Lecture 3: Selection on Observables

6 Selection Problem and Assignment Mechanism

6.1 The Selection Problem in Program Evaluation

What can we learn (identify) from the observed outcomes, Y_i ? **Comparing mean of Y_i by T_i**

$$\begin{aligned} E[Y_i|T_i=1]-E[Y_i|T_i=0] &= E[Y_{1i}|T_i=1]-E[Y_{0i}|T_i=1]+E[Y_{0i}|T_i=1]-E[Y_{0i}|T_i=0] \\ &= \underbrace{E[Y_{1i}-Y_{0i}|T_i=1]}_{\text{ATE on the treated (ATT)}}+\underbrace{E[Y_{0i}|T_i=1]-E[Y_{0i}|T_i=0]}_{\text{Selection Bias}} \end{aligned}$$

When can we identify the ATE?

6.2 Assignment Mechanism

6.2.1 Random Assignment – "Gold Standard" Solution: $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp T_i$

$$E[Y_i|T_i=1]-E[Y_i|T_i=0] = E[Y_{1i}-Y_{0i}|T_i=1]+\{E[Y_{0i}|T_i=1]-E[Y_{0i}|T_i=0]\} \tag{1}$$

$$= E[Y_{1i}-Y_{0i}] \equiv ATE \tag{2}$$

– By definition of random assignment, this identifies "Average Treatment Effect"

6.2.2 Selection on Observables: $(Y_{1i}, Y_{0i}) \not\perp\!\!\!\perp T_i$ but $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp T_i|\mathbf{X}_i$

$$\begin{aligned} E[Y_i|T_i=1,\mathbf{X}_i]-E[Y_i|T_i=0,\mathbf{X}_i] &= E[Y_{1i}-Y_{0i}|T_i=1,\mathbf{X}_i]+\{E[Y_{0i}|T_i=1,\mathbf{X}_i]-E[Y_{0i}|T_i=0,\mathbf{X}_i]\} \\ &= E[Y_{1i}-Y_{0i}|\mathbf{X}_i] \equiv ATE(\mathbf{X}_i) \end{aligned}$$

Then, $ATE \equiv E[Y_{1i}-Y_{0i}] = E_X\Big[E[Y_{1i}-Y_{0i}|\mathbf{X}_i]\Big]$

Example : Switching regression model for *potential* outcomes

$$Y_{0i} = \alpha_0 + g_0(\mathbf{X}_i) + U_{0i}, \quad \text{if } T_i = 0 \tag{3}$$

$$Y_{1i} = \alpha_1 + g_1(\mathbf{X}_i) + U_{1i}, \quad \text{if } T_i = 1 \tag{4}$$

If $g_0(\mathbf{X}_i) = \mathbf{X}_i'\beta_0$, $g_1(\mathbf{X}_i) = \mathbf{X}_i'\beta_1$, and $\beta_0 = \beta_1$, then $ATE \equiv E[Y_{1i}-Y_{0i}] = \alpha_1 - \alpha_0 \equiv \theta$

Can we identify θ using the regression model for the *observed* outcome:

$$Y_i = \alpha_0 + \mathbf{X}_i'\beta_0 + (\alpha_1-\alpha_0)T_i + \{U_{0i} + (U_{1i}-U_{0i})T_i\} \tag{5}$$

For now, assume no *unobserved* heterogeneity: $U_{1i} = U_{0i}$

$$Y_i = \alpha_0 + \theta \cdot T_i + \mathbf{X}_i'\beta_0 + U_{0i} \tag{6}$$

1. Random Assignment:

$$\begin{aligned} E[Y_{1i} - Y_{0i}] &= E[Y_i|T_i = 1] - E[Y_i|T_i = 0] \\ &= \Big\{\alpha_0 + \theta + E[\mathbf{X}_i'|T_i = 1]\beta_0 + E[U_{0i}|T_i = 1]\Big\} - \Big\{\alpha_0 + E[\mathbf{X}_i'|T_i = 0]\beta_0 + E[U_{0i}|T_i = 0]\Big\} \\ &= \theta + \Big\{E[\mathbf{X}_i'|T_i = 1] - E[\mathbf{X}_i'|T_i = 0]\Big\}\beta_0 + \Big\{E[U_{0i}|T_i = 1] - E[U_{0i}|T_i = 0]\Big\} = \theta \end{aligned}$$

   Indirect test of random assignment : $\overline{\mathbf{X}}_1 \approx \overline{\mathbf{X}}_0, \quad \forall X_k$

2. Selection on Observables:

$$\begin{aligned} E[Y_{1i} - Y_{0i}] &= E\Big[E[Y_i|\mathbf{X}_i, T_i = 1] - E[Y_i|\mathbf{X}_i, T_i = 0]\Big] \\ &= E\Big[\theta + \Big\{E[U_{0i}|\mathbf{X}_i, T_i = 1] - E[U_{0i}|\mathbf{X}_i, T_i = 0]\Big\}\Big] = \theta \end{aligned}$$

7 Selection on Observables

Selection on Observables: random assignment conditional on observables

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp T_i|\mathbf{X}_i \tag{7}$$

T_i is independent of potential outcomes <u>conditional on observable</u> characteristics $\mathbf{X}_i$:

- Only sources of bias are due to $\mathbf{X}_i$ (observables)
  - No selection once we conditioned on $\mathbf{X}_i$
- Ex. "kitchen sink" regression? – i.e., use as many $\mathbf{X}_i$ as possible?
  - Problem: Data mining, arbitrary specification ⇒ may lead to O.V.B.
- The selection on observables concerned mostly about the "incorrect functional form"

