Conditional Independence
- Matching and Regression

A&P, p. 64: Good controls are variables that we can think of as having been fixed at the time the regressor of interest was determined.

Definition of ATE and ATOT conditional on X

Let X denote a (vector of) observable explanatory variables

$$\begin{aligned} ATE(X) &= E[Y_1 - Y_0 \mid X] \\ &= E[Y_1 \mid X] - E[Y_0 \mid X] \end{aligned}$$

$$\begin{aligned} ATOT(X) &= E[Y_1 - Y_0 \mid D = 1, X] \\ &= E[Y_1 \mid D = 1, X] - E[Y_0 \mid D = 1, X] \end{aligned}$$

Treatment effects for individuals with characteristics X.

Conditional and unconditional ATE and ATOT

Law of iterated expectations:
$E_X[E(Y|X)] = E(Y)$

Average Treatment Effect

$$ATE = E(Y_1) - E(Y_0) = E_X[ATE(X)]$$

$$= \int ATE(x) \cdot f_X(x)\, dx$$

$$= \int \{E[Y_1 \mid X = x] - E[Y_0 \mid X = x]\} \cdot f_x(x)\, dx$$

Average Treatment Effect on the Treated

$$ATOT = E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1)$$

$$= E_{X|D=1}[ATOT(X) \mid D = 1] = \int ATOT(x) \cdot f_{X|D=1}(x)\, dx$$

$$= \int \{E[Y_1 \mid D = 1, X = x] - E[Y_0 \mid D = 1, X = x]\} \cdot f_{X|D=1}(x)\, dx$$

Law of iterated expectations:
$E_X[E(Y|X)] = E(Y)$

ATE(x), ATOT(x) and Regression Notation

$$Y_1 = E[Y_1 \mid X] + U_1 = \mu_1(X) + U_1$$
$$Y_0 = E[Y_0 \mid X] + U_0 = \mu_0(X) + U_0$$

$$\begin{aligned} \mathrm{ATE}(X) &= \mathrm{E}[Y_1 - Y_0 \mid X] = \mathrm{E}[Y_1 \mid X] - \mathrm{E}[Y_0 \mid X] \\ &= \mu_1(X) - \mu_0(X) \end{aligned}$$

$$\begin{aligned} \mathrm{ATOT}(X) &= \mathrm{E}[Y_1 - Y_0 \mid D = 1, X] = \mathrm{E}[Y_1 \mid D = 1, X] - \mathrm{E}[Y_0 \mid D = 1, X] \\ &= \mu_1(X) - \mu_0(X) + \mathrm{E}[U_1 - U_0 \mid D = 1, X] \\ &= \mu_1(X) - \mu_0(X) + \mathrm{E}[U_1 \mid D = 1, X] - \mathrm{E}[U_0 \mid D = 1, X] \end{aligned}$$

ATE(x), ATOT(x) and Regression Notation

Special Case: Linear relationships

$$Y_1 = E[Y_1 \mid X] + U_1 = \mu_1(X) + U_1 = \alpha_1 + \boldsymbol{\beta}_1' \mathbf{X} + U_1$$

$$Y_0 = E[Y_0 \mid X] + U_0 = \mu_0(X) + U_0 = \alpha_0 + \boldsymbol{\beta}_0' \mathbf{X} + U_0$$

$$\mathrm{ATE}(X) = \mu_1(X) - \mu_0(X) = (\alpha_1 - \alpha_0) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)' \mathbf{X}$$

$$\begin{aligned}\mathrm{ATOT}(X) &= \mu_1(X) - \mu_0(X) + E[U_1 - U_0 \mid D = 1, X] \\ &= \mathrm{ATE}(X) + E[U_1 - U_0 \mid D = 1, X] \\ &= (\alpha_1 - \alpha_0) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)' \mathbf{X} + E[U_1 - U_0 \mid D = 1, X]\end{aligned}$$

ATE(x), ATOT(x) and Regression Notation

As a single equation:

$$Y = Y_1 \cdot D + Y_0 \cdot (1 - D)$$
$$= \mu_0(X) + [\mu_1(X) - \mu_0(X)] \cdot D + U_0 + (U_1 - U_0) \cdot D$$

Special case: Linear relationships

$$Y = \mu_0(X) + [\mu_1(X) - \mu_0(X)] \cdot D + U_0 + (U_1 - U_0) \cdot D$$
$$= \alpha_0 + \boldsymbol{\beta}_0' \mathbf{X} + [(\alpha_1 - \alpha_0) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)' \mathbf{X}] \cdot D + U_0 + (U_1 - U_0) \cdot D$$

The Conditional Independence Assumption

Conditional Independence Assumption (CIA):

(Y_1, Y_0) are indepedent from D, **conditional on X**

Formally: $(Y_1, Y_0) \perp D \mid X$

This implies **Mean Independence**:

$$E[Y_1 \mid D = 1, X] = E[Y_1 \mid D = 0, X] = E[Y_1 \mid X]$$

$$E[Y_0 \mid D = 0, X] = E[Y_0 \mid D = 1, X] = E[Y_0 \mid X]$$

ATE and ATOT under conditional mean independence

Average Treatment Effect

$$\begin{aligned} ATE &= E(Y_1) - E(Y_0) \\ &= E_X[E(Y_1 \mid X) - E(Y_0 \mid X)] \\ &= E_X[E(Y_1 \mid D=1, X) - E(Y_0 \mid D=0, X)] \end{aligned}$$

Law of iterated expectations (holds in general)

requires mean independence

Average Treatment Effect on the Treated

$$\begin{aligned} ATOT &= E(Y_1 \mid D=1) - E(Y_0 \mid D=1) \\ &= E_{X\mid D=1}[E(Y_1 \mid D=1, X) - E(Y_0 \mid D=1, X)] \\ &= E_{X\mid D=1}[E(Y_1 \mid D=1, X) - E(Y_0 \mid D=0, X)] \\ &= E(Y_1 \mid D=1) - E_{X\mid D=1}[E(Y_0 \mid D=0, X)] \end{aligned}$$

ATE and ATOT under conditional mean independence

$$E[Y_1 \mid D=1, X] = E[Y_1 \mid D=0, X] = E[Y_1 \mid X]$$

$$E[Y_0 \mid D=0, X] = E[Y_0 \mid D=1, X] = E[Y_0 \mid X]$$

This implies

$$ATE(X) = ATOT(X)$$

$$E[Y_1 \mid X] - E[Y_0 \mid X] = E[Y_1 \mid D=1, X] - E[Y_0 \mid D=1, X]$$

However, in general ATE ≠ ATOT. Why?

$$\int ATE(x) \cdot f_X(x)\, dx \neq \int ATOT(x) \cdot f_{X \mid D=1}(x)\, dx$$

Distributions of X in population and in D=1 group may differ

For ATOT, a weaker version of CIA is sufficient

$$ATOT = E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1)$$

The 1st term can simply be estimated by $\overline{Y}_{1|1} = \frac{1}{n_1} \sum_{\forall i \text{ with } D=1} Y_{1i}$

Using the law of iterated expectations for the 2nd term:

$$E(Y_0 \mid D = 1) = E_{X|D=1}[E(Y_0 \mid D = 1, X)]$$

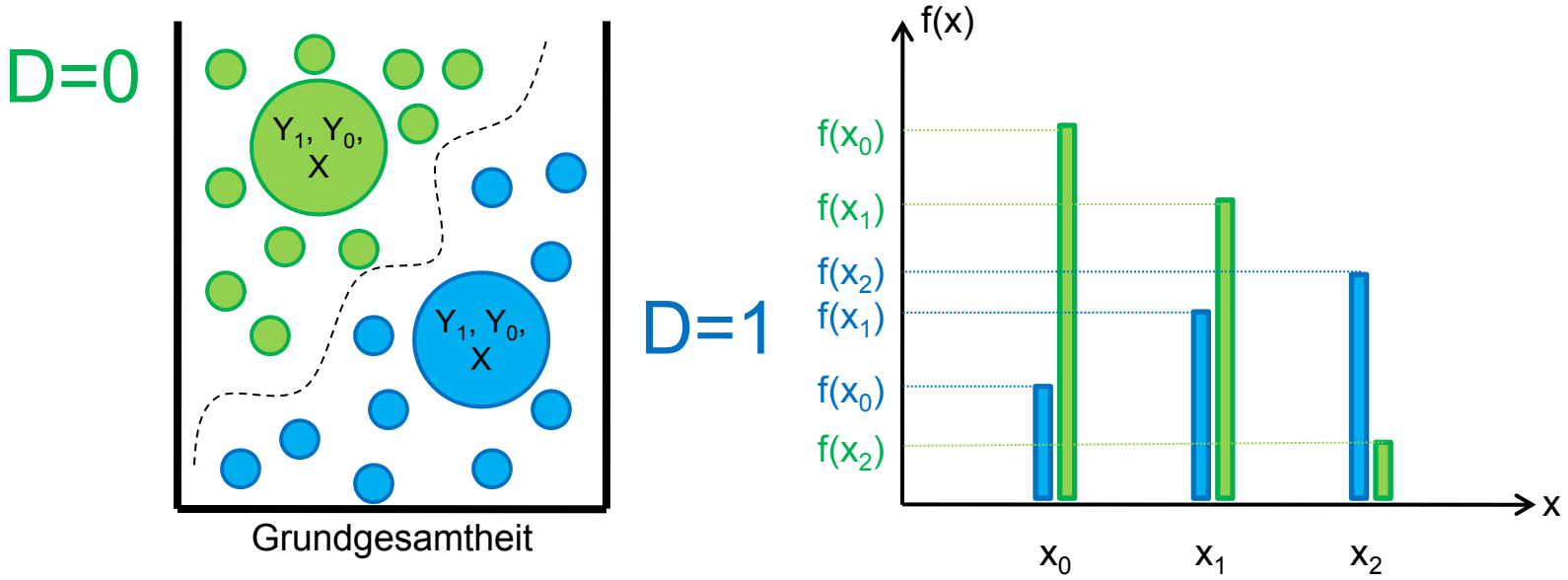
Hence, for ATOT **mean independence** for $\mathbf{Y_0}$ is sufficient

$$E(Y_0 \mid D = 1, X) = E(Y_0 \mid D = 0, X)$$

Implying:

$$ATOT = E(Y_1 \mid D = 1) - E_{X|D=1}[E(Y_0 \mid D = 0, X)]$$

Visualization of LIE and ATOT



$$ATOT = E(Y_1 \mid D=1) - \sum_{x_0,x_1,x_2} E(Y_0 \mid D=0,x) \cdot f(x \mid D=1)$$

$$= E(Y_1 \mid D=1) - \left[\begin{array}{l} E(Y_0 \mid D=0,x_0) \cdot f(x_0 \mid D=1) + \\ E(Y_0 \mid D=0,x_1) \cdot f(x_1 \mid D=1) + \\ E(Y_0 \mid D=0,x_2) \cdot f(x_2 \mid D=1) \end{array} \right]$$

Which X? (I)

CIA: $(Y_1, Y_0) \perp D \mid X = x; \quad \forall x \in \chi$

This will be satisfied, if X contains **all** variables, **influencing** potential outcomes (Y_1, Y_0) and selection (D) into treatment.

The variables in X are `**pre-treatment**', i.e. they may not be affected by receiving (or not receiving) the treatment.

$$ATOT = E(Y_1 \mid D = 1) - E_{X|D=1}\left[E(Y_0 \mid D = 0, X)\right]$$

$$E_{X|D=1}\left[E(Y_0 \mid D = 0, X)\right] = \int E\left[Y_0 \mid D = 1, X = x\right] \cdot f_{X|D=1}(x)\, dx$$

Here, we use distribution of X among D=1, to get their counterfactual in the absence of treatment. Won't work if distrribution of X is altered by treatment.

Conditional Independence Assumption (CIA):

(Y_1, Y_0) are indepedent from D, **conditional on X**

Formally: $(Y_1, Y_0) \perp D \mid X$

Hence

$$f[Y_1 \mid D = 1, X] = f[Y_1 \mid D = 0, X] = f[Y_1 \mid X]$$
$$f[Y_0 \mid D = 0, X] = f[Y_0 \mid D = 1, X] = f[Y_0 \mid X]$$

and in particular

$$E[Y_1 \mid D = 1, X] =$$
$$E[Y_0 \mid D = 0, X] =$$

This identifying assumption can not be directly tested because it involves counterfactual distributions

Conditional Independence Assumption (CIA):

(Y_1, Y_0) are indepedent from D, **conditional on X**

Formally: $(Y_1, Y_0) \perp D \mid X$

Independence is symmetric. That is, D is independent from (Y_1, Y_0):

$$P[D = 1 \mid Y_1, Y_0, X] = P[D = 1 \mid X]$$

or equivalently

$$P[D = 1 \mid U_1, U_0, X] = P[D = 1 \mid X]$$

„Selection on the observables“

CIA revisited (3)

$$P[D = 1 \mid Y_1, Y_0, X] = P[D = 1 \mid X]$$

$$P[D = 1 \mid U_1, U_0, X] = P[D = 1 \mid X]$$

Intuitively, this assumes that, conditioning on observable covariates, we can take assignment to treatment to have been random and that, in particular, unobservables play no role in the treatment assignment; comparing two individuals with the same observable characteristics, one of whom was treated and one of whom was not, is like comparing those two individuals in a randomized experiment.

