Matching

Lingguo Cheng

Nanjing University

2025.Fall

Introduction

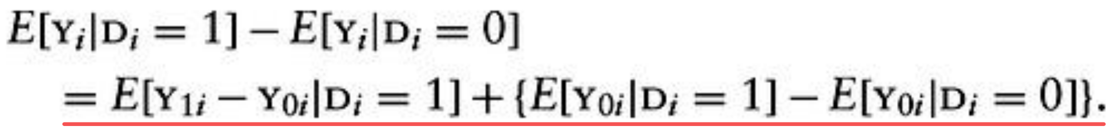
- Matching as a strategy to control for covariates is typically motivated by the CIA.
- Matching estimators are appealingly simple: at the bottom, matching amounts to covariate-specific treatment-control comparisons, weighted together to produce a single overall average treatment effect.
- An attractive feature of matching strategies is that they are typically accompanied by an explicit statement of the *conditional independence assumption* required to give matching estimates a causal interpretation.

Regression Meets Matching

- Matching and regression are both control strategies. Since the core assumption underlying causal inference is the same for the two strategies, it’s worth asking whether or to what extent matching really differs from regression.
- Our view is that regression can be motivated as a particular sort of weighted matching estimator, and therefore the differences between regression and matching estimates are unlikely to be of major empirical importance.
- To flesh out this idea, it helps to look more deeply into the mathematical structure of the matching and regressions estimands, that is, the population quantities that these methods attempt to estimate.

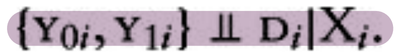
What’s the difference between regression and matching

- In the case of the military service example. veteran status, D_i , Y_{1i} and Y_{0i} to denote potential outcomes.



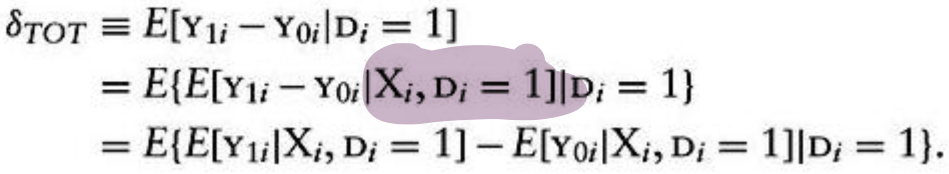
$$\begin{aligned} &E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\ &\quad \underline{= E[Y_{1i} - Y_{0i}|D_i = 1] + \{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]\}.} \end{aligned}$$

- The CIA in this context says that



$$\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp D_i|X_i.$$

- Given the CIA, selection bias disappears after conditioning on X_i, so the effect of treatment on the treated can be constructed by iterating expectations over X_i:



$$\begin{aligned} \delta_{TOT} &\equiv E[Y_{1i} - Y_{0i}|D_i = 1] \\ &= E\{E[Y_{1i} - Y_{0i}|X_i, D_i = 1]|D_i = 1\} \\ &= E\{E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 1]|D_i = 1\}. \end{aligned}$$

- By virtue of the CIA,

$$E[Y_{0i}|X_i, D_i = 0] = E[Y_{0i}|X_i, D_i = 1].$$

- Therefore

$$\begin{aligned}\delta_{TOT} &= E\{E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 0]|D_i = 1\} \\ &= E[\delta_X|D_i = 1],\end{aligned}$$

- Where

$$\delta_X \equiv E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0],$$

is the difference in mean earnings by veteran status at each value of X_i. At a particular value, say $X_i = x$, we write δ_x.

- In the discrete case, the *matching estimand* can be written

$$E[Y_{1i} - Y_{0i}|D_i = 1] = \sum_x \delta_x P(X_i = x|D_i = 1),$$

where $P(X_i = x | D_i = 1)$ is the probability mass function for X_i given $D_i = 1$. X_i takes on values determined by all possible combinations of the variables it contains.

Matching Estimator of the ATE

- we can just as easily construct the unconditional average treatment effect

$$\delta_{ATE} = E\{E[\text{Y}_{1i}|\text{X}_i, \text{D}_i = 1] - E[\text{Y}_{0i}|\text{X}_i, \text{D}_i = 0]\}$$

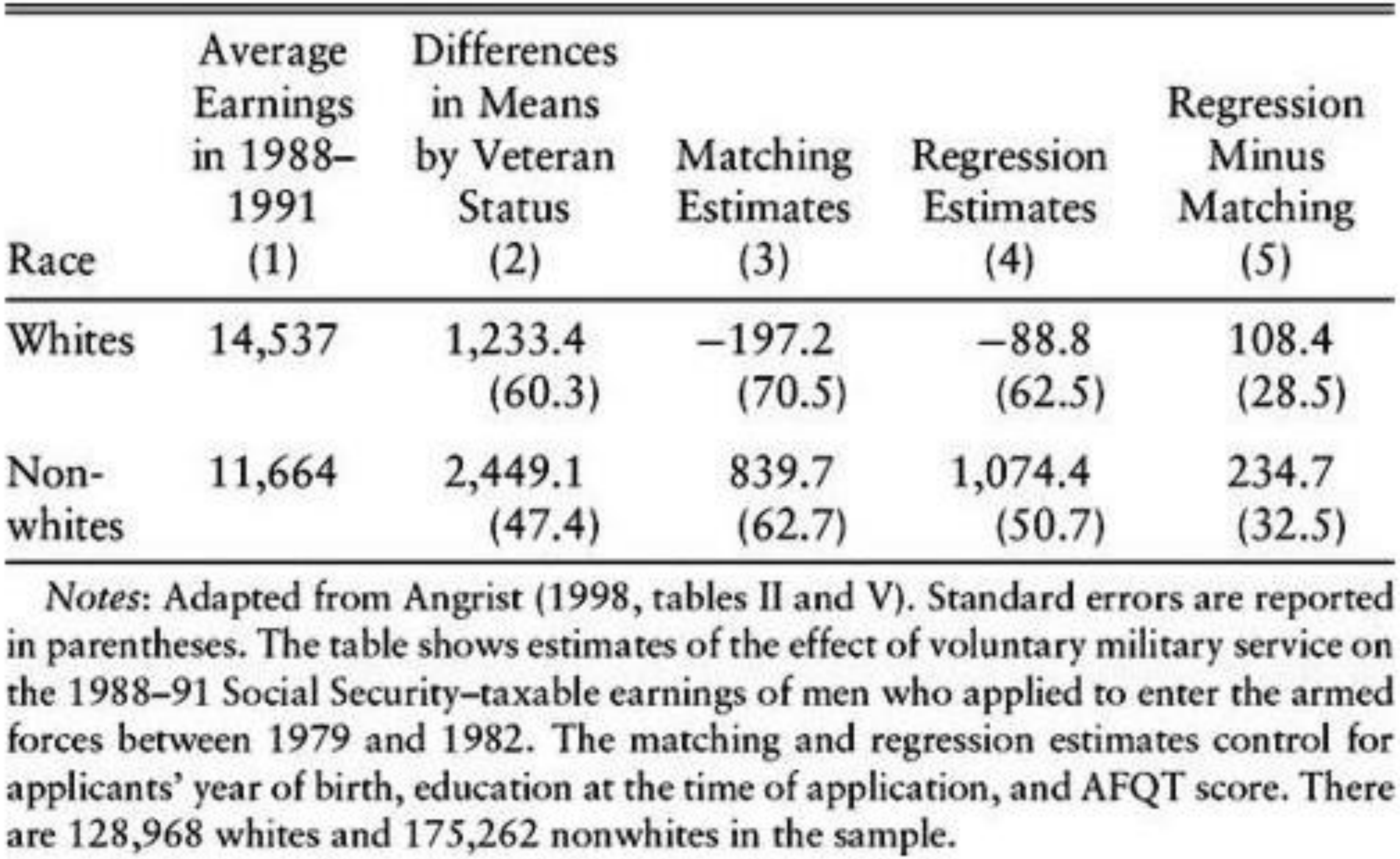
$$= \sum_x \delta_x P(\text{X}_i = x) = E[\text{Y}_{1i} - \text{Y}_{0i}].$$

- This is the expectation of δ_X using the marginal distribution of X_i instead of the distribution among the treated. δ_{TOT} tells us how much the typical soldier gained or lost as a consequence of military service, while δ_{ATE} tells us how much the typical applicant to the military gained or lost (since the Angrist, 1998, population consists of applicants.)