7.1 Regression Analogy

1. Linear Regression: $Y_i = \alpha + \theta \cdot T_i + \mathbf{X}_i'\beta + U_i, \quad E[U_i|T_i, \mathbf{X}_i'\beta] = 0$
   - Advantage: low dimension – just control for linear function, $\mathbf{X}_i'\beta$ using OLS
   - Disadvantage: if $\mathbf{X}_i'\beta$ is misspecified, then O.V.B.
2. Nonlinear Regression: $Y_i = \alpha + \theta \cdot T_i + g(\mathbf{X}_i) + U_i, \quad E[U_i|T_i, g(\mathbf{X}_i)] = 0$
   - Disadvantage: high dimension – $g(\mathbf{X}_i)$ may include polynomials and interactions

Approach: Multivariate Matching and Propensity Score

7.2 Multivariate Matching

- Basic Idea : If we have same (identical) individuals based on $\mathbf{X}_i$ (i.e., match the treated with the untreated individual having exactly same $\mathbf{X}_i$), then the form of $g(\mathbf{X}_i)$ does not matter

[Figure HERE]

- For each treatment observation, match control case with "identical" $\mathbf{X}_i$. At each stratum defined by $\mathbf{X}$, need treated and untreated individuals (”overlap” assumption)

$$0 < \Pr(T_i = 1|\mathbf{X}_i) < 1 \tag{8}$$

- Rosenbaum and Rubin (1983) refer to the combination of two assumptions, (7) and (8), as "*strongly ignorable*" treatment assignment

- If the strong ignorability holds, then the matching estimator identifies the ATE (and ATT)

Example of matching estimators: ATT (Heckman, Ichimura, and Todd, 1997, 1998)

$$\widehat{E}[Y_{1i} - Y_{0i}|T_i = 1] = \frac{1}{N_1} \sum_{i \in \{T_i=1\}} \left[Y_{1i} - \sum_{j \in \{T_j=0\}} W(i,j) Y_{0j} \right] \tag{9}$$

where N_i is the number of treated individuals; and $W(i,j)$ is the weight given to the jth observation in the control group, such that $\sum_{j \in \{T_i=0\}} W(i,j) = 1$ and that $0 \leq W(i,j) \leq 1$

(i) Nearest-neighbor matching

$$W(i,j) = \begin{cases} 1, & \text{if } j \in A_i \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

$$A_i = \{j | \min_j ||\mathbf{X}_i - \mathbf{X}_j||\} \tag{11}$$

where $||\cdot||$ is a distance metric (e.g., Mahalanobis metric: $||\cdot|| = (\mathbf{X}_i - \mathbf{X}_j)'\Sigma_X^{-1}(\mathbf{X}_i - \mathbf{X}_j)$, where Σ_X is the variance-covariance matrix of $\mathbf{X}$)

(ii) Caliper matching

$$A_i = \{j | ||\mathbf{X}_i - \mathbf{X}_j|| < \varepsilon\} \tag{12}$$

where ε is a pre-specified tolerance.

(iii) Kernel matching

$$W(i,j) = \frac{K\left(\frac{\mathbf{X}_j - \mathbf{X}_i}{h}\right)}{\sum_{j=1}^{N_0} K\left(\frac{\mathbf{X}_j - \mathbf{X}_i}{h}\right)} \tag{13}$$

where K is a kernel function and h is a bandwidth parameter.

Caveat

- *Curse of Dimensionality*: fitting flexible functional form with K argument ($\dim(\mathbf{X}_i) = K$) $\Rightarrow$ computational burden (N^K)
- Common support (overlap) problem: for each treated, need to match at least one control unit $\Rightarrow$ especially difficult to find the matched unit as ($\dim(\mathbf{X}_i) = K$) gets bigger

How to reduce the dimensionality and remove bias due to $\mathbf{X}_i$?

7.3 Propensity Score

The propensity score is the conditional probability of being treated given $\mathbf{X}_i$

$$\Pr(T_i = 1|X_i) \equiv p(\mathbf{X}_i) \equiv p_i \tag{14}$$

Propensity Score Theorem (Rosenbaum and Rubin, 1983)
If T_i is independent of potential outcomes conditional on $\mathbf{X}_i$, then T_i is independent of potential outcomes conditional on the propensity score, $p(\mathbf{X}_i)$:

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp T_i|\mathbf{X}_i \Longrightarrow (Y_{0i}, Y_{1i}) \perp\!\!\!\perp T_i|p(\mathbf{X}_i) \tag{15}$$

<u>Proof</u>: It is sufficient to show that $\Pr\left(T_i = 1|Y_{1i}, Y_{0i}, p(\mathbf{X}_i)\right) = \Pr\left(T_i = 1|p(\mathbf{X}_i)\right)$