Dehejia and Wahba (2002)

CIA revisited

$$P[D = 1 | Y_1, Y_0, X] = P[D = 1 | X]$$

$$P[D = 1 | U_1, U_0, X] = P[D = 1 | X]$$

"From an economic standpoint, this assumption rules out selection on the basis of unobservables (U_1,U_0) that may be partially known to people taking training but are unknown to the observing economist. … It defines an implicit model that assumes that agents do not enter the program on the basis of gains unobserved by analysts."

(Heckman, Lalonde and Smith)

Propensity Score

$$P[D = 1 | Y_1, Y_0, X] = P[D = 1 | X] = P(X)$$

$$P[D = 1 | U_1, U_0, X] = P[D = 1 | X] = P(X)$$

(When) Is the CIA plausible?

The conditional independence assumption that motivates the use of regression and matching is most plausible when researchers have extensive knowledge of the process determining treatment status. An example in this spirit is the Angrist (1998) study of the effect of voluntary military service on the civilian earnings of soldiers after discharge. …. The CIA seems plausible in this context because soldiers are selected on the basis of a few well-documented criteria related to age, schooling, and test scores and because the control group also applied to enter the military.

Joshua Angrist (The New Palgrave, 2008)

Common Support

$$ATE = E(Y_1) - E(Y_0)$$
$$= E_X[E(Y_1 \mid D=1, X) - E(Y_0 \mid D=0, X)]$$

Using LIE and CIA

E[·] is w.r.to f(X)

Need to compute these for all X with f(X)>0

$$ATOT = E(Y_1 \mid D=1) - E(Y_0 \mid D=1)$$
$$= E_{X|D=1}[E(Y_1 \mid D=1, X) - E(Y_0 \mid D=0, X)]$$
$$= E(Y_1 \mid D=1) - E_{X|D=1}[E(Y_0 \mid D=0, X)]$$

Using LIE and CIA

E[·] is w.r.to f(X|D=1)

Need to compute this for all X with f(X|D=1)>0

But what if, say, P(D=1|X)=0 for some X?

➔Can estimate ATE and ATOT only over "**common support**"

Common Support (II)

Common support: all values of X for which

$$P(D = 1 | X) < 1$$

alternatively:

all values of X for which f(X|D=1)>0 and f(X|D=0)>0

Hence, with this approach, we can aim only for ATE and ATOT over the common support

$$ATE_{CS} = E[Y_1 - Y_0 \mid 0 < P(D = 1 | X) < 1]$$

$$ATOT_{CS} = E[Y_1 - Y_0 \mid D = 1, P(D = 1 | X) < 1]$$

Exploiting CIA
- Matching

Estimating ATOT using CIA

Estimation by **Stratification**

Suppose X is discrete and can take on only the following values: $\{x_1,\ldots,x_k,\ldots,x_K\}$

Example: $X = [X_1 \quad X_2]$

		X_2 = 1.5	2	3.5
X_1	7	x_1	x_2	x_3
	8.5	x_4	x_5	x_6
	9	x_7	x_8	x_9

Each combination of values of X_1 and of X_2 forms a „cell“ and (in a slight abuse of notation) a value of X (viewed as the collection of X_1 and of X_2). Hence, the possible outcomes of X are $x_1, x_2, \ldots, x_9$

Estimation by **Stratification**

Suppose X is discrete and can take on only the following values: $\{x_1,...,x_k,...,x_K\}$

Let N_{1k} denote the number of treatment group observations in the population with $X=x_k$

and let n_{1k} and n_{0k} be similarly defined sample ferquencies for treatment and control group members.

Estimating ATOT using CIA

Estimation by **Stratification**

X discrete with $\{x_1,\ldots,x_k,\ldots,x_K\}$

and N_{1k}, n_{1k} and n_{0k} be # of obs. with $X=x_k$ in resp. group

		X_2		
		1.5	2	3.5
X_1	7	x_1	x_2	x_3
	8.5	x_4	x_5	x_6
	9	x_7	x_8	x_9

		X_2		
		1.5	2	3.5
X_1	7	N_{11}	N_{12}	N_{13}
	8.5	N_{14}	N_{15}	N_{16}
	9	N_{17}	N_{18}	N_{19}

population frequencies in treatment group

		X_2		
		1.5	2	3.5
X_1	7	n_{11}	n_{12}	n_{13}
	8.5	n_{14}	n_{15}	n_{16}
	9	n_{17}	n_{18}	n_{19}

sample frequencies in treatment group

Estimating ATOT using CIA

Estimation by **Stratification**

if N_{1k}s are not known, use n_{1k}s instead

$$\hat{A}TOT_{Str} = \sum_{k=1}^{K} \frac{\delta_k \cdot N_{1k}}{\sum_{k=1}^{K} \delta_k \cdot N_{1k}} \cdot \left[\bar{Y}_{1k} - \bar{Y}_{0k} \right]$$

sum over all cells (all outcomes of X)

$$\bar{Y}_{1k} = \frac{1}{n_{1k}} \sum_{i:D_i=1 \cap X_i=x_k} Y_{1i} \qquad \bar{Y}_{0k} = \frac{1}{n_{0k}} \sum_{i:D_i=0 \cap X_i=x_k} Y_{0i}$$

$$\delta_k = I\left[n_{1k} > 0, n_{0k} > 0\right]$$

δ_k is an indicator function (=1 if argument is true, 0 otherwise)

Estimating ATOT using CIA

Estimation by **Stratification**

Note how

$$\hat{A}TOT_{Str} = \sum_{k=1}^{K} \frac{\delta_k \cdot N_{1k}}{\sum_{k=1}^{K} \delta_k \cdot N_{1k}} \cdot \left[\overline{Y}_{1k} - \overline{Y}_{0k} \right]$$

is „imitating“ the population version of ATOT under CIA

$$ATOT = E_{X|D=1}\left[E(Y_1 \mid D=1, X) - E(Y_0 \mid D=0, X)\right]$$

Indicator function δ_k is enforcing common support.

Weights $\delta_k \cdot N_{1k} \Big/ \sum_{k=1}^{K} \delta_k \cdot N_{1k}$

correspond to f(X|D=1).

Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants

Joshua D. Angrist
Econometrica, Vol. 66, No. 2. (Mar., 1998), pp. 249-288.

Angrist (1998)

What is the labor-market value of service in the volunteer military?

Many econometric studies have compared the earnings of veterans and nonveterans.

The proper interpretation of results from such studies is unclear, however, because veterans are both self-selected and screened by the military.

The problem of selection bias plagues almost all evaluation research outside of randomized trials.

This paper presents evidence from two new strategies for estimating the effect of voluntary military service on the earnings and employment status of veterans.

His strategies are Matching (covered now) and IV (covered later)

Angrist (1998)

Arguments for his Matching Strategy (1)

First, comparisons by veteran status are restricted to a sample of applicants to the military, only about half of whom actually enlist. Nonenlisting applicants probably provide a better control group for veterans than conventional cross-section samples because, like veterans, applicants have indicated a strong interest in military service.

Moreover, the data analyzed here contain information on most of the characteristics used by the military to screen applicants. The selection bias induced by military screening can therefore be eliminated using regression techniques or by matching on the covariates used in the screening process.

Angrist (1998)

The data

Administrative data from US military
+ Earnings data from Social Security Administration

The military data: information on applicants and entrants to the military for each fiscal year.
Information at the time of application:

- basic demographic variables,
- physical examination results,
- test scores.

SSA keeps track of the earnings of all workers covered by Social Security

The data

Random sample from military data is matched to SSA earnings histories. Limited to

- men aged 17-22 who applied during 1976-82,
- had valid sex and race codes,
- data on Armed Forces Qualification Test scores
- at least a 9th grade education
- but no more than a 4-year college degree.

Target population: 2.2 million white men and 900,000 nonwhite men.

The data

Target population:
2.2 million white men and 900,000 nonwhite men.

Matched sample: 697,944 applicants with Social Security earnings for each year from 1974 through 1991. Used for descriptive analysis

Estimates of causal effect based on **restricted*** **sample** of applicants

- who applied from 1979-82,
- with AFQT scores in groups III and IV.
- contains 128,968 whites and 175,262 nonwhites.

* the restrictions are motivated by imposed to aid IV estimation

More about the data (from a cohort perspective)

TABLE I

APPLICANT POPULATION AND SAMPLE

Race	Application Year 1976	1977	1978	1979	1980	1981	1982
A. *Population*[a]							
White	339.5	286.9	235.9	253.1	348.6	387.3	309.8
Percent veteran[b]	53	52	54	55	53	49	52
Nonwhite	128.6	114.8	103.6	119.5	134.3	149.3	112.5
Percent veteran	44	46	50	46	41	36	43
B. *Sample*[c]							
White	49.2	46.5	40.0	39.4	52.9	57.9	47.3
Percent veteran	56	53	55	57	54	50	53
Nonwhite	50.9	48.1	44.6	51.9	57.0	63.7	48.7
Percent veteran	49	49	52	49	44	38	45

[a] The population is as in Angrist (1993a, Table 4), excluding those with less than a 9th grade education at the time of application. Numbers reported are thousands.

[b] Veterans are applicants identified as entrants to the military within two years following application.

More about the data:

"The typical applicant in the matched sample was

- aged 18-20 at the time he applied,
- had an 11th or 12th grade education,
- and scored in the lower to middle range of the AFQT scale.
- roughly 30 percent of applicants in the sample were aged 18 when they applied to the military, 25 percent were aged 19, and 16 percent were aged 20.
- A total of 40 percent of applicants in the sample were high school graduates, 4 percent were GED certified, and 34 percent had completed 11^{th} grade only. Out of nearly 700,000 applicants in the sample, only 739 were college graduates."

What is Y?

Two outcomes: earnings and employment status.

We focus on his results on earnings

"The primary purpose of this paper is to estimate the impact of military service on the earnings of veterans."

$$E[Y_1 - Y_0 \mid D = 1] = E[Y_1 \mid D = 1] - E[Y_0 \mid D = 1].$$

This tells us whether, on average, veterans benefited or suffered from military service.

Angrist (1998)

"Simple comparisons by veteran status can be used to estimate $E[Y_1 - Y_0 \mid D = 1]$. Because the sample used here includes only applicants to the military, these comparisons control for differences between veterans and nonveterans that originate in the decision to apply to the military."

Angrist (1998)

"The sample is restricted to applicants with AFQT scores in the middle range….
In 1979, 67 percent of white applicants and 78 percent of nonwhite applicants had AFQT scores in categories III and IV, corresponding to the 10th through 64th percentiles of the AFQT reference population."

"In the sample of 1979-82 applicants with AFQT scores in categories III and IV, veterans earned more than nonveterans in every year in which they applied to the military. This can be seen in Figure 2, which plots earning profiles by veteran status and application year."