- Proof:

$$
\begin{aligned}
E[Y_{1i} - Y_{0i}] &= E\{E[(Y_{1i} - Y_{0i}) \mid D_i]\} \\
&= P(D_i = 1)E[Y_{1i} - Y_{0i} \mid D_i = 1] + P(D_i = 0)E[Y_{1i} - Y_{0i} \mid D_i = 0] \\
&= P(D_i = 1) \sum_{x \in \{D_i = 1\}_{i=1\ldots N}} \delta_X P(X_i = x \mid D_i = 1) + P(D_i = 0) \sum_{x \in \{D_i = 0\}_{i=1\ldots N}} \delta_X P(X_i = x \mid D_i = 0) \\
&= \sum_{x \in \{D_i = 1\}_{i=1\ldots N}} \delta_X P(X_i = x, D_i = 1) + \sum_{x \in \{D_i = 0\}_{i=1\ldots N}} \delta_X P(X_i = x, D_i = 0) \\
&= \sum_x \delta_X P(X_i = x)
\end{aligned}
$$

TABLE 3.3.1
Uncontrolled, matching, and regression estimates of the effects of voluntary military service on earnings



Race	Average Earnings in 1988–1991 (1)	Differences in Means by Veteran Status (2)	Matching Estimates (3)	Regression Estimates (4)	Regression Minus Matching (5)
Whites	14,537	1,233.4 (60.3)	−197.2 (70.5)	−88.8 (62.5)	108.4 (28.5)
Non-whites	11,664	2,449.1 (47.4)	839.7 (62.7)	1,074.4 (50.7)	234.7 (32.5)

Notes: Adapted from Angrist (1998, tables II and V). Standard errors are reported in parentheses. The table shows estimates of the effect of voluntary military service on the 1988–91 Social Security–taxable earnings of men who applied to enter the armed forces between 1979 and 1982. The matching and regression estimates control for applicants' year of birth, education at the time of application, and AFQT score. There are 128,968 whites and 175,262 nonwhites in the sample.

What is the regression estimand?

- Regression estimates of the effect of voluntary military service, controlling for the same set of covariates that were used to construct the matching estimates. These are estimates of δ_R in the equation

$$Y_i = \sum_x d_{ix}\alpha_x + \delta_R D_i + e_i,$$

- Where $d_{ix} = 1\ [X_i = x]$ is a dummy variable that indicates $X_i = x$, α_x is a regression effect for $X_i = x$, and δ_r is the regression estimand.
- Note that this regression model allows a separate parameter for every value taken on by the covariates. This model can therefore be said to be *saturated-in-x_i*, since it includes a parameter for every value of x_i. It is not fully saturated, however, because there is a single additive effect for di with no $D_i \cdot X_i$ interactions.

Regression Estimator as a sort of matching estimator

- The reason the regression and matching estimates are similar is that regression, too, can be seen as a sort of matching estimator:
- The regression estimand differs from the matching estimands only in the weights used to combine the covariate-specific effects, δ_x, into a single average effect.
- In particular, while matching ***uses the distribution of covariates among the treated*** to weight covariate-specific estimates into an estimate of the effect of treatment on the treated, regression produces a ***variance-weighted*** average of these effects.

- To see this, start by using the regression anatomy formula to write the coefficient on D_i in the regression of Y_i on X_i and D_i as

$$\delta_R = \frac{Cov(Y_i, \tilde{D}_i)}{V(\tilde{D}_i)} = \frac{E[(D_i - E[D_i|X_i])Y_i]}{E[(D_i - E[D_i|X_i])^2]}$$
$$= \frac{E\{(D_i - E[D_i|X_i])E[Y_i|D_i, X_i]\}}{E[(D_i - E[D_i|X_i])^2]}.$$

- The second equality in this set of expressions uses the fact that saturating the model in X_i means $E[D_i\,|X_i\,]$ is linear. Hence, $\tilde{D}_i$, which is defined as the residual from a regression of D_i on X_i , is the difference between D_i and $E[D_i|X_i]$. The third equality uses the fact that the regression of Y_i on D_i and X_i is the same as the regression of Y_i on $E[Y_i|D_i,X_i]$

$$E[Y_i|D_i, X_i] = E[Y_i|D_i = 0, X_i] + \delta_X D_i,$$

- This gives

$$E\{(D_i - E[D_i|X_i])E[Y_i|D_i, X_i]\}$$
$$= E\{(D_i - E[D_i|X_i])E[Y_i|D_i = 0, X_i]\}$$
$$+ E\{(D_i - E[D_i|X_i])D_i\delta_X\}.$$

- The first term on the right-hand side is zero because $E[Y_i|D_i = 0,X_i]$ is a function of X_i only and is therefore uncorrelated with $(D_i - E[D_i\,|X_i\,])$. Similarly, the second term simplifies to

$$E\{(D_i - E[D_i|X_i])D_i\delta_X\} = E\{(D_i - E[D_i|X_i])^2\delta_X\}.$$

- Then, the regression estimand can be shown as

$$\begin{aligned}\delta_R &= \frac{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2\delta_X]}{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2]}\\ &= \frac{E\{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2|\mathrm{X}_i]\delta_X\}}{E\{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2|\mathrm{X}_i]\}} = \frac{E[\sigma_D^2(\mathrm{X}_i)\delta_X]}{E[\sigma_D^2(\mathrm{X}_i)]}\end{aligned}$$

Where

$$\sigma_D^2(\mathrm{X}_i) \equiv E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2|\mathrm{X}_i]$$

is the conditional variance of D_i given X_i . This establishes that the regression model, produces a treatment-variance weighted average of δ_X .

- Because the regressor of interest, D_i, is a dummy variable, one last step can be taken. In this case, , so

$$\delta_R = \frac{\sum_x \delta_x[P(\mathrm{D}_i = 1|\mathrm{X}_i = x)(1 - P(\mathrm{D}_i = 1|\mathrm{X}_i = x))]P(\mathrm{X}_i = x)}{\sum_x [P(\mathrm{D}_i = 1|\mathrm{X}_i = x)(1 - P(\mathrm{D}_i = 1|\mathrm{X}_i = x))]P(\mathrm{X}_i = x)}.$$

- This shows that the regression estimand weights each covariate-specific treatment effect by *$[P(X_i = x|D_i = 1)(1-P(X_i = x|D_i = 1))]P(X_i = x)$*.

- In contrast, the *matching estimand* for the effect of treatment on the treated can be written:

$$E[Y_{1i} - Y_{0i} | D_i = 1] = \sum_x \delta_x P(X_i = x | D_i = 1)$$

$$= \frac{\sum_x \delta_x P(D_i = 1 | X_i = x) P(X_i = x)}{\sum_x P(D_i = 1 | X_i = x) P(X_i = x)},$$

Using the fact that

$$P(X_i = x | D_i = 1) = \frac{P(D_i = 1 | X_i = x) \cdot P(X_i = x)}{P(D_i = 1)}.$$

- The treatment-on-the-treated estimand puts the most weight on covariate cells containing those who are most likely to be treated.

  In contrast, regression puts the most weight on covariate cells where the conditional variance of treatment status is largest. As a rule, treatment variance is maximized when, in other words, for cells where there are equal numbers of treated and control observations.

Remember: Regression Estimand

$$\delta_R = \frac{\sum_x \delta_x [P(D_i = 1 | X_i = x)(1 - P(D_i = 1 | X_i = x))] P(X_i = x)}{\sum_x [P(D_i = 1 | X_i = x)(1 - P(D_i = 1 | X_i = x))] P(X_i = x)}.$$

- Neither the regression nor the covariate-matching estimands give any weight to covariate cells that do not contain both treated and control observations.
- Consider a value of x_i, say x^*, where either no one is treated or everyone is treated. Then, $\delta_x{}^*$ is undefined, while the regression weights, $[p(D_i = 1|X_i = x^*)(1 - p(D_i = 1|X_i = x^*))]$, are zero.
- In the language of the econometric literature on matching, with saturated control for covariates both the regression and matching estimands impose *common support*, that is, they are limited to covariate values where both treated and control observations are found.

II. Control for Covariates Using the Propensity Score

- **Theorem: The Propensity Score Theorem**

Suppose the CIA holds such that . $\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp D_i | X_i$. *Then* $\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp D_i | p(X_i)$.

Where P(X_i) is the propensity score, defined *as* $p(X_i) \equiv E[D_i | X_i] = P[D_i = 1 | X_i]$

Proof. It's enough to show that $P[D_i = 1 | Y_{ji}, p(X_i)]$ does not depend on Y_{ji} for $j = 0, 1$:

$$\begin{aligned} P[D_i = 1 | Y_{ji}, p(X_i)] &= E[D_i | Y_{ji}, p(X_i)] \\ &= E\{E[D_i | Y_{ji}, p(X_i), X_i] | Y_{ji}, p(X_i)\} \\ &= E\{E[D_i | Y_{ji}, X_i] | Y_{ji}, p(X_i)\} \\ &= E\{E[D_i | X_i] | Y_{ji}, p(X_i)\}, \text{ by the CIA.} \end{aligned}$$

But $E\{E[D_i | X_i] | Y_{ji}, p(X_i)\} = E\{p(X_i) | Y_{ji}, p(X_i)\}$, which is clearly just $p(X_i)$.

Several comments

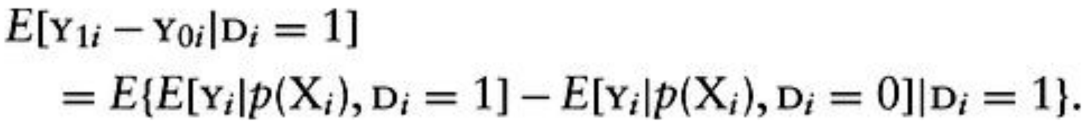
- The propensity score theorem says that if potential outcomes are independent of treatment status conditional on a multivariate covariate vector X_i, then potential outcomes are independent of treatment status conditional on a scalar function of covariates, *the propensity score*, defined as $p(X_i) \equiv E[D_i|X_i] = P[D_i = 1|X_i]$.