Since $\Pr(T_i = 1|p(\mathbf{X}_i)) = E[T_i|p(\mathbf{X}_i)] = E\Big[E[T_i|\mathbf{X}_i]\Big|p(\mathbf{X}_i)\Big] = E[p(\mathbf{X}_i)|p(\mathbf{X}_i)] = p(\mathbf{X}_i)$,
it is sufficient to show that $\Pr\left(T_i = 1|Y_{1i}, Y_{0i}, p(\mathbf{X}_i)\right) = p(\mathbf{X}_i)$

$$\begin{aligned}
\Pr(T_i = 1|Y_{0i}, Y_{1i}, p(\mathbf{X}_i)) &= E[T_i|Y_{0i}, Y_{1i}, p(\mathbf{X}_i)] = E\Big[E[T_i|Y_{0i}, Y_{1i}, p(\mathbf{X}_i), \mathbf{X}_i]\Big|Y_{0i}, Y_{1i}, p(\mathbf{X}_i)\Big] \\
&= E\Big[E[T_i|Y_{0i}, Y_{1i}, \mathbf{X}_i]\Big|Y_{0i}, Y_{1i}, p(\mathbf{X}_i)\Big] = E\Big[E[T_i|\mathbf{X}_i]\Big|Y_{0i}, Y_{1i}, p(\mathbf{X}_i)\Big] \\
&= E\Big[p(\mathbf{X}_i)\Big|Y_{0i}, Y_{1i}, p(\mathbf{X}_i)\Big] = p(\mathbf{X}_i)
\end{aligned}$$

<u>**Use of propensity score**</u>

- Idea : Since T_i is binary, $E(T_i|\mathbf{X}_i)$ and $Var(T_i|\mathbf{X}_i)$ determined by $p(\mathbf{X}_i)$, that is, $p(\mathbf{X}_i)$ is sufficient statistics for the relationship between T_i and $\mathbf{X}_i \Rightarrow T_i \perp\!\!\!\perp \mathbf{X}_i|p(\mathbf{X}_i)$
- It reduces dimensionality by controlling just for single index $p(\mathbf{X}_i)$ that balances $\mathbf{X}_i \Rightarrow$ useful "<u>descriptive tool</u>" (dimension-reduction tool) in practice.
- Adjusts for selection bias due to $\mathbf{X}_i$ in a 2-step way
  1. Estimate $p(\mathbf{X}_i)$ (e.g., by logit model)
  2. Estimate average treatment effects controlling for $\widehat{p}(\mathbf{X}_i)$
     Ex. regression, matching, subclassfication, and weighting based on $\widehat{p}(\mathbf{X}_i)$

8 Propensity Score Methods

8.1 Estimate $\widehat{p}(\mathbf{X}_i)$ by logit

$$p(\mathbf{X}_i) \equiv \Pr(T_i = 1|\mathbf{X}_i) = \frac{e^{h(\mathbf{X}_i)}}{1 + e^{h(\mathbf{X}_i)}} \tag{16}$$

– Issue: the functional form of $p(\mathbf{X}_i)$ – i.e., functional form of $h(\mathbf{X}_i)$ in the logit

Goal: Balance of $\mathbf{X}_i$ between T and C group conditional on $p(\mathbf{X}_i)$, $\underline{\mathbf{X}_i \perp\!\!\!\perp T_i|p(\mathbf{X}_i)}$

Overlap in $\widehat{p}(\mathbf{X}_i)$ between T and C group $\Rightarrow$ Overlap in $\mathbf{X}_i$ (Balance of $\mathbf{X}_i$)

8.1.1 "Algorithm" for estimating $p(\mathbf{X}_i)$

Rosenbaum and Rubin (1983, 1984)

1) Using parsimonious logit, estimate $\widehat{p}(\mathbf{X}_i)$

2) Stratify data into quintiles of the distribution of $\widehat{p}(\mathbf{X}_i)$ – i.e., five equal-sized blocks

3) Test $\overline{\mathbf{X}}_1 = \overline{\mathbf{X}}_0$ within each block (t-test)

   i) If X_k are "balanced" in each block, then STOP
   e.g., stop when fail to reject $\overline{X}_{1k} = \overline{X}_{0k}$ for over 90% of t-tests within a block

   ii) If X_k are not balanced in certain block, then divide that block into 2 sub-blocks and re–evaluate (t-test)

   iii) If X_k are not balanced in all blocks, then generalize the specification of $\widehat{p}(\mathbf{X}_i)$ (i.e., add polynomial and/or interaction of X_k) and re–evaluate

8.1.2 Assessing overlap in $\widehat{p}(\mathbf{X}_i)$

Ex. Box–Plot (or histogram) – distribution of $\widehat{p}(\mathbf{X}_i)$ between treatment and control group

graph box command in STATA.

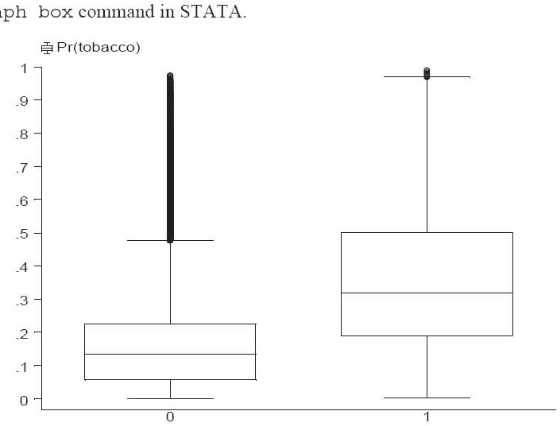


Figure 1: Box-Plot of $\widehat{p}(\mathbf{X}_i)$ for control and treatment group

- Amount of "overlap" in the plot $\approx$ similarity of $\mathbf{X}_i$ in treatment and control groups
- A lot of overlap $\Rightarrow$ very little selection on observable $\mathbf{X}_i$ (good research design)
- Little overlap $\Rightarrow$ pure selection on observable $\mathbf{X}_i$ (bad design)

  extrapolating across *non-comparable* population
- What does the Box-Plot look like if the treatment is randomly assigned?