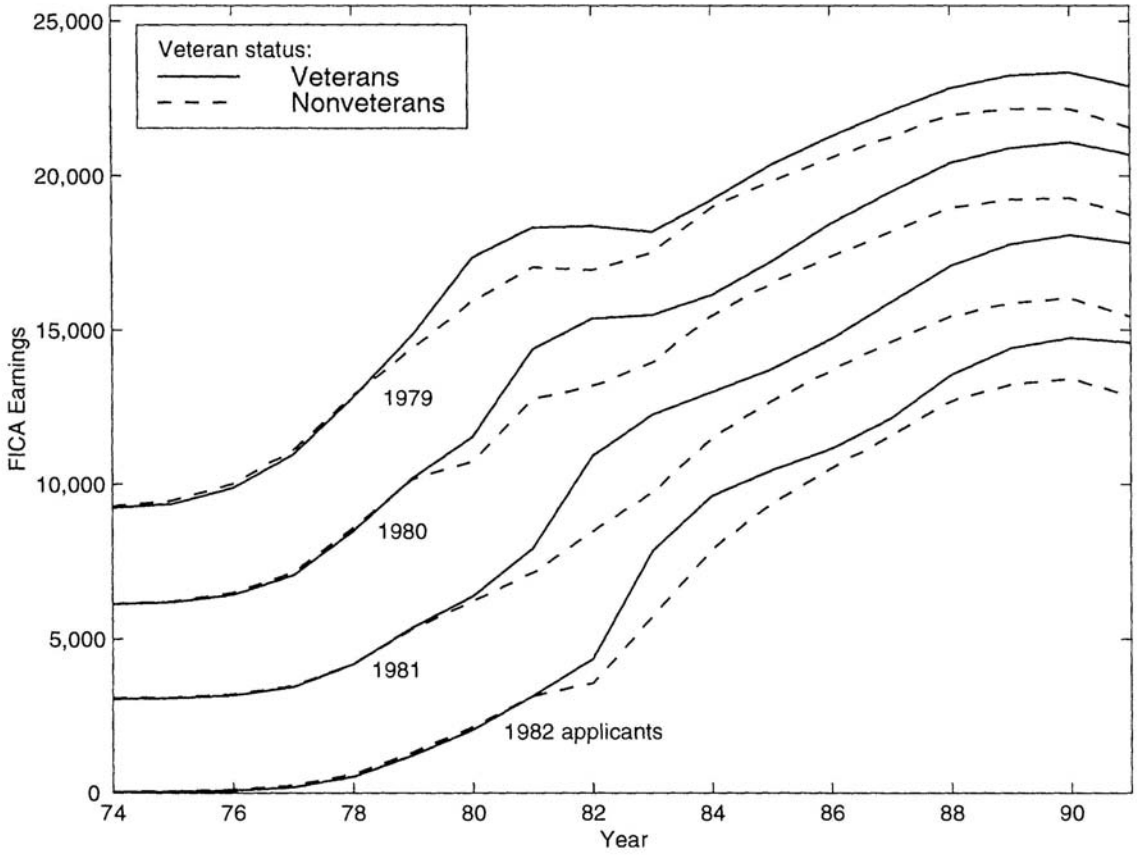


FIGURE 2.—Earnings profiles by veteran status and application year for men who applied 1979–82, with AFQT scores in categories III and IV. The plot shows the actual earnings of men who applied in 1982, earnings + $3,000 for men who applied in 1981, earnings + $6,000 for men who applied in 1980, and earnings + $9,000 for men who applied in 1979.

Angrist (1998)

First the **raw differences**

"Differences in earnings by veteran status are reported with standard errors in columns 2 and 6 of Table 11, separately by race.

Because the sample is so large, all of the post-1978 differences are very precisely measured and significantly different from zero. Some of the earlier small differences are significant as well.

The veteran earnings gap reached a peak of 1500 dollars for whites and 2900 dollars for nonwhites in 1982-83, and remained substantial through the end of the sample period.

The fact that pre-application-year differences are small tends to support the interpretation of the veteran/nonveteran contrast as an unbiased estimate of $E[Y_1-Y_0|D=1]$.
In Section 4, however, I show that these simple contrasts are misleading.

	Whites		Nonwhites	
Year	Mean (1)	Difference in Means[c] (2)	Mean (5)	Difference in Means (6)
A. *Earnings*[a]				
74	182.7	−26.1	157.2	−4.9
		(7.0)		(4.4)
75	237.9	−41.4	216.9	−.6
		(6.3)		(4.5)
76	473.4	−47.9	413.6	−14.5
		(8.1)		(6.4)
77	1012.9	−7.1	820.9	−13.0
		(11.3)		(9.1)
78	2147.1	40.3	1677.9	58.1
		(16.7)		(13.4)
79	3560.7	188.0	2797.0	340.3
		(21.0)		(16.2)
80	4709.0	572.9	3932.2	1154.3
		(23.4)		(18.0)
81	6226.0	855.5	5218.8	1920.0
		(27.2)		(20.7)
82	7200.6	1508.5	6150.2	2917.1
		(30.3)		(23.4)
83	8398.1	1390.5	7221.1	2889.9
		(34.4)		(27.0)
84	9874.2	652.8	8377.2	2202.9
		(39.5)		(30.5)
85	10972.7	469.8	9306.8	1955.5
		(44.6)		(34.4)
86	12004.5	543.7	10106.2	1881.3
		(50.4)		(38.7)
87	13045.7	663.9	10833.0	2050.1
		(54.6)		(41.8)
88	14136.1	904.3	11480.1	2175.0
		(58.3)		(44.9)
89	14716.1	1169.1	11751.4	2379.1
		(61.0)		(47.6)
90	14886.1	1300.8	11904.3	2483.6
		(63.0)		(49.4)
91	14407.9	1559.6	11518.7	2758.8
		(64.6)		(50.8)

Recall: sample members applied during 1976-82.

What do the numbers mean? "To save space, differences in the table are for earnings averaged across application cohorts".

Arguments for his Matching Strategy (2)

Because the sample is confined to applicants, comparisons of earnings by veteran status such as those in Table II control for veterans' decisions to apply to the military.

On the other hand, veterans are carefully selected by the military on the basis of personal characteristics, like schooling and test scores, that are clearly related to future earnings. This fact motivates the matching estimator.

It is worth mentioning again, however, that the modest pre-application provide little evidence of selection bias. Of course, part of the problem with the use of such early comparisons as a specification check is that earnings or labor force participation as a teenager may not be related to earnings potential as an adult.

What is X?

In practice, the observed covariates take on values in the set of all possible combinations of

- race,
- application year,
- schooling at the time of application,
- Armed Forces Qualification Test (AFQT) score group
- year of birth.

466 possible values of X for whites and 429 possible values of X for nonwhites

Is there a common support problem?

„In practice, it can happen that some population cells where both treatment and control observations are available nevertheless remain unrepresented in a random sample. In this study, however, the sample was drawn conditional on X. Therefore, sample observations on both veterans and nonveterans are necessarily available wherever the population probability of treatment is neither zero nor one“

Matching estimator: Stratification

$$\hat{A}TOT_{Str} = \sum_{k=1}^{K} \frac{\delta_k \cdot N_{1k}}{\sum_{k=1}^{K} \delta_k \cdot N_{1k}} \cdot \left[\overline{Y}_{1k} - \overline{Y}_{0k} \right]$$

sum over all cells
(all outcomes of X)

$$\overline{Y}_{1k} = \frac{1}{n_{1k}} \sum_{i:D_i=1 \cap X_i=x_k} Y_{1i} \qquad \overline{Y}_{0k} = \frac{1}{n_{0k}} \sum_{i:D_i=0 \cap X_i=x_k} Y_{0i} \qquad \delta_k = I\left[n_{1k} > 0, n_{0k} > 0\right]$$

Micro data of 697644 persons from Military applicants data base can be linked via SSN to Social Security Earnings records. Angrist only obtained cell-wise information: counts, average and standard deviation of earnings.

[7] As noted above, the sample design implies that $\delta_k = 1[n_{1k} > 0, n_{0k} > 0]$ equals the population indicator $1[N_{1k} > 0, N_{0k} > 0]$. In practice, however, a few cells are missing because of the confidentiality edit.

Angrist (1998)

"Matching estimates (averaged over application cohorts) of veteran effects suggest that the simple comparisons of earnings by veteran status overestimate the effect of military service on earnings and employment.

For whites, they range from a high of only 783 dollars in 1982 to a low of -557 dollars in 1986. Standard errors for these estimates are less than 60 dollars. It is negative in every year after 1983 except 1991.

Effects for nonwhites are much larger although they are also substantially smaller than the corresponding simple comparisons. The largest estimate is 2,186 dollars in 1982 and the smallest is 708 dollars in 1988.

The 1991 estimate of 1,026 dollars for nonwhites is less than 9 percent of nonwhite's average 1991 FICA earnings. The 1991 estimate for whites is about 30 dollars and is not statistically different from zero.

	Whites				Nonwhites			
Year	Mean (1)	Difference in Means[c] (2)	Controlled Contrast (3)	Regression Estimates (4)	Mean (5)	Difference in Means (6)	Controlled Contrast (7)	Regression Estimates (8)
A. *Earnings*[a]								
74	182.7	−26.1	−14.0	−13.0	157.2	−4.9	−2.0	−3.9
		(7.0)	(9.2)	(9.4)		(4.4)	(6.0)	(5.8)
75	237.9	−41.4	−14.2	−12.0	216.9	−.6	−17.1	−15.2
		(6.3)	(7.6)	(7.8)		(4.5)	(6.0)	(5.5)
76	473.4	−47.9	−14.8	−12.7	413.6	−14.5	−33.3	−30.2
		(8.1)	(9.0)	(9.3)		(6.4)	(8.0)	(7.4)
77	1012.9	−7.1	−8.6	−9.4	820.9	−13.0	−56.0	−51.3
		(11.3)	(12.3)	(12.2)		(9.1)	(11.1)	(10.0)
78	2147.1	40.3	−23.5	−22.4	1677.9	58.1	−53.6	−42.5
		(16.7)	(18.1)	(17.2)		(13.4)	(16.1)	(14.1)
79	3560.7	188.0	−8.4	−11.2	2797.0	340.3	119.1	122.3
		(21.0)	(23.2)	(21.6)		(16.2)	(20.1)	(17.2)
80	4709.0	572.9	178.0	175.9	3932.2	1154.3	741.6	738.5
		(23.4)	(27.2)	(24.6)		(18.0)	(23.4)	(19.5)
81	6226.0	855.5	249.5	249.9	5218.8	1920.0	1299.9	1318.5
		(27.2)	(32.4)	(29.1)		(20.7)	(28.2)	(23.1)
82	7200.6	1508.5	783.3	782.4	6150.2	2917.1	2186.0	2210.1
		(30.3)	(36.4)	(32.5)		(23.4)	(32.0)	(26.0)
83	8398.1	1390.5	588.8	601.5	7221.1	2889.9	2103.8	2142.3
		(34.4)	(41.1)	(36.6)		(27.0)	(36.7)	(29.8)
84	9874.2	652.8	−235.7	−198.5	8377.2	2202.9	1333.0	1428.9
		(39.5)	(46.9)	(41.7)		(30.5)	(41.4)	(33.4)
85	10972.7	469.8	−521.3	−459.6	9306.8	1955.5	932.3	1059.2
		(44.6)	(52.6)	(46.8)		(34.4)	(46.2)	(37.3)
86	12004.5	543.7	−557.3	−491.7	10106.2	1881.3	720.9	872.3
		(50.4)	(59.0)	(52.5)		(38.7)	(51.2)	(41.6)
87	13045.7	663.9	−548.0	−464.3	10833.0	2050.1	751.0	925.0
		(54.6)	(63.9)	(56.8)		(41.8)	(55.2)	(44.8)
88	14136.1	904.3	−415.5	−311.7	11480.1	2175.0	708.2	923.7
		(58.3)	(68.2)	(60.6)		(44.9)	(59.5)	(48.1)
89	14716.1	1169.1	−248.6	−136.3	11751.4	2379.1	799.7	1031.9
		(61.0)	(71.2)	(63.2)		(47.6)	(62.7)	(50.9)
90	14886.1	1300.8	−154.5	−53.2	11904.3	2483.6	824.9	1064.0
		(63.0)	(73.6)	(65.2)		(49.4)	(65.4)	(52.7)
91	14407.9	1559.6	29.8	146.2	11518.7	2758.8	1026.1	1277.9
		(64.6)	(75.6)	(66.9)		(50.8)	(67.2)	(54.3)

General Matching Estimator

$$\hat{ATOT}_{GM} = \frac{1}{n_1} \sum_{i=1}^{n_1} \left\{ Y_{1i} - \hat{E}\left(Y_0 \mid D = 0, X_i\right) \right\}$$

where $$\hat{E}\left(Y_0 \mid D = 0, X_i\right) = \sum_{j=1}^{n_0} W_{i,j} \cdot Y_{0j}$$

where the weights sum to one: $\sum_{j \in I_0} W_{i,j} = 1$

Example

$$W_{i,j} = \begin{cases} 1 \text{ if } j \text{ is the (unique) nearest neighbor of } i \\ 0 \text{ otherwise} \end{cases}$$

How observations are matched

Several approaches exist to match treatment group observation *i* with control group observation(s) :

- **k-Nearest-Neighbor**-Matching (k may be one)
- **Caliper Matching**:
  use all comparison group observations within a specified radius (“caliper”)
- **Blockwise Matching**
- **Kernel Matching**

We can match with or without replacement.