- The propensity score theorem says that you need only control for covariates that affect the probability of treatment. But it also says something more: the only covariate you really need to control for is the probability of treatment itself.

- In practice, the propensity score theorem is usually used for estimation in two steps:
  - First, $p(x_i)$ is estimated using some kind of parametric model, say, logit or probit.
  - Then estimates of the effect of treatment are computed either by matching on the estimated score from this first step or using a weighting scheme described below

Two lines of implementing matching estimator

- Direct propensity score matching works in the same way as covariate matching except that we match on the score instead of the covariates directly. By the propensity score theorem and the CIA,



$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} | \mathrm{D}_i = 1]$$
$$= E\{E[\mathrm{Y}_i | p(\mathrm{X}_i), \mathrm{D}_i = 1] - E[\mathrm{Y}_i | p(\mathrm{X}_i), \mathrm{D}_i = 0] | \mathrm{D}_i = 1\}.$$

- Estimates of the effect of treatment on the treated can therefore be obtained by stratifying on an estimate of p(X$_i$) and substituting conditional sample averages for expectations or by matching each treated observation to controls with similar values of the propensity score (both of these approaches were used by Dehejia and Wahba, 1999).

** Nonparametric weighting estimator (NOT REQUIRED)

A model-based or nonparametric estimate of E[Y_i|p(X_i),D_i] (Heckman, Ichimura, and Todd, 1998)

Estimates of the average treatment effect from the sample analog of {Newey (1990) and Robins, Mark, and Newey (1992)}

ATE

$$E[Y_{1i}-Y_{0i}] = E\left[\frac{Y_iD_i}{p(X_i)} - \frac{Y_i(1-D_i)}{1-p(X_i)}\right]$$
$$= E\left[\frac{(D_i - p(X_i))Y_i}{p(X_i)(1-p(X_i))}\right].$$

Proof：

$$E[\tfrac{Y_iD_i}{p(X_i)}] \overset{LIE}{=} E\{E[\tfrac{Y_iD_i}{p(X_i)}|X_i]\};\ E[\tfrac{Y_iD_i}{p(X_i)}|X_i] = \frac{E[Y_i|D_i=1,X_i]p(X_i)}{p(X_i)} = E[Y_{1i}|D_i = 1, X_i] = E[Y_{1i}|X_i]$$

$$= E\left\{E\left[\frac{Y_iD_i}{p(X_i)}\Big|D_i, X_i\right]\Big|X_i\right\} = E\left\{E\left[\frac{Y_{1i}}{p(X_i)}\Big|D_i=1, X_i\right]p(X_i)\Big|X_i\right\}$$

(NOT REQUIRED)

- We can similarly calculate the effect of treatment on the treated from the sample analog of:

ATT

$$E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_i=1]=E\left[\frac{(\mathrm{D}_i-p(\mathrm{X}_i))\mathrm{Y}_i}{(1-p(\mathrm{X}_i))P(\mathrm{D}_i=1)}\right].$$

- The idea that you can correct for nonrandom sampling via weighting by the reciprocal of the probability of selection dates back to Horvitz and Thompson (1952). Of course, to make this approach feasible, and for the resulting estimates to be consistent, we need a consistent estimator of $p(X_i)$.

(NOT REQUIRED)

- Consider again the regression estimand, δ_R, for the population regression of Y_i on D_i, controlling for a saturated model for covariates. This estimand can be written

$$\delta_R = \frac{E[(D_i - p(X_i))Y_i]}{E[p(X_i)(1-p(X_i))]}.$$

和 OLS 的 δ_R 一样. $\delta_R = \frac{cov(\cdots)}{var(\cdots)} = \frac{E[(D_i - p)Y_i]}{E[p(1-p)]}$

- Hirano, Imbens, and Ridder (2003)： the class of weighted average estimands

$$E\left\{g(X_i)\left[\frac{Y_i D_i}{p(X_i)} - \frac{Y_i(1-D_i)}{(1-p(X_i))}\right]\right\}$$

where $g(X_i)$ is a known weighting function.

- For the average treatment effect, set $g(X_i) = 1$;
- for the effect on the treated, set $g(X_i) = \frac{p(X_i)}{P(D_i=1)}$;
- for regression, set (NOT REQUIRED)

$$g(X_i) = \frac{p(X_i)(1-p(X_i))}{E[p(X_i)(1-p(X_i))]}.$$

Estimate $p(X_i)$

- A big question here is how best to model and estimate $p(X_i)$, or how much smoothing or stratification to use when estimating $E[Y_i|p(X_i), D_i]$, especially if the covariates are continuous.

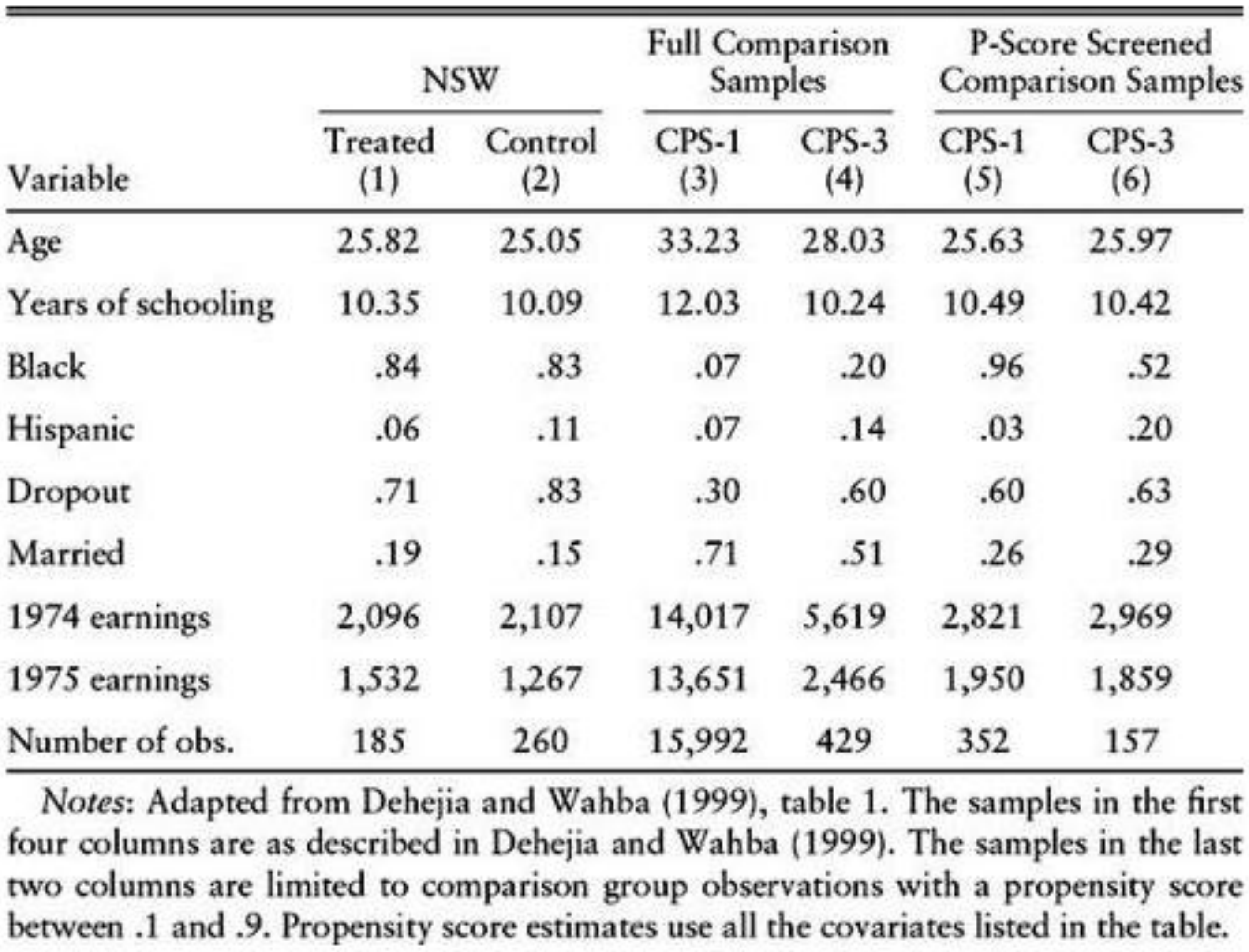
- The regression analog of this question is how to parameterize the control variables (e.g., polynomials or main effects and interaction terms if the covariates are coded as discrete).

- The answer to this is inherently application-specific. A growing empirical literature suggests that a logit model for the propensity score with a few polynomial terms in continuous covariates works well in practice, though this cannot be a theorem, and inevitably, some experimentation will be required (see, e.g., Dehejia and Wahba, 1999).