8.2 Estimate Average Treatment Effects controlling for $\widehat{p}(\mathbf{X}_i)$

8.2.1 Graphical Analysis - Most general (informative) use of $\widehat{p}(\mathbf{X}_i)$

Nonparametric estimation of

$$E[Y_i|\widehat{p}_i, T_i = 0] \quad \text{and} \quad E[Y_i|\widehat{p}_i, T_i = 1] \tag{17}$$

- Can estimate these by bivariate nonparametric regression (e.g., kernel regression, local linear regression - lowess in STATA command)
- More transparently, we can calculate means of outcome over 100 (or more) equal-sized cells of $\widehat{p}_i$, separately for treatment and controls, and plot against $\widehat{p}_i$
- Bias can be summarized by single index $\widehat{p}_i$ (i.e., different slope) $\Rightarrow$ No dimensionality problem
- Useful when N_1 and N_0 are large

  Example: maternal smoking vs birth weight

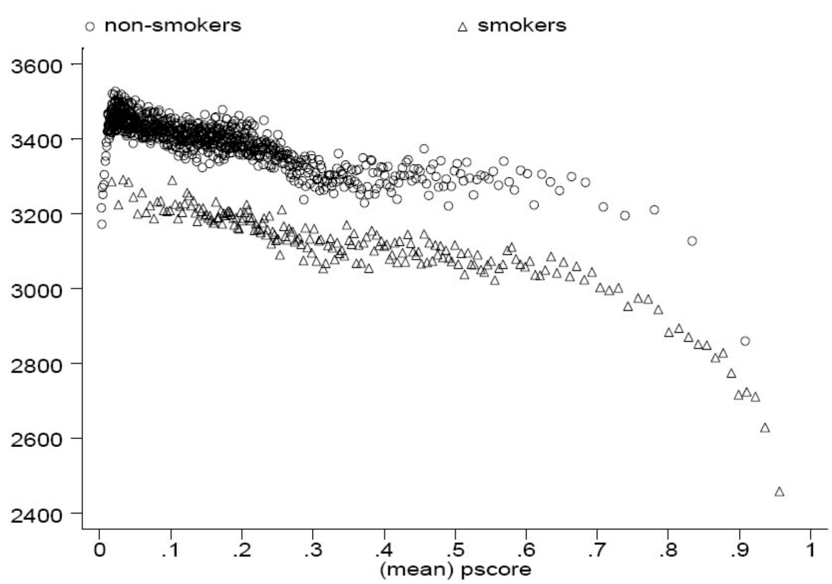


Figure 2: Average infant birth weight of smoking and non-smoking women by $\widehat{p}(\mathbf{X}_i)$

Another Example: Wage vs Union status

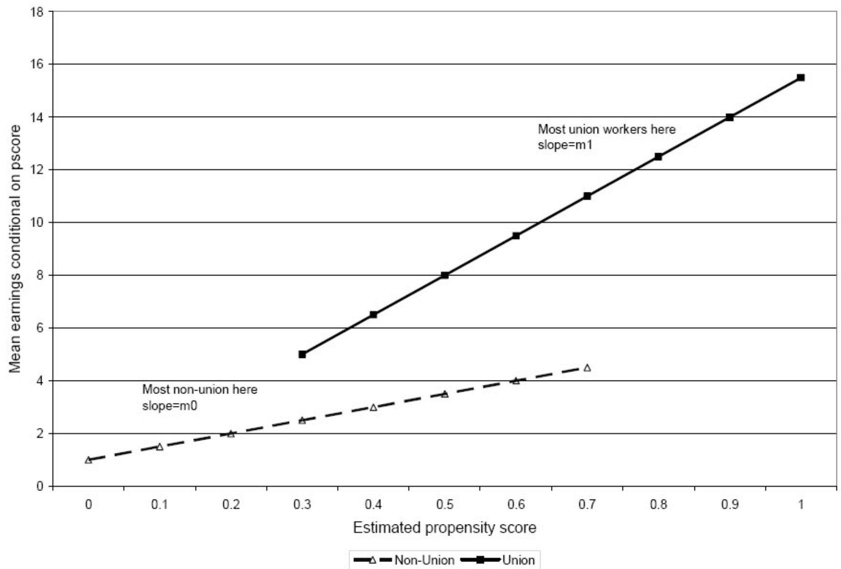


Figure 3: Mean earnings of union and non-union workers by $\widehat{p}(\mathbf{X}_i)$

a. Slope (m_0 and m_1): selection on observables
   ($m_1 - m_0$) gives differential selection into 2 sectors (union vs. non-union)
b. Bias in unadjusted union wage gap $\left(\overline{Y}_1 - \overline{Y}_0\right)$ because $m_1, m_0 \neq 0$
c. $\left(\overline{Y}_1 - \overline{Y}_0 \middle| \text{ fixed } \widehat{p}_i\right)$: union wage gap adjusted for selection on $\mathbf{X}_i$
d. If treatments are R.A. *conditional on observables*, it gives constant T.E. at each fixed $\widehat{p}_i$
   $\Rightarrow$ parallel lines (Not in the union example above, while parallel in the smoking example)
e. "Unrestricted" description of selection process and heterogeneity in treatment effects with the probability of selection only on *observables*

8.2.2 Regression Analogy

$$Y_i = \alpha + \theta \cdot T_i + \delta_1 \widehat{p}_i + \delta_2 T_i \left(\widehat{p}_i - \overline{\widehat{p}}\right) + U_i, \quad \overline{\widehat{p}} = \frac{1}{N}\sum_i \widehat{p}_i \tag{18}$$