Kernel Estimator of ATOT („Kernel Matching“)

$$ATOT = E(Y_1 \mid D = 1) - E_{X \mid D=1}[E(Y_0 \mid D = 0, X)]$$

$$\hat{ATOT}_K = \frac{1}{n_1} \sum_{i=1}^{n_1} \left\{ Y_{1i} - \hat{E}(Y_0 \mid D = 0, X_i) \right\}$$

$$\hat{ATOT} = \frac{1}{n_1} \sum_{i=1}^{n_1} \{ Y_{1i} - \sum_{j=1}^{n_0} \frac{K\left(\frac{X_i - X_j}{h} \right)}{\sum_{l=1}^{n_0} K\left(\frac{X_i - X_l}{h} \right)} \cdot Y_{0j} \}$$

Kernel Estimator of ATOT („Kernel Matching“)

$$ATOT = E(Y_1 \mid D=1) - E_{X|D=1}\left[E(Y_0 \mid D=0, X)\right]$$

$$\hat{ATOT}_K = \frac{1}{n_1}\sum_{i=1}^{n_1}\left\{Y_{1i} - \hat{E}(Y_0 \mid D=0, X_i)\right\}$$

$$\hat{ATOT} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left\{Y_{1i} - \sum_{j=1}^{n_0}\frac{K\left(\frac{X_i - X_j}{h}\right)}{\sum_{l=1}^{n_0}K\left(\frac{X_i - X_l}{h}\right)} \cdot Y_{0j}\right\}$$

Recall: **Kernel Estimator** of E[Y|X=x]

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right)} = \sum_{i=1}^{n} \frac{K\left(\frac{x - X_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right)} Y_i$$

$$= \sum_{i=1}^{n} W_{h,i}(x) Y_i$$

In our specific application

$$\hat{E}(Y_0 \mid D = 0, X_i) = \sum_{j=1}^{n_0} \frac{K\left(\frac{X_i - X_j}{h}\right)}{\sum_{l=1}^{n_0} K\left(\frac{X_i - X_l}{h}\right)} \cdot Y_{0j}$$

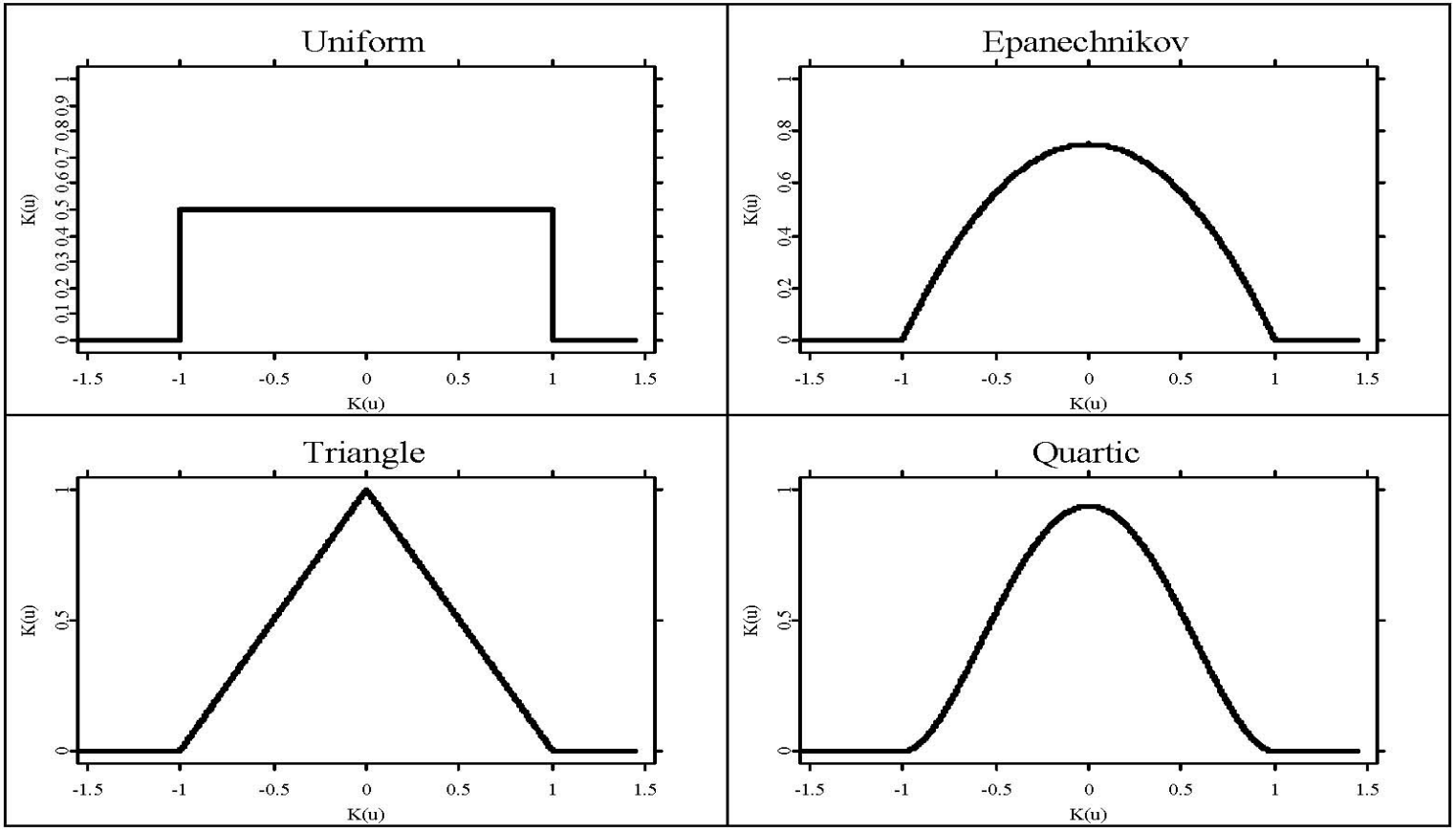
Kernel function

$$u = \frac{x - X_i}{h}$$

I = Indicator function

Kernel	**K(u)**
Uniform	$\frac{1}{2} \cdot \mathbf{I}(\lvert u \rvert \le 1)$
Triangle	$(1 - \lvert u \rvert) \cdot \mathbf{I}(\lvert u \rvert \le 1)$
Epanechnikov	$\frac{3}{4}(1 - u^2) \cdot \mathbf{I}(\lvert u \rvert \le 1)$
Quartic	$\frac{15}{16}(1 - u^2)^2 \cdot \mathbf{I}(\lvert u \rvert \le 1)$
Triweight	$\frac{35}{32}(1 - u^2)^3 \cdot \mathbf{I}(\lvert u \rvert \le 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$
Cosinus	$\frac{4}{\pi} \cos\left(\frac{\pi}{2}u\right) \cdot \mathbf{I}(\lvert u \rvert \le 1)$

Kernel functions



$$E(Y \mid X) = E(Y \mid X_1, .., X_K) = m(x)$$

Multivariate Nadaraya-Watson-Kernel Estimator

$$\hat{m}_H(x_1, x_2 \ldots, x_k) = \sum_{i=1}^{n} \frac{K\left(\frac{X_{1i} - x_1}{h_1}, \frac{X_{2i} - x_1}{h_2}, \ldots, \frac{X_{ki} - x_1}{h_k}\right)}{\sum_{j=1}^{n} K\left(\frac{X_{1j} - x_1}{h_1}, \frac{X_{2j} - x_1}{h_2}, \ldots, \frac{X_{kj} - x_1}{h_k}\right)} \cdot Y_i$$

Suffers from the **curse of dimensionality**

Definition Propensity Score:

The `Propensity Score' is the conditional probability of receiving the treatment (or the fraction of the population with charactersitics X belonging to the treatment group)

$$P(D = 1 | X_1, ..., X_k) = P(X) = E(D | X)$$

- It maps the multidemsional X (which is shorthand for $X_1, X_2, \ldots, X_k$) into the scalar $0 \leq P(X) \leq 1$
- If we knew $P(X_i)$ for each individual i, then we could look for matches with respect to this scalar (rather than the multidimensional X)

Is this justified?

Matching on the Propensity Score (II)

Is matching on the propensity score justified?

Yes! If $\left(Y^1, Y^0\right) \perp D \mid X \Rightarrow \left(Y^1, Y^0\right) \perp D \mid P(X)$

If independence of the potential outcomes from the treatment holds conditional on X, then it also holds conditional on P(X). That is, CIA for X implies CIA for P(X)!
In terms of the means:

If
$$E[Y_1 \mid D=1, X] = E[Y_1 \mid D=0, X] = E[Y_1 \mid X]$$
$$E[Y_0 \mid D=0, X] = E[Y_0 \mid D=1, X] = E[Y_0 \mid X]$$

then
$$E[Y_1 \mid D=1, P(X)] = E[Y_1 \mid D=0, P(X)] = E[Y_1 \mid P(X)]$$
$$E[Y_0 \mid D=0, P(X)] = E[Y_0 \mid D=1, P(X)] = E[Y_0 \mid P(X)]$$

Matching on the Propensity Score (III)

$$\left(Y^1, Y^0\right) \perp D \mid X \Rightarrow \left(Y^1, Y^0\right) \perp D \mid P(X)$$

Proof:

What we need to show is that

$$\text{if } P\left[D = 1 \mid Y^1, Y^0, X\right] = P\left[D = 1 \mid X\right]$$

$$P\left[D = 1 \mid Y^1, Y^0, P(X)\right] = P\left[D = 1 \mid P(X)\right]$$

Preliminary result: „Balancing property of P(X)“

$$D \perp X \mid P(X)$$

$$\text{i.e. } P\left(D = 1 \mid X, P(X)\right) = P\left(D = 1 \mid P(X)\right)$$

$$\text{or } f\left(X \mid D = 1, P(X)\right) = f\left(X \mid D = 0, P(X)\right)$$

Matching on the Propensity Score (IV)

$$P(D=1|X_1,...,X_k)=P(X)=E(D|X)$$

$$D \perp X|P(X)$$

Proof: show that $P(D=1|X,P(X))=P(D=1|P(X))$

$$\begin{aligned}P(D=1|X,P(X)) &= E(D|X,P(X))=E(D|X)=P(D=1|X)\\ &= P(X)\end{aligned}$$

$$\begin{aligned}P[D=1|P(X)] &= E[D|P(X)]=E\{E[D|X,P(X)]|P(X)\}\\ &= E[P(X)|P(X)]\\ &= P(X)\end{aligned}$$

$$\Rightarrow \quad f(X|D=1,P(X))=f(X|D=0,P(X))$$

Numerical Example:

			f(X)	P(D=1|X)
	1	100	0.5	0.9
X	2	40	0.2	0.9
	3	20	0.1	0.2
	4	40	0.2	0.4
		200	1	

Is $P(D=1|X,P(X)) = P(D=1|P(X))$?

$$P(D=1|X=1,P(X)=0.9) = P(D=1|X=1, X=1 \cup X=2)$$
$$= P(D=1|X=1) = P(1) = 0.9$$
$$P(D=1|X=2,P(X)=0.9) = P(D=1|X=2, X=1 \cup X=2)$$
$$= P(D=1|X=2) = P(2) = 0.9$$

Numerical Example:

			f(X)	P(D=1|X)	D=1	D=0	f(X|D=1)	f(X|D=0)
	1	100	0.5	0.9	90	10	0.616	0.185
X	2	40	0.2	0.9	36	4	0.247	0.074
	3	20	0.1	0.2	4	16	0.027	0.296
	4	40	0.2	0.4	16	24	0.110	0.444
		200	1		146	54	1	

Is $f(X \mid D=1, P(X)) = f(X \mid D=0, P(X))$?

$$
\begin{aligned}
f(X=1 \mid D=1, P(X)=0.9) &= P(X=1 \mid D=1, P(X)=0.9) \\
&= P(X=1 \mid D=1, X=1 \cup X=2) \\
&= \frac{P(X=1 \cap (X=1 \cup X=2) \mid D=1)}{P(X=1 \cup X=2 \mid D=1)} \\
&= \frac{P(X=1 \mid D=1)}{P(X=1 \mid D=1) + P(X=2 \mid D=1)} \\
&= \frac{0.616}{0.616 + 0.247} = 0.714
\end{aligned}
$$

Numerical Example:

			f(X)	P(D=1|X)	D=1	D=0	f(X|D=1)	f(X|D=0)
	1	100	0.5	0.9	90	10	0.616	0.185
X	2	40	0.2	0.9	36	4	0.247	0.074
	3	20	0.1	0.2	4	16	0.027	0.296
	4	40	0.2	0.4	16	24	0.110	0.444
		200	1		146	54	1	

Is $f(X \mid D=1, P(X)) = f(X \mid D=0, P(X))$?

$$
\begin{aligned}
f(X=1 \mid D=0, P(X)=0.9) &= P(X=1 \mid D=0, P(X)=0.9) \\
&= P(X=1 \mid D=0, X=1 \cup X=2) \\
&= \frac{P(X=1 \cap (X=1 \cup X=2) \mid D=0)}{P(X=1 \cup X=2 \mid D=0)} \\
&= \frac{P(X=1 \mid D=0)}{P(X=1 \mid D=0) + P(X=2 \mid D=0)} \\
&= \frac{0.185}{0.185 + 0.074} = 0.714
\end{aligned}
$$

