# Propensity Score Matching

Rajeev Dehejia

New York University

# Adjustment for observables in observational studies (Conditioning)

1. Subclassification
2. Matching
3. Regression
4. Propensity Score Methods

# Recall the goal and challenge

- We are in the universe of selection on observables:

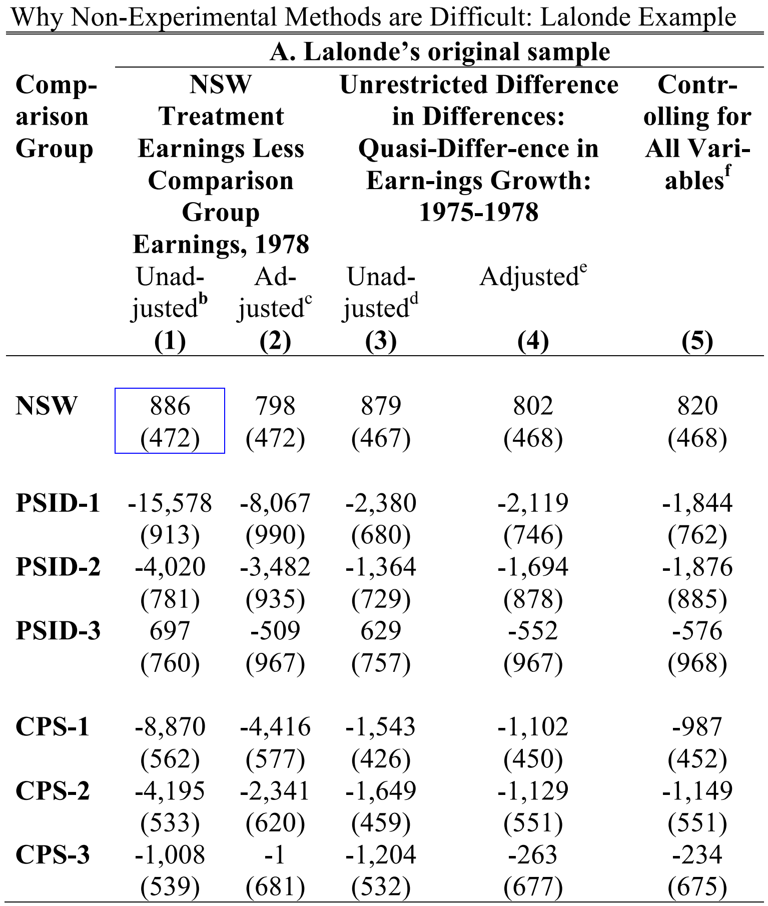
$$\{Y_{i1},\, Y_{i0} \perp\!\!\!\perp T_i\} \,|\, X_i.$$

- The problem is that to estimate

$$\tau\Big|_{T_i=1} = E_X[E(Y_{i1} \,|\, T_i = 1, X_i)] - E_X[E(Y_{i0} \,|\, T_i = 1, X_i)]$$

when **the treatment and comparison groups are very different**, and when *X* is high dimensional.

# Why is this hard? Lalonde example

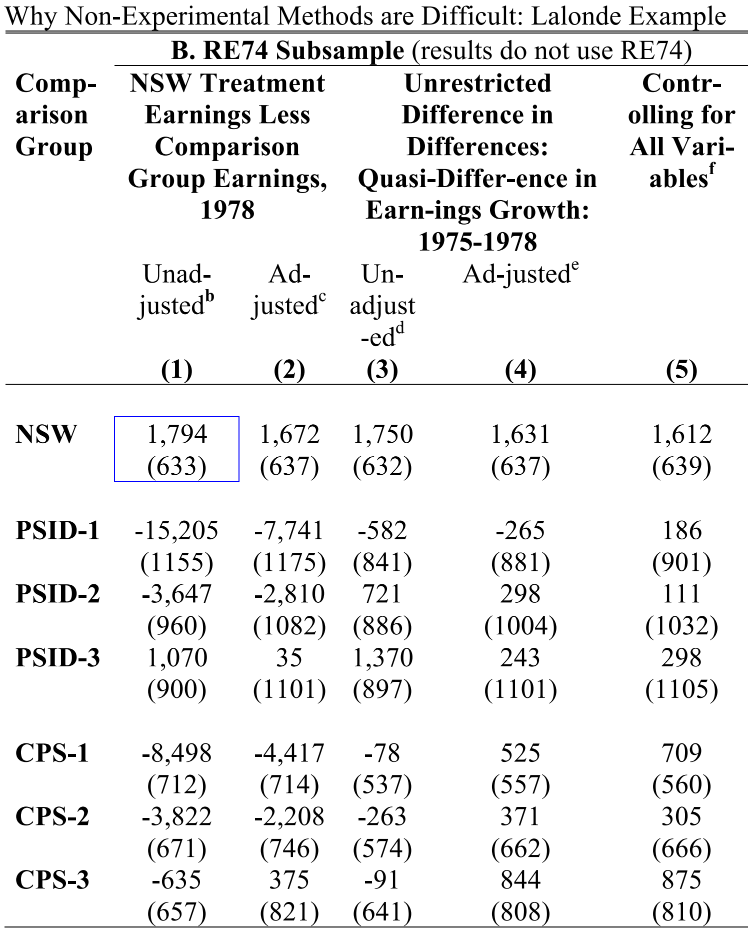


Why Non-Experimental Methods are Difficult: Lalonde Example

|  | A. Lalonde's original sample |  |  |  |  |
|---|---|---|---|---|---|
| **Comparison Group** | **NSW Treatment Earnings Less Comparison Group Earnings, 1978** |  | **Unrestricted Difference in Differences: Quasi-Differ-ence in Earn-ings Growth: 1975-1978** |  | **Controlling for All Variables[f]** |
|  | Unadjusted[b] | Adjusted[c] | Unadjusted[d] | Adjusted[e] |  |
|  | **(1)** | **(2)** | **(3)** | **(4)** | **(5)** |
| **NSW** | 886<br>(472) | 798<br>(472) | 879<br>(467) | 802<br>(468) | 820<br>(468) |
| **PSID-1** | -15,578<br>(913) | -8,067<br>(990) | -2,380<br>(680) | -2,119<br>(746) | -1,844<br>(762) |
| **PSID-2** | -4,020<br>(781) | -3,482<br>(935) | -1,364<br>(729) | -1,694<br>(878) | -1,876<br>(885) |
| **PSID-3** | 697<br>(760) | -509<br>(967) | 629<br>(757) | -552<br>(967) | -576<br>(968) |
| **CPS-1** | -8,870<br>(562) | -4,416<br>(577) | -1,543<br>(426) | -1,102<br>(450) | -987<br>(452) |
| **CPS-2** | -4,195<br>(533) | -2,341<br>(620) | -1,649<br>(459) | -1,129<br>(551) | -1,149<br>(551) |
| **CPS-3** | -1,008<br>(539) | -1<br>(681) | -1,204<br>(532) | -263<br>(677) | -234<br>(675) |

PSID-1: All male household heads under age 55 who did not classify themselves as retired in 1975.
PSID-2: Selects from PSID-1 all men who were not working when surveyed in the spring of 1976.
PSID-3: Selects from PSID-2 all men who were not working in 1975.
CPS-1: All CPS males under age 55.
CPS-2: Selects from CPS-1 all males who were not working when surveyed in March 1976.
CPS-3: Selects from CPS-2 all the unemployed males in 1976 whose income in 1975 was below the poverty level.

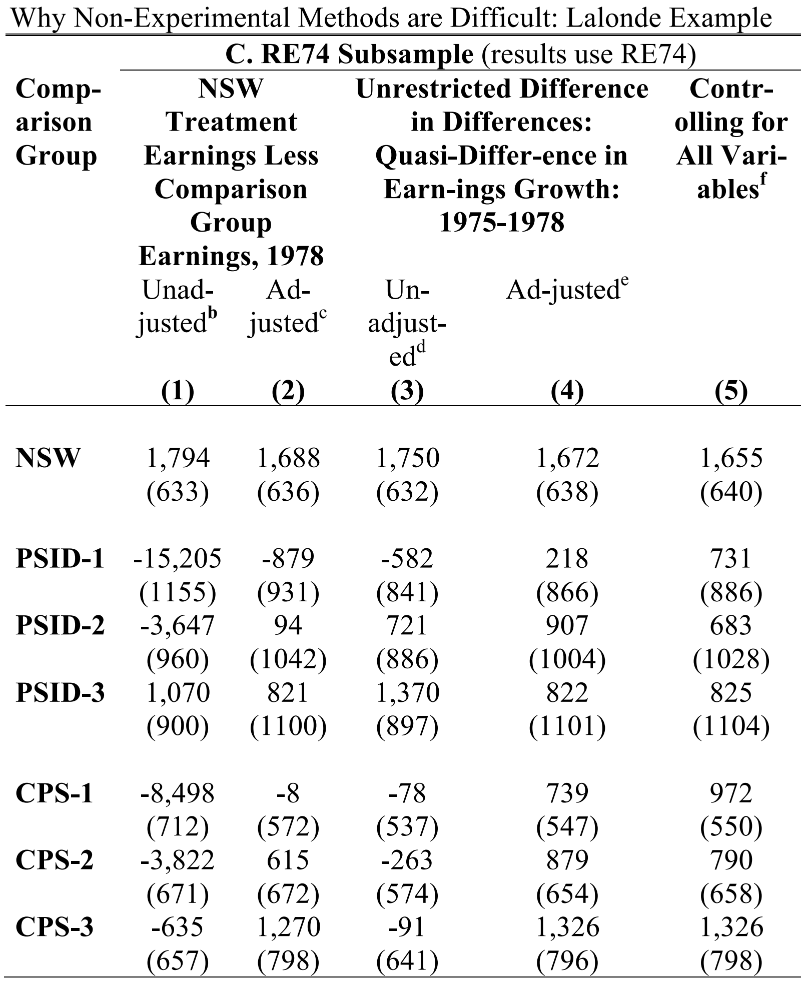
# Why is this hard? Lalonde example



Why Non-Experimental Methods are Difficult: Lalonde Example

| | **B. RE74 Subsample** (results do not use RE74) | | | | |
|---|---|---|---|---|---|
| **Comparison Group** | **NSW Treatment Earnings Less Comparison Group Earnings, 1978** | | **Unrestricted Difference in Differences: Quasi-Differ-ence in Earn-ings Growth: 1975-1978** | | **Controlling for All Variables**[f] |
| | Unadjusted[b] | Adjusted[c] | Unadjusted[d] | Ad-justed[e] | |
| | **(1)** | **(2)** | **(3)** | **(4)** | **(5)** |
| **NSW** | 1,794<br>(633) | 1,672<br>(637) | 1,750<br>(632) | 1,631<br>(637) | 1,612<br>(639) |
| **PSID-1** | -15,205<br>(1155) | -7,741<br>(1175) | -582<br>(841) | -265<br>(881) | 186<br>(901) |
| **PSID-2** | -3,647<br>(960) | -2,810<br>(1082) | 721<br>(886) | 298<br>(1004) | 111<br>(1032) |
| **PSID-3** | 1,070<br>(900) | 35<br>(1101) | 1,370<br>(897) | 243<br>(1101) | 298<br>(1105) |
| **CPS-1** | -8,498<br>(712) | -4,417<br>(714) | -78<br>(537) | 525<br>(557) | 709<br>(560) |
| **CPS-2** | -3,822<br>(671) | -2,208<br>(746) | -263<br>(574) | 371<br>(662) | 305<br>(666) |
| **CPS-3** | -635<br>(657) | 375<br>(821) | -91<br>(641) | 844<br>(808) | 875<br>(810) |

# Why is this hard? Lalonde example



Why Non-Experimental Methods are Difficult: Lalonde Example

| | **C. RE74 Subsample** (results use RE74) | | | | |
|---|---|---|---|---|---|
| **Comparison Group** | **NSW Treatment Earnings Less Comparison Group Earnings, 1978** | | **Unrestricted Difference in Differences: Quasi-Differ-ence in Earn-ings Growth: 1975-1978** | | **Controlling for All Variables**[f] |
| | Unadjusted[b] | Adjusted[c] | Unadjusted[d] | Ad-justed[e] | |
| | **(1)** | **(2)** | **(3)** | **(4)** | **(5)** |
| **NSW** | 1,794<br>(633) | 1,688<br>(636) | 1,750<br>(632) | 1,672<br>(638) | 1,655<br>(640) |
| **PSID-1** | -15,205<br>(1155) | -879<br>(931) | -582<br>(841) | 218<br>(866) | 731<br>(886) |
| **PSID-2** | -3,647<br>(960) | 94<br>(1042) | 721<br>(886) | 907<br>(1004) | 683<br>(1028) |
| **PSID-3** | 1,070<br>(900) | 821<br>(1100) | 1,370<br>(897) | 822<br>(1101) | 825<br>(1104) |
| **CPS-1** | -8,498<br>(712) | -8<br>(572) | -78<br>(537) | 739<br>(547) | 972<br>(550) |
| **CPS-2** | -3,822<br>(671) | 615<br>(672) | -263<br>(574) | 879<br>(654) | 790<br>(658) |
| **CPS-3** | -635<br>(657) | 1,270<br>(798) | -91<br>(641) | 1,326<br>(796) | 1,326<br>(798) |

# Why is this hard? Cochran

Cochran (1965) wrote:

> If the original x-distributions [of comparison groups] diverge widely, none of the methods [e.g., regression adjustment] can be trusted to remove all, or nearly all, the bias. This discussion brings out the importance of finding comparison groups in which the initial difference among the distributions of the distributing variables are small.