- In the union example above, $\delta_1 = m_0$, $\delta_2 = (m_1 - m_0)$
- Restrictive linear specification in $\widehat{p}_i$: prone to misspecification
  plim $\widehat{\theta}_{OLS} = \theta$ iff $E[Y_{ji}|T_{ji}, \widehat{p}_i]$ are linear in $\widehat{p}_i$ – can test by including polynomials of $\widehat{p}_i$
- In fact, one can use nonparametric estimator (e.g., series estimator) – see Hahn (1998)
- Use bootstrap to calculate standard errors (generated regressor $\widehat{p}_i$)
- Again, it controls for selection bias only on *observables*

Example. Maternal smoking, birth weight, and the propensity score for smoking during pregnancy

```
. reg bweight tobacco, robust

Linear regression                                      Number of obs =  496677
                                                       F(  1,496675) =18946.80
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0385
                                                       Root MSE      =  577.17

------------------------------------------------------------------------------
             |               Robust
      dbirwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tobacco |  -284.8479   2.069403  -137.65   0.000    -288.9038   -280.7919
       _cons |   3423.683   .9117554  3755.05   0.000     3421.896     3425.47
------------------------------------------------------------------------------

. reg bweight tobacco pscore tobacco_pscore, robust

Linear regression                                      Number of obs =  496677
                                                       F(  3,496673) = 8068.09
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0510
                                                       Root MSE      =  573.43

--------------------------------------------------------------------------------
               |               Robust
       bweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
       tobacco |  -195.1582   2.376122   -82.13   0.000    -199.8153    -190.501
        pscore |   -359.974   6.921381   -52.01   0.000    -373.5397   -346.4083
tobacco_pscore |  -92.36724    11.0625    -8.35   0.000    -114.0494    70.68509
         _cons |   3482.924   1.375986  2531.22   0.000     3480.227    3485.621
--------------------------------------------------------------------------------
```

Should have corrected for standard error in estimated/generated regressor (pscore) – e.g., bootstrap

8.2.3 Subclassification on $\widehat{p}(\mathbf{X}_i)$

1. Stratify sample into G blocks based on $\widehat{p}_i$

   – How many blocks? $\Rightarrow$ apply the same algorithm previously used to estimate $\widehat{p}(\mathbf{X}_i)$

2. In each block $g = 1, \ldots, G$, estimate the mean difference in outcome variable ($\widehat{\theta}_g$)

$$\widehat{\theta}_g = \overline{Y}_{1g} - \overline{Y}_{0g} = \frac{\sum_{i\in g} T_i Y_i}{\sum_{i\in g} T_i} - \frac{\sum_{i\in g}(1-T_i)Y_i}{\sum_{i\in g}(1-T_i)} \tag{19}$$

   One can use the regression-adjusted estimator $\widehat{\theta}_g$: $Y_{ig} = \theta_g T_{ig} + \mathbf{X}'_{ig}\beta + U_{ig}$

3. Estimate average treatment effect (ATE) and average treatment effect on the treated (ATT)

$$\widehat{\theta}_{ATE} = \sum_{g=1}^{G}\left(\frac{N_{1g}+N_{0g}}{N}\right)\widehat{\theta}_g, \qquad \widehat{\theta}_{ATT} = \sum_{g=1}^{G}\left(\frac{N_{1g}}{N_1}\right)\widehat{\theta}_g, \tag{20}$$

8.2.4 Use $\widehat{p}(\mathbf{X}_i)$ as Individual Weights

Horvitz and Thompson (1952); Hahn (1998); Hirano, Imbens, and Ridder (2003)

(1) Average Treatment Effect: $E[Y_{1i} - Y_{0i}]$
Note that

$$E\left[\frac{T_i Y_i}{p(\mathbf{X}_i)}\right] = E\left[\frac{T_i Y_{1i}}{p(\mathbf{X}_i)}\right] = E\left[E\left[\frac{T_i Y_{1i}}{p(\mathbf{X}_i)}\right]\middle|\mathbf{X}_i\right] = E\left[\frac{E[T_i|\mathbf{X}_i]E[Y_{1i}|\mathbf{X}_i]}{p(\mathbf{X}_i)}\right] = E[E[Y_{1i}|\mathbf{X}_i] = E[Y_{1i}]$$

and similarly

$$E\left[\frac{(1-T_i)Y_i}{1-p(\mathbf{X}_i)}\right] = E[Y_{0i}]$$

Thus, an obvious estimator of ATE with the known propensity score $p(\mathbf{X}_i)$

$$\widetilde{E}[Y_{1i} - Y_{0i}] \equiv \widetilde{\theta}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{T_i Y_i}{p(\mathbf{X}_i)} - \frac{(1-T_i)Y_i}{1-p(\mathbf{X}_i)}\right] \tag{21}$$

Estimate $p(\mathbf{X}_i)$ flexibly (Hirano, Imbens, and Ridder, 2003)

$$\widehat{E}[Y_{1i} - Y_{0i}] \equiv \widehat{\theta}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{T_i Y_i}{\widehat{p}(\mathbf{X}_i)} - \frac{(1-T_i)Y_i}{1-\widehat{p}(\mathbf{X}_i)}\right] \tag{22}$$

