Lecture 4: Propensity Score, Part 1

$$y_i = \theta T_i + X_i'\beta + \varepsilon_i$$

T_i is a <u>binary</u> (0-1) treatment with <u>homogenous</u> treatment effects:

$$T_i = \begin{cases} 1 & \text{if treated} \\ 0 & \text{not treated} \end{cases}$$

↳ There's nothing in between! No being "half-way" treated. You are either treated or not. 0 or 1.

↳ $\theta_i = \theta \quad \forall \; i$ (although we'll consider heterogenous treatment effects later).

i has two potential outcomes:

y_{0i} if $T_i = 0$

y_{1i} if $T_i = 1$

$$\boxed{\theta_i = y_{1i} - y_{0i}} \Rightarrow \boxed{E(y_{1i} - y_{0i}) = E(\theta_i) = \theta}$$

$$(\bar{y}_1 - \bar{y}_0)$$

<u>Problem:</u>
For individual i, observe either y_{1i} <u>or</u> y_{0i}, but not both! Unobserved counterfactual.

What's the "gold standard" solution?

- Random assignment of T_i

$\Rightarrow \{y_{1i}, y_{0i}\} \perp\!\!\!\perp T_i$

- Control group (y_{0i}) identifies correct counterfactual as $N \to \infty$

$$\bar{y}_1 - \bar{y}_0 \xrightarrow{p} \theta \quad \text{as} \quad N \to \infty$$

- Indirect test of random assignment:

$\bar{X}_1 \approx \bar{X}_0 \; \forall \; x_{ik}$

Linear model:

$$y_i = \theta T_i + X_i'\beta + \varepsilon_i \quad \text{where we assume } E(\varepsilon_i \cdot T_i) = 0$$

Comparing mean of y_i conditional on T_i:

$$E(y_i|T_i=1) - E(y_i|T_i=0) = E(y_{1i}|T_i=1) - E(y_{0i}|T_i=0)$$

$$= E(y_{1i}|T_i=1) - E(y_{0i}|T_i=1) + E(y_{0i}|T_i=1) - E(y_{0i}|T_i=0)$$

$$= \underbrace{E(y_{1i} - y_{0i}|T_i=1)}_{= ATE} + \underbrace{\left[E(y_{0i}|T_i=1) - E(y_{0i}|T_i=0)\right]}_{=0 \text{ if } T_i \text{ randomly assigned}}$$

$= \theta$ + selection bias if T_i not randomly assigned.

Example: Birthweight and Smoking.

$$E(y_{1i} - y_{0i} \mid T_i = 1) + [E(y_{0i} \mid T_i = 1) - E(y_{0i} \mid T_i = 0)]$$

$$= -210\,g + [2950\,g - 3000\,g]$$

$$= -210\,g + [-50\,g]$$

$$= -260\,g$$

The counterfactual BW for women who smoke if they didn't smoke: 2950 g

Why lower? Because they're drinking, not going to prenatal visits, etc. Their BW is lower than for women who don't smoke (3000g).

- Random assignment conditional on observables:

- $\{y_{1i}, y_{0i}\} \perp T_i \mid X_i$

↳ T_i is <u>independent</u> of potential outcomes <u>conditional</u> on X_i.

$= E(y_{1i} - y_{0i} \mid T_i = 1, X_i) + \underbrace{\left[E(y_{0i} \mid T_i = 1, X_i) - E(y_{0i} \mid T_i = 0, X_i)\right]}$

$= \quad 0 \quad$ $= 0$ if only source of bias before was due to X_i and now we've removed this bias.

$\neq 0$ if omitted variables or misspecification of $g(X_i)$.

$g(X_i) = X_i'\beta$ is only one (linear) possibility. What if not the right one?

- Matching: Univariate Case
  - For each treatment observation, match control case with "identical" X_i.
    (Problem if X_i, T_i are collinear $\Rightarrow$ impossible to match).
  - Using the matched pairs, run a regression controlling for "pair identifier" fixed effects.

```
cd "C:\Users\lfeler1\Documents\Applied Econometrics Course\Notes\Weeks 4-6"
clear
use smoking2

**This is the regression we would run if smoking was randomly assigned**
reg dbirwt tobacco


**Check if smoking randomly assigned**
sort tobacco
ttest dmage, by(tobacco)
ttest dmeduc, by(tobacco)
ttest dmar, by(tobacco)
ttest nprevist, by(tobacco)
ttest alcohol, by(tobacco)
ttest anemia, by(tobacco)
ttest mblack, by(tobacco)
    *So no, it does not look like smoking is randomly assigned!*


**Assume smoking is randomly assigned conditional on observables**
reg dbirwt tobacco dmage dmeduc dmar dlivord nprevist dfage dfeduc anemia diabete ///
phyper alcohol drink foreignb plural deadkids mblack motherr mhispan fblack fotherr
fhispan first

estat imtest, white
    *So reject homo in favor of heteroskedasticity

reg dbirwt tobacco dmage dmeduc dmar dlivord nprevist dfage dfeduc anemia diabete ///
phyper alcohol drink foreignb plural deadkids mblack motherr mhispan fblack fotherr
fhispan first, robust

estat ovtest
    *This regresses y on x y-hat^2 y-hat^3 y-hat^4 and jointly tests that the coeffs on
y-hat^2 y-hat^3 y-hat^4 are zero.
    *We cannot reject that the model as no omitted variables.


*Let's try matching on... education.*
sort dmeduc
areg dbirwt tobacco, absorb(dmeduc) robust
```

- Propensity Score

- What is it? An index (one variable) constructed out of all the X_K's, and possibly their squares, cubics, and interactions.