Issues on asymptotic efficiency

- First, from the point of view of asymptotic efficiency, there is usually a cost to matching on the propensity score instead of full covariate matching. We can get lower asymptotic standard errors by matching on any covariate that explains outcomes, whether or not it turns up in the propensity score. Hahn’s (1998)
- A philosophical argument is that the propensity score rightly focuses researcher attention on models for treatment assignment, something about which we may have reasonably good information, instead of the typically more complex and mysterious process of determining outcomes
- Angrist and Hahn (2004): though there is no asymptotic efficiency gain from the use of estimators based on the propensity score, there will often be a gain in precision in finite samples.
- Hirano, Imbens, and Ridder (2003): Even though estimates of treatment effects based on a known propensity score are inefficient, for models with continuous covariates, a Horvitz-Thompson-type weighting estimator is efficient when the weighting scheme uses a nonparametric estimate of the score.

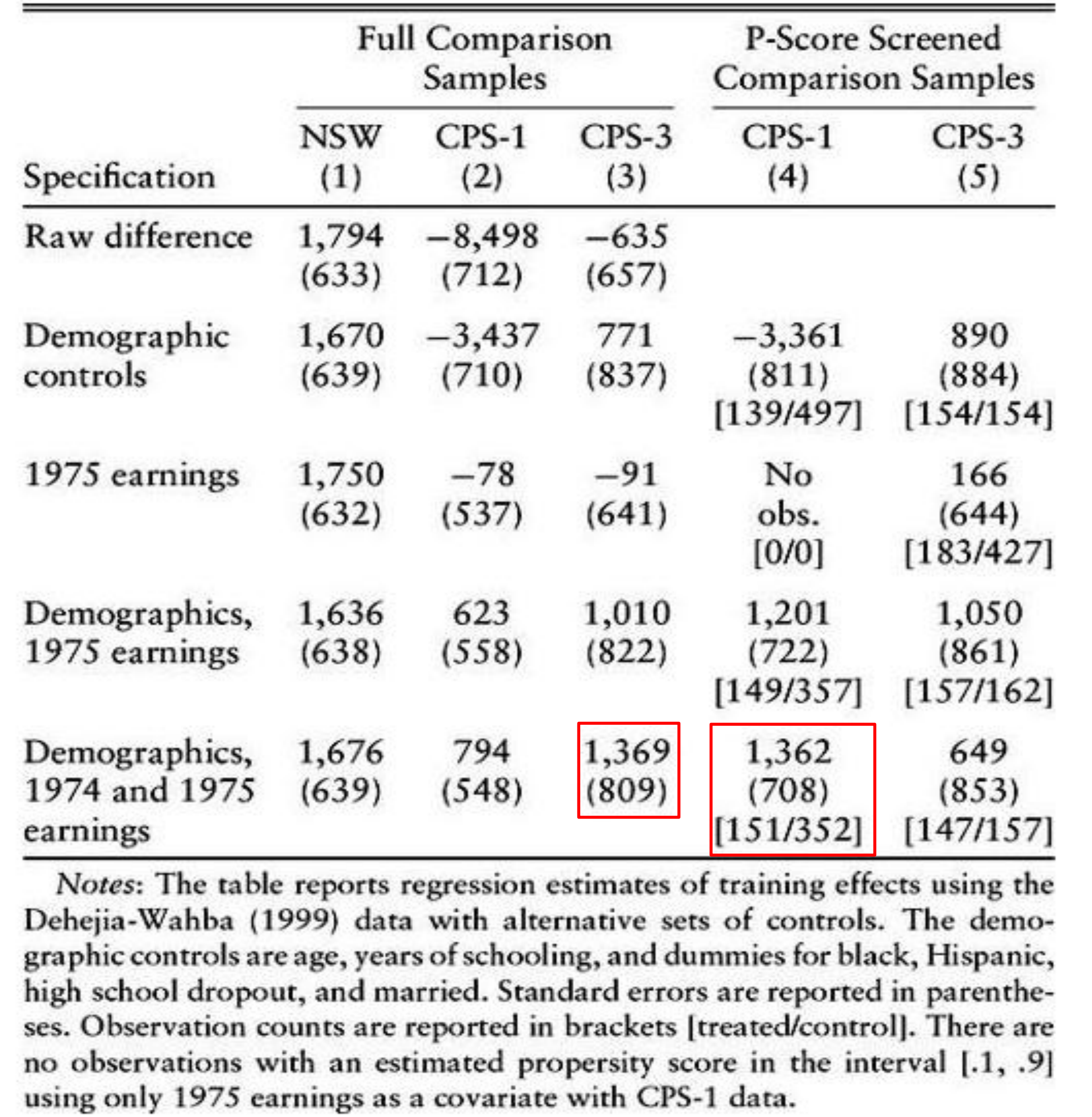
TABLE 3.3.2 Covariate means in the NSW and observational control samples



	NSW		Full Comparison Samples		P-Score Screened Comparison Samples	
Variable	Treated (1)	Control (2)	CPS-1 (3)	CPS-3 (4)	CPS-1 (5)	CPS-3 (6)
Age	25.82	25.05	33.23	28.03	25.63	25.97
Years of schooling	10.35	10.09	12.03	10.24	10.49	10.42
Black	.84	.83	.07	.20	.96	.52
Hispanic	.06	.11	.07	.14	.03	.20
Dropout	.71	.83	.30	.60	.60	.63
Married	.19	.15	.71	.51	.26	.29
1974 earnings	2,096	2,107	14,017	5,619	2,821	2,969
1975 earnings	1,532	1,267	13,651	2,466	1,950	1,859
Number of obs.	185	260	15,992	429	352	157

Notes: Adapted from Dehejia and Wahba (1999), table 1. The samples in the first four columns are as described in Dehejia and Wahba (1999). The samples in the last two columns are limited to comparison group observations with a propensity score between .1 and .9. Propensity score estimates use all the covariates listed in the table.

TABLE 3.3.3 Regression estimates of NSW training effects using alternative controls



	Full Comparison Samples			P-Score Screened Comparison Samples	
Specification	NSW (1)	CPS-1 (2)	CPS-3 (3)	CPS-1 (4)	CPS-3 (5)
Raw difference	1,794 (633)	−8,498 (712)	−635 (657)		
Demographic controls	1,670 (639)	−3,437 (710)	771 (837)	−3,361 (811) [139/497]	890 (884) [154/154]
1975 earnings	1,750 (632)	−78 (537)	−91 (641)	No obs. [0/0]	166 (644) [183/427]
Demographics, 1975 earnings	1,636 (638)	623 (558)	1,010 (822)	1,201 (722) [149/357]	1,050 (861) [157/162]
Demographics, 1974 and 1975 earnings	1,676 (639)	794 (548)	1,369 (809)	1,362 (708) [151/352]	649 (853) [147/157]

Notes: The table reports regression estimates of training effects using the Dehejia-Wahba (1999) data with alternative sets of controls. The demographic controls are age, years of schooling, and dummies for black, Hispanic, high school dropout, and married. Standard errors are reported in parentheses. Observation counts are reported in brackets [treated/control]. There are no observations with an estimated propersity score in the interval [.1, .9] using only 1975 earnings as a covariate with CPS-1 data.

Implementing Matching: (Subclassification) Estimator

1. *$(Y^1,Y^0) \perp D \mid X$* (conditional independence)

2. $0 < Pr(D = 1 \mid X) < 1$ with probability one (common support)

$$
\begin{aligned}
E[Y^1 - Y^0 \mid X] &= E[Y^1 - Y^0 \mid X, D = 1] \\
&= E[Y^1 \mid X, D = 1] - E[Y^0 \mid X, D = 0] \\
&= E[Y \mid X, D = 1] - E[Y \mid X, D = 0]
\end{aligned}
$$