- Weighting treatment unit by $1/\widehat{p}(\mathbf{X}_i)$ and control unit by $1/\left(1-\widehat{p}(\mathbf{X}_i)\right)$. Intuition?
- The weights do not necessarily add up to 1 $\Rightarrow$ normalizing the weights so they sum to 1:

$$\widehat{\theta}^{IPW}_{ATE} = \left[\sum_{i=1}^{N}\frac{T_i Y_i}{\widehat{p}(\mathbf{X}_i)}\Big/\sum_{i=1}^{N}\frac{T_i}{\widehat{p}(\mathbf{X}_i)}\right] - \left[\sum_{i=1}^{N}\frac{(1-T_i)Y_i}{1-\widehat{p}(\mathbf{X}_i)}\Big/\sum_{i=1}^{N}\frac{(1-T_i)}{1-\widehat{p}(\mathbf{X}_i)}\right] \tag{23}$$

  a.k.a. Inverse Probability Weighting (IPW) estimator

```
[STATA] reg Y T [pw=ate_weight], robust
```

(2) Average Treatment Effect on the Treated: $E(Y_{1i} - Y_{0i}|T_i = 1)$

Note that $(1-T_i)Y_i = (1-T_i)Y_{0i} \Rightarrow E[T_i Y_{0i}] = E[Y_{0i}] - E[(1-T_i)Y_i]$

$$\begin{aligned} E[Y_{1i} - Y_{0i}|T_i = 1] &= \frac{E[T_i Y_{1i}] - E[T_i Y_{0i}]}{\Pr(T_i = 1)} = \frac{E[T_i Y_i] - E[Y_{0i}] + E[(1-T_i)Y_i]}{\Pr(T_i = 1)} = \frac{E[Y_i] - E[Y_{0i}]}{\Pr(T_i = 1)} \\ &= \frac{E[Y_i] - E\left[\frac{(1-T_i)Y_i}{1-p(\mathbf{X}_i)}\right]}{\Pr(T_i = 1)} = \frac{E\left[T_i Y_i - \frac{p(\mathbf{X}_i)(1-T_i)Y_i}{1-p(\mathbf{X}_i)}\right]}{\Pr(T_i = 1)} \end{aligned}$$

Thus,

$$\widehat{E}(Y_{1i} - Y_{0i}|T_i = 1) \equiv \widehat{\theta}_{ATT} = \frac{1}{N}\left(\frac{N_1}{N}\right)^{-1}\sum_{i=1}^{N}\left[T_iY_i - \frac{\widehat{p}(\mathbf{X}_i)(1-T_i)Y_i}{1-\widehat{p}(\mathbf{X}_i)}\right] \tag{24}$$

- Weighting treatment unit by 1 and control unit by $\widehat{p}(\mathbf{X}_i)/\left(1-\widehat{p}(\mathbf{X}_i)\right)$
- Again, normalizing the weights so they sum to 1:

$$\widehat{\theta}_{ATT}^{IPW} = \left[\frac{1}{N_1}\sum_{i=1}^{N}T_iY_i\right] - \left[\sum_{i=1}^{N}\frac{\widehat{p}(\mathbf{X}_i)(1-T_i)Y_i}{1-\widehat{p}(\mathbf{X}_i)}\Bigg/\sum_{i=1}^{N}\frac{\widehat{p}(\mathbf{X}_i)(1-T_i)}{1-\widehat{p}(\mathbf{X}_i)}\right] \tag{25}$$

```
[STATA] reg Y T [pw=att_weight], robust
```

Example. ATE and ATT for birth weight using $\widehat{p}(\mathbf{X}_i)$ as individual weights

(a) Average Treatment Effect (ATE)

```
. reg bweight tobacco [pw=ate_weight], robust
(sum of wgt is   9.9410e+05)

Linear regression                                      Number of obs =  496677
                                                       F(  1,496675) = 3846.82
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0291
                                                       Root MSE      =  578.38

------------------------------------------------------------------------------
             |               Robust
     bweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tobacco |  -200.2889   2.069403  -137.65   0.000    -288.9038   -280.7919
       _cons |   3404.804   1.161227  2932.08   0.000     3402.528     3407.08
------------------------------------------------------------------------------
```

(b) Average Treatment Effect on the Treated (ATT)

```
. reg bweight tobacco [pw=att_weight], robust
(sum of wgt is   2.0810e+05)

Linear regression                                      Number of obs =  496677
                                                       F(  1,496675) = 2784.34
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0244
                                                       Root MSE      =  616.51

------------------------------------------------------------------------------
             |               Robust
     bweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tobacco |  -195.2046   3.699382   -52.77   0.000    -202.4553   -187.9539
       _cons |    3334.04   3.199109  1042.18   0.000      3327.77     3340.31
------------------------------------------------------------------------------
```