- Why? Reduces multi-dimensional X_K into one dimension! Makes matching possible.

↳ Propensity Score Theorem:

If $\{y_{1i}, y_{0i}\} \perp\!\!\!\perp T_i | X_i$, then

$\{y_{1i}, y_{0i}\} \perp\!\!\!\perp T_i | p(X_i)$.

↳ where $p_i \equiv Pr(T_i = 1 | X_i) = E(T_i | X_i) \equiv p(X_i)$

Probability of treatment conditional on X_i.

- Now just control or match for single index $p(X_i)$ rather than all X_K's.

$\rightsquigarrow$ How?

(1) Estimate propensity score, $\hat{p}(x_i)$ such that it balances X_k's.

(2) Estimate $\theta \equiv$ ATE by controlling for $\hat{p}(x_i)$.

Step (1): Estimate propensity score, $\hat{p}(x_i)$.

- $Pr(T_i = 1 | X_i) = \dfrac{e^{h(x_i)}}{1 + e^{h(x_i)}}$

- $h(x_i)$ contains linear and possibly higher order terms and interactions.
  $\rightsquigarrow$ include enough terms so that Treatments and Controls with similar $\hat{p}(x_i)$ have similar X_k's.

"Algorithm" for estimating $p(x_i)$

(1) start with parsimonious logit $\rightarrow$ estimate $\hat{p}(x_i)$

(2) stratify data into 5 blocks of $\hat{p}(x_i)$

(3) test $\bar{X}_1 = \bar{X}_0$ for all K within each block, using t-test of significant differences in sample means.

(a) if X_k's "balanced" in each block, STOP.

(b) if X_k's not balanced in some blocks, divide block into 2 blocks

and reevaluate.

(c) if X_K's not balanced in <u>all</u> blocks, add interaction and/or polynomial of X_K's to logit and reevaluate.

<u>Goal</u>: Balance X_K's in Treatment and Control Groups in each block.

→ overlap in $\hat{p}(x_i)$ for $T_i = 1$ and $T_j = 0$ implies overlap in X_i's.

<u>Stopping Rule</u>: Stop when fail to reject $\bar{X}_{1K} = \bar{X}_{0K}$ for over 90% of t-tests within a block.

```
cd "C:\Users\lfeler1\Documents\Applied Econometrics Course\Notes\Weeks 4-6"
clear
use smoking2
set seed 1000
sample 20

pscore tobacco  dmage dmeduc dmar dlivord nprevist dfage dfeduc anemia diabete phyper
alcohol ///
drink foreignb plural deadkids mblack motherr mhispan fblack fotherr fhispan first, ///
logit pscore(phat1) blockid(block1) numblo(5) detail



pscore tobacco  dmage dmeduc dmar dlivord nprevist dfage dfeduc anemia diabete phyper
alcohol ///
drink foreignb plural deadkids mblack motherr mhispan fblack fotherr fhispan first, ///
logit pscore(phat2) blockid(block2) numblo(5)


gen dmeduc2=dmeduc^2
gen dmageXdmeduc=dmage*dmeduc

pscore tobacco  dmage dmeduc dmar dlivord nprevist dfage dfeduc anemia diabete phyper
alcohol ///
drink foreignb plural deadkids mblack motherr mhispan fblack fotherr fhispan first ///
dmeduc2 dmageXdmeduc, ///
logit pscore(phat3) blockid(block3) numblo(15)


gen dmeducXdfeduc=dmeduc*dfeduc

pscore tobacco  dmage dmeduc dmar dlivord nprevist dfage dfeduc anemia diabete phyper
alcohol ///
drink foreignb plural deadkids mblack motherr mhispan fblack fotherr fhispan first ///
dmeduc2 dmageXdmeduc dmeducXdfeduc, ///
logit pscore(phat4) blockid(block4) numblo(15)


gen mblackXfblack=mblack*fblack

pscore tobacco  dmage dmeduc dmar dlivord nprevist dfage dfeduc anemia diabete phyper
alcohol ///
drink foreignb plural deadkids mblack motherr mhispan fblack fotherr fhispan first ///
dmeduc2 dmageXdmeduc dmeducXdfeduc mblackXfblack, ///
logit pscore(phat5) blockid(block5) numblo(20)

****This last one just made things worse, so go back***

graph box phat4, by(tobacco)


```