*In other words, try to compare apples to apples as much as possible…*

# Why is this hard? Lalonde revisited

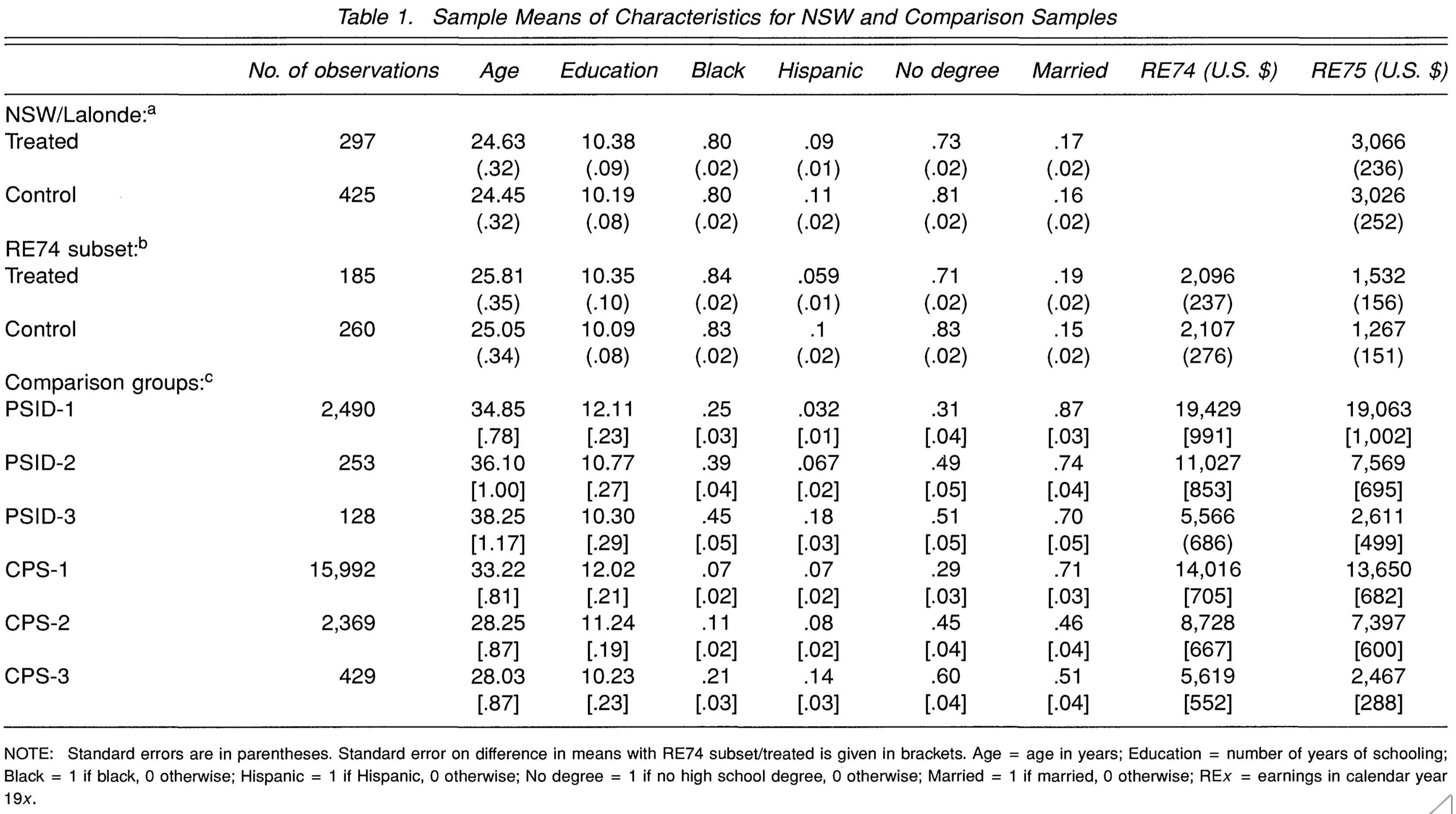


*Table 1. Sample Means of Characteristics for NSW and Comparison Samples*

| | *No. of observations* | *Age* | *Education* | *Black* | *Hispanic* | *No degree* | *Married* | *RE74 (U.S. $)* | *RE75 (U.S. $)* |
|---|---|---|---|---|---|---|---|---|---|
| NSW/Lalonde:[a] | | | | | | | | | |
| Treated | 297 | 24.63 | 10.38 | .80 | .09 | .73 | .17 | | 3,066 |
| | | (.32) | (.09) | (.02) | (.01) | (.02) | (.02) | | (236) |
| Control | 425 | 24.45 | 10.19 | .80 | .11 | .81 | .16 | | 3,026 |
| | | (.32) | (.08) | (.02) | (.02) | (.02) | (.02) | | (252) |
| RE74 subset:[b] | | | | | | | | | |
| Treated | 185 | 25.81 | 10.35 | .84 | .059 | .71 | .19 | 2,096 | 1,532 |
| | | (.35) | (.10) | (.02) | (.01) | (.02) | (.02) | (237) | (156) |
| Control | 260 | 25.05 | 10.09 | .83 | .1 | .83 | .15 | 2,107 | 1,267 |
| | | (.34) | (.08) | (.02) | (.02) | (.02) | (.02) | (276) | (151) |
| Comparison groups:[c] | | | | | | | | | |
| PSID-1 | 2,490 | 34.85 | 12.11 | .25 | .032 | .31 | .87 | 19,429 | 19,063 |
| | | [.78] | [.23] | [.03] | [.01] | [.04] | [.03] | [991] | [1,002] |
| PSID-2 | 253 | 36.10 | 10.77 | .39 | .067 | .49 | .74 | 11,027 | 7,569 |
| | | [1.00] | [.27] | [.04] | [.02] | [.05] | [.04] | [853] | [695] |
| PSID-3 | 128 | 38.25 | 10.30 | .45 | .18 | .51 | .70 | 5,566 | 2,611 |
| | | [1.17] | [.29] | [.05] | [.03] | [.05] | [.05] | (686) | [499] |
| CPS-1 | 15,992 | 33.22 | 12.02 | .07 | .07 | .29 | .71 | 14,016 | 13,650 |
| | | [.81] | [.21] | [.02] | [.02] | [.03] | [.03] | [705] | [682] |
| CPS-2 | 2,369 | 28.25 | 11.24 | .11 | .08 | .45 | .46 | 8,728 | 7,397 |
| | | [.87] | [.19] | [.02] | [.02] | [.04] | [.04] | [667] | [600] |
| CPS-3 | 429 | 28.03 | 10.23 | .21 | .14 | .60 | .51 | 5,619 | 2,467 |
| | | [.87] | [.23] | [.03] | [.03] | [.04] | [.04] | [552] | [288] |

NOTE: Standard errors are in parentheses. Standard error on difference in means with RE74 subset/treated is given in brackets. Age = age in years; Education = number of years of schooling; Black = 1 if black, 0 otherwise; Hispanic = 1 if Hispanic, 0 otherwise; No degree = 1 if no high school degree, 0 otherwise; Married = 1 if married, 0 otherwise; RE*x* = earnings in calendar year 19*x*.

# Why is this hard? Recall Abadie-Imbens result

A standard technique for estimating the treatment effect is to matching on the covariates *X.*

## Idea

In the above expression we are conditioning on *X*. So if we can somehow match units with the same (similar) *X*, then within these cells we can directly estimate $\mathrm{E}(Y_{i1}|T_i=1,X_i)$ and

$\mathrm{E}(Y_{i0}|T_i=1,X_i)$.

## Problem

Theorem (Abadie-Imbens): If there are two or more continuous covariates, then the simple matching estimator is asymptotically biased (not $N^{-1/2}$ consistent).

# An obvious, efficient, but impractical solution

Hahn (1998) shows that if we estimate E($Y_{i1}$|$T_i$=1,$X_i$) and E($Y_{i0}$|$T_i$=1,$X_i$) non-parametrically, the semi-parametric efficiency bound is achieved.

## Problem

This involves estimating two non-parametric equations. When $X$ is high-dimensional and has many continuous variables (typical application – 10 variables, 5 of them continuous) this is very difficult computationally.

- E.g., local linear regression like we used in RD would not be practical

# Propensity score methods

Propensity score methods help to control for sample election bias, under the assumption of selection on observables.

Experience suggests that dealing with the problem of sample selection bias is difficult.

- Extrapolation and linear regression regression is unreliable under many circumstances.
- Parametric selection models are not robust.
- Instruments are not always available.
- Higher-order non-parametric regression is difficult.

Propensity score methods offer a solution to this problem, when selection is on observable variables.

# Propensity score methods: General setup

- The treatment may be an explicit or organized program (such as labor training, a medical trial, an educational experiment) or may be a policy variation (such as certain subjects being exposed to a new set of regulations).

- Treatment group: exposed to the variation of phenomenon of interest. Denote treatment-exposed individuals, $i$, in this group with $T_i$=1.

- Control group: not exposed to the treatment; exposed to an alternative, baseline, or status quo program. Denote individuals, $i$, in this group with $T_i$=0.

# Propensity score methods

## Definition

Propensity score is defined as the **selection probability conditional on the confounding variables**: $p(X) = P(T = 1|X)$

## Identification Assumption

1. $(Y_1,Y_0) \perp\!\!\!\perp T|X$ *(selection on observables)*
2. $0 < \Pr(T=1|X) < 1$ *(common support)*

## Identification Result (Rosenbaum and Rubin, 1983)

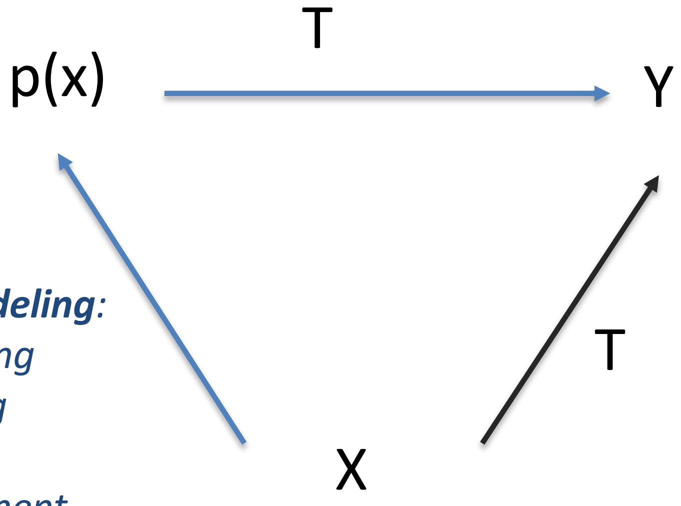
*Selection on observables implies:*

$$(Y_1,Y_0) \perp\!\!\!\perp T \mid X \Rightarrow (Y_1,Y_0) \perp\!\!\!\perp T \mid p(X)$$

→ *conditioning on the propensity score is enough to have independence between the treatment indicator and the potential outcomes*

→ *substantial dimension reduction in the matching variables!*

# Propensity Score Matching: Graphical Representation



***Propensity score modeling**: control for confounding variables by modeling relationship between covariates and treatment assignment*

***Standard regression adjustment:** control for confounding variables adjustment by modeling relationship between covariates and outcome*

# Propensity score methods

## Proof.

Assume that $(Y_1,Y_0) \perp\!\!\!\perp T|X$. Then:

$$\begin{aligned}
\mathrm{P}(T = 1|Y_1, Y_0, p(X)) &= E[T|Y_1, Y_0, p(X)] \\
&= E\{E_X[E(T|\mathrm{Y}_1, Y_0, X)]|\mathrm{Y}_1, Y_0, p(X)\} \\
&= E\{E_X[E(T|\mathrm{X})]|\mathrm{Y}_1, Y_0, p(X)\} \\
&= E\{E_X[p(X)]|\mathrm{Y}_1, Y_0, p(X)\} \\
&= p(X)
\end{aligned}$$

Using a similar argument, we obtain

$$\begin{aligned}
\mathrm{P}(T = 1|p(X)) &= E[T|p(X)] = E[E[T|\mathrm{X}]|p(X)] \\
&= E[p(X)|p(X)] = p(X)
\end{aligned}$$

→ $\mathrm{P}(T = 1|Y_1, Y_0, p(X)) = \mathrm{P}(T = 1|p(X))$

→ $(Y_1,Y_0) \perp\!\!\!\perp T|p(X)$

# Matching on the propensity score

Corollary

*If* $(Y_1, Y_0) \perp\!\!\!\perp T | X$*, then*

$$E[Y_1 - Y_0 | p(X)] = E[Y | T = 1, p(X)] - E[Y | T = 0, p(X)]$$

Suggests a two-step procedure to estimate causal effects under selection on observables:

1. Estimate the propensity score $p(X) = P(T = 1 | X)$ (e.g., using logit or probit regression)
2. Do matching or subclassification on the estimated propensity score

Hugely popular method to estimate treatment effects. However, valid method to calculate standard errors not known until (very) recently.

# Propensity score: balancing property

Because the propensity score, $p(X)$ is a function of $X$:

$$\Pr(T = 1 \mid p(X)) = \Pr(T = 1 \mid X)$$
$$= p(X)$$

→ conditional on $p(X)$, the probability that $T = 1$ does not depend on $X$ beyond $p(X)$. *Think back to the exclusion restriction from IV….*

→ $T$ and $X$ are independent conditional on $p(X)$:

$$T \perp\!\!\!\perp X \mid p(X).$$

So we obtain the **balancing property** of the propensity score:

$$P(X \mid T = 1,\ p(X)) = P(X \mid T = 0,\ p(X)),$$

*conditional on the propensity score, the distribution of the covariates is the same for treated and non-treated.*

We can use this to check if our estimated propensity score actually produces balance:

$$P\left(X \middle| T = 1, \hat{p}\left(X\right)\right) = P\left(X \middle| T = 0, \hat{p}\left(X\right)\right)$$

# How to do it: using balancing to estimate *p*(*X*)

1. Start with a parsimonious logit specification to estimate the score.
2. Sort data according to estimated propensity score (ranking from lowest to highest).
3. Stratify all observations such that estimated propensity scores within a stratum for treated and comparison units are close (no significant difference); e.g., start by dividing observations into strata of equal score range (0-0.2,...,0.8-1).

| Unit ID | Earnings | College Grad | ***+ k descriptive characteristics*** |
|---|---|---|---|
| 1 | $100 | 0 | *Several fixed observables such as gender, age, parents' education, etc. for each unit* |
| 2 | $300 | 1 | |
| 3 | $150 | 1 | |
| 4 | $50 | 0 | |
| 5 | $75 | 1 | |
| 6 | $225 | 0 | |

Install one of a few stata programs, such as **pscore**, to run the logit reg, estimate the propensity scores, and create a propensity score variable by which you can sort the data

| Unit ID | Earnings | College Grad | ***Propensity Score*** | ***Quartile*** |
|---|---|---|---|---|
| 2 | $300 | 1 | *0.98* | *1* |
| 4 | $50 | 0 | *0.91* | *1* |
| 5 | $75 | 1 | *0.61* | *2* |
| 6 | $225 | 0 | *0.59* | *2* |
| 1 | $100 | 0 | *0.40* | *3* |
| 3 | $150 | 1 | *0.39* | *3* |

# How to do it: using balancing to estimate *p*(*X*)

**Statistical test:** for all covariates, differences in means across treated and comparison units within each stratum are not significantly different from zero.

- ✓ If covariates are balanced between treated and comparison observations for all strata, stop.
- ✓ If covariates are not balanced for some stratum, divide the stratum into finer strata and re-evaluate.
- ✓ If a covariate is not balanced for many strata, modify the logit by adding interaction terms and/or higher-order terms of the covariate and re-evaluate.

# How to do it: using balancing to estimate $p(X)$

✓ The difference in the means of the propensity scores in the two groups being compared must be small (e.g., the means must be less than half a standard deviation apart), unless the situation is benign in the sense that: (a) the distributions of the covariates in both groups are nearly symmetric, (b) the distributions of the covariates in both groups have nearly the same variances, and (c) the sample sizes are approximately the same.

✓ The ratio of the variances of the propensity score in the two groups must be close to one (e.g., 1/2 or 2 are far too extreme)

✓ The ratio of the variances of the residuals of the covariates after adjusting for the propensity score must be close to one (e.g., 1/2 or 2 are far too extreme)

# How to do it: controversies with balancing

- Imai, King, and Stuart (2007) argue that traditional hypothesis testing is inappropriate to test balance. And in principle they are right:
  - Balance is a property of the sample at hand, not a population, so not clear that starting hypothesis tests based on sampling apply.
  - Mechanically reducing sample size (matching, throwing observations away even at random) reduces precision, so t-stats come down.
- Their alternative is to look at the matched sample:
  - Compare means, relative to standard deviations.
  - QQ plots (differences of quantiles of the the samples).
  - Useful, but does render the decision of whether balance has been achieved as judgmental.