8.2.5 More on Using $\widehat{p}(\mathbf{X}_i)$ as Individual Weights

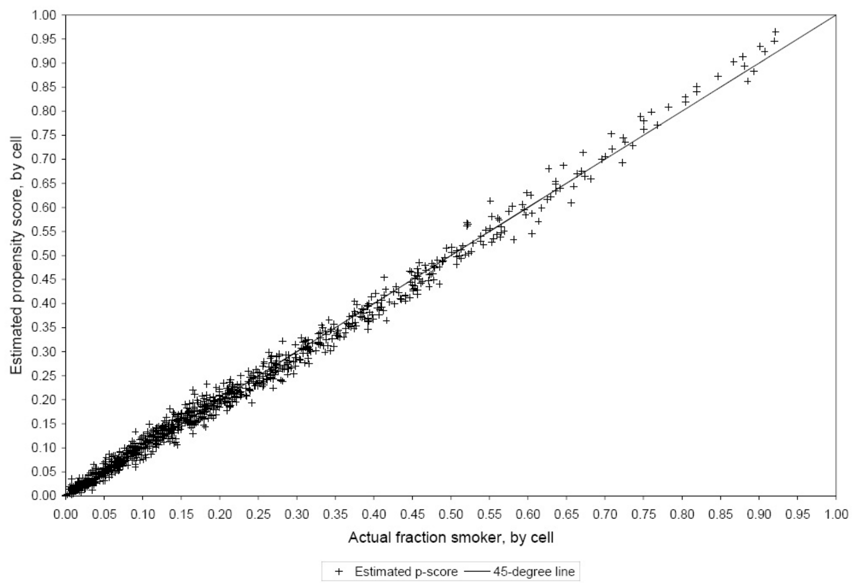
- "Efficient" if use estimated $\widehat{p}(\mathbf{X}_i)$ rather than true $p(\mathbf{X}_i)$ (Hirano, Imbens, and Ridder, 2003)
- Efficient but "sensitive to misspecification" of $\widehat{p}(\mathbf{X}_i) \Rightarrow$ Bias

  – If $\widehat{p}(\mathbf{X}_i) > p(\mathbf{X}_i) \Rightarrow$ too much weight; if $\widehat{p}(\mathbf{X}_i) < p(\mathbf{X}_i) \Rightarrow$ too little weight
- More sensitive for controls with high $\widehat{p}(\mathbf{X}_i)$.

  Ex. Suppose the true $p(\mathbf{X}_i) = .95 \Rightarrow$ IPW for the control unit is 20 (ATE) and 19 (ATT)

  If we overestimate $\widehat{p}(\mathbf{X}_i) = .98 \Rightarrow \widehat{\text{IPW}}$ for the control unit is 50 (ATE) and 49 (ATT)

  $\Rightarrow \widehat{p}(\mathbf{X}_i)$ may be poor approximation at high $p(\mathbf{X}_i)$ (few control units, etc.)
- Checking the sensitivity of $\widehat{p}(\mathbf{X}_i)$? Plot average estimated p-score ($\overline{\widehat{p}}(\mathbf{X}_i)$) against actual fraction smoker for 1,000 equal sized cells of the estimated p-score (about 500 obs. per cell)



- Busso, DiNardo, and McCrary (2013): when overlap is good, the p–score weighting estimator almost always outperforms matching estimators

Other Applications of IPW

- Use IPW to correct for *sample* selection (if sample selection is on the observables) instead of Heckman's selection-correction model (Will cover later)
- Use IPW to derive *counterfactual* <u>*distributions*</u> (DiNardo, Fortin, and Lemieux, 1996)

  What would the wage distribution of non-union members have been if they had same characteristics ($\mathbf{X}_i$) as union members $\Rightarrow$ reweigh the non-union members by $\frac{\widehat{p}(\mathbf{X}_i)}{1-\widehat{p}(\mathbf{X}_i)}$

8.2.6 Propensity Score Weighting and Regression

- "Doubly-Robust" estimator: combination of propensity score weighting and regression (Robins and Rotnitzky, 1995; Rotnitzky, Robins, and Scharfstein, 1998)
- Idea: the combined estimator for ATE is consistent as long as the propensity score or the regression functions are specified correctly
- Implement the following Weighted Least Square (WLS) estimation

$$Y_i = \alpha + \theta \cdot T_i + (\mathbf{X}_i - \overline{\mathbf{X}})'\beta + U_i \tag{26}$$

  with weights equal to

$$\omega_i = \frac{T_i}{\widehat{p}(\mathbf{X}_i)} + \frac{1 - T_i}{1 - \widehat{p}(\mathbf{X}_i)} \tag{27}$$

- The estimator for ATT?
- Application: Hirano and Imbens (2001)

8.2.7 Propensity Score Matching

Matching on the propensity score p_i instead of covariates $\mathbf{X}_i$ – e.g., the nearest-neighbor matching:

$$W(i,j) = \begin{cases} 1, & \text{if } j \in A_i \\ 0, & \text{otherwise} \end{cases} \tag{28}$$

$$A_i = \{j | \min_j ||p_i - p_j||\} \tag{29}$$

Very refined version of Graphical Analysis (3.2.1) or Subclassification (3.2.3)

In summary,

propensity score methods "decompose" dimensionality problem into 2 parts:

1. Selection equation $p(\mathbf{X}_i)$: allow for higher order terms
2. Outcome equation: control for $p(\mathbf{X}_i)$

Could misspecify either equation. It is NOT Research Design, just a Descriptive Tool !!

9 Applications: Evaluation of Job Training Program

- Ashenfelter (1974, 1978): the only way to get credible estimates of training's impact ⇒ randomized experiment

  Nonexperimental methods: 1) unstable estimates and 2) don't replicate experimental results

- Problem of nonexperimental methods: Non-random selection into the program

  Ex. Training program in 1964 under the Manpower Development and Training Act (MDTA)