Numerical Example:

			f(X)	P(D=1|X)	D=1	D=0	f(X|D=1)	f(X|D=0)
	1	100	0.5	0.9	90	10	0.616	0.185
X	2	40	0.2	0.9	36	4	0.247	0.074
	3	20	0.1	0.2	4	16	0.027	0.296
	4	40	0.2	0.4	16	24	0.110	0.444
		200	1		146	54	1	

In the same way, it can be shown that

$$f(X=2\,|\,D=1,P(X)=0.9)=f(X=2\,|\,D=0,P(X)=0.9)$$
$$=0.286$$

Hence, $f(X\,|\,D=1,P(X)=0.9)=f(X\,|\,D=0,P(X)=0.9)$ for all x

			f(X)	P(D=1|X)	D=1	D=0	f(X|D=1)	f(X|D=0)	f(X|D=1,P(X)=0.9)	f(X|D=0,P(X)=0.9)
	1	100	0.5	0.9	90	10	0.616	0.185	0.714	0.714
X	2	40	0.2	0.9	36	4	0.247	0.074	0.286	0.286
	3	20	0.1	0.2	4	16	0.027	0.296		
	4	40	0.2	0.4	16	24	0.110	0.444		
		200	1		146	54	1			

Numerical Example:

			f(X)	P(D=1|X)	D=1	D=0	f(X|D=1)	f(X|D=0)	f(X|D=1,P(X)=0.9)	f(X|D=0,P(X)=0.9)
	1	100	0.5	0.9	90	10	0.616	0.185	0.714	0.714
X	2	40	0.2	0.9	36	4	0.247	0.074	0.286	0.286
	3	20	0.1	0.2	4	16	0.027	0.296		
	4	40	0.2	0.4	16	24	0.110	0.444		
		200	1		146	54	1			

$$f(X \mid D=1, P(X)=0.9) = f(X \mid D=0, P(X)=0.9) \quad \text{for all } x$$

Propensity score matching combines groups with different values of Z but the same values of $\Pr(D=1|Z)$. To see why this works, consider two groups, one with $Z = Z_1$ and the other with $Z = Z_2$, but where $\Pr(D=1|Z=Z_1) = \Pr(D=1|Z=Z_2)$. Combining these groups in the matching works because they will have the same relative proportions in the $D=0$ and $D=1$ populations precisely because they have the same probabilty of participation. As a result, any difference in $E(Y_0)$

Smith and Todd (2003)

Matching on the Propensity Score (V)

$$\left(Y^1, Y^0\right) \perp D \mid X \Rightarrow \left(Y^1, Y^0\right) \perp D \mid P(X)$$

Proof: show that $P\left[D = 1 \mid Y^1, Y^0, P(X)\right] = P\left[D = 1 \mid P(X)\right]$

$$\begin{aligned}
P\left[D = 1 \mid Y^1, Y^0, P(X)\right] &= E\left[D \mid Y^1, Y^0, P(X)\right] \\
&= E\left\{E\left[D \mid X, Y^1, Y^0, P(X)\right] \mid Y^1, Y^0, P(X)\right\} \\
&= E\left\{E\left[D \mid X, Y^1, Y^0\right] \mid Y^1, Y^0, P(X)\right\} \\
&= E\left\{E\left[D \mid X\right] \mid Y^1, Y^0, P(X)\right\} \\
&= E\left[P(X) \mid Y^1, Y^0, P(X)\right] = P(X)
\end{aligned}$$

But $P\left[D = 1 \mid P(X)\right] = P(X)$

$$\Rightarrow \quad P\left[D = 1 \mid Y^1, Y^0, P(X)\right] = P\left[D = 1 \mid P(X)\right]$$

Matching on the Propensity Score (VI)

$$\left(Y^1, Y^0\right) \perp D \mid P(X)$$

This implies:

$$E[Y_1 \mid D = 1, P(X)] = E[Y_1 \mid D = 0, P(X)] = E[Y_1 \mid P(X)]$$
$$E[Y_0 \mid D = 0, P(X)] = E[Y_0 \mid D = 1, P(X)] = E[Y_0 \mid P(X)]$$

$$\begin{aligned} ATE &= E(Y_1) - E(Y_0) = E_{P(X)}\left[E(Y_1 \mid P(X)) - E(Y_0 \mid P(X))\right] \\ &= E_{P(X)}\left[E(Y_1 \mid D = 1, P(X)) - E(Y_0 \mid D = 0, P(X))\right] \end{aligned}$$

$$\begin{aligned} ATOT &= E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1) \\ &= E_{P(X)\mid D=1}\left[E(Y_1 \mid D = 1, P(X)) - E(Y_0 \mid D = 1, P(X))\right] \\ &= E_{P(X)\mid D=1}\left[E(Y_1 \mid D = 1, P(X)) - E(Y_0 \mid D = 0, P(X))\right] \\ &= E(Y_1 \mid D = 1) - E_{P(X)\mid D=1}\left[E(Y_0 \mid D = 0, P(X))\right] \end{aligned}$$

Numerical Example:

			f(X)	P(D=1|X)	D=1	D=0	f(X|D=1)	f(X|D=0)	E[Y1|X]	E[Y0|X]	ATE(X)	ATE(X)*f(X|D=1)
	1	100	0.5	0.9	90	10	0.616	0.185	4	2	2	1.23287671
X	2	40	0.2	0.9	36	4	0.247	0.074	6	5	1	0.24657534
	3	20	0.1	0.2	4	16	0.027	0.296	2	3	-1	-0.02739726
	4	40	0.2	0.4	16	24	0.110	0.444	1	4	-3	-0.32876712
		200	1		146	54	1					1.12328767

$$ATOT = E_{X|D=1}\left[E(Y_1 \mid D=1, X) - E(Y_0 \mid D=0, X)\right]$$

Based on CIA

$$E[Y_1 \mid D=1, X] = E[Y_1 \mid D=0, X] = E[Y_1 \mid X]$$

$$E[Y_0 \mid D=0, X] = E[Y_0 \mid D=1, X] = E[Y_0 \mid X]$$

Do we get the same answer from $(Y^1, Y^0) \perp D \mid P(X)$

$$E[Y_1 \mid D=1, P(X)] = E[Y_1 \mid D=0, P(X)] = E[Y_1 \mid P(X)]$$

$$E[Y_0 \mid D=0, P(X)] = E[Y_0 \mid D=1, P(X)] = E[Y_0 \mid P(X)]$$

and

$$ATOT = E_{P(X)|D=1}\left[E(Y_1 \mid D=1, P(X)) - E(Y_0 \mid D=0, P(X))\right]$$

Numerical Example:

			f(X)	P(D=1|X)	D=1	D=0	f(X|D=1)	f(X|D=0)	E[Y1|X]	E[Y0|X]	ATE(X)	ATE(X)*f(X|D=1)
	1	100	0.5	0.9	90	10	0.616	0.185	4	2	2	1.23287671
X	2	40	0.2	0.9	36	4	0.247	0.074	6	5	1	0.24657534
	3	20	0.1	0.2	4	16	0.027	0.296	2	3	-1	-0.02739726
	4	40	0.2	0.4	16	24	0.110	0.444	1	4	-3	-0.32876712
		200	1		146	54	1					1.12328767

We need to calculate

$$ATOT = E_{P(X)|D=1}\left[E(Y_1 \mid D=1, P(X)) - E(Y_0 \mid D=0, P(X))\right]$$

using

$$E[Y_1 \mid D=1, P(X)] = E[Y_1 \mid D=0, P(X)] = E[Y_1 \mid P(X)]$$

$$E[Y_0 \mid D=0, P(X)] = E[Y_0 \mid D=1, P(X)] = E[Y_0 \mid P(X)]$$

		f(P(X)|D=1)	f(P(X)|D=0)	E[Y1|P(X)]	E[Y0|P(X)]	ATE(P(X))	
	0.2	0.027	0.296	2	3	-1	-0.02739726
P(X)	0.4	0.110	0.444	1	4	-3	-0.32876712
	0.9	0.863	0.259	4.571	2.857	1.714	1.47945205
		1	1				1.12328767

$$f(P(X)=0.9 \mid D=1) = P(X=1 \cup X=2 \mid D=1)$$
$$= P(X=1 \mid D=1) + P(X=2 \mid D=1)$$
$$= 0.616 + 0.247 = 0.863$$

$$E[Y_1 \mid P(X)=0.9, D=1] = E[Y_1 \mid X=1, D=1] \cdot f(X=1 \mid D=1, P(X)=0.9)$$
$$+ E[Y_1 \mid X=2, D=1] \cdot f(X=2 \mid D=1, P(X)=0.9)$$
$$= 4 \cdot 0.741 + 6 \cdot 0.286 = 4.571$$

Propensity Score

needs to estimated in practice

Probit-Modell:

$$P(X) = P(D = 1 | X_1, \ldots, X_k) = \Phi(\beta_0 + \beta_1 \cdot X_1, \ldots, \beta_k \cdot X_k)$$

Logit Model:

$$P(X) = P(D = 1 | X_1, \ldots, X_k) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \cdot X_1, \ldots, \beta_k \cdot X_k))}$$

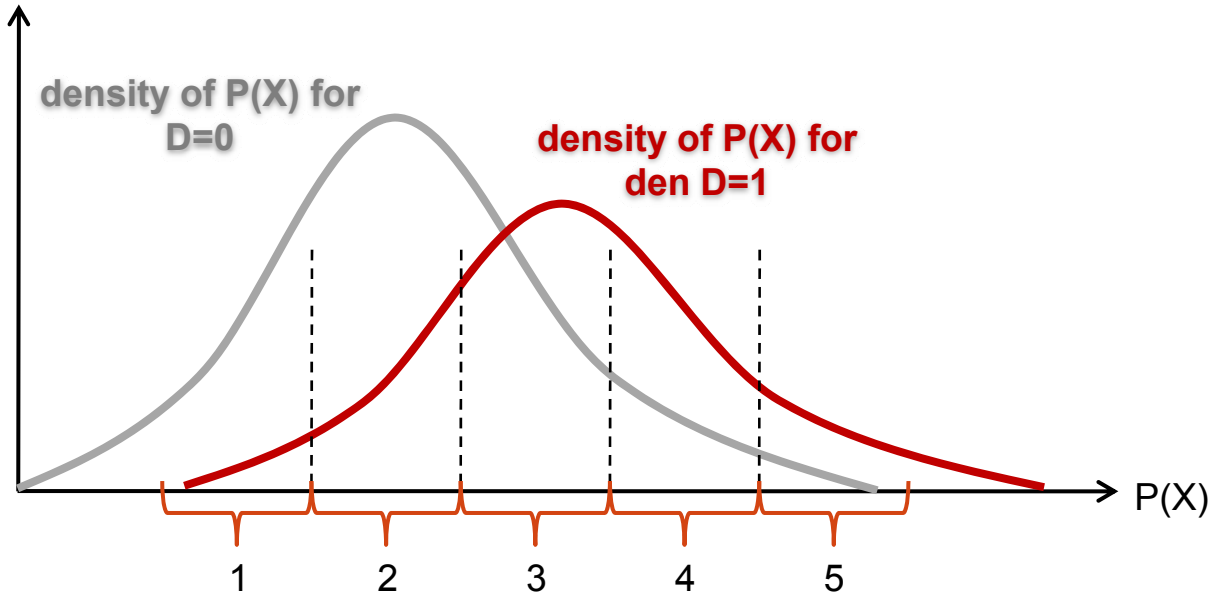
How matching is performed

In Propensity Score Matching we (again) have several alternative ways in which matching can be performed

- Blockwise Matching
- k-nearest neighbor matching
- Kernel matching
- Caliper Matching

Warning: The following formulas pretend that Propensity Score is known. In practice, it needs to be estimated

Blockwise Matching on the Propensity Score



For the **1st block:**

$$\hat{ATOT}_1^S = \frac{1}{n_{11}} \cdot \sum_{i \in I(1)} Y_{1i} - \frac{1}{n_{01}} \cdot \sum_{j \in I(1)} Y_{0j}$$

Average of Y among D = 0 in 1st Propensity Score-block

all treatment group members in 1st block

For all Q blocks:

$$\hat{ATOT}^S = \sum_{q=1}^{Q} \hat{ATOT}_q^S \cdot \frac{\sum_{i \in I(q)} D_i}{\sum_{\forall i} D_i}$$

Weighing with the importance of a block, i.e. the fraction of treatment group members in a block

Variance of this estimator

$$Var\left(\hat{ATOT}^S\right) = \frac{1}{n_1} \cdot \left[Var(Y_{i1}) + \sum_{q=1}^{Q} \frac{n_{q1}}{n_1} \cdot \frac{n_{q1}}{n_{q0}} \cdot Var(Y_{j0}) \right]$$

This assumes that the Propensity Score is known.