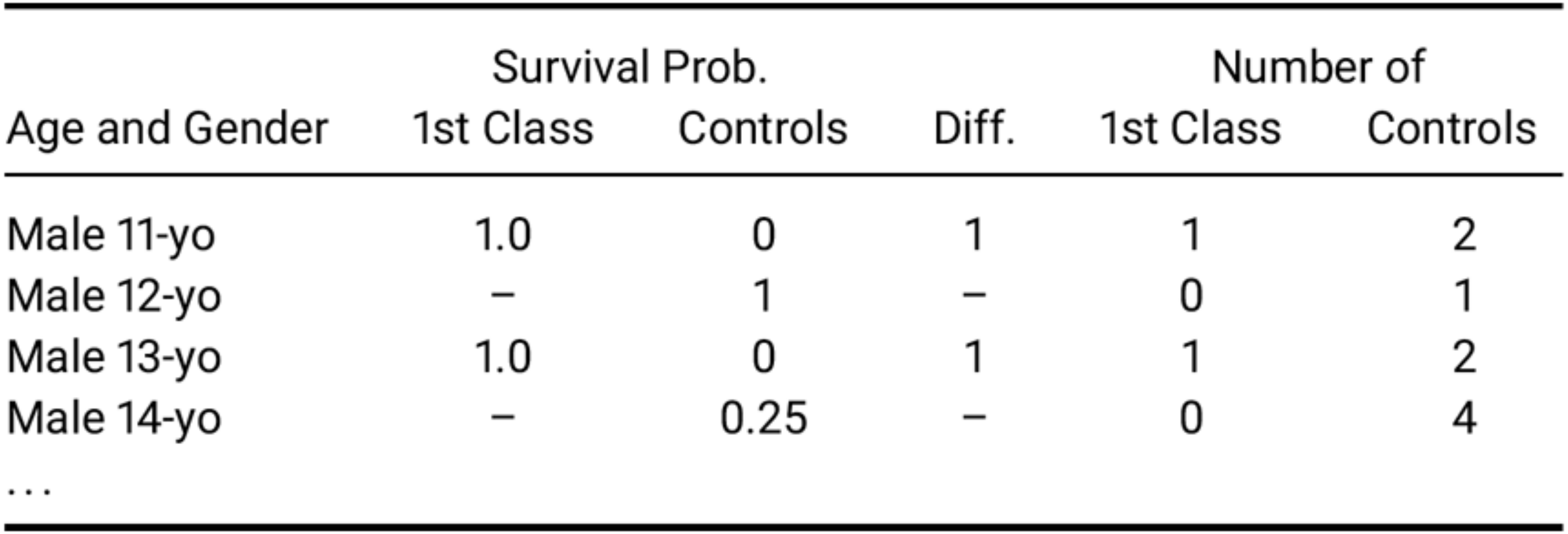
$$\widehat{\delta_{ATE}} = \int \Big(E[Y \mid X, D = 1] - E[Y \mid X, D = 0]\Big) d\Pr(X)$$

$$\widehat{\delta}_{ATT} = \sum_{k=1}^{K} \Big(\underset{\uparrow \atop (D_i=1,\, X_i=x_k)}{\overline{Y}^{1,k}} - \overset{(D_i=0,\, X_i=x_k) \atop \downarrow}{\overline{Y}^{0,k}}\Big) \times \Big(\frac{N_T^k}{N_T}\Big) \xleftarrow{\text{sub.}} \text{fraction.}$$

Subclassification exercise

- *Titanic* data set
- Using the data set on the *Titanic*, we calculate a simple difference in mean outcomes (SDO)
- Next we use subclassification weighting to control for these confounders
- Here are the steps that will entail:
  - 1. Stratify the data into four groups: young males, young females, old males, and old females.
  - 2. Calculate the difference in survival probabilities for each group.
  - 3. Calculate the number of people in the non-first-class groups and divide by the total number of non-first-class population. These are our strata-specific weights.
  - 4. Calculate the weighted average survival rate using the strata weights.

Curse of dimensionality



	Survival Prob.			Number of	
Age and Gender	1st Class	Controls	Diff.	1st Class	Controls
Male 11-yo	1.0	0	1	1	2
Male 12-yo	–	1	–	0	1
Male 13-yo	1.0	0	1	1	2
Male 14-yo	–	0.25	–	0	4
. . .					

Exact Matching

- A simple matching estimator

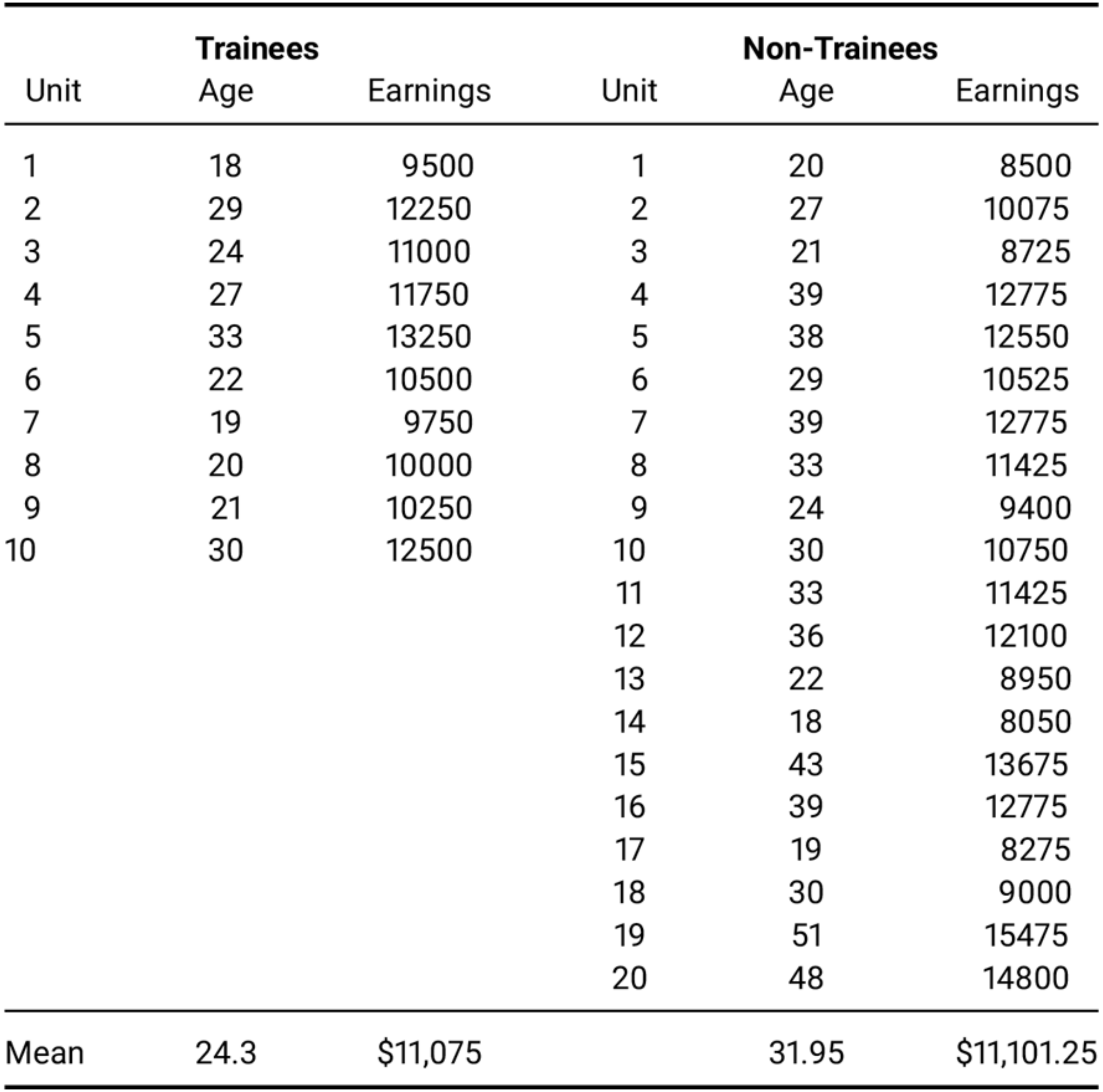
$$\widehat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$$

- where $Y_j({}_i)$ is the *j*th unit matched to the *i*th unit based on the *j*th being "closest to" the *i*th unit for some *X* covariate.

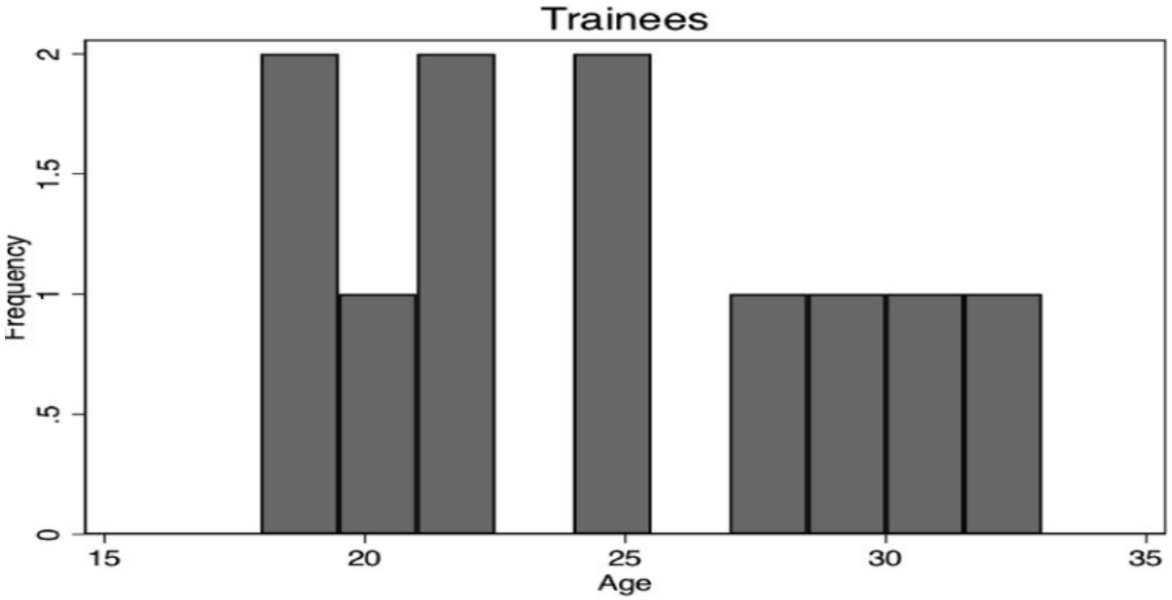
$$\widehat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \left[\frac{1}{M} \sum_{m=1}^{M} Y_{j_m(1)} \right] \right)$$

- Training example with exact matching (Stata illustration)

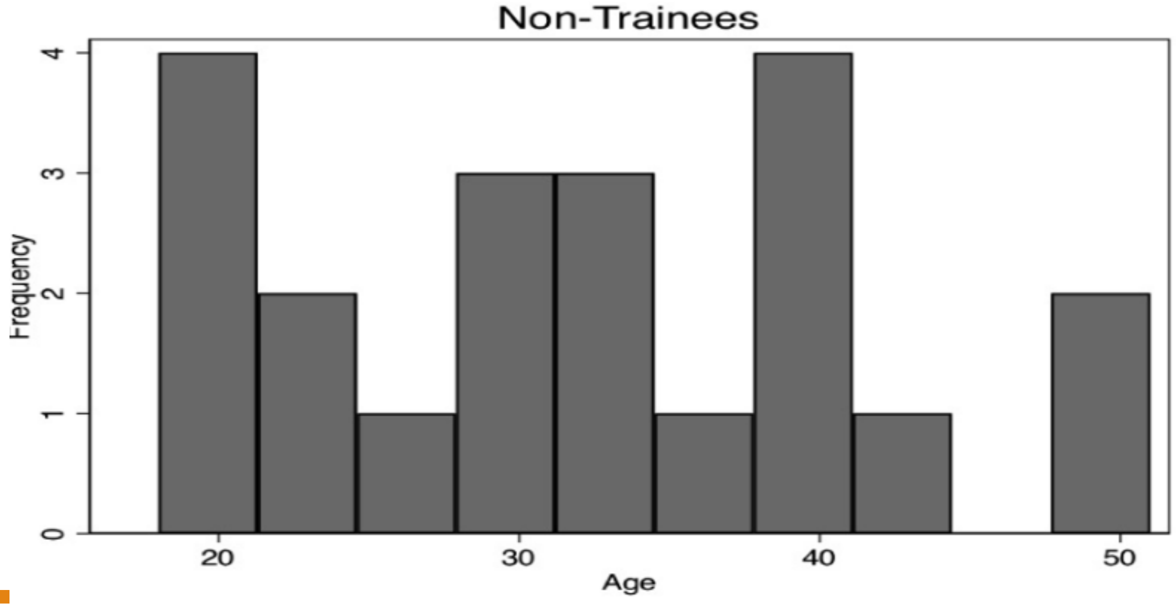
Training example with exact matching.



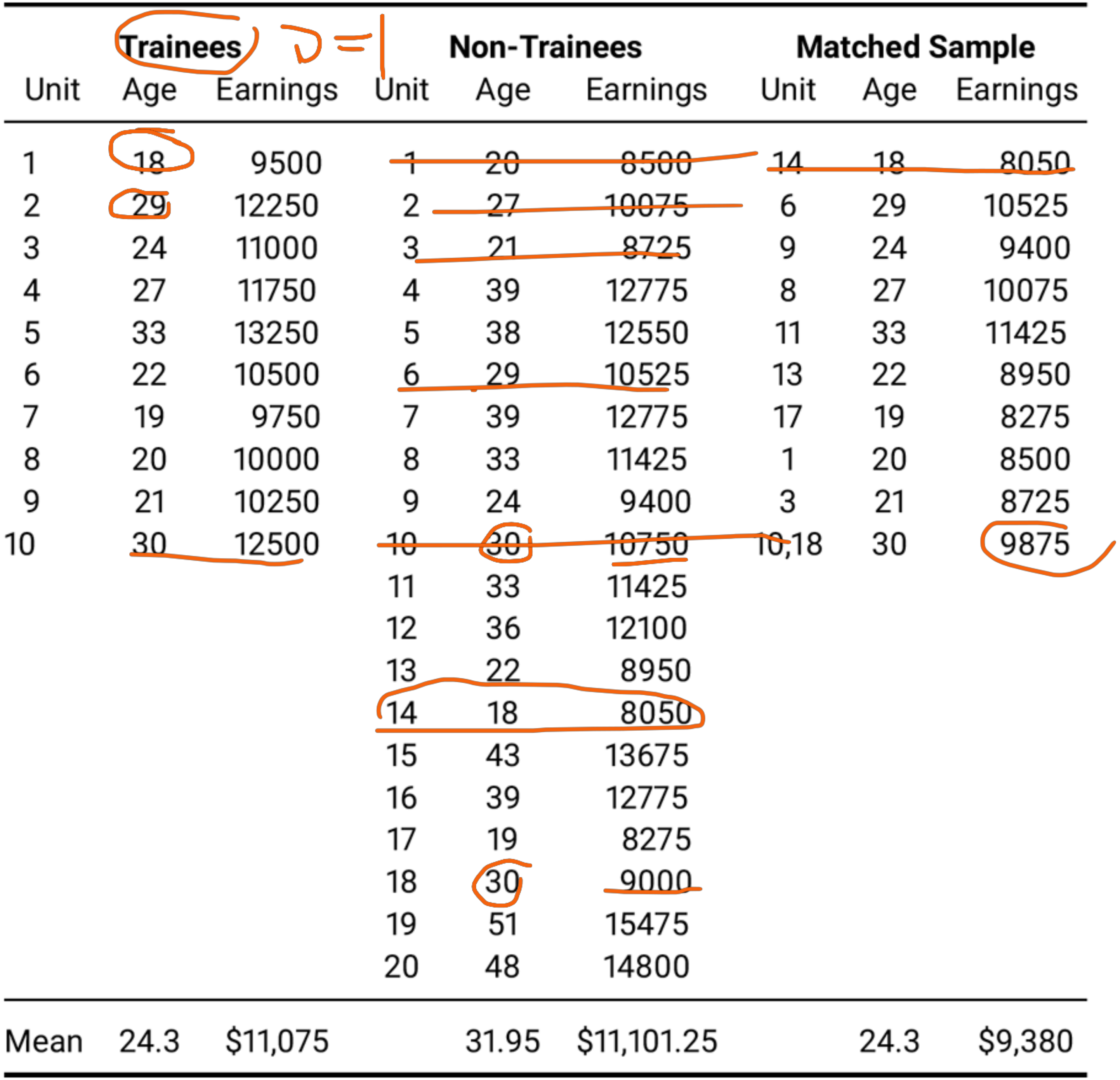
	Trainees			Non-Trainees	
Unit	Age	Earnings	Unit	Age	Earnings
1	18	9500	1	20	8500
2	29	12250	2	27	10075
3	24	11000	3	21	8725
4	27	11750	4	39	12775
5	33	13250	5	38	12550
6	22	10500	6	29	10525
7	19	9750	7	39	12775
8	20	10000	8	33	11425
9	21	10250	9	24	9400
10	30	12500	10	30	10750
			11	33	11425
			12	36	12100
			13	22	8950
			14	18	8050
			15	43	13675
			16	39	12775
			17	19	8275
			18	30	9000
			19	51	15475
			20	48	14800
Mean	24.3	$11,075		31.95	$11,101.25


Trainees
Frequency
0
.5
1
1.5
2
15
20
25
30
35
Age

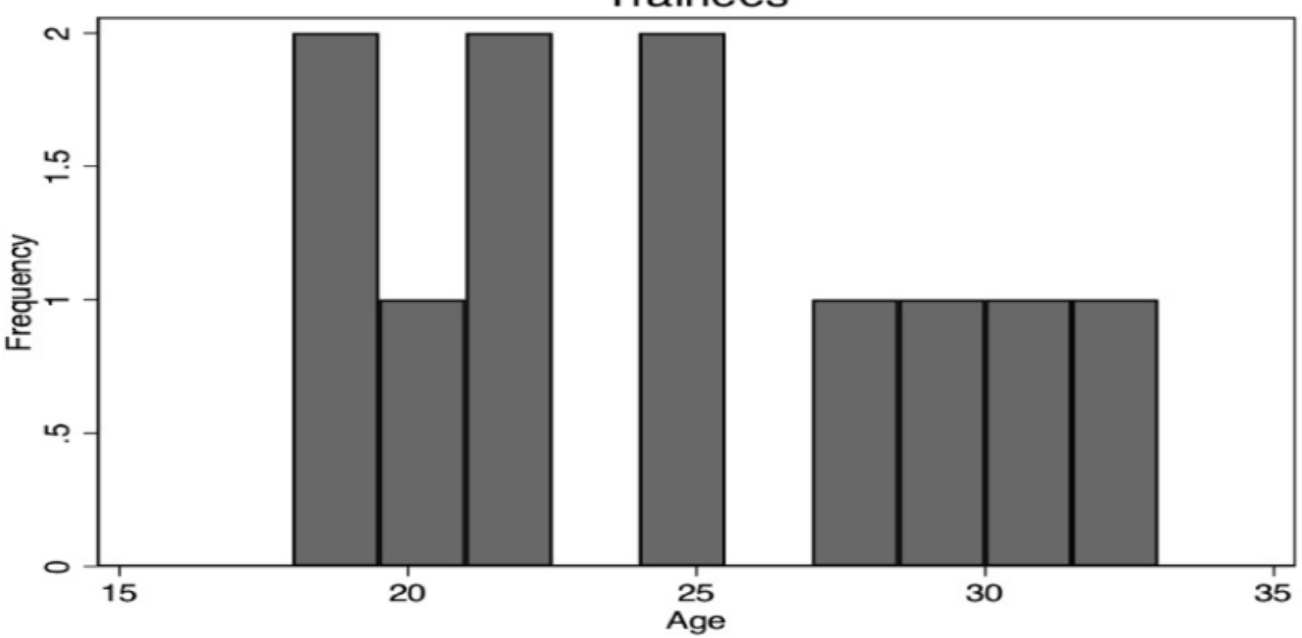
(a)

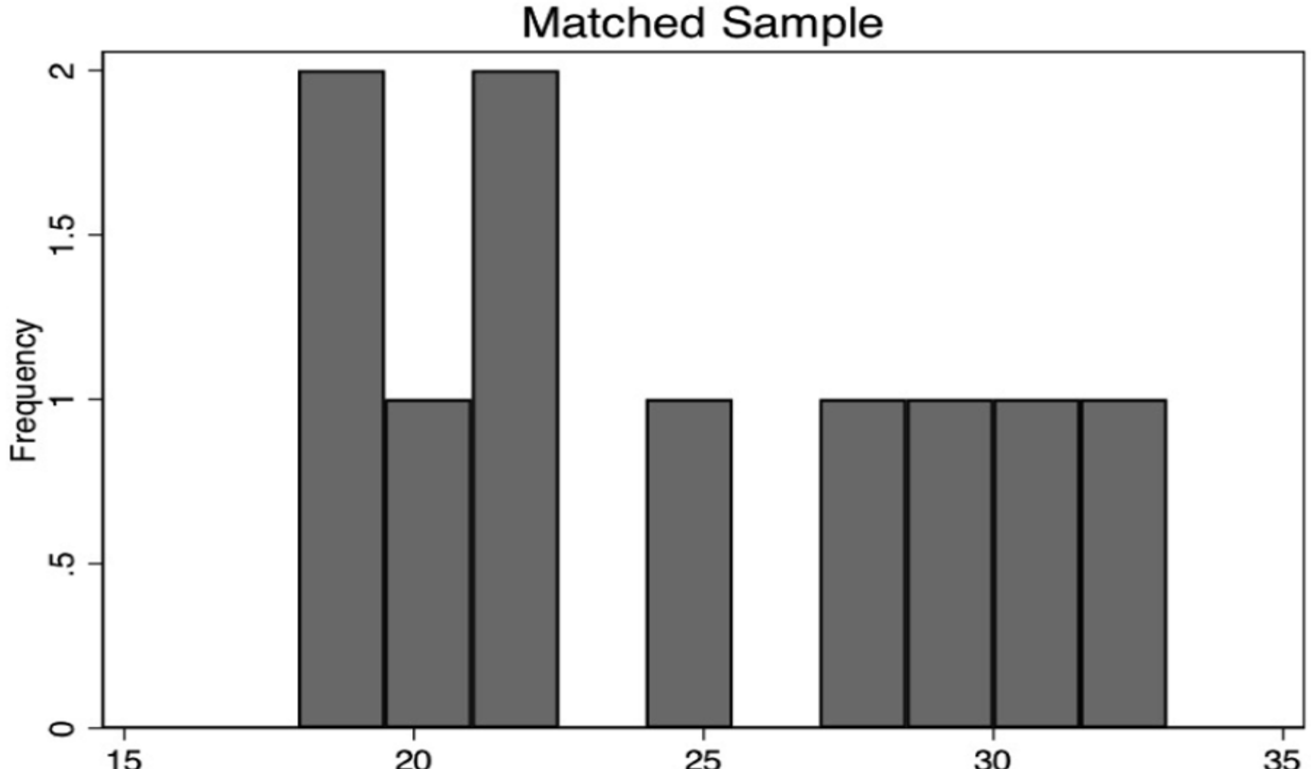

Non-Trainees
Frequency
0
1
2
3
4
20
30
40
50
Age

(b)



	Trainees D=1		Non-Trainees			Matched Sample		
Unit	Age	Earnings	Unit	Age	Earnings	Unit	Age	Earnings
1	18	9500	~~1~~	~~20~~	~~8500~~	~~14~~	~~18~~	~~8050~~
2	29	12250	~~2~~	~~27~~	~~10075~~	6	29	10525
3	24	11000	~~3~~	~~21~~	~~8725~~	9	24	9400
4	27	11750	4	39	12775	8	27	10075
5	33	13250	5	38	12550	11	33	11425
6	22	10500	~~6~~	~~29~~	~~10525~~	13	22	8950
7	19	9750	7	39	12775	17	19	8275
8	20	10000	8	33	11425	1	20	8500
9	21	10250	9	24	9400	3	21	8725
10	30	12500	~~10~~	~~30~~	~~10750~~	10,18	30	9875
			11	33	11425			
			12	36	12100			
			13	22	8950			
			14	18	8050			
			15	43	13675			
			16	39	12775			
			17	19	8275			
			18	30	9000			
			19	51	15475			
			20	48	14800			
Mean	24.3	$11,075		31.95	$11,101.25		24.3	$9,380


Trainees
Frequency
0
.5
1
1.5
2
15
20
25
30
35
Age


Matched Sample
Frequency
0
.5
1
1.5
2
15
20
25
30
35

Approximate Matching

what if you couldn’t find another unit with that exact same value?

Nearest neighbor covariate matching.

What does it mean for one unit’s covariate to be “close” to someone else’s? Furthermore, what does it mean when there are multiple covariates with measurements in multiple dimensions?

What is the distance

Euclidean distance

$$\begin{aligned}||X_i - X_j|| &= \sqrt{(X_i - X_j)'(X_i - X_j)} \\ &= \sqrt{\sum_{n=1}^{k}(X_{ni} - X_{nj})^2}\end{aligned}$$

the normalized Euclidean distance

$$||X_i - X_j|| = \sqrt{(X_i - X_j)'\widehat{V}^{-1}(X_i - X_j)}$$

$$\widehat{V}^{-1} = \begin{pmatrix} \widehat{\sigma}_1^2 & 0 & \dots & 0 \\ 0 & \widehat{\sigma}_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \widehat{\sigma}_k^2 \end{pmatrix}$$

the *Mahalanobis* distance

$$||X_i - X_j|| = \sqrt{(X_i - X_j)' \widehat{\Sigma}_X^{-1} (X_i - X_j)}$$

where the weighting is the sample variance-covariance matrix of X.

Implementation of Propensity Score Matching

Model Choice

It makes little difference between choosing between logit or Probit. Probit is preferred to by economists. The linear probability model is not recommended for its well-known shortcomings.

Multiple treatment case (Imbens, 2000; Lechner,2001): choice of model becomes important.

Variables Choice

Matching strategy builds on the CIA, which requires choosing a set of variables X that credibly satisfy this condition. The better and more informative the data are, the easier it is to credibly justify the CIA and the matching procedure.

- △ Include all variables that affect both participation decision and outcomes. And only the variables that influence simultaneously the participation decision and the outcome variable should be included. Smith and Todd (2005), Sianesi (2004)

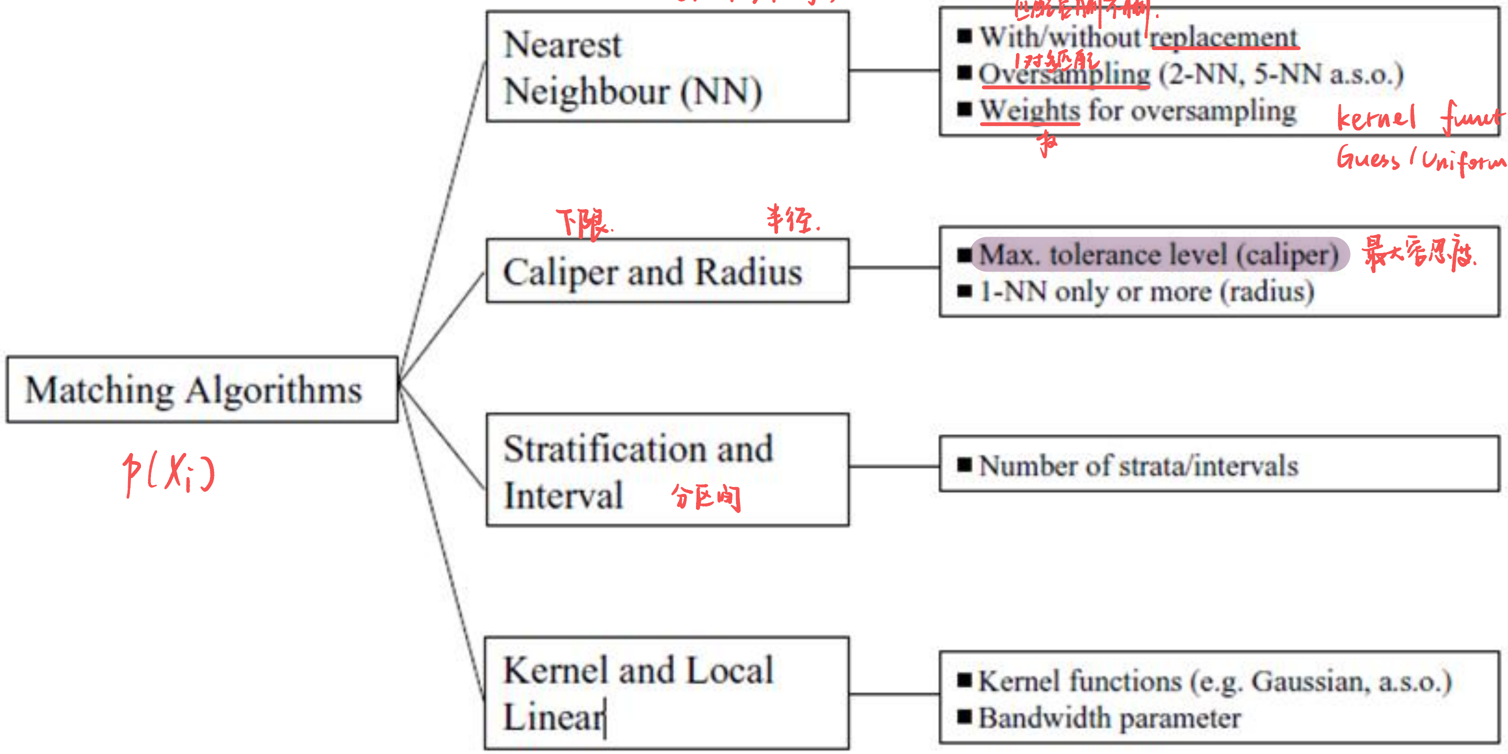
不受D取值影响.

- △ Only variables that are unaffected by participation (or the anticipation of it) should be included in the model.
- Heckman et al. (1999): the data for participants and nonparticipants should stem from the same data sources (e.g. the same questionnaire)
- It should also be clear that "too good" data is not helpful either. If P(X)=0 or P(X)=1 for some value of X, then we cannot use matching conditional on those X values to estimate a treatment effect, because persons with such characteristics either always or never receive treatment, hence the common support assumption is violated.

e) In case of uncertainty of the proper specification, whether it is better to include too many rather than too few variables. Bryson, Dorsett, and Purdon(2002) gave two reasons why over-parameterized model should be avoided. [1] Exacerbate the support problem; [2] increase variance. Confirmed by Augurzky and Schmidt (2000).

f) Including instruments variables is a bad idea. 不要放IV、

Choosing a Matching Algorithm



Matching Algorithms

$p(X_i)$

Nearest Neighbour (NN)

min $d(P(X_i), P(X_j))$

- With/without replacement
- Oversampling (2-NN, 5-NN a.s.o.)
- Weights for oversampling

匹配后删不删.

1对多匹配

权

kernel function

Guess / Uniform / ...

Caliper and Radius

下限. 半径.

- Max. tolerance level (caliper)
- 1-NN only or more (radius)

最大容忍度.

Stratification and Interval 分区间

- Number of strata/intervals

Kernel and Local Linear

Local Linear

- Kernel functions (e.g. Gaussian, a.s.o.)
- Bandwidth parameter

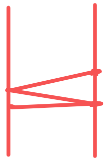


Figure 2. Different Matching Algorithms.

Nearest neighbor matching (NN)

- This is the most common form of matching in the statistical literature. For the treatment on the treated effect parameter, each treated person is matched to the nearest untreated person in terms of propensity score.

- Matching with replacement or not

That is, should a given untreated observation form the counterfactual for more than one treated observation?

  - Matching without replacement can yield very bad matches if the number of comparison (D=0) observations comparable to the treated observations is small.
  - Matching without replacement keeps variance low at the cost of potential bias, while matching with replacement keeps bias low at the cost of a larger variance.
  - Matching without replacement is order-dependent. (Hansen, 2004, JASA, "optimal matching")

How many nearest neighbors?

- There is a tradeoff between bias and variance: matching just one nearest neighbor minimizes bias as, hopefully, all matches are close matches.
- Matching using additional nearest neighbors increases the bias, as the marginal observations are necessarily worse matches, but decreases the variance because more information is being used to construct the counterfactual for each treated person.

Caliper and Radius Matching

- NN matching faces the risk of bad matches if the closest neighbor is far away. This can be avoided by imposing a tolerance level on the maximum propensity score distance (caliper). Hence, caliper matching is one form of imposing a common support condition.
- Applying caliper matching means that an individual from the comparison group is chosen as a matching partner for a treated individual that lies within the caliper ('propensity range') and is closest in terms of propensity score. (**Caliper+1 NN**)--the variance of the estimates increases
- Dehejia and Wahba (2002: radius matching.

Use not only the NN within each caliper but all of the comparison members within the caliper. A benefit of this approach is that it uses only as many comparison units as are available within the caliper and, therefore, allows for the usage of extra (fewer) units when good matches are (not) available. Hence, it shares the attractive feature of oversampling mentioned above but avoids the risk of bad matches. (**Caliper+ Radius**)

Choice of calipers, kernels

Stratification or Interval Matching

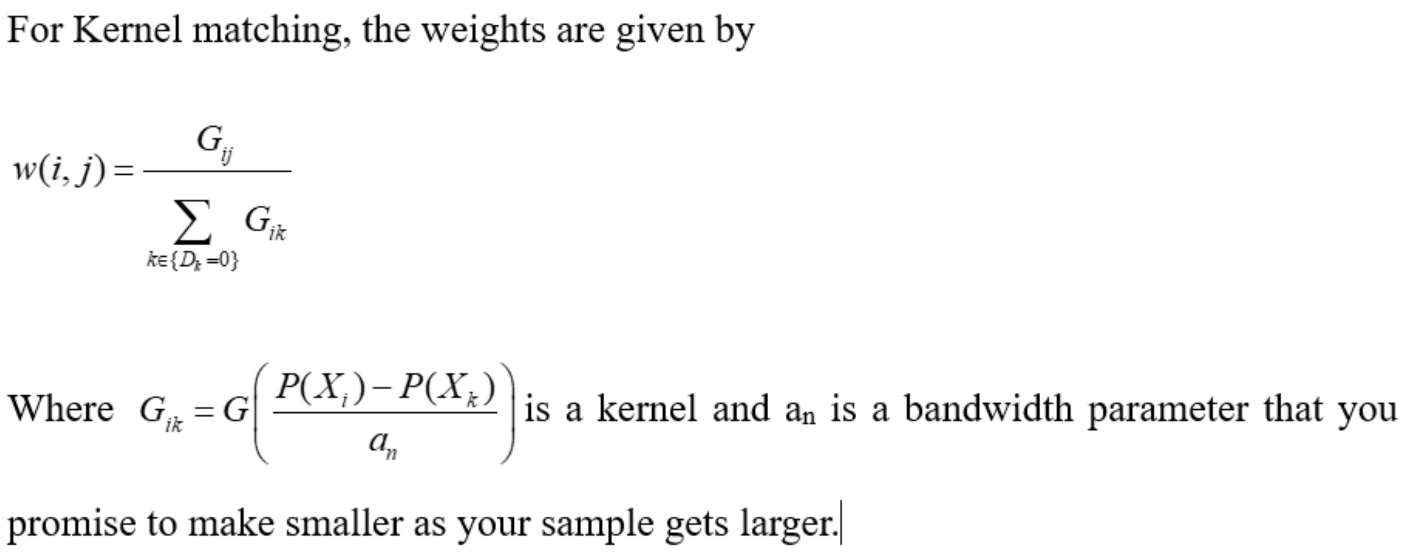
- Dividing the range of propensity score (with common support) into a fixed number of intervals
- Taking the difference between the mean outcomes of the treated and untreated units in each interval, get the impact within each intervals (strata)
- A weighted average of the interval-specific impacts based on the distribution of the treated units among the intervals estimates the ATT. A weighted average using all of the observations estimates the ATE.

Kernel matching

Kernel matching (KM) and local linear matching (LLM) use weighted averages of (nearly) all individuals in the control group to construct the counterfactual outcome.

Two issues must be taken into consideration:

Choice of kernel: normal, triangle, and Epanechnikov kernel



For Kernel matching, the weights are given by

$$w(i,j) = \frac{G_{ij}}{\sum_{k \in \{D_k = 0\}} G_{ik}}$$

Where $G_{ik} = G\left(\frac{P(X_i) - P(X_k)}{a_n}\right)$ is a kernel and a_n is a bandwidth parameter that you promise to make smaller as your sample gets larger.

Local linear or local polynomial matching

- Smith and Todd (2005), *Journal of Econometrics*
- The difference between KM and LLM is that the latter includes in addition to the intercept a linear term in the propensity score of a treated individual. This is an advantage whenever comparison group observations are distributed asymmetