holds, then $\pi_i = e(x_i)$ and $\pi_j = e(x_j)$. Thus, matching on the propensity scores, $e(x_i)$ and $e(x_j)$, will bring us back to the world of a paired randomized experiment
- ▶ In other words, **IF** the condition holds, then

$$D \perp y^0, y^1 | e(x_i)$$

so we can make valid claims of the **causal** effect of D on the outcome y .

- ▶ But even if $\pi_i \neq e(x_i)$ it might be useful to use propensity scores
- ▶ For example, say you have one observed covariate x_i . You might run a regression of the form

$$y_i = \alpha + \tau D_i + \beta x_i + \epsilon_i$$

- ▶ But, even if x_i is the *only* covariate that needs to be controlled to render the treatment assignment ignorable, you might have the wrong functional form, i.e., it might be that you have to control for x_i^2 , x_i^4 , or $\log(x_i)$, etc.
- ▶ If you use propensity scores, on the other hand, the **whole distribution** across the treatment and the control group will be the same (balancing property), so you won't misspecify the functional form

How to match?

There are many different ways to match individuals after propensity scores are estimated. For example,

1. stratifying by propensity scores
2. one-to-one matching
3. one-to-many matching
4. many-to-many matching
5. caliper matching
6. kernel matching
7. Mahalanobis distance matching
8. etc...

Weighting by propensity scores

- ▶ Another way to use propensity scores is to use them as weights in regressions
- ▶ We weight each observation by the **inverse** of the propensity to receive the treatment: if i 's propensity score is $e(x_i)$ then the weight for this observation will be

$$w_i = \begin{cases} \frac{1}{e(x_i)}, & \text{if } D_i = 1 \\ \frac{1}{1-e(x_i)}, & \text{if } D_i = 0 \end{cases}$$

- ▶ This produces an unbiased estimate of the ATE under the assumption that
 1. the treatment assignment is ignorable conditional on the propensity scores, and
 2. SUTVA holds

Why (only for the *extremely* curious)?

Suppose we are interested in the SATE and so the potential outcomes are fixed. Further, assume that given the propensity scores $e(x_i)$, the treatment assignment is ignorable. Then, we have

$$E \left[\frac{y_i D_i}{e(x_i)} \middle| x_i \right] = \frac{1}{e(x_i)} E[y_i^1 D_i | x_i] = \frac{1}{e(x_i)} y_i^1 E[D_i | x_i] = y_i^1.$$

Thus, by the property of iterated expectations, it follows that

$$E \left[\frac{y_i D_i}{e(x_i)} \right] = y_i^1$$

Similarly,

$$E \left[\frac{y_i(1 - D_i)}{1 - e(x_i)} \right] = y_i^0.$$

Thus,

$$E \left[\sum_{i=1}^n \left(\frac{y_i D_i}{e(x_i)} - \frac{y_i(1 - D_i)}{1 - e(x_i)} \right) \right] = \sum_{i=1}^n (y_i^1 - y_i^0)$$

i.e., the total difference in the potential outcomes.

A natural estimator, called the Horvitz-Thompson (HT) estimator, for the sample average treatment effect would be therefore,

$$\hat{\tau}_{HT} = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i D_i}{e(x_i)} - \frac{y_i(1 - D_i)}{1 - e(x_i)} \right).$$

Note that this estimator is different from the usual Hajek estimator, i.e.,

$$\hat{\tau}_H = \frac{\sum_{i=1}^n w_i (y_i D_i - y_i(1 - D_i))}{\sum_{i=1}^n w_i}$$

where

$$w_i = \begin{cases} \frac{1}{e(x_i)} & \text{if } D_i = 1 \\ \frac{1}{1 - e(x_i)} & \text{if } D_i = 0 \end{cases}$$

which is used if you specify “survey weights” in a STATA or R (If you plug-in the weights, you’ll see that the only difference is the denominator which is n for the HT estimator and $\sum_i w_i$ for the Hajek estimator).

What is their difference? Consider the sum $\sum_i y_i^1/n$, which is estimated by $n^{-1} \sum_{i=1}^n \frac{D_i y_i}{e(x_i)}$ when using the HT estimator. We have shown above that this estimator is unbiased. The Hajek estimator is, on the other hand,

$$\frac{\sum_{i=1}^n w_i D_i y_i}{\sum_{i=1}^n w_i} = \sum_{i=1}^n \frac{D_i y_i}{e(x_i)} / \sum_{i=1}^n \frac{D_i}{e(x_i)}.$$

We know that the numerator is unbiased for $\sum_i y_i^1$, so let us look at the denominator for the i th term:

$$\mathbb{E} \left[\frac{D_i}{e(x_i)} \right] = \mathbb{E} \left[\mathbb{E} \left[\frac{D_i}{e(x_i)} \mid x_i \right] \right] = \mathbb{E} \left[\frac{\mathbb{E}[D_i | x_i]}{e(x_i)} \right] = 1.$$

Thus,

$$\mathbb{E} \left[\sum_{i=1}^n \frac{D_i}{e(x_i)} \right] = n.$$

The same holds for the untreated units, showing that the denominator of the Hajek estimator is unbiased for the sample size.

In short, the Hajek estimator is a ratio of two unbiased estimators. The problem is that, in general, for two random variables V and W ,

$$E\left[\frac{W}{V}\right] \neq \frac{E[W]}{E[V]}.$$

Thus, in general, the Hajek estimator will be biased. The main reason STATA uses the Hajek estimator is because its sampling variability is smaller than that of the HT estimator, and the bias in the estimator tends to be small with a reasonably large sample size.

For example, if $e(x_i)$ is extremely large or small, the HT estimator would explode: as $e(x_i) \rightarrow 0$, then $1/e(x_i) \rightarrow \infty$; if $e(x_i) \rightarrow 1$, on the other hand, then $1/(1 - e(x_i)) \rightarrow \infty$. Dividing through by the sum of the weights, therefore, gives the Hajek estimator more stability.