K-nearest-neighbor matching

$$\hat{ATOT}^{M} = \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} \left[Y_{1i} - \sum_{j \in C(i)} W_{ij} \cdot Y_{0j} \right]$$

$$= \frac{1}{n_1} \cdot \left[\sum_{i=1}^{n_1} Y_{1i} - \sum_{i=1}^{n_1} \sum_{j \in C(i)} W_{ij} \cdot Y_{0j} \right]$$

$$= \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} Y_{1i} - \frac{1}{n_1} \cdot \sum_{j=1}^{n0} W_{j} \cdot Y_{0j}$$

$$\text{where } W_{ij} = \frac{1}{n_{0i}} \text{ if } j \in C(i), \text{ otherwise } W_{ij} = 0$$

control group members belonging to the k-nearest neighbors of treatment group observation i (in terms of Propensity Scores)

Variance of the k-NN estimator

$$\mathrm{Var}\left(\hat{\mathrm{ATOT}}^{M}\right) = \frac{1}{(n_1)^2}\cdot\left[\sum_{i=1}^{n_1}\mathrm{Var}(Y_{i1}) + \sum_{j=1}^{n_0}(W_j)^2\,\mathrm{Var}(Y_{j0})\right]$$

$$= \frac{1}{(n_1)^2}\cdot\left[n_1\cdot\mathrm{Var}(Y_{i1}) + \sum_{j=1}^{n_0}(W_j)^2\,\mathrm{Var}(Y_{j0})\right]$$

$$= \frac{1}{n_1}\cdot\mathrm{Var}(Y_{i1}) + \frac{1}{(n_1)^2}\cdot\sum_{j=1}^{n_0}(W_j)^2\,\mathrm{Var}(Y_{j0})$$

Kernel Matching on Propensity Score

$$\hat{ATOT}^{K} = \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} \{ Y_{i1} - \frac{\sum_{j=1}^{n_0} Y_{j0} \cdot K\left(\frac{P(X_i) - P(X_j)}{h} \right)}{\sum_{k=1}^{n_0} K\left(\frac{P(X_i) - P(X_k)}{h} \right)} \}$$

$$= \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} \{ Y_{i1} - \sum_{j=1}^{n_0} Y_{j0} \cdot W_{ij} \}$$

$$== \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} \{ Y_{i1} - \hat{Y}_{i0} \}$$

For each control variable X:

Standardized Bias prior to Matching

$$SB_{before} = 100 \cdot \frac{(\overline{X}_1 - \overline{X}_0)}{\sqrt{0{,}5 \cdot [Var_1(X) + Var_0(X)]}}$$

Standardized Bias after Matching

$$SB_{after} = 100 \cdot \frac{(\overline{X}_{1M} - \overline{X}_{0M})}{\sqrt{0{,}5 \cdot [Var_{1M}(X) + Var_{0M}(X)]}}$$

Dehejia and Wahba (1999,2002)

Rajeev H. Dehejia and Sadek Wahba

"Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs“
Journal of the American Statistical Association, Vol. 94, Number 448 (December 1999), pp. 1053-1062

PROPENSITY SCORE-MATCHING METHODS FOR NONEXPERIMENTAL CAUSAL STUDIES
The Review of Economics and Statistics, February 2002, 84(1): 151–161

Example: Causal Effect of the NSW program

National Supported Work (NSW) Demonstration project

- Aim: provide work experience (6-18 months) for individuals with economic and social problems
- Four target groups:
  - Women on Aid to Families with Dependent Children (AFDC)
  - former addicts
  - former offenders
  - young school dropouts.
- Randomized treatment:
  Candidates were selected on the basis of eligibility criteria and then randomly assigned to the program.
- Outcome:
  Real annual earnings in 1978

LaLonde (AER 1986):

"Evaluating the Econometric Evaluations of Training Programs with Experimental Data"

1. Estimate causal effect by comparing outcomes of experimental treatment and control group

2. Use survey data to form a non-experimental control group

3. Using experimental treatment group + nonexperimental control group + econometrics (OLS, IV, etc.): can experimental estimate of causal effect be replicated?

Lalonde: "This comparison shows that many of the econometric procedures do not replicate the experimentally determined results, and it suggests that researchers should be aware of the potential for specification errors in other nonexperimental evaluations."

Heckman & Hotz (1989): A reanalysis of the NSW data reveals that a simple testing procedure eliminates the range of nonexperimental estimators at variance with the experimental estimates of program impact

Dehejia and Wahba (1999, 2002)

D&W reanalyze the NSW data and find:

the right nonexperimental method (Propensity Score matching) is getting the right answer!

Sample:

"The data we use are a subsample of the data used in LaLonde (1986). Using the LaLonde male sample of 297 treated and 425 control units, we exclude the observations for which (pre-treatment) earnings in 1974 could not be obtained, thus arriving at a reduced sample of 185 treated observations and 260 control observations."

Dehejia and Wahba (2002)

TABLE 1.—SAMPLE MEANS AND STANDARD ERRORS OF COVARIATES FOR MALE NSW PARTICIPANTS

National Supported Work Sample (Treatment and Control)		
	Dehejia-Wahba Sample	
Variable	Treatment	Control
Age	25.81 (0.52)	25.05 (0.45)
Years of schooling	10.35 (0.15)	10.09 (0.1)
Proportion of school dropouts	0.71 (0.03)	0.83 (0.02)
Proportion of blacks	0.84 (0.03)	0.83 (0.02)
Proportion of Hispanic	0.06 (0.017)	0.10 (0.019)
Proportion married	0.19 (0.03)	0.15 (0.02)
Number of children	0.41 (0.07)	0.37 (0.06)
No-show variable	0 (0)	n/a
Month of assignment (Jan. 1978 = 0)	18.49 (0.36)	17.86 (0.35)
Real earnings 12 months before training	1,689 (235)	1,425 (182)
Real earnings 24 months before training	2,096 (359)	2,107 (353)
Hours worked 1 year before training	294 (36)	243 (27)
Hours worked 2 years before training	306 (46)	267 (37)
Sample size	185	260

Dehejia and Wahba (1999, 2002)

Nonexperimenta control-group for matching analysis are constructed from

CPS:

The Current Population Survey is a monthly survey of households conducted by the Bureau of Census for the Bureau of Labor Statistics.

PSID:

The Panel Study of Income Dynamics (PSID), begun in 1968, is a longitudinal study of a representative sample of U.S. individuals (men, women, and children) and the family units in which they reside.

Dehejia and Wahba (1999)

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]

A Simple Algorithm for Estimating the Propensity Score

1. Start with a parsimonious logit specification
2. Sort data according to estimated propensity score
3. Stratify observations such that estimated propensity scores within a stratum for treated and comparison units are close
4. Statistical test: for all covariates, differences in means across treated and comparison units within each stratum are not significantly different from zero.
   a) If covariates are balanced between treated and comparison observations for all strata, stop.
   b) If covariates are not balanced for some stratum, divide the stratum into finer strata and reevaluate.
   c) 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 reevaluate.

Propensity Score Estimation via Logit Model:

$$P(X) = P(D = 1 \mid X_1, \ldots, X_k) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_k \cdot X_k))}$$

$$= F(\beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_k \cdot X_k)$$

$$\hat{P}(X_i) = F(\hat{\beta}_0 + \hat{\beta}_1 \cdot X_{1i} + \ldots + \hat{\beta}_k \cdot X_{ki}) \quad \text{where } \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k \text{ are MLEs}$$

X-variables: age, age^2, educ, educ^2, married no degree, black, Hispanic, RE74, RE75, RE74^2, RE75^2,U74, U75, U74*black (for model with PSID control group).

How good is the matching?

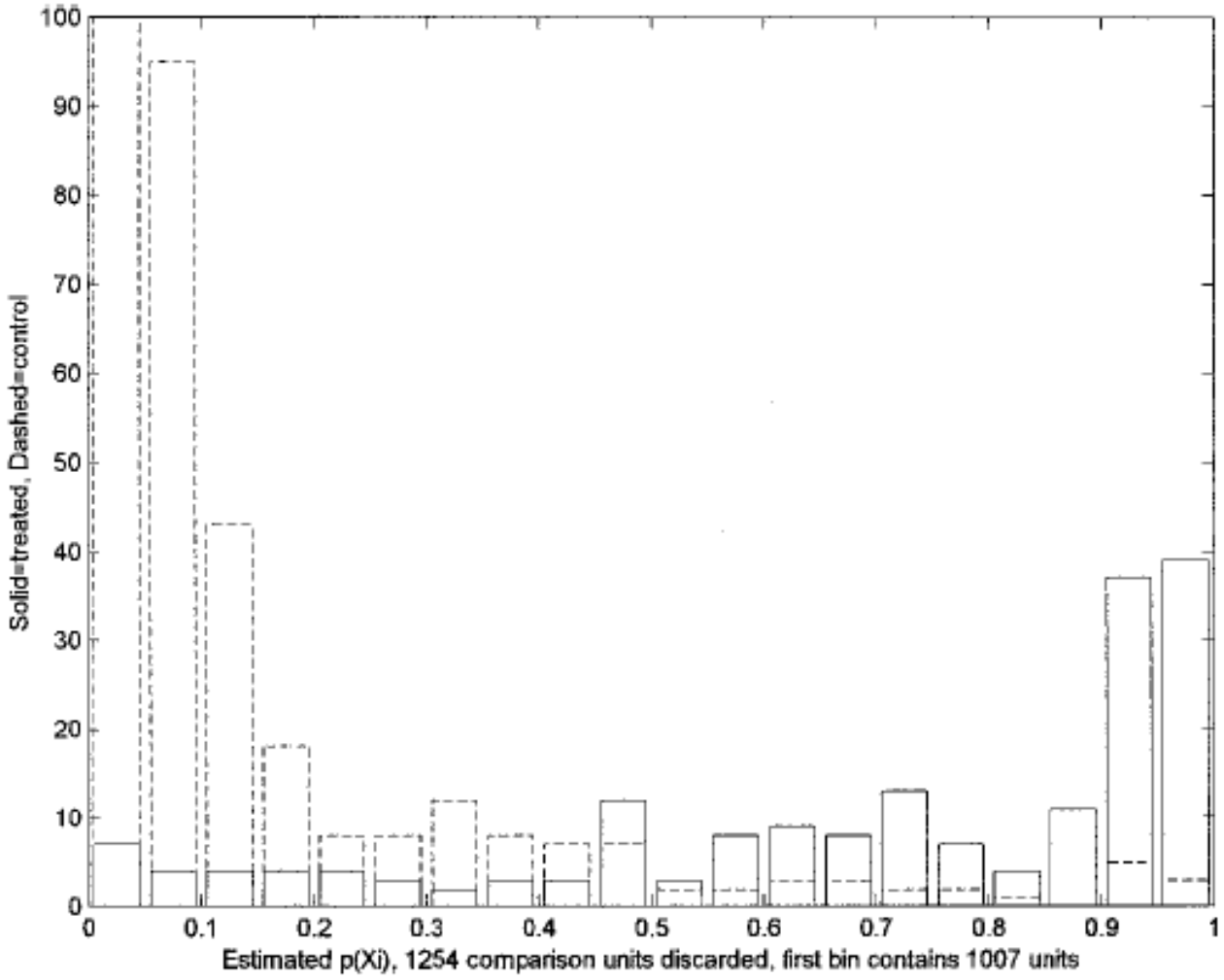
11168 (of 15992) of the CPS obs and 1254 (of 2490) of the PSID obs have estimated propensity scores less than the minimum estimated propensity score for the treated units.

Most of the remaining comparison obs (4398 for the CPS and 1007 for the PSID) are in the first bin of sorted propensity scores.

In the NSW-PSID sample, many of the upper bins have far more treated units than comparison units.

“Hence, it is clear that very few of the comparison units are comparable to the treated units.“

FIGURE 2.—HISTOGRAM OF ESTIMATED PROPENSITY SCORE, NSW AND PSID



Dehejia and Wahba (2002)

“Three issues arise in implementing matching:

1. whether or not to match with replacement,
2. how many comparison units to match to each treated unit
3. which matching method to choose.

1. Whether or not to match with replacement?

Matching with replacement minimizes the propensity score distance between the matched comparison units and the treatment unit: each treatment unit can be matched to the nearest comparison unit, even if a comparison unit is matched more than once. This is beneficial in terms of bias reduction.