# How to do it: stratification

1. **Sort** treatment and comparison observations based on the estimated propensity score.
2. **Drop** comparison observations whose estimated p($X$) is lower than the min(p($X_i$)) or max(p($X_i$)) (for $i$ treated) (called *common support or overlap*).
3. **Check:** If there is a range where p($X_i$) for treated observations is less than min(p($X_i$)) or more likely greater than max(p($X_i$)), then the non-overlap problem is fatal – we can’t estimate the average treatment effect.
4. **Create strata** (bins, sub-classes) on p($X_i$) such that within each sub-class the distribution of ($X_i$, p($X_i$)) is balanced across the treated and comparison groups.
5. Within each stratum **compute** E($Y_i$| p($X_i$), $T_i$=1) – E($Y_i$| p($X_i$), $T_i$=0), and **weight** these by the number of treated observations in each stratum to estimate the average treatment effect for the treated (corresponds to a piece-wise linear functional form).

# Stratification: Dehejia-Wahba Example

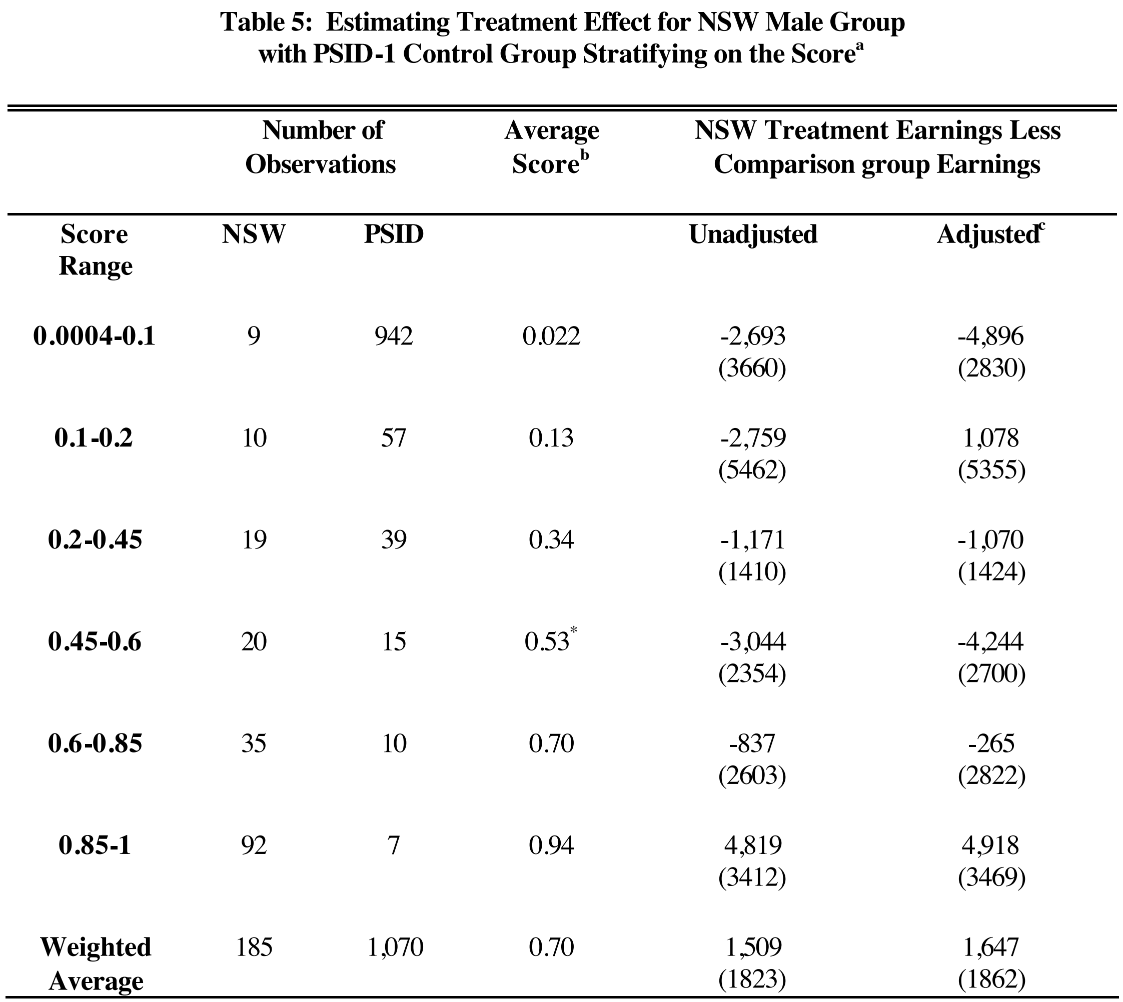


**Table 5: Estimating Treatment Effect for NSW Male Group with PSID-1 Control Group Stratifying on the Score[a]**

| | Number of Observations | | Average Score[b] | NSW Treatment Earnings Less Comparison group Earnings | |
|---|---|---|---|---|---|
| **Score Range** | **NSW** | **PSID** | | **Unadjusted** | **Adjusted[c]** |
| **0.0004-0.1** | 9 | 942 | 0.022 | -2,693<br>(3660) | -4,896<br>(2830) |
| **0.1-0.2** | 10 | 57 | 0.13 | -2,759<br>(5462) | 1,078<br>(5355) |
| **0.2-0.45** | 19 | 39 | 0.34 | -1,171<br>(1410) | -1,070<br>(1424) |
| **0.45-0.6** | 20 | 15 | 0.53* | -3,044<br>(2354) | -4,244<br>(2700) |
| **0.6-0.85** | 35 | 10 | 0.70 | -837<br>(2603) | -265<br>(2822) |
| **0.85-1** | 92 | 7 | 0.94 | 4,819<br>(3412) | 4,918<br>(3469) |
| **Weighted Average** | 185 | 1,070 | 0.70 | 1,509<br>(1823) | 1,647<br>(1862) |

# Stratification: Dehejia-Wahba example

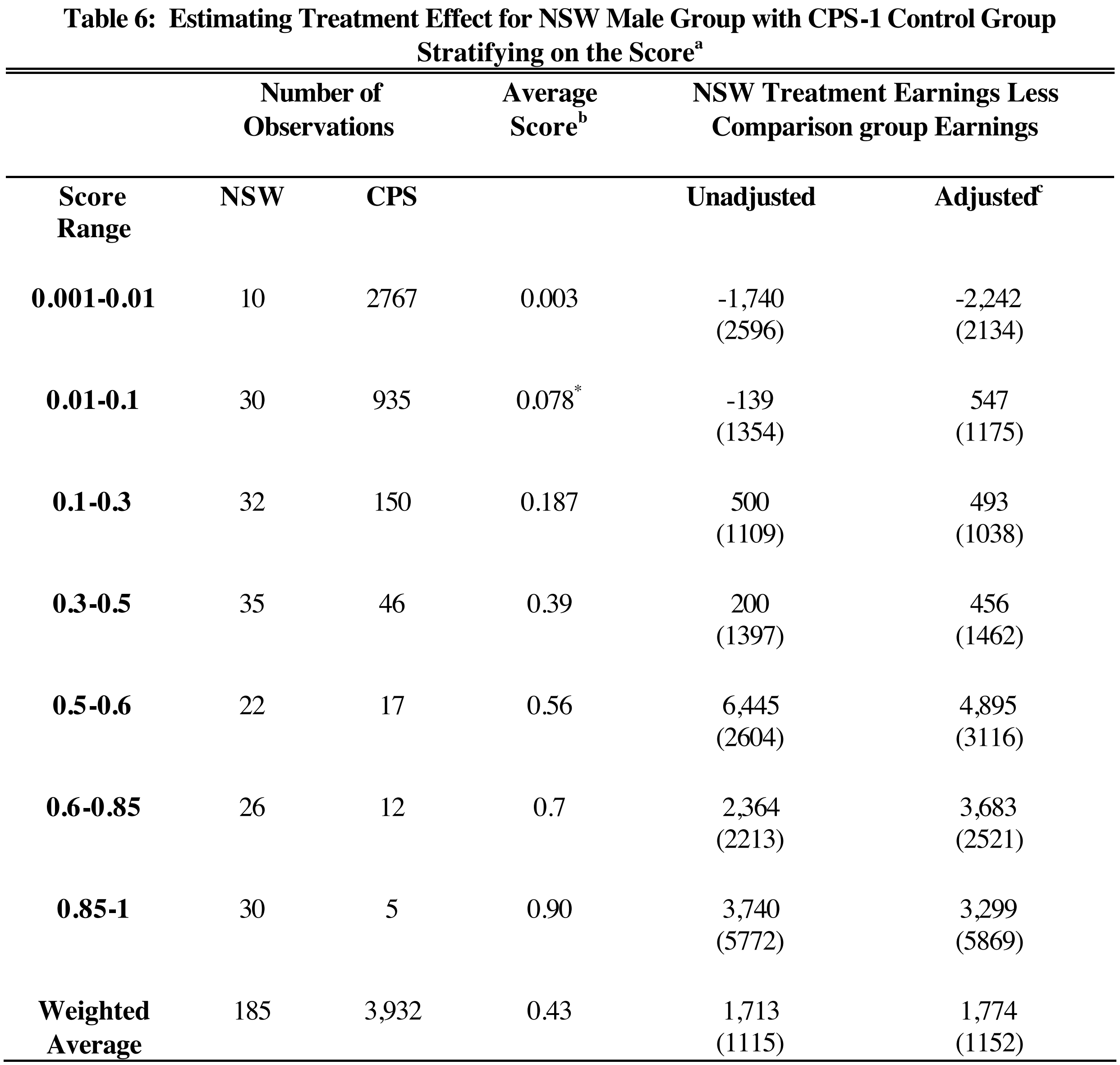


**Table 6: Estimating Treatment Effect for NSW Male Group with CPS-1 Control Group Stratifying on the Score[a]**

| | Number of Observations | | Average Score[b] | NSW Treatment Earnings Less Comparison group Earnings | |
|---|---|---|---|---|---|
| **Score Range** | **NSW** | **CPS** | | **Unadjusted** | **Adjusted[c]** |
| **0.001-0.01** | 10 | 2767 | 0.003 | -1,740<br>(2596) | -2,242<br>(2134) |
| **0.01-0.1** | 30 | 935 | 0.078* | -139<br>(1354) | 547<br>(1175) |
| **0.1-0.3** | 32 | 150 | 0.187 | 500<br>(1109) | 493<br>(1038) |
| **0.3-0.5** | 35 | 46 | 0.39 | 200<br>(1397) | 456<br>(1462) |
| **0.5-0.6** | 22 | 17 | 0.56 | 6,445<br>(2604) | 4,895<br>(3116) |
| **0.6-0.85** | 26 | 12 | 0.7 | 2,364<br>(2213) | 3,683<br>(2521) |
| **0.85-1** | 30 | 5 | 0.90 | 3,740<br>(5772) | 3,299<br>(5869) |
| **Weighted Average** | 185 | 3,932 | 0.43 | 1,713<br>(1115) | 1,774<br>(1152) |

# Stratification: Dehejia-Wahba example

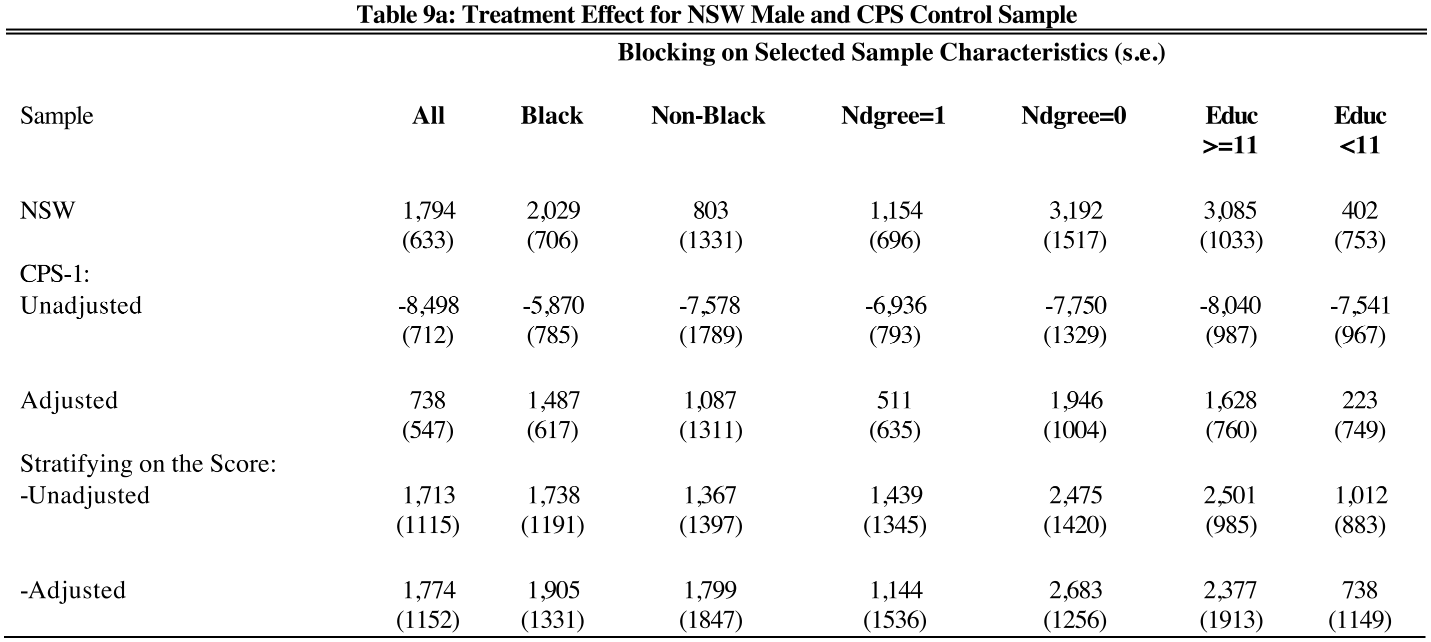


**Table 9a: Treatment Effect for NSW Male and CPS Control Sample**

| | **Blocking on Selected Sample Characteristics (s.e.)** | | | | | | |
|---|---|---|---|---|---|---|---|
| Sample | **All** | **Black** | **Non-Black** | **Ndgree=1** | **Ndgree=0** | **Educ >=11** | **Educ <11** |
| NSW | 1,794<br>(633) | 2,029<br>(706) | 803<br>(1331) | 1,154<br>(696) | 3,192<br>(1517) | 3,085<br>(1033) | 402<br>(753) |
| CPS-1: | | | | | | | |
| Unadjusted | -8,498<br>(712) | -5,870<br>(785) | -7,578<br>(1789) | -6,936<br>(793) | -7,750<br>(1329) | -8,040<br>(987) | -7,541<br>(967) |
| Adjusted | 738<br>(547) | 1,487<br>(617) | 1,087<br>(1311) | 511<br>(635) | 1,946<br>(1004) | 1,628<br>(760) | 223<br>(749) |
| Stratifying on the Score: | | | | | | | |
| -Unadjusted | 1,713<br>(1115) | 1,738<br>(1191) | 1,367<br>(1397) | 1,439<br>(1345) | 2,475<br>(1420) | 2,501<br>(985) | 1,012<br>(883) |
| -Adjusted | 1,774<br>(1152) | 1,905<br>(1331) | 1,799<br>(1847) | 1,144<br>(1536) | 2,683<br>(1256) | 2,377<br>(1913) | 738<br>(1149) |