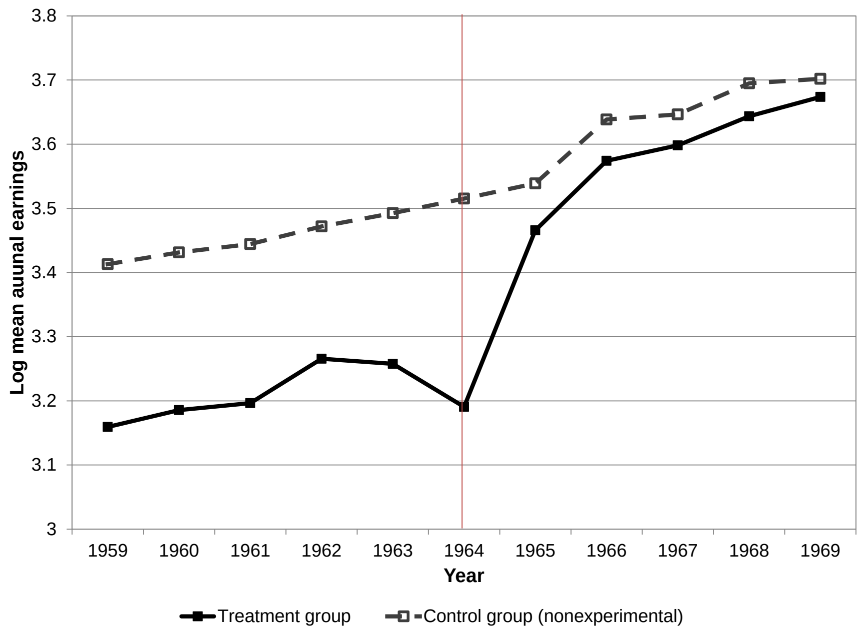


Figure 4: "Ashenfelter's Dip" (Ashenfelter, 1978)

- Differences in Y_{it} before training program: disadvantaged people select into the program
- Differences in earnings dynamics:

  trainees have decline in Y_{it} right before they select the program (a.k.a., "Ashenfelter's dip");

  administrators were told to pick these types, and those with negative shocks will participate

  ⇒ Even if no program, income would rise (mean reversion/feedback problem)

  ⇒ Difference-in-Differences estimators are biased

  Solution? **GET MORE DATA**

  i) with long-horizon longitudinal data, take time window far away from the negative shock

  ii) also useful to plot the series of outcomes for T & C groups to see if trends are *parallel*

  AND need to model the selection rule (i.e., depending on pre-training earnings)

LaLonde (1986): Experimental vs Nonexperimental Approaches

- National Supported Work (NSW) program between Jan. 1976 and July 1977
- Experimental Estimates
  - Difference in post-training (1979) earnings: \$851 (AFDC women) and \$886 (men)
  - Insensitive (robust) to econometric procedures
- Nonexperimental Estimates
  - Nonexperimental control groups from PSID and CPS
  - Apply regression, difference-in-differences, and selection correction methods
  - Estimates vary widely – in many cases not even close to the experimental estimates

Dehejia and Wahba (1999)

- Claim: LaLonde (1986) only allows $\mathbf{X}_i$ to enter restrictively in regression equation
- Apply <u>propensity score methods</u> : regression (including $\widehat{p}_i^2$), subclassification, and matching
  - Use nonexperimental control groups from PSID and CPS
  - Argue that nonexperimental estimates are close to the experimental benchmark (Table 3)
- Note: small number of covariates, don't show fit of models, and thus, didn't give non-experimental econometric methods fair shake
  - Nonexperimental control groups are totally different from the treatment group (Table 1)
  - Almost no overlap in $\widehat{p}_i$ (Figure 1 and 2) $\Rightarrow$ poor design
  - Huge standard errors (Table 3)
  - Specification search $\Rightarrow$ hardly replicable (different people get different estimates)
  - Need to know experimental estimates to assess nonexperimental ones

Smith and Todd (2005)

- Claim: Dehejia and Wahba (1999) propensity score estimates are very sensitive to control variables and the different control groups
- Apply <u>difference-in-differences matching</u> to eliminate time-invariant biases due to:

  i) difference in geographic location between treatment and control group;

  ii) difference in the measurement of the outcome variable (i.e., earnings)
- Difference-in-difference matching performs better

<u>Next</u>: Selection on Unobservables

References

Ashenfelter, Orley. 1974. "The Effect of Manpower Training on Earnings: Preliminary Result." In *Proceedings of the Twenty-Seventh Annual Win- ter Meetings of the Industrial Relations Research Association*, edited by J. Stern and B. Dennis. Madison, WI: Industrial Relations Research Association.

———. 1978. "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics* 60 (1):47–57.

Busso, Matias, John DiNardo, and Justin McCrary. 2013. "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators." Working Paper.

Dehejia, Rajeev H. and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94 (448):1053–1062.

DiNardo, John., Nicole M. Fortin, and Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." *Econometrica* 64 (5):1001–1044.

Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66 (2):315–331.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1998. "Matching As an Econometric Evaluation Estimator." *Review of Economic Studies* 65 (2):261–294.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. "Matching As an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64 (4):605–654.

Hirano, Keisuke and Guido W. Imbens. 2001. "Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization." *Health Services and Outcomes Research Methodology* 2 (3):259–278.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71 (4):1161–1189.

Horvitz, Daniel G. and Donovan J. Thompson. 1952. "A Generalization of Sampling Without Replacement From a Finite Universe." *Journal of the American Statistical Association* 47 (260):663–685.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76 (4):604–620.

Robins, James M. and Andrea Rotnitzky. 1995. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data." *Journal of the American Statistical Association* 90 (429):122–129.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1):41–55.

———. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79 (387):516–524.

Rotnitzky, Andrea, James M. Robins, and Daniel O. Scharfstein. 1998. "Semiparametric Regression for Repeated Outcomes with Nonignorable Nonresponse." *Journal of the American Statistical Association* 93 (444):1321–1339.

Smith, Jeffrey A. and Petra E. Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (1-2):305–353.