1. Whether or not to match with replacement?

"In contrast, by **matching without replacement**, when there are few comparison units similar to the treatedunits, we may be forced to match treated units to comparison units that are quite different in terms of the estimated propensity score. This increases bias, but it could improve the precision of the estimates. An additional complication of matching without replacement is that the results are potentially sensitive to the order in which the treatment units are matched."

2. How many comparison units to match?

"By using **a single comparison unit** for each treatment unit, we ensure the smallest propensity-score distance between the treatment and comparison units.

By using **more comparison units**, one increases the precision of the estimates, but at the cost of increased bias. One method of selecting a set of comparison units is the nearest-neighbor method, which selects the *m* comparison units whose propensity scores are closest to the treated unit in question."

3. Which matching method to choose?

One method of selecting a set of comparison units is the **nearest-neighbor method**, which selects the *m* comparison units whose propensity scores are closest to the treated unit in question.

Another method is **caliper matching**, which uses all of the comparison units within a predetermined propensity score radius (or "caliper"). A benefit of caliper matching is that it uses only as many comparison units as are available within the calipers, allowing for the use of extra (fewer) units when good matches are (not)available."

Dehejia and Wahba (2002)

Empirical results

TABLE 2.—SAMPLE CHARACTERISTICS AND ESTIMATED IMPACTS FROM THE NSW AND CPS SAMPLES

Control Sample	No. of Observations	Mean Propensity Score[A]	Age	School	Black	Hispanic	No Degree	Married	RE74	RE75	U74	U75	Treatment Effect (Diff. in Means)	Regression Treatment Effect
NSW	185	0.37	25.82	10.35	0.84	0.06	0.71	0.19	2095	1532	0.29	0.40	1794[B]	1672[C]
													(633)	(638)
Full CPS	15992	0.01	33.23	12.03	0.07	0.07	0.30	0.71	14017	13651	0.88	0.89	−8498	1066
		(0.02)[D]	(0.53)	(0.15)	(0.03)	(0.02)	(0.03)	(0.03)	(367)	(248)	(0.03)	(0.04)	(583)[E]	(554)
Without replacement:														
Random								0.22	2305	1687	0.37	0.51	1559	1651
								(0.04)	(495)	(341)	(0.05)	(0.05)	(733)	(709)
Low to high								0.22	2286	1687	0.37	0.51	1605	1681
								(0.04)	(495)	(341)	(0.05)	(0.05)	(730)	(704)
High to low								0.22	2305	1687	0.37	0.51	1559	1651
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(733)	(709)
With replacement:														
Nearest neighbor	119	0.37	25.36	10.31	0.84	0.06	0.69	0.17	2407	1516	0.35	0.49	1360	1375
		(0.03)	(1.04)	(0.31)	(0.06)	(0.04)	(0.07)	(0.06)	(727)	(506)	(0.07)	(0.07)	(913)	(907)
Caliper, $\delta = 0.00001$	325	0.37	25.26	10.31	0.84	0.07	0.69	0.17	2424	1509	0.36	0.50	1119	1142
		(0.03)	(1.03)	(0.30)	(0.06)	(0.04)	(0.07)	(0.06)	(845)	(647)	(0.06)	(0.06)	(875)	(874)
Caliper, $\delta = 0.00005$	1043	0.37	25.29	10.28	0.84	0.07	0.69	0.17	2305	1523	0.35	0.49	1158	1139
		(0.02)	(1.03)	(0.32)	(0.05)	(0.04)	(0.06)	(0.06)	(877)	(675)	(0.06)	(0.60)	(852)	(851)
Caliper, $\delta = 0.0001$	1731	0.37	25.19	10.36	0.84	0.07	0.69	0.17	2213	1545	0.34	0.50	1122	1119
		(0.02)	(1.03)	(0.31)	(0.05)	(0.04)	(0.06)	(0.06)	(890)	(701)	(0.06)	(0.06)	(850)	(843)

the treated units are ranked (from lowest to highest or highest to lowest propensity score, or randomly).

Variables: Age, age of participant; School, number of school years; Black, 1 if black, 0 otherwise; Hisp, 1 if Hispanic, 0 otherwise; No degree, 1 if participant had no school degrees, 0 otherwise; Married, 1 if married, 0 otherwise; RE74, real earnings (1982US$) in 1974; RE75, real earnings (1982US$) in 1975; U74, 1 if unemployed in 1974, 0 otherwise; U75, 1 if unemployed in 1975, 0 otherwise; and RE78, real earnings (1982US$) in 1978.

(A) The propensity score is estimated using a logit of treatment status on: Age, Age^2, Age^3, School, School^2, Married, No degree, Black, Hisp, RE74, RE75, U74, U75, School · RE74.

(B) The treatment effect for the NSW sample is estimated using the experimental control group.

(C) The regression treatment effect controls for all covariates linearly. For matching with replacement, weighted least squares is used, where treatment units are weighted at 1 and the weight for a control is the number of times it is matched to a treatment unit.

(D) The standard error applies to the difference in means between the matched and the NSW sample, except in the last two columns, where the standard error applies to the treatment effect.

(E) Standard errors for the treatment effect and regression treatment effect are computed using a bootstrap with 500 replications.

Empirical results

TABLE 2.—SAMPLE CHARACTERISTICS AND ESTIMATED IMPACTS FROM THE NSW AND CPS SAMPLES

Control Sample	No. of Observations	Mean Propensity Score[A]	Age	School	Black	Hispanic	No Degree	Married	RE74	RE75	U74	U75	Treatment Effect (Diff. in Means)	Regression Treatment Effect
NSW	185	0.37	25.82	10.35	0.84	0.06	0.71	0.19	2095	1532	0.29	0.40	1794[B]	1672[C]
													(633)	(638)
Full CPS	15992	0.01	33.23	12.03	0.07	0.07	0.30	0.71	14017	13651	0.88	0.89	−8498	1066
		(0.02)[D]	(0.53)	(0.15)	(0.03)	(0.02)	(0.03)	(0.03)	(367)	(248)	(0.03)	(0.04)	(583)[E]	(554)
Without replacement:														
Random	185	0.32	25.26	10.30	0.84	0.06	0.65	0.22	2305	1687	0.37	0.51	1559	1651
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(733)	(709)
Low to high	185	0.32	25.23	10.28	0.84	0.06	0.66	0.22	2286	1687	0.37	0.51	1605	1681
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(730)	(704)
High to low	185	0.32	25.26	10.30	0.84	0.06	0.65	0.22	2305	1687	0.37	0.51	1559	1651
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(733)	(709)
With replacement:														
Nearest neighbor	119	0.37	25.36	10.31	0.84	0.06	0.69	0.17	2407	1516	0.35	0.49	1360	1375
		(0.03)	(1.04)	(0.31)	(0.06)	(0.04)	(0.07)	(0.06)	(727)	(506)	(0.07)	(0.07)	(913)	(907)
Caliper, $\delta = 0.00001$	325	0.37	25.26	10.31	0.84	0.07	0.69	0.17	2424	1509	0.36	0.50	1119	1142
		(0.03)	(1.03)	(0.30)	(0.06)	(0.04)	(0.07)	(0.06)	(845)	(647)	(0.06)	(0.06)	(875)	(874)
Caliper, $\delta = 0.00005$	1043	0.37	25.29	10.28	0.84	0.07	0.69	0.17	2305	1523	0.35	0.49	1158	1139
		(0.02)	(1.03)	(0.32)	(0.05)	(0.04)	(0.06)	(0.06)	(877)	(675)	(0.06)	(0.60)	(852)	(851)
Caliper, $\delta = 0.0001$	1731	0.37	25.19	10.36	0.84	0.07	0.69	0.17	2213	1545	0.34	0.50	1122	1119
		(0.02)	(1.03)	(0.31)	(0.05)	(0.04)	(0.06)	(0.06)	(890)	(701)	(0.06)	(0.06)	(850)	(843)

Variables: Age, age of participant; School, number of school years; Black, 1 if black, 0 otherwise; Hisp, 1 if Hispanic, 0 otherwise; No degree, 1 if participant had no school degrees, 0 otherwise; Married, 1 if married, 0 otherwise; RE74, real earnings (1982US$) in 1974; RE75, real earnings (1982US$) in 1975; U74, 1 if unemployed in 1974, 0 otherwise; U75, 1 if unemployed in 1975, 0 otherwise; and RE78, real earnings (1982US$) in 1978.

(A) The propensity score is estimated using a logit of treatment status on: Age, Age^2, Age^3, School, School^2, Married, No degree, Black, Hisp, RE74, RE75, U74, U75, School · RE74.

(B) The treatment effect for the NSW sample is estimated using the experimental control group.

(C) The regression treatment effect controls for all covariates linearly. For matching with replacement, weighted least squares is used, where treatment units are weighted at 1 and the weight for a control is the number of times it is matched to a treatment unit.

(D) The standard error applies to the difference in means between the matched and the NSW sample, except in the last two columns, where the standard error applies to the treatment effect.

(E) Standard errors for the treatment effect and regression treatment effect are computed using a bootstrap with 500 replications.

Conclusions

It is something of an irony that the data that we use were originally employed by LaLonde (1986) to demonstrate the failure of standard nonexperimental methods in accurately estimating the treatment effect. Using matching methods on both of his samples, we are able to replicate the experimental benchmark …. in a setting in which the treated group differs substantially from the pool of potential comparison units. The method is able to pare the large comparison group down to the relevant comparisons “

Smith and Todd

Jeffrey A. Smith and Petra E. Todd "Does matching overcome LaLonde's critique of nonexperimental estimators?"
Journal of Econometrics 125 (2005) 305–353

"This paper applies cross-sectional and longitudinal propensity score matching estimators to data from the National Supported Work (NSW) Demonstration that have been previously analyzed by LaLonde (1986) and Dehejia and Wahba (1999, 2002).

We find that estimates of the impact of NSW based on propensity score matching are highly sensitive to both the set of variables included in the scores and the particular analysis sample used in the estimation.

Among the estimators we study, the difference-in-differences matching estimator performs the best. ….. Our analysis demonstrates that while propensity score matching is a potentially useful econometric tool, it does not represent a general solution to the evaluation problem.

Heckman, Ichimura, Smith and Todd (1996)

„ Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluationmethod”
Proc Natl Acad Sci U S A. 1996 November 12; 93(23): 13416–13420.

Using data from a recent social experiment, … We find that matching based on the propensity score eliminates some but not all of the measured selection bias, with the remaining bias still a substantial fraction of the estimated impact.

We find that the support of the distribution of propensity scores for the comparison group is typically only a small portion of the support for the participant group.

For values outside the common support, it is impossible to reliably estimate the effect of program participation using matching methods.

Exploiting CIA
- Regression

Recall

Average Treatment Effect

$$\begin{aligned} ATE &= E(Y_1) - E(Y_0) \\ &= E_X[E(Y_1 \mid X) - E(Y_0 \mid X)] \\ &= E_X[E(Y_1 \mid D=1, X) - E(Y_0 \mid D=0, X)] \end{aligned}$$

Law of iterated expectations (holds in general)

requires mean independence

Average Treatment Effect on the Treated

$$\begin{aligned} ATOT &= E(Y_1 \mid D=1) - E(Y_0 \mid D=1) \\ &= E_{X\mid D=1}[E(Y_1 \mid D=1, X) - E(Y_0 \mid D=1, X)] \\ &= E_{X\mid D=1}[E(Y_1 \mid D=1, X) - E(Y_0 \mid D=0, X)] \\ &= E(Y_1 \mid D=1) - E_{X\mid D=1}[E(Y_0 \mid D=0, X)] \end{aligned}$$

Special Case: Linear relationships

$$Y_1 = E[Y_1 \mid X] + U_1 = \mu_1(X) + U_1 = \alpha_1 + \boldsymbol{\beta}_1' \mathbf{X} + U_1$$

$$Y_0 = E[Y_0 \mid X] + U_0 = \mu_0(X) + U_0 = \alpha_0 + \boldsymbol{\beta}_0' \mathbf{X} + U_0$$

$$\mathrm{ATE}(X) = \mu_1(X) - \mu_0(X) = (\alpha_1 - \alpha_0) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)' \mathbf{X}$$

$$\begin{aligned}\mathrm{ATOT}(X) &= \mu_1(X) - \mu_0(X) + \mathrm{E}[U_1 - U_0 \mid D = 1, X] \\ &= \mathrm{ATE}(X) + \mathrm{E}[U_1 - U_0 \mid D = 1, X] \\ &= (\alpha_1 - \alpha_0) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)' \mathbf{X} + \mathrm{E}[U_1 - U_0 \mid D = 1, X]\end{aligned}$$

Under the CIA assumption

$$E[Y_1 \mid D=1, X] = E[Y_1 \mid D=0, X] = E[Y_1 \mid X]$$
$$E[Y_0 \mid D=0, X] = E[Y_0 \mid D=1, X] = E[Y_0 \mid X]$$

$$E[U_1 \mid D=1, X] = E[U_1 \mid D=0, X] = E[U_1 \mid X] = 0$$
$$E[U_0 \mid D=0, X] = E[U_0 \mid D=1, X] = E[U_0 \mid X] = 0$$

$$\begin{aligned} ATOT(X) &= (\alpha_1 - \alpha_0) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)' \mathbf{X} + E[U_1 - U_0 \mid D=1, X] \\ &= (\alpha_1 - \alpha_0) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)' \mathbf{X} = ATE(X) \end{aligned}$$

CIA **and** Linearity (II)

Under the CIA assumption

$$\begin{aligned} ATOT(X) &= (\alpha_1 - \alpha_0) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)'\mathbf{X} + E[U_1 - U_0 \mid D = 1, X] \\ &= (\alpha_1 - \alpha_0) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)'\mathbf{X} = ATE(X) \end{aligned}$$

In general, $ATE = E_X[ATE(X)]$ $\quad ATOT = E_{X|D=1}[ATOT(X)]$

Under CIA and linearity

$$\hat{A}TE_{Reg} = \frac{1}{N}\sum_{i=1}^{N}\left[(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)'\mathbf{X}_i\right]$$

$$\hat{A}TOT_{Reg} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left[(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)'\mathbf{X}_i\right]$$

Alternatively for ATOT $\quad \hat{A}TOT_{Reg} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left[Y_{1i} - (\hat{\alpha}_0 + \hat{\boldsymbol{\beta}}_0'\mathbf{X}_i)\right]$

CIA **and** Linearity (III)

$$\hat{\text{ATE}}_{\text{Reg}} = \frac{1}{N}\sum_{i=1}^{N}\left[(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)'\mathbf{X}_i\right]$$

$$\hat{\text{ATOT}}_{\text{Reg}} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left[(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)'\mathbf{X}_i\right] \qquad \hat{\text{ATOT}}_{\text{Reg}} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left[Y_{1i} - (\hat{\alpha}_0 + \hat{\boldsymbol{\beta}}_0'\mathbf{X}_i)\right]$$

These estimators use the data more efficiently than the nonparametric matching estimators. Indeed, if we further assume that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0$ then we can use OLS to estimate

$$Y_i = \alpha_0 + (\alpha_1 - \alpha_0)D_i + \boldsymbol{\beta}'\mathbf{X}_i + U_i$$

The OLS estimator of the coefficient of D is a consistent estimator of ATOT and ATE.

CIA **and** Linearity (III)

$$\hat{\text{ATE}}_{\text{Reg}} = \frac{1}{N}\sum_{i=1}^{N}\left[(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)'\mathbf{X}_i\right]$$

$$\hat{\text{ATOT}}_{\text{Reg}} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left[(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)'\mathbf{X}_i\right]$$

However, these estimators may rely heavily on extrapolation.

Counterfactuals are obtained straight from linearity

$$\hat{\text{ATOT}}_{\text{Reg}} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left[Y_{1i} - (\hat{\alpha}_0 + \hat{\boldsymbol{\beta}}_0'\mathbf{X}_i)\right]$$

Example (continued)

Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants

Joshua D. Angrist
Econometrica, Vol. 66, No. 2. (Mar., 1998), pp. 249-288.

Angrist's **almost saturated regression** estimates

status. In the sample of men who applied in 1979–82, with AFQT scores in groups III or IV, and earnings data in 1991 for both veterans and nonveterans, there are 466 possible values of X for whites and 429 possible values of X for nonwhites.[28] As before, index these possible values by $k = 1, \ldots, K$. Observations on the dependent variable can then be written $\bar{y}_{Dk}$, denoting average earnings for men with veteran status D and covariate-combination k.

Regression estimates of the effect of military service are based on the following model, estimated separately, for each calendar year and race group:

$$\bar{y}_{Dk} = \beta_k + \alpha_r D + \bar{\epsilon}_{Dk}, \tag{17}$$

where β_k is an effect for covariate-combination k, α_r is a veteran effect, and $\bar{\epsilon}_{Dk}$ is an error term that is orthogonal to D and X by definition of β_k and α_r. Estimates of α_r, denoted $\hat{\alpha}_r$, were computed by weighted least squares using population cell counts (N_{Dk}) as weights. This weighting scheme, when applied to

Angrist (1998)

"Difference in Means"
= Naïve estimates

"Controlled Contrasts"
= Matching via Stratification

„Regression Estimates"
= Almost saturated Regression
(see equation (17) on prev. slide)

	Whites				Nonwhites			
Year	Mean (1)	Difference in Means[c] (2)	Controlled Contrast (3)	Regression Estimates (4)	Mean (5)	Difference in Means (6)	Controlled Contrast (7)	Regression Estimates (8)
A. *Earnings*[a]								
74	182.7	−26.1	−14.0	−13.0	157.2	−4.9	−2.0	−3.9
		(7.0)	(9.2)	(9.4)		(4.4)	(6.0)	(5.8)
75	237.9	−41.4	−14.2	−12.0	216.9	−.6	−17.1	−15.2
		(6.3)	(7.6)	(7.8)		(4.5)	(6.0)	(5.5)
76	473.4	−47.9	−14.8	−12.7	413.6	−14.5	−33.3	−30.2
		(8.1)	(9.0)	(9.3)		(6.4)	(8.0)	(7.4)
77	1012.9	−7.1	−8.6	−9.4	820.9	−13.0	−56.0	−51.3
		(11.3)	(12.3)	(12.2)		(9.1)	(11.1)	(10.0)
78	2147.1	40.3	−23.5	−22.4	1677.9	58.1	−53.6	−42.5
		(16.7)	(18.1)	(17.2)		(13.4)	(16.1)	(14.1)
79	3560.7	188.0	−8.4	−11.2	2797.0	340.3	119.1	122.3
		(21.0)	(23.2)	(21.6)		(16.2)	(20.1)	(17.2)
80	4709.0	572.9	178.0	175.9	3932.2	1154.3	741.6	738.5
		(23.4)	(27.2)	(24.6)		(18.0)	(23.4)	(19.5)
81	6226.0	855.5	249.5	249.9	5218.8	1920.0	1299.9	1318.5
		(27.2)	(32.4)	(29.1)		(20.7)	(28.2)	(23.1)
82	7200.6	1508.5	783.3	782.4	6150.2	2917.1	2186.0	2210.1
		(30.3)	(36.4)	(32.5)		(23.4)	(32.0)	(26.0)
83	8398.1	1390.5	588.8	601.5	7221.1	2889.9	2103.8	2142.3
		(34.4)	(41.1)	(36.6)		(27.0)	(36.7)	(29.8)
84	9874.2	652.8	−235.7	−198.5	8377.2	2202.9	1333.0	1428.9
		(39.5)	(46.9)	(41.7)		(30.5)	(41.4)	(33.4)
85	10972.7	469.8	−521.3	−459.6	9306.8	1955.5	932.3	1059.2
		(44.6)	(52.6)	(46.8)		(34.4)	(46.2)	(37.3)
86	12004.5	543.7	−557.3	−491.7	10106.2	1881.3	720.9	872.3
		(50.4)	(59.0)	(52.5)		(38.7)	(51.2)	(41.6)
87	13045.7	663.9	−548.0	−464.3	10833.0	2050.1	751.0	925.0
		(54.6)	(63.9)	(56.8)		(41.8)	(55.2)	(44.8)
88	14136.1	904.3	−415.5	−311.7	11480.1	2175.0	708.2	923.7
		(58.3)	(68.2)	(60.6)		(44.9)	(59.5)	(48.1)
89	14716.1	1169.1	−248.6	−136.3	11751.4	2379.1	799.7	1031.9
		(61.0)	(71.2)	(63.2)		(47.6)	(62.7)	(50.9)
90	14886.1	1300.8	−154.5	−53.2	11904.3	2483.6	824.9	1064.0
		(63.0)	(73.6)	(65.2)		(49.4)	(65.4)	(52.7)
91	14407.9	1559.6	29.8	146.2	11518.7	2758.8	1026.1	1277.9
		(64.6)	(75.6)	(66.9)		(50.8)	(67.2)	(54.3)

Angrist (1998)

Comparing matching and regression estimates:

fects by the proportion of veterans at each value of X. In practice, the regression and matching estimates are almost identical through 1984. This can be seen in Table II, which reports $\hat{\alpha}_r$ as well as $\hat{\alpha}_c$ for each year. In contrast with the 1974–84 results, however, regression estimates for each year after 1984 are larger than the corresponding matching estimates. The largest difference is

Why are they different after 1984? Because they put different weights on conditional contrasts (see also next slide)

race groups. The matching estimator gives the small covariate-specific estimates for men with high probabilities of service the most weight, while the larger covariate-specific estimates for men with low probability of service are given less weight. The regression estimator, in contrast, gives more weight to covariate-specific estimates where the probability of military service conditional on covariates is close to one-half. This leads to a higher overall treatment effect.

Matching vs. OLS-Regression

For the simple case of a binary X-variable Angrist shows that

“the difference between matching and OLS is in the nature of the weights (of contrasts) at values of x where both veterans and nonveterans are observed.

Matching weights each of the underlying treatment effects by P[D= 1 I X] P [X] ,whereas OLS regression weights each of the underlying treatment effects by P[D= 1 IX](1 -P[D= 1 I X])P [X].

In other words, the weights underlying matching are proportional to the probability of veteran status at each value of the covariates while the weights underlying OLS regression are proportional to the variance of veteran status at each value of the covariates.“
➔Matching puts more weights on ATOT(x) if x values are important in treatment group, OLS on those x with a larger variance of D.

Matching vs. Regression

Angrist & Pischke:

“We believe regression should be the starting point for most empirical projects… The first reason why we don’t find ourselves on the propensity score bandwagon are practical: there are many details to be filled in when implementing propensity score matching, such as how to model the score and how to do inference; these are details not yet standardized.”

Matching vs. Regression

Angrist & Pischke (p.86):

“Moreover, ..there isn’t very much theoretical daylight between regression and propensity score weighting. If the regression model for covariates is fairly flexible, say, close to saturated, regression can be seen as a type of propensity score weighting, so the difference is mostly in the implementation.”

Heckman Nobel Lecture - MICRODATA, HETEROGENEITY AND THE EVALUATION OF PUBLIC POLICY

"If there were no unobservables, or if fortuitously conditioning on *X* eliminated mean differences in unobservables, as is assumed by statisticians who advocate the method of matching, then the selection bias term vanishes. Yet the poor fit of most microdata equations suggests that the assumption of no unobservables is unacceptable. Reliance on matching is an act of faith."

CIA : $E[U_0 \mid D=0, X] = E[U_0 \mid D=1, X] = E[U_0 \mid X] = 0$

"In the debate (about the plausibility of the CIA) it has been argued that agents' optimizing behavior precludes choices being independent of the potential outcomes. This seems an unduly narrow view. In response I will offer three arguments for considering these assumptions….."

The first is a statistical, data-descriptive motivation. A natural starting point in the evaluation of any program is a comparison of average outcomes for treated and control units. A logical next step is to adjust any difference in average outcomes for differences in exogenous background characteristics (exogenous in the sense of not being affected by the treatment). Such an analysis may not lead to the final word on the efficacy of the treatment, but its absence would seem difficult to rationalize in a serious attempt to understand the evidence regarding the effect of the treatment.

A second argument is that almost any evaluation of a treatment involves comparisons of units who received the treatment with units who did not. The question is typically not whether such a comparison should be made, but rather which units should be compared, that is, which units best represent the treated units had they not been treated. Economic theory can help in classifying variables into those that need to be adjusted for versus those that do not, on the basis of their role in the decision process (for example, whether they enter the utility function or the constraints). Given that, the unconfoundedness assumption merely asserts that all variables that need to be adjusted for are observed by the researcher. This is an empirical question, and not one that should be controversial as a general principle. It is clear that settings where some of these covariates are not observed will require strong assumptions to allow for identification. Such assumptions include instrumental variables settings where some covariates are assumed to be independent of the potential outcomes. Absent those assumptions, typically only bounds can be identified (as in Manski, 1990, 2003).

A third, related argument is that even when agents choose their treatment optimally, two agents with the same values for observed characteristics may differ in their treatment choices without invalidating the unconfoundedness assumption if the difference in their choices is driven by differences in unobserved characteristics that are themselves unrelated to the outcomes of interest. The plausibility of this will depend critically on the exact nature of the optimization process faced by the agents. In particular it may be important that the objective of the decision maker is distinct from the outcome that is of interest to the evaluator. For example,