# Stratification: Dehejia-Wahba example

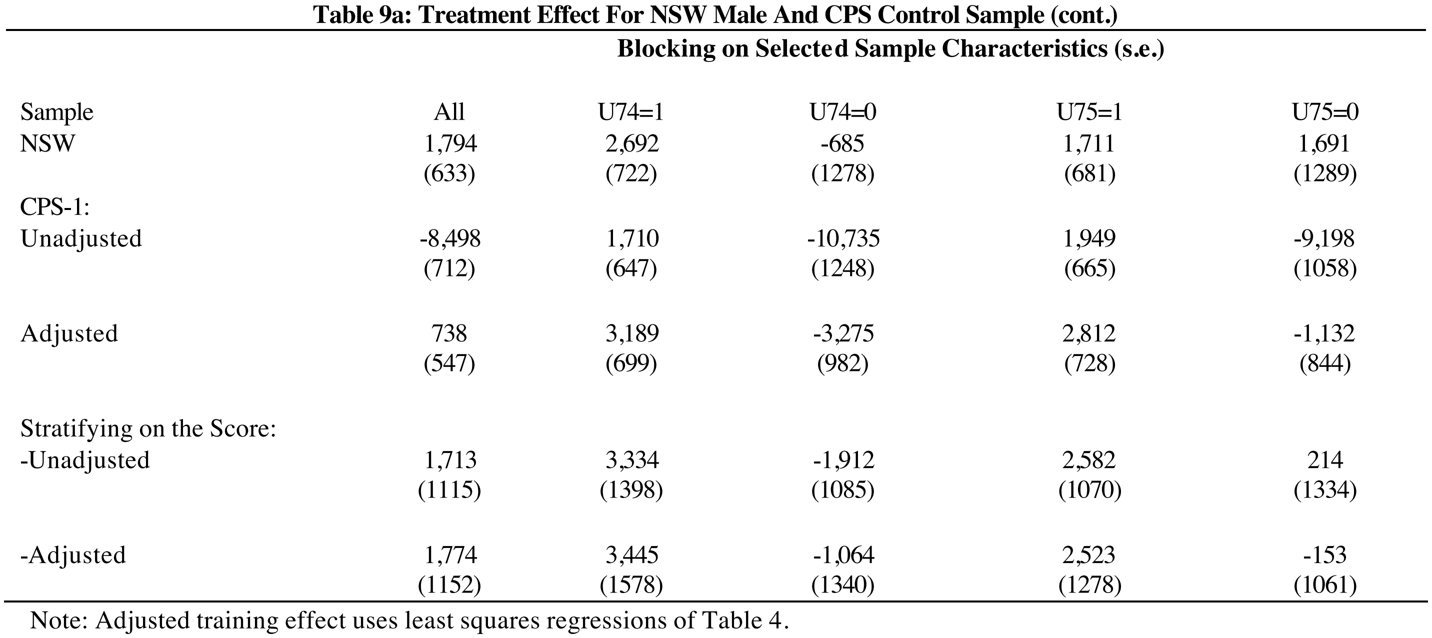


**Table 9a: Treatment Effect For NSW Male And CPS Control Sample (cont.)**

| | **Blocking on Selected Sample Characteristics (s.e.)** | | | | |
|---|---|---|---|---|---|
| Sample | All | U74=1 | U74=0 | U75=1 | U75=0 |
| NSW | 1,794 | 2,692 | -685 | 1,711 | 1,691 |
| | (633) | (722) | (1278) | (681) | (1289) |
| CPS-1: | | | | | |
| Unadjusted | -8,498 | 1,710 | -10,735 | 1,949 | -9,198 |
| | (712) | (647) | (1248) | (665) | (1058) |
| Adjusted | 738 | 3,189 | -3,275 | 2,812 | -1,132 |
| | (547) | (699) | (982) | (728) | (844) |
| Stratifying on the Score: | | | | | |
| -Unadjusted | 1,713 | 3,334 | -1,912 | 2,582 | 214 |
| | (1115) | (1398) | (1085) | (1070) | (1334) |
| -Adjusted | 1,774 | 3,445 | -1,064 | 2,523 | -153 |
| | (1152) | (1578) | (1340) | (1278) | (1061) |

Note: Adjusted training effect uses least squares regressions of Table 4.

# Stratification: Dehejia-Wahba example

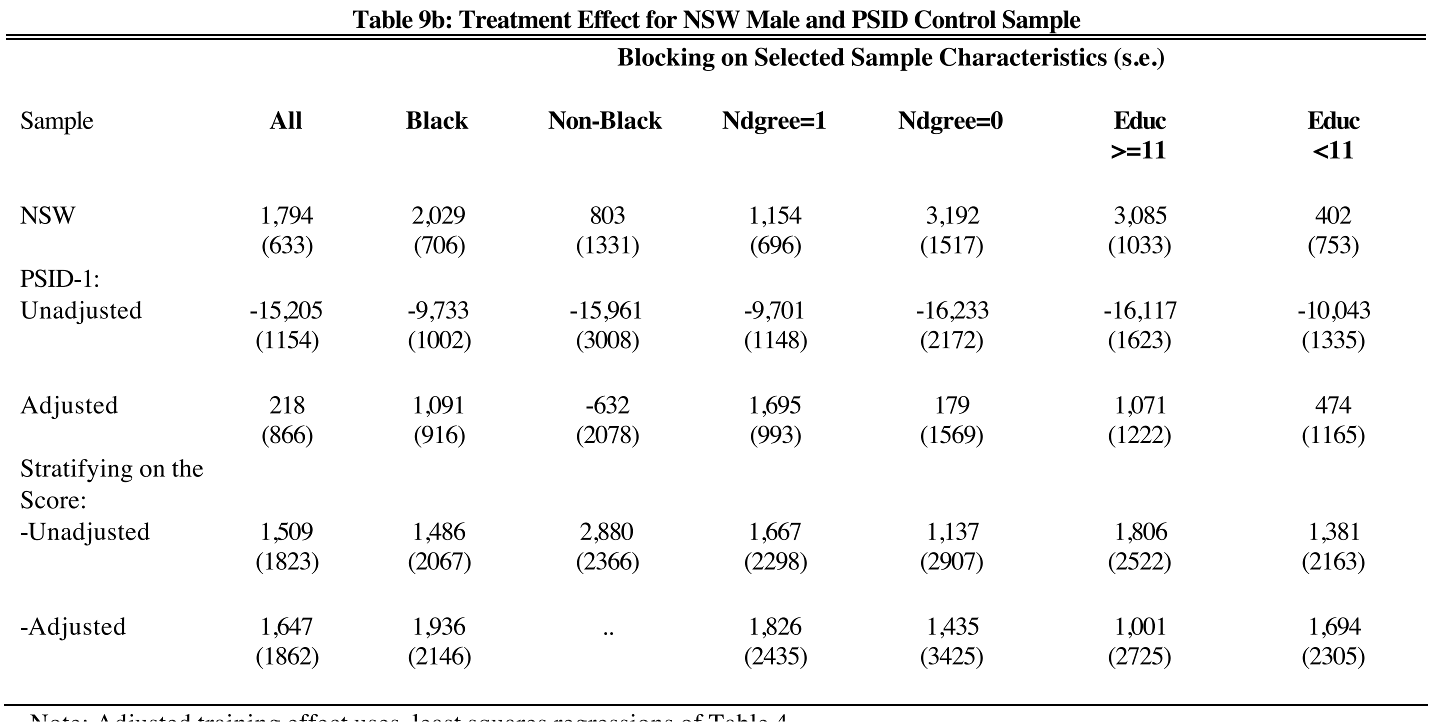


**Table 9b: Treatment Effect for NSW Male and PSID Control Sample**

| | **Blocking on Selected Sample Characteristics (s.e.)** | | | | | | |
|---|---|---|---|---|---|---|---|
| Sample | **All** | **Black** | **Non-Black** | **Ndgree=1** | **Ndgree=0** | **Educ >=11** | **Educ <11** |
| NSW | 1,794<br>(633) | 2,029<br>(706) | 803<br>(1331) | 1,154<br>(696) | 3,192<br>(1517) | 3,085<br>(1033) | 402<br>(753) |
| PSID-1:<br>Unadjusted | -15,205<br>(1154) | -9,733<br>(1002) | -15,961<br>(3008) | -9,701<br>(1148) | -16,233<br>(2172) | -16,117<br>(1623) | -10,043<br>(1335) |
| Adjusted | 218<br>(866) | 1,091<br>(916) | -632<br>(2078) | 1,695<br>(993) | 179<br>(1569) | 1,071<br>(1222) | 474<br>(1165) |
| Stratifying on the Score:<br>-Unadjusted | 1,509<br>(1823) | 1,486<br>(2067) | 2,880<br>(2366) | 1,667<br>(2298) | 1,137<br>(2907) | 1,806<br>(2522) | 1,381<br>(2163) |
| -Adjusted | 1,647<br>(1862) | 1,936<br>(2146) | .. | 1,826<br>(2435) | 1,435<br>(3425) | 1,001<br>(2725) | 1,694<br>(2305) |

Note: Adjusted training effect uses least squares regressions of Table 4

# Stratification: Dehejia-Wahba example

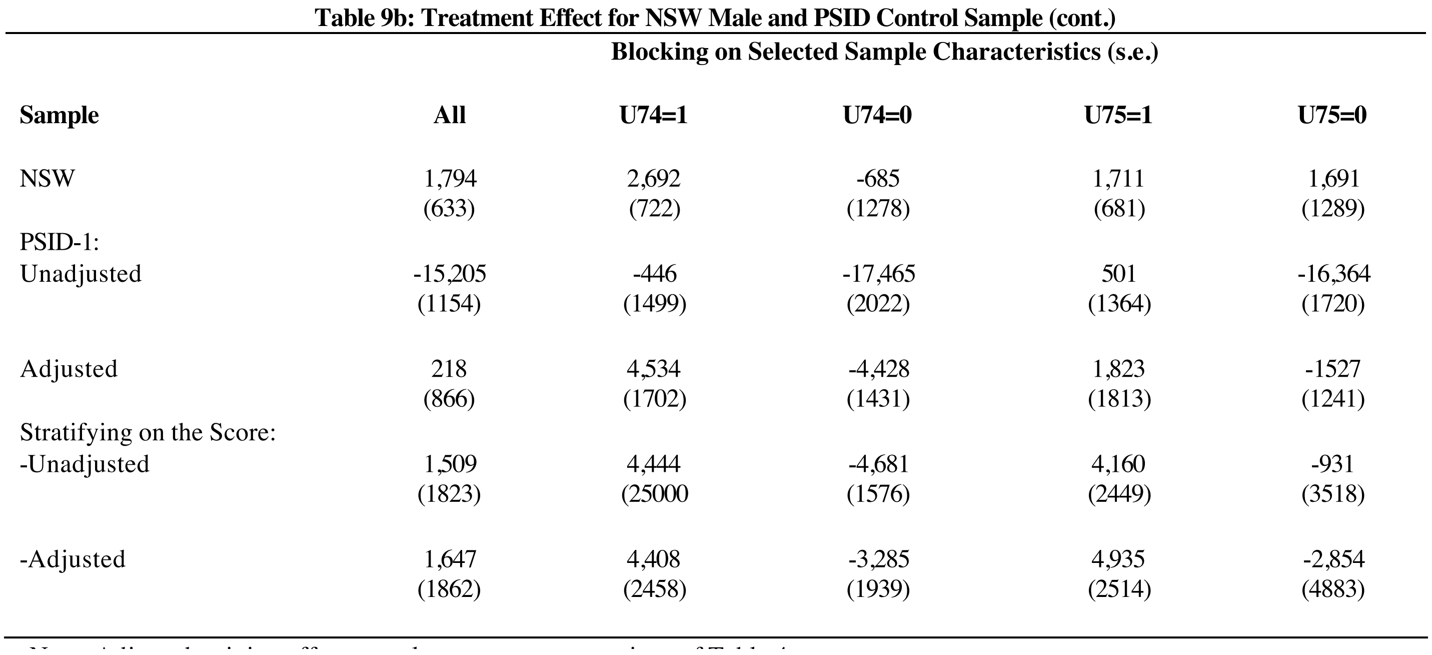


**Table 9b: Treatment Effect for NSW Male and PSID Control Sample (cont.)**

| | **Blocking on Selected Sample Characteristics (s.e.)** | | | | |
|---|---|---|---|---|---|
| **Sample** | **All** | **U74=1** | **U74=0** | **U75=1** | **U75=0** |
| NSW | 1,794<br>(633) | 2,692<br>(722) | -685<br>(1278) | 1,711<br>(681) | 1,691<br>(1289) |
| PSID-1: | | | | | |
| Unadjusted | -15,205<br>(1154) | -446<br>(1499) | -17,465<br>(2022) | 501<br>(1364) | -16,364<br>(1720) |
| Adjusted | 218<br>(866) | 4,534<br>(1702) | -4,428<br>(1431) | 1,823<br>(1813) | -1527<br>(1241) |
| Stratifying on the Score: | | | | | |
| -Unadjusted | 1,509<br>(1823) | 4,444<br>(25000 | -4,681<br>(1576) | 4,160<br>(2449) | -931<br>(3518) |
| -Adjusted | 1,647<br>(1862) | 4,408<br>(2458) | -3,285<br>(1939) | 4,935<br>(2514) | -2,854<br>(4883) |

# How to do it: Propensity score matching

Matching on the propensity score is essentially a weighting scheme, which determines what weights are placed on comparison units when computing the estimated treatment effect:

$$\hat{\tau}\Big|_{T=1} = \frac{1}{|N|}\sum_{i\in N}\left(Y_i - \frac{1}{|J(i)|}\sum_{j\in J(i)} Y_j\right)$$

where *N* is the treatment group, |*N*| the number of units in the treatment group, *J*(*i*) is the set of comparison units matched to treatment unit *i*.

## Issues:

? Whether or not to match with **replacement**

? How many **comparison units** to match to each treated unit

? Which **matching method** to choose

# Propensity score matching: Dehejia-Wahba example

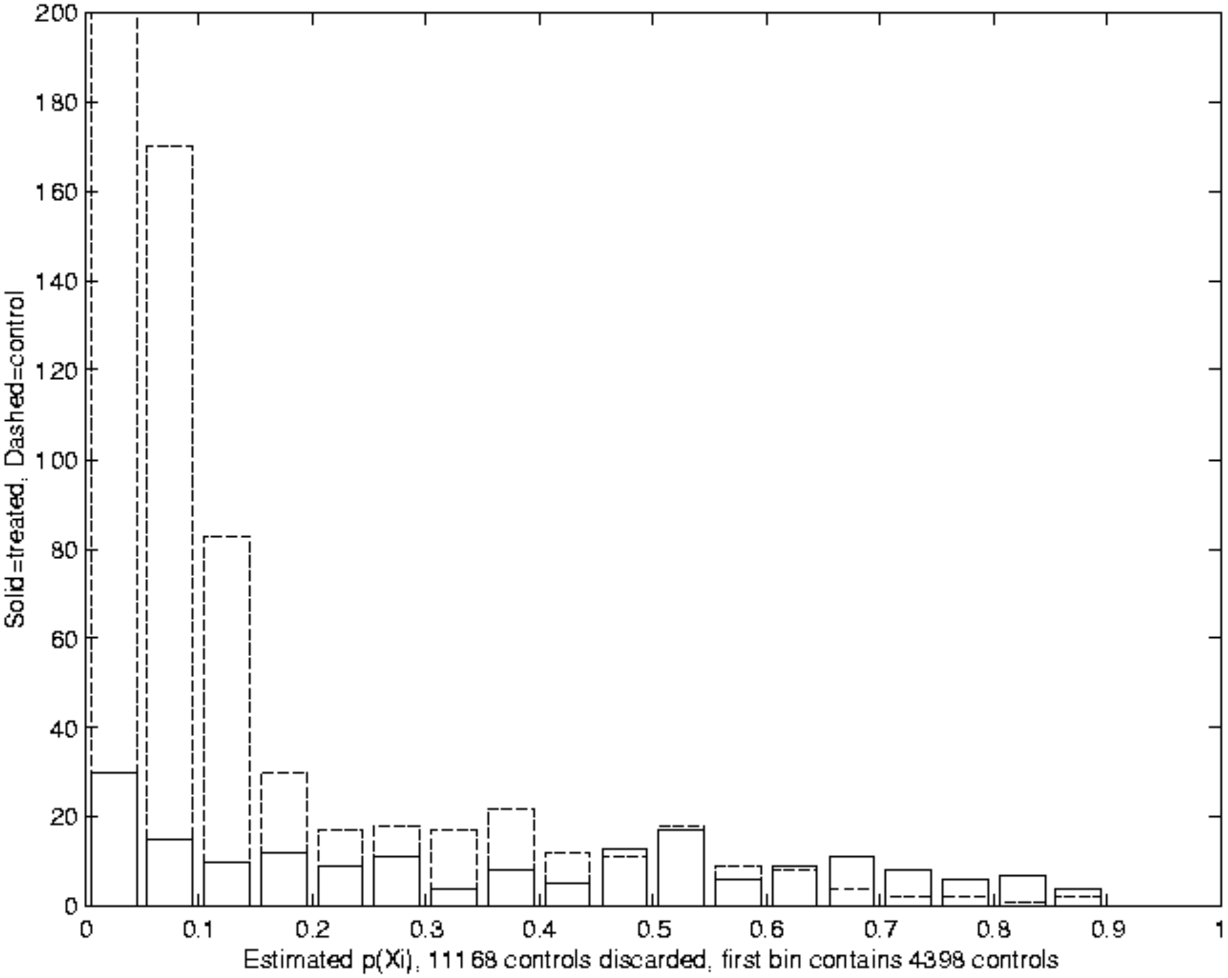


Figure 1: Histogram of Estimated Propensity Score, NSW and CPS

# Propensity score matching: Dehejia-Wahba example

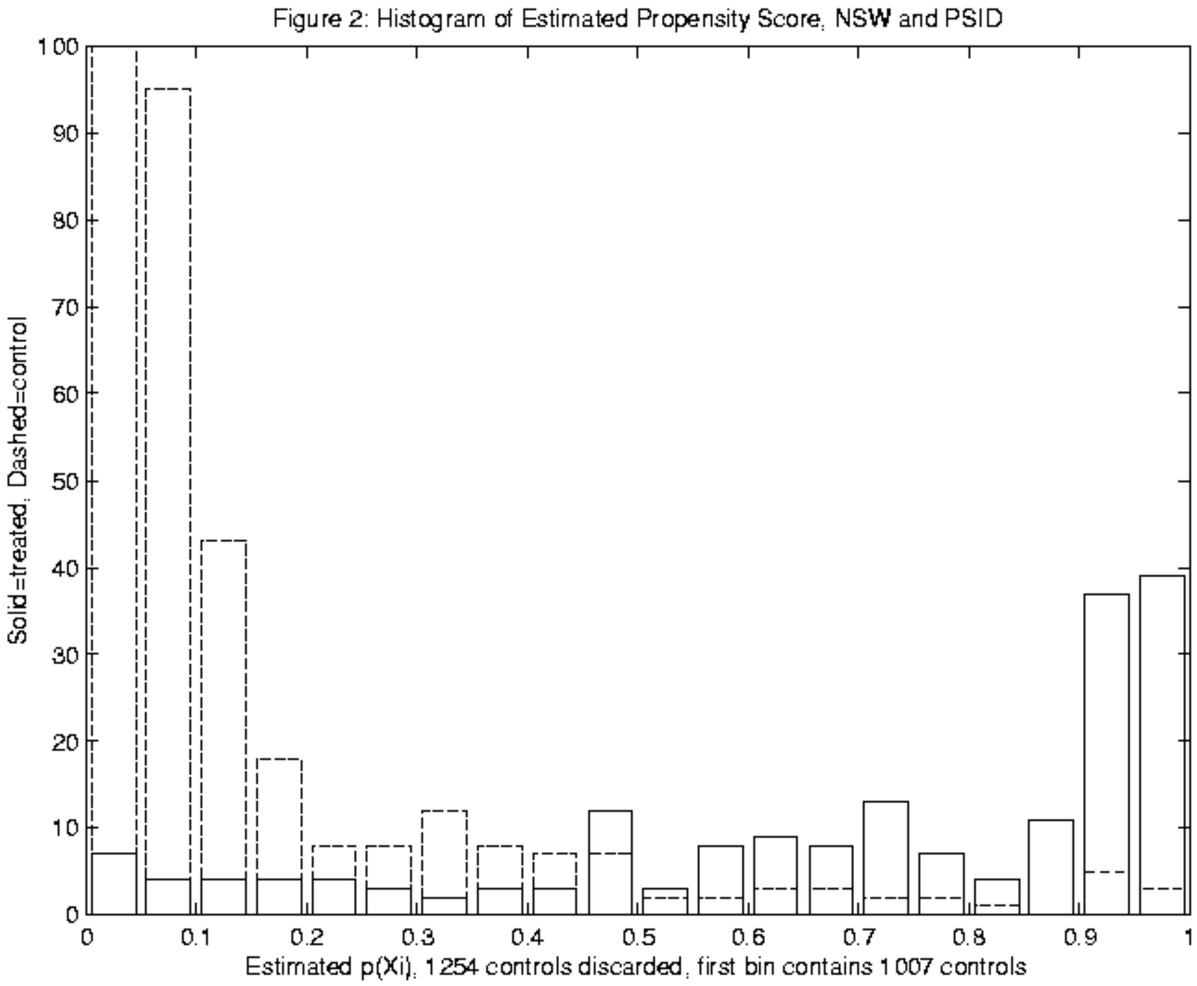


Figure 2: Histogram of Estimated Propensity Score, NSW and PSID

# Propensity score matching: Dehejia-Wahba example

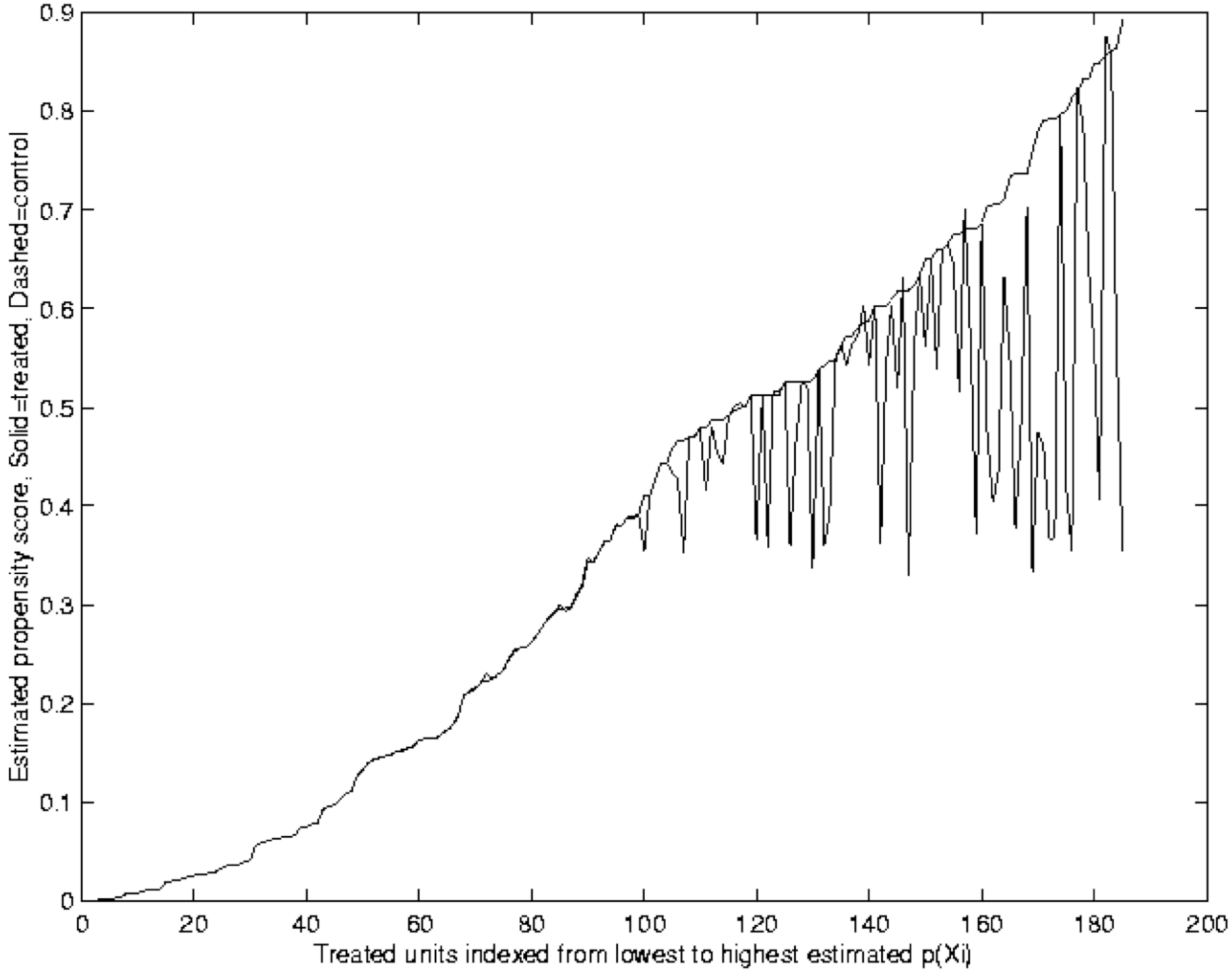


Figure 3: Propensity Score for Treated and Matched Control Units, Random

# Propensity score matching: Dehejia-Wahba example

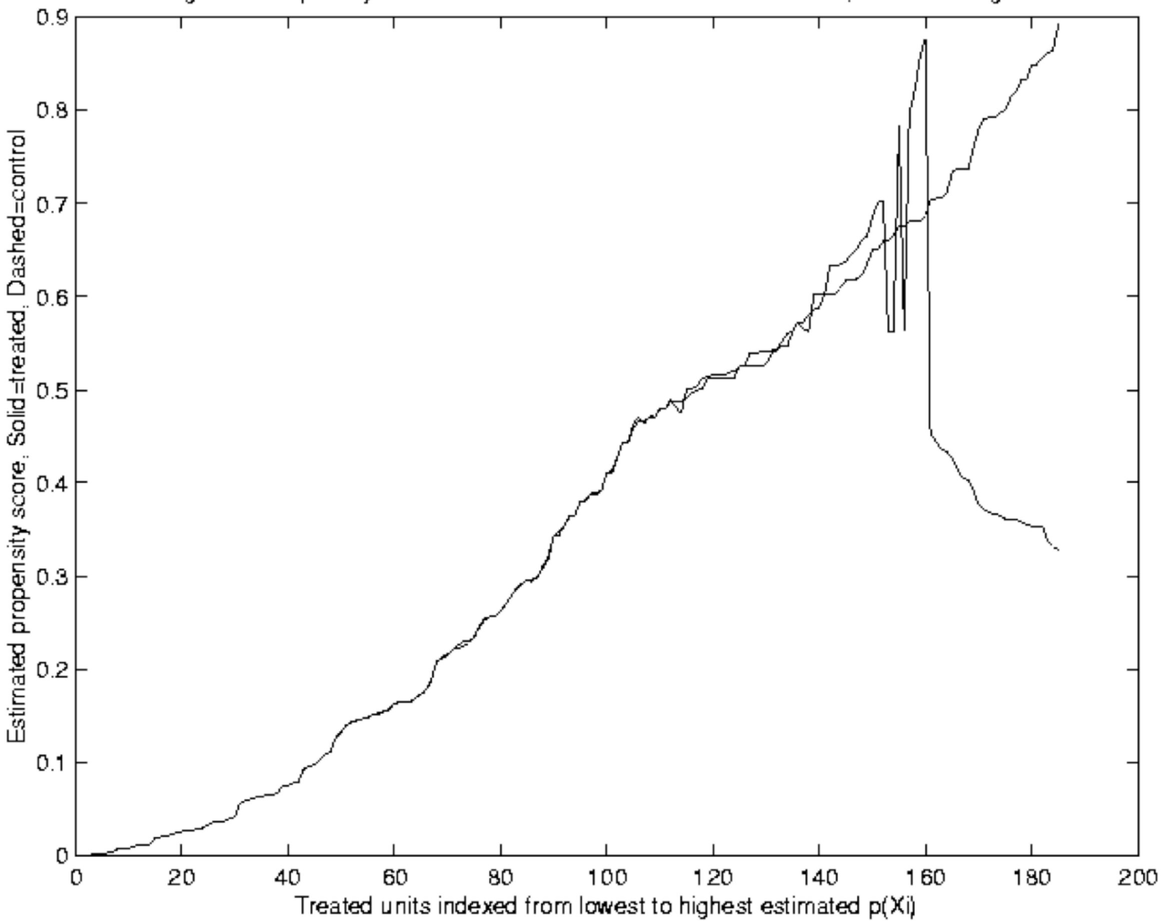
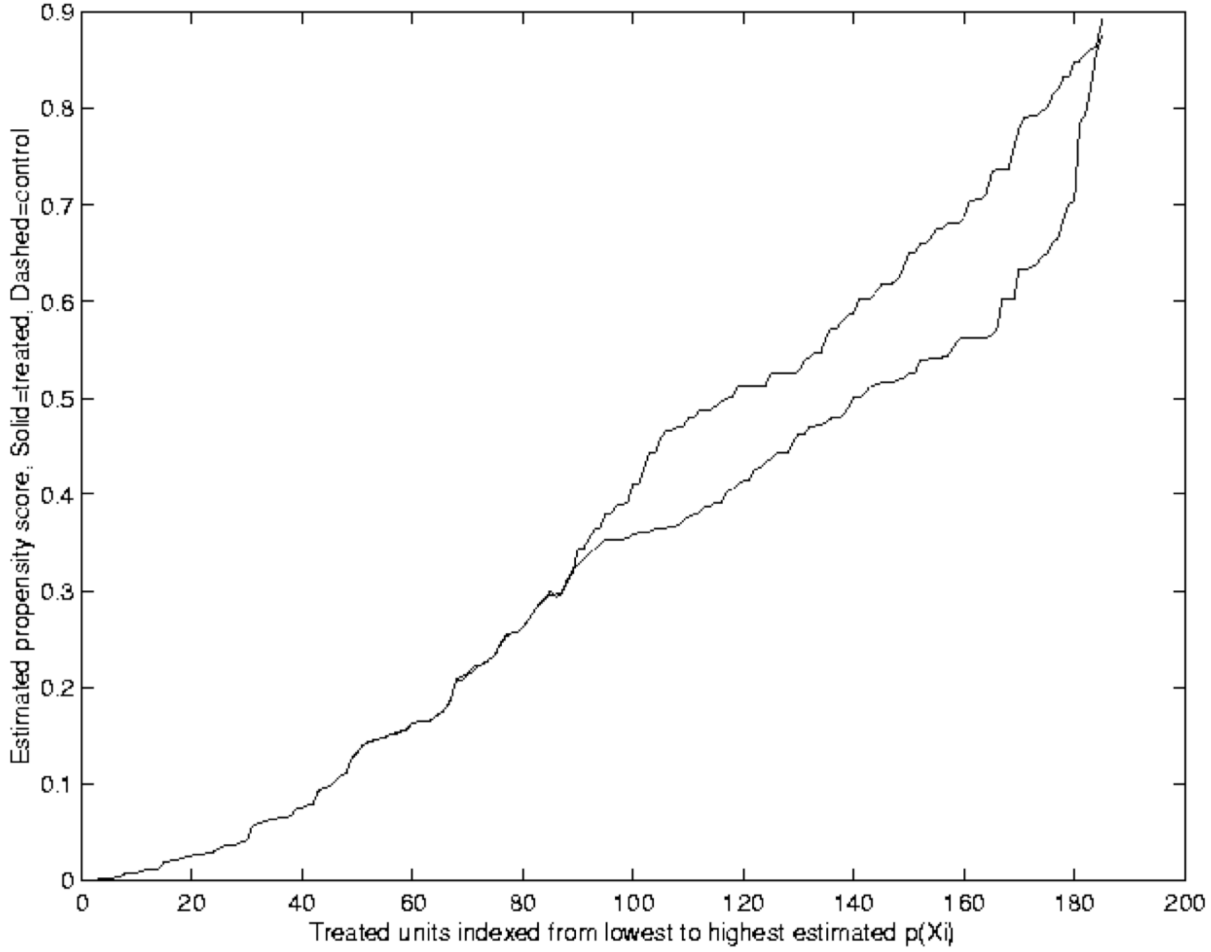


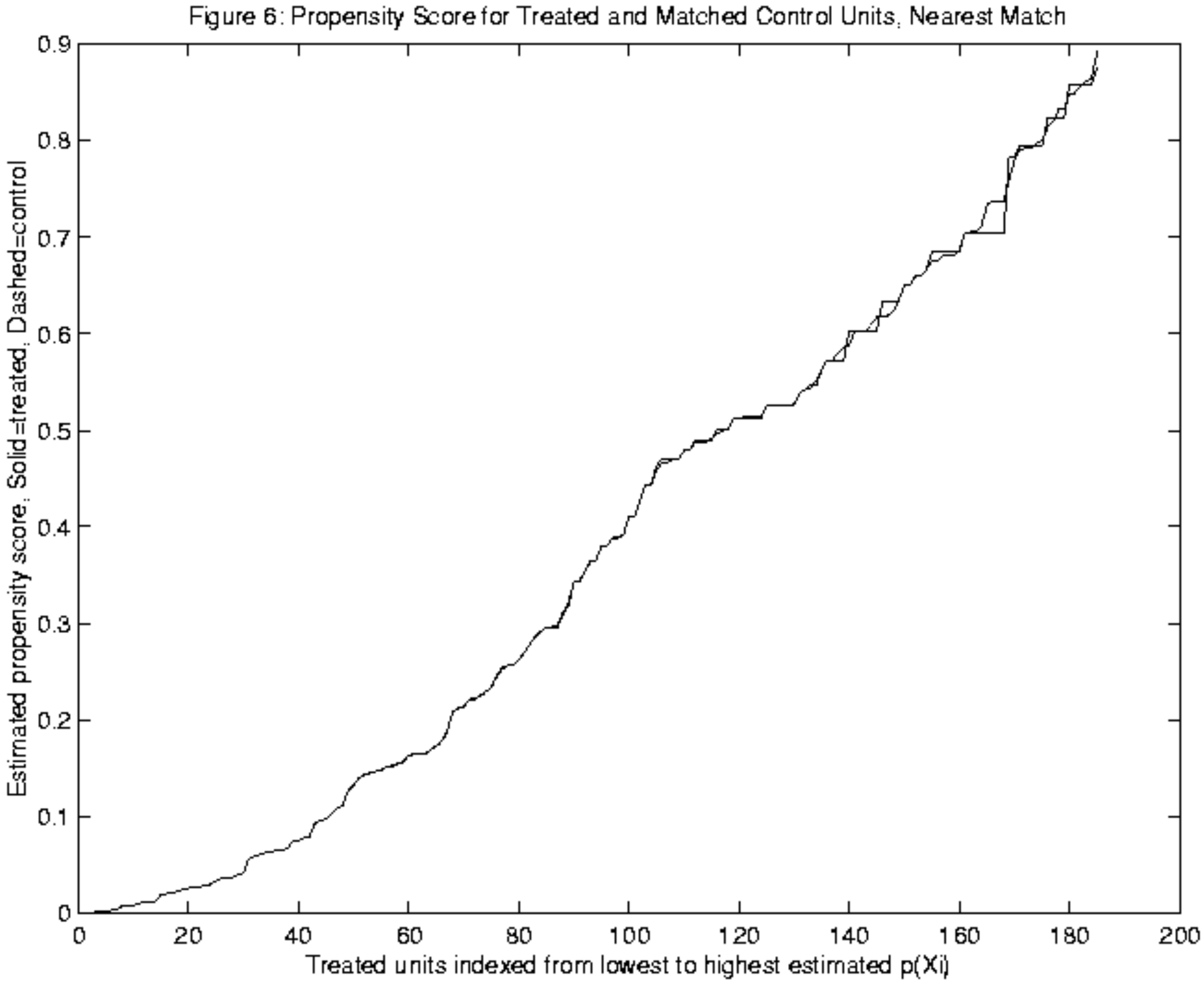
Figure 4: Propensity Score for Treated and Matched Control Units, Lowest to Highest

# Propensity score matching: Dehejia-Wahba example

Figure 5: Propensity Score for Treated and Matched Control Units, Highest to Lowest



# Propensity score matching: Dehejia-Wahba example



# Propensity score matching: Dehejia-Wahba example

Sample Means of Characteristics for Matched Control Samples

| **Matched Samples** | No. of Obs. | Age | School | Black | Hisp | Nodegree | Married | RE74 US$ | RE75 US$ |
|---|---|---|---|---|---|---|---|---|---|
| **NSW** | 185 | 25.81 | 10.35 | 0.84 | 0.06 | 0.71 | 0.19 | 2,096 | 1,532 |
| **MPSID-1** | 56 | 26.39 | 10.62 | 0.86 | 0.02 | 0.55 | 0.15 | 1,794 | 1,126 |
| | | [2.56] | [0.63] | [0.13] | [0.06] | [0.13] | [0.12] | [1406] | [1146] |
| **MPSID-2** | 49 | 25.32 | 11.10 | 0.89 | 0.02 | 0.57 | 0.19 | 1,599 | 2,225 |
| | | [2.63] | [0.83] | [0.14] | [0.08] | [0.16] | [0.16] | [1905] | [1228] |
| **MPSID-3** | 30 | 26.86 | 10.96 | 0.91 | 0.01 | 0.52 | 0.25 | 1,386 | 1,863 |
| | | [2.97] | [0.84] | [0.13] | [0.08] | [0.16] | [0.16] | [1680] | [1494] |
| **MCPS-1** | 119 | 26.91 | 10.52 | 0.86 | 0.04 | 0.64 | 0.19 | 2,110 | 1,396 |
| | | [1.25] | [0.32] | [0.06] | [0.04] | [0.07] | [0.06] | [841] | [563] |
| **MCPS-2** | 87 | 26.21 | 10.21 | 0.85 | 0.04 | 0.68 | 0.20 | 1,758 | 1,204 |
| | | [1.43] | [0.37] | [0.08] | [0.05] | [0.09] | 0.08 | [896] | [661] |
| **MCPS-3** | 63 | 25.94 | 10.69 | 0.87 | 0.06 | 0.53 | 0.13 | 2,709 | 1,587 |
| | | [1.68] | [0.48] | [0.09] | [0.06] | [0.10] | [0.09] | [1285] | [760] |

# Issues with matching

! When matching with replacement, standard non-parametric bootstrap standard errors are not valid

! Options are to use new asymptotic standard errors or to use the *m* of *n* bootstrap or subsampling

! Also note that bootstrap *does* work for matching without replacement

# Weighting on the propensity score

## Proposition

*If $Y_1, Y_0 \perp\!\!\!\perp T \mid X$, then*

$$\alpha_{ATE} = E\left[Y_1 - Y_0\right] = E\left[Y \cdot \frac{T - p(X)}{p(X) \cdot \left(1 - p(X)\right)}\right]$$

$$\alpha_{ATET} = E\left[Y_1 - Y_0 \middle| T = 1\right] = \frac{1}{\Pr\left(T = 1\right)} \cdot E\left[Y \cdot \frac{T - p(X)}{1 - p(X)}\right]$$

## Proof.

$$E\left[Y \frac{T - p(X)}{p(X)\left(1 - p(X)\right)} \middle| X\right] = E\left[\frac{Y}{p(X)} \middle| X, T = 1\right] p(X) + E\left[\frac{-Y}{1 - p(X)} \middle| X, T = 0\right]\left(1 - p(X)\right)$$

$$= E\left[Y \middle| X, T = 1\right] - E\left[Y \middle| X, T = 0\right]$$

*The results follow from integration over P(X) and P(X|T = 1)*

# Weighting on the propensity score
## *Why this works pt. 1*

**Population:** 50 men and 50 women.

**Outcome of interest:** All men have earnings of $10 and all women have earnings of $12

**Sampling:** For some reason you randomly sample men at probability 1/10 (from the male sample or 1/20 from the full sample) and women at probability 1/2 (from the female sample or ¼ from the full sample).

**Sample:** Your sample consists of 5 men (earnings 10) and 25 women (earnings 12).

**Population mean *$\mu$*:** You know the true population mean is ½ 10 + ½ 12 = 11.

**Sample mean $\bar{x}$:** A naïve (wrong) mean would be (10 x 5 + 25 x 12) / 30 = 11.83, because women are overrepresented in the sample.

**Adjustment:** How can you recover the true mean from the sample? Weight by the probability of sampling:

$$(5*10/(1/10) + 25*12 /(1/2))/(5/10+25/2) =11$$

# Weighting on the propensity score

## *Why this works pt. 2*

More generally, consider a population of *N* units, made up of *k* types (e.g., covariate values) of $n_\mathrm{k}$ units each. Sample of $k$=1…$K$ units each of category with probality $p_k$=p($n_k$)=1/$n_k$.

Expectation of the sample mean as the sample gets bigger:

$$\frac{E\left(\Sigma\frac{x_i}{p_i}\right)}{\left(\Sigma\frac{1}{p_i}\right)} \rightarrow \frac{E(\sum n_k x_k)}{(\sum n_k)} \rightarrow \frac{E(\sum n_k x_k)}{N}$$

**What this means:** weighting each observation by its probability of selection adjusts its weight so it is in proportion to the overall popuation.

# Weighting on the propensity score

## *Why this works pt. 3*

$$\alpha_{ATET} = E\left[Y_1 - Y_0 \middle| T = 1\right] = \frac{1}{\Pr\left(T = 1\right)} \cdot E\left[Y \cdot \frac{T - p\left(X\right)}{1 - p\left(X\right)}\right]$$

If T=1, weight = $\frac{1}{\Pr\left(T = 1\right)}$.

If T=0, weight = $\frac{1}{\Pr\left(T = 1\right)} \frac{p\left(X\right)}{1 - p\left(X\right)} = \frac{1}{\Pr\left(T = 1\right)} \frac{1}{(1 - p\left(X\right)) / p\left(X\right)}$

For the treated group (T=1), each observation gets equal weight and all observations together are reweighted by probability of being in the treated group.

For the comparison group (T=0), each observation is weighted (up) by the probability of being in the comparison group relative to the treatment group – and again the whole is reweighted to match the probability of being the treated group.

# Weighting on the propensity score

## *Why this works pt. 4*

$$\alpha_{ATET} = E\left[Y_1 - Y_0 \middle| T=1\right] = \frac{1}{\Pr\left(T=1\right)} \cdot E\left[Y \cdot \frac{T-p\left(X\right)}{1-p\left(X\right)}\right]$$

$$= \frac{1}{\Pr\left(T=1\right)} \cdot \Pr(T=1) E\left[Y \cdot \frac{T-p\left(X\right)}{1-p\left(X\right)} | T=1\right] + \frac{1}{\Pr\left(T=1\right)} \cdot (1-\Pr(T=1)) E\left[Y \cdot \frac{T-p\left(X\right)}{1-p\left(X\right)} | T=0\right]$$

$$= E\left[Y_1 | T=1\right] - \frac{1}{\Pr\left(T=1\right)} \cdot (1-\Pr(T=1)) E\left[Y \cdot \frac{p\left(X\right)}{1-p\left(X\right)} | T=0\right]$$

$$= E\left[Y_1 | T=1\right] - \frac{1}{\Pr\left(T=1\right)} \cdot (1-\Pr(T=1)) E\left[E_X\left[Y_0 \frac{p\left(X\right)}{1-p\left(X\right)} | X, T=0\right] | T=0\right]$$

$$= E\left[Y_1 | T=1\right] - \frac{1}{\Pr\left(T=1\right)} \cdot (1-\Pr(T=1)) E\left[\frac{p\left(T=1\right)}{1-p\left(T=1\right)} E_X\left[Y_0 | X, T=1\right] | T=0\right]$$

$$= E\left[Y_1 | T=1\right] - \frac{1}{\Pr\left(T=1\right)} \cdot (1-\Pr(T=1)) \frac{p\left(T=1\right)}{1-p\left(T=1\right)} E\left[Y_0 | X, T=1\right]$$

$$= E\left[Y_1 | T=1\right] - E\left[Y_0 | T=1\right]$$

**What's the intuition**?

- ✓ **For *$Y_1$*** (potential outcomes for treated group) it's a simple reweighting argument
- ✓ **For *$Y_0$*** (potential outcomes for comparison group) first you reweight to match the full comparison population, and then reweight the comparison to match the treated population.

# Weighting on the propensity score

$$\alpha_{ATE} = E\left[Y_1 - Y_0\right] = E\left[Y \cdot \frac{T - p(X)}{p(X)\cdot\left(1 - p(X)\right)}\right]$$

$$\alpha_{ATET} = E\left[Y_1 - Y_0 \middle| T = 1\right] = \frac{1}{\Pr(T = 1)} \cdot E\left[Y \cdot \frac{T - p(X)}{1 - p(X)}\right]$$

The analogy principle suggests a two-step estimator:

1. Estimate the propensity score: $\hat{p}(X)$
2. Use estimated score to produce analogous estimators:

$$\hat{\alpha}_{ATE} = \frac{1}{N}\sum_{i=1}^{N} Y_i \cdot \frac{T_i - \hat{p}(X_i)}{\hat{p}(X_i)\cdot\left(1 - \hat{p}(X_i)\right)}$$

$$\hat{\alpha}_{ATET} = \frac{1}{N_1}\sum_{i=1}^{N} Y_i \cdot \frac{T_i - \hat{p}(X_i)}{\left(1 - \hat{p}(X_i)\right)}$$

# Weighting on the propensity score

$$\hat{\alpha}_{ATE}=\frac{1}{N}\sum_{i=1}^{N}Y_i\cdot\frac{T_i-\hat{p}\left(X_i\right)}{\hat{p}\left(X_i\right)\cdot\left(1-\hat{p}\left(X_i\right)\right)}$$

$$\hat{\alpha}_{ATET}=\frac{1}{N_1}\sum_{i=1}^{N}Y_i\cdot\frac{T_i-\hat{p}\left(X_i\right)}{\left(1-\hat{p}\left(X_i\right)\right)}$$

Standard errors:

✓ We need to adjust the s.e.'s for first-step estimation of *p*(*X*)

✓ Parametric first-step: Newey & McFadden (1994)

✓ Non-parametric first-step: Newey (1994)

✓ Or bootstrap the entire two-step procedure!

# Weighting on the propensity score

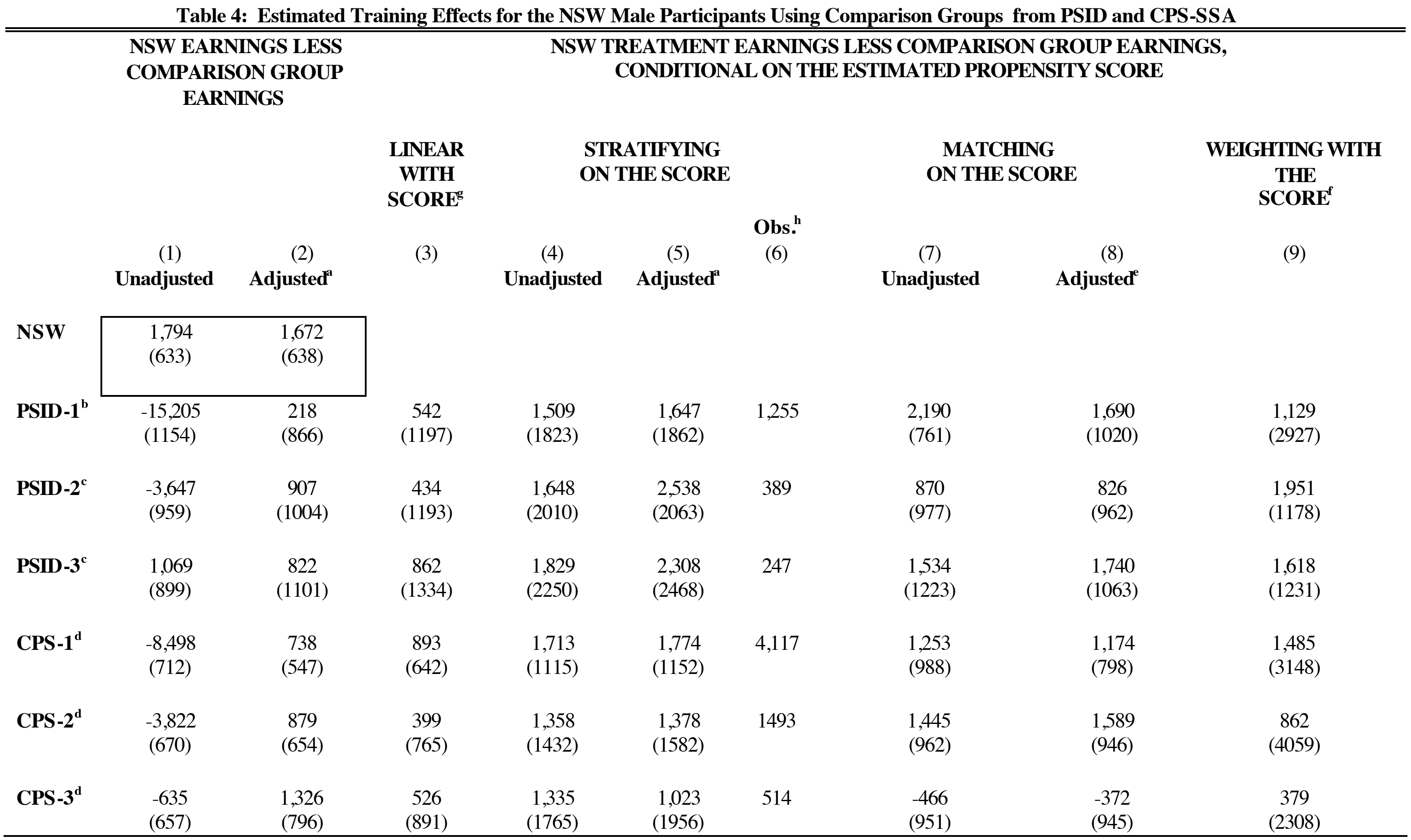


Table 4: Estimated Training Effects for the NSW Male Participants Using Comparison Groups from PSID and CPS-SSA

| | NSW EARNINGS LESS COMPARISON GROUP EARNINGS | | NSW TREATMENT EARNINGS LESS COMPARISON GROUP EARNINGS, CONDITIONAL ON THE ESTIMATED PROPENSITY SCORE | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | LINEAR WITH SCORE[g] | STRATIFYING ON THE SCORE | | Obs.[h] | MATCHING ON THE SCORE | | WEIGHTING WITH THE SCORE[f] |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| | Unadjusted | Adjusted[a] | | Unadjusted | Adjusted[a] | | Unadjusted | Adjusted[e] | |
| **NSW** | 1,794<br>(633) | 1,672<br>(638) | | | | | | | |
| **PSID-1**[b] | -15,205<br>(1154) | 218<br>(866) | 542<br>(1197) | 1,509<br>(1823) | 1,647<br>(1862) | 1,255 | 2,190<br>(761) | 1,690<br>(1020) | 1,129<br>(2927) |
| **PSID-2**[c] | -3,647<br>(959) | 907<br>(1004) | 434<br>(1193) | 1,648<br>(2010) | 2,538<br>(2063) | 389 | 870<br>(977) | 826<br>(962) | 1,951<br>(1178) |
| **PSID-3**[c] | 1,069<br>(899) | 822<br>(1101) | 862<br>(1334) | 1,829<br>(2250) | 2,308<br>(2468) | 247 | 1,534<br>(1223) | 1,740<br>(1063) | 1,618<br>(1231) |
| **CPS-1**[d] | -8,498<br>(712) | 738<br>(547) | 893<br>(642) | 1,713<br>(1115) | 1,774<br>(1152) | 4,117 | 1,253<br>(988) | 1,174<br>(798) | 1,485<br>(3148) |
| **CPS-2**[d] | -3,822<br>(670) | 879<br>(654) | 399<br>(765) | 1,358<br>(1432) | 1,378<br>(1582) | 1493 | 1,445<br>(962) | 1,589<br>(946) | 862<br>(4059) |
| **CPS-3**[d] | -635<br>(657) | 1,326<br>(796) | 526<br>(891) | 1,335<br>(1765) | 1,023<br>(1956) | 514 | -466<br>(951) | -372<br>(945) | 379<br>(2308) |

# Weighting on the propensity score

- Traditional concern is that misspecification of estimated propensity score can lead to extreme weights.
- Buso, DiNardo, and McCrary argue that weighting performs well in small samples.
- Argue for normalizing the weight (standard option in most regression packages) as a way to deal with this.
- Can also be applied to diff-in-diffs estimators (Abadie 2005).

# Check sensitivity to *p*(*X*) specification

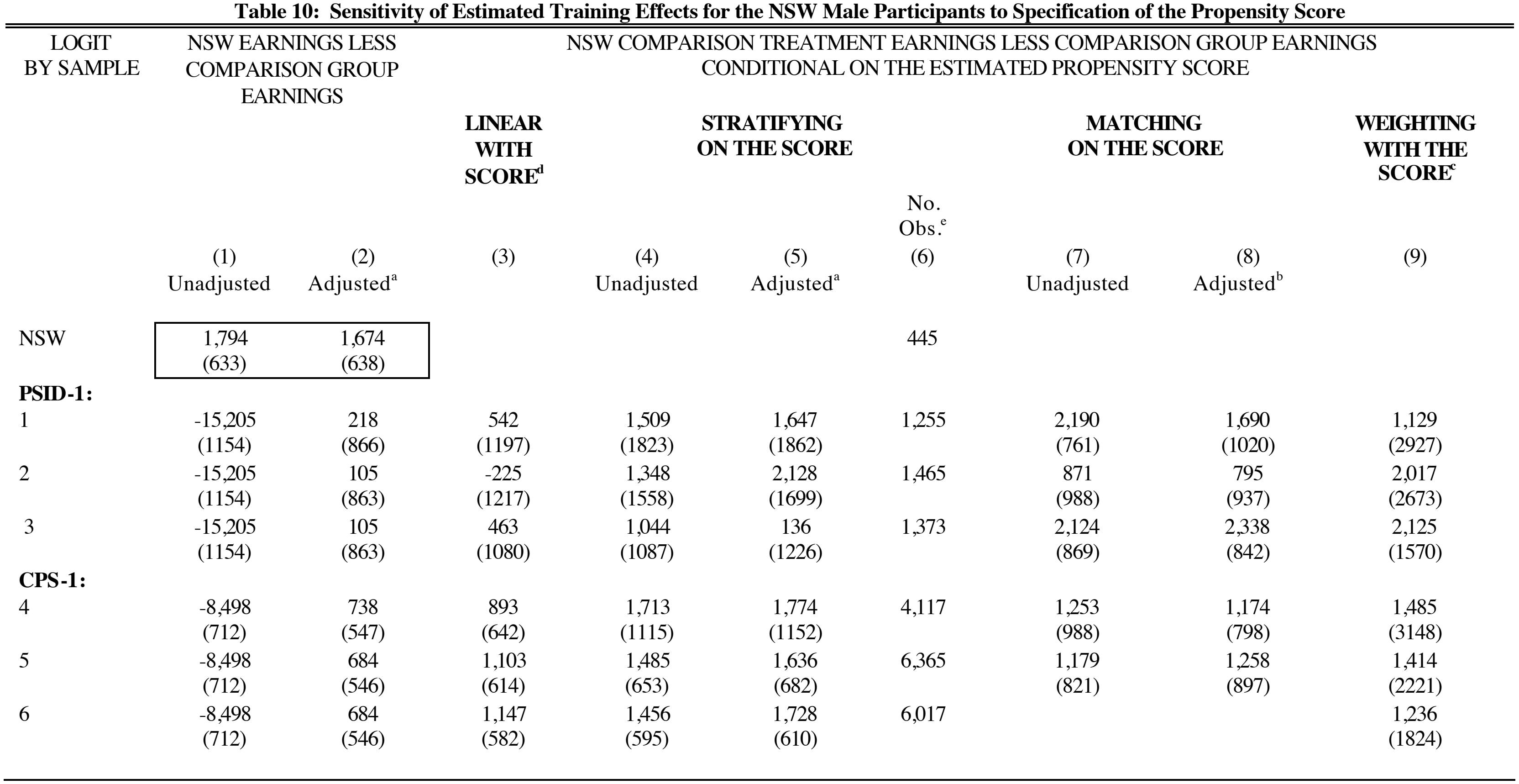


**Table 10: Sensitivity of Estimated Training Effects for the NSW Male Participants to Specification of the Propensity Score**

| LOGIT BY SAMPLE | NSW EARNINGS LESS COMPARISON GROUP EARNINGS | | NSW COMPARISON TREATMENT EARNINGS LESS COMPARISON GROUP EARNINGS CONDITIONAL ON THE ESTIMATED PROPENSITY SCORE | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | **LINEAR WITH SCORE[d]** | **STRATIFYING ON THE SCORE** | | | **MATCHING ON THE SCORE** | | **WEIGHTING WITH THE SCORE[c]** |
| | (1)<br>Unadjusted | (2)<br>Adjusted[a] | (3) | (4)<br>Unadjusted | (5)<br>Adjusted[a] | No. Obs.[e]<br>(6) | (7)<br>Unadjusted | (8)<br>Adjusted[b] | (9) |
| NSW | 1,794<br>(633) | 1,674<br>(638) | | | | 445 | | | |
| **PSID-1:** | | | | | | | | | |
| 1 | -15,205<br>(1154) | 218<br>(866) | 542<br>(1197) | 1,509<br>(1823) | 1,647<br>(1862) | 1,255 | 2,190<br>(761) | 1,690<br>(1020) | 1,129<br>(2927) |
| 2 | -15,205<br>(1154) | 105<br>(863) | -225<br>(1217) | 1,348<br>(1558) | 2,128<br>(1699) | 1,465 | 871<br>(988) | 795<br>(937) | 2,017<br>(2673) |
| 3 | -15,205<br>(1154) | 105<br>(863) | 463<br>(1080) | 1,044<br>(1087) | 136<br>(1226) | 1,373 | 2,124<br>(869) | 2,338<br>(842) | 2,125<br>(1570) |
| **CPS-1:** | | | | | | | | | |
| 4 | -8,498<br>(712) | 738<br>(547) | 893<br>(642) | 1,713<br>(1115) | 1,774<br>(1152) | 4,117 | 1,253<br>(988) | 1,174<br>(798) | 1,485<br>(3148) |
| 5 | -8,498<br>(712) | 684<br>(546) | 1,103<br>(614) | 1,485<br>(653) | 1,636<br>(682) | 6,365 | 1,179<br>(821) | 1,258<br>(897) | 1,414<br>(2221) |
| 6 | -8,498<br>(712) | 684<br>(546) | 1,147<br>(582) | 1,456<br>(595) | 1,728<br>(610) | 6,017 | | | 1,236<br>(1824) |

# Check sensitive to including RE74

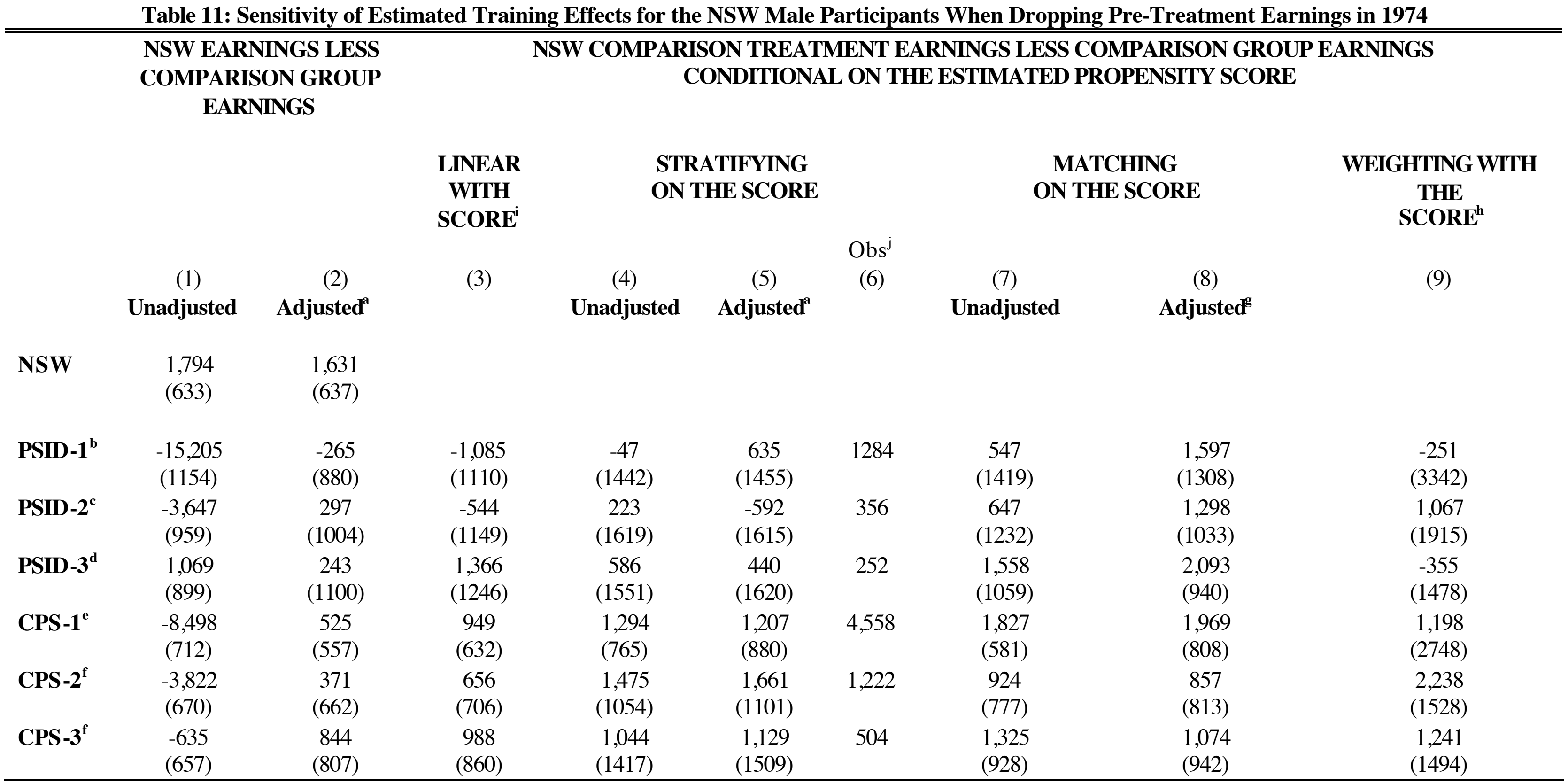


Table 11: Sensitivity of Estimated Training Effects for the NSW Male Participants When Dropping Pre-Treatment Earnings in 1974

| | NSW EARNINGS LESS COMPARISON GROUP EARNINGS | | NSW COMPARISON TREATMENT EARNINGS LESS COMPARISON GROUP EARNINGS CONDITIONAL ON THE ESTIMATED PROPENSITY SCORE | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | LINEAR WITH SCORE[i] | STRATIFYING ON THE SCORE | | | MATCHING ON THE SCORE | | WEIGHTING WITH THE SCORE[h] |
| | | | | | | Obs[j] | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| | Unadjusted | Adjusted[a] | | Unadjusted | Adjusted[a] | | Unadjusted | Adjusted[g] | |
| **NSW** | 1,794 | 1,631 | | | | | | | |
| | (633) | (637) | | | | | | | |
| **PSID-1[b]** | -15,205 | -265 | -1,085 | -47 | 635 | 1284 | 547 | 1,597 | -251 |
| | (1154) | (880) | (1110) | (1442) | (1455) | | (1419) | (1308) | (3342) |
| **PSID-2[c]** | -3,647 | 297 | -544 | 223 | -592 | 356 | 647 | 1,298 | 1,067 |
| | (959) | (1004) | (1149) | (1619) | (1615) | | (1232) | (1033) | (1915) |
| **PSID-3[d]** | 1,069 | 243 | 1,366 | 586 | 440 | 252 | 1,558 | 2,093 | -355 |
| | (899) | (1100) | (1246) | (1551) | (1620) | | (1059) | (940) | (1478) |
| **CPS-1[e]** | -8,498 | 525 | 949 | 1,294 | 1,207 | 4,558 | 1,827 | 1,969 | 1,198 |
| | (712) | (557) | (632) | (765) | (880) | | (581) | (808) | (2748) |
| **CPS-2[f]** | -3,822 | 371 | 656 | 1,475 | 1,661 | 1,222 | 924 | 857 | 2,238 |
| | (670) | (662) | (706) | (1054) | (1101) | | (777) | (813) | (1528) |
| **CPS-3[f]** | -635 | 844 | 988 | 1,044 | 1,129 | 504 | 1,325 | 1,074 | 1,241 |
| | (657) | (807) | (860) | (1417) | (1509) | | (928) | (942) | (1494) |

# Regression estimators

Can we use regression estimators?

1. Least squares as an approximation
2. Least squares as a weighting scheme
3. Estimators of average treatment effects based on nonparametric regression

# 1. Least squares as an approximation

$(Y_1,Y_0) \perp\!\!\!\perp T|X$ implies that the conditional expectation $E[Y|T, X]$ can be interpreted as a conditional causal response function:

$$E[Y|T = 1, X] = E[Y_1|T = 1, \mathrm{X}] = \mathrm{E}[Y_1|X],$$

$$E[Y|T = 0, X] = E[Y_0|T = 0, \mathrm{X}] = \mathrm{E}[Y_0|X]$$

So $E[Y|T, X]$ provides average potential responses with and without the treatment.

The functional form of $E[Y|T, X]$ is typically unknown, but Ordinary Least Squares provides a well-defined approximation.

# 2. Least squares as a weighting scheme

Suppose that the covariates take on a finite number of values *K*. Then, from the subclassification section we know that:

$$\alpha_{ATE} = \sum_{k=1}^{K} \left( E\left[Y \middle| T = 1, X = x^{k}\right] - E\left[Y \middle| T = 0, X = x^{k}\right] \right) \Pr\left(X = x^{k}\right)$$

Now, we suppose that you run a regression that is **saturated** in *X*: a regression including one dummy variable, $Z^k$, for each possible value of *X*:

$$Y = \alpha T + \sum_{k=1}^{K} Z^{k} \beta_{k} + u,$$

where

$$Z^{k} = \begin{cases} 1 & \text{if } X = x^{k} \\ 0 & \text{if } X \neq x^{k} \end{cases}$$

# 2. Least squares as a weighting scheme

It can be shown that the coefficient $\hat{\alpha}_{\text{OLS}}$ of the saturated regression converges to:

$$\alpha_{\text{OLS}} = \sum_{k=1}^{K} \left( E\left[ Y \middle| T = 1, X = x^k \right] - E\left[ Y \middle| T = 0, X = x^k \right] \right) w_k$$

where

$$w_k = \frac{\text{var}\left(T \middle| X = x^k\right)\Pr\left(X = x^k\right)}{\sum_{k=1}^{K} \text{var}\left(T \middle| X = x^k\right)\Pr\left(X = x^k\right)}.$$

- Strata *k* with a higher

$$\text{var}(T|X = x^k) = \Pr(T|X = X^k)(1\text{-}\Pr(T|X = X^k))$$

(i.e. propensity scores close to 0.5) receive higher weight. Strata with propensity score close to 0 or 1 receive lower weight.

- OLS down-weights strata where the average causal effects are less precisely estimated.

# 3. Estimators based on nonparametric regression

$$\alpha_{ATE} = \int \left( E\left[Y \middle| X, T=1\right] - E\left[Y \middle| X, T=0\right] \right) dp\left(X\right)$$

$$\alpha_{ATET} = \int \left( E\left[Y \middle| X, T=1\right] - E\left[Y \middle| X, T=0\right] \right) dp\left(X \middle| T=1\right)$$

This suggests the following estimators:

$$\hat{\alpha}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} \left( \hat{\mu}_1\left(X_i\right) - \hat{\mu}_0\left(X_i\right) \right)$$

$$\hat{\alpha}_{ATET} = \frac{1}{N_1} \sum_{Ti=1}^{N} \left( \hat{\mu}_1\left(X_i\right) - \hat{\mu}_0\left(X_i\right) \right),$$

where $\hat{\mu}_1(\;)$ and $\hat{\mu}_0(\;)$ are nonparametric estimators of $E[Y|X, T = 1]$ and $E[Y|X, T = 0]$, respectively.

But estimating these regressions nonparametrically is difficult if the dimension of $X$ is large (curse of dimensionality again!).

# 3. Estimators based on nonparametric regression

It is enough, however, to adjust for the propensity score. So we can proceed in two steps:

1. Estimate the propensity score, $\hat{p}(X)$.
2. Calculate:

$$\hat{\alpha}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{V}_1\left(\hat{p}\left(X_i\right)\right) - \hat{V}_0\left(\hat{p}\left(X_i\right)\right)\right)$$

$$\hat{\alpha}_{ATET} = \frac{1}{N_1}\sum_{i=1}^{N}\left(\hat{V}_1\left(\hat{p}\left(X_i\right)\right) - \hat{V}_0\left(\hat{p}\left(X_i\right)\right)\right),$$

where $\hat{v}_1(\ )$ and $\hat{v}_0(\ )$ are nonparametric estimators of $E[Y|p(X), T = 1]$ and $E[Y|p(X), T = 0]$, respectively.

# Issues: what about efficiency?

**Question:** How can matching be efficient, if we are throwing away information?

**Answer:** It's not.

**Solution:**

1. **Kernel matching** (Heckman, Ichimura, and Todd): Match each treatment unit to all the comparison unit using a kernel weight, which uses a smoother set of weights on the comparison units (comparison units that are very far away get very low, rather than no, weight).
2. **Weighting by the propensity score**: Hirano, Imbens, and Ridder prove that weighting by the estimated propensity score is efficient. (Problem: efficiency pre-supposes a non-parametric series estimator of the propensity score. How is that any easier than a non-parametric series estimate of the outcome equation?)

# Assessing unconfoundedness: multiple control groups

Suppose we have a three-valued indicator $T_i \in \{0, -1, 1\}$ for the groups (e.g., ineligibles, eligible nonparticipants and participants), with the treatment indicator equal to

$W_i$ = 1 $\{T_i$ = 1$\}$ so that

$$Y_i = \begin{cases} Y_i(0) & \text{if } T_i \in \{-1, 0\} \\ Y_i(1) & \text{if } T_i = 1. \end{cases}$$

Suppose we extend the unconfoundedness assumption to independence of the potential outcomes and the three-valued group indicator given covariates

$$Y_i(0),\ Y_i(1) \perp\!\!\!\perp T_i \,|\, X_i$$

# Assessing unconfoundedness: multiple control groups

Now a testable implication is

$$Y_i(0) \perp\!\!\!\perp 1\{T_i = 0\} \,\Big|\, X_i, T_i \in \{-1,0\}.$$

and thus

$$Y_i \perp\!\!\!\perp 1\{T_i = 0\} \,\Big|\, X_i, T_i \in \{-1,0\}.$$

An implication of this independence condition is being tested by the tests discussed above – should be no difference at Y in across the two non-treatment groups. Whether this test has much bearing on the unconfoundedness assumption depends on whether the extension of the assumption is plausible given unconfoundedness itself.

# Basic Procedure

✓ **Define “closeness”:** the distance measure that will be used to determine whether one individual is a good match for another. (e.g. propensity score ranges, Mahalanobis distance, exact matching…)

✓ Implement a **matching method** (e.g. nearest neighbor, with or without replacement…)

✓ **Assess the quality of the matching**: typically looking at balance on covariates (e.g. graph density of propensity scores by treatment, control, matched, unmatched; test for statistically significant differences in mean variance between groups)

✓ **Analyze the data**: run your regression (reweight to get ATE or ATET; you should still control for covariates within subclassifications!)

# Conclusions

(1) Propensity score methods succeed in reducing the dimensionality problem of dealing with observables.

(2) Practically speaking, when we have a large comparison group, and we are looking for the best matches to the treatment group, this method helps us to zoom in on the right subset of comparison units.

(3) By reducing dimensionality, it becomes very easy to allow for a flexible functional form.

(4) Economists are guilty of using functional forms that are “excessively” linear and expecting thee functional forms to successfully extrapolate between very different groups.

(5) This method only works if we observe all (enough) of the variables that determine the selection process.

(6) In training programs, earnings seem to be crucial.

# Current Research and Discussion

- For health policy folks: Brown DW, DeSantis SM, Greene TJ, et al. A novel approach for propensity score matching and stratification for multiple treatments: Application to an electronic health record-derived study. Stat Med. 2020
- For international monetary policy folks: "The effectiveness of macroprudential policies and capital controls against volatile capital inflows" from the Bank for International Settlements (June 2020)
- For machine learning and/or labor econ folks: Goller, Daniel & Lechner, Michael & Moczall, Andreas & Wolff, Joachim, 2019. "Does the Estimation of the Propensity Score by Machine Learning Improve Matching Estimation? The Case of Germany's Programmes for Long Term Unemployed,"  Labour Economics. August 2020

# Additional Resources

- #EconTwitter! Prof. Nick Huntington-Klein has another animated graphic for matching
- Some accessible info on conducting PSM in R (if of interest, not needed for this course)
- Very high-level overview of the PSM concept from the Urban Institute
- Prof. Elizabeth Stuart's (Johns Hopkins School of Public Health) presentation on PSM for the 2011 Society of Prevention Research conference (workshop slides)

# During Recitation…

- ✓ Discuss main skills and questions related to PS4
- ✓ Discussion of progress on replication exercise
- ✓ Chat about how propensity score matching is evolving and is being used in applied research in industry (example from energy regulation)
- ✓ Review of readings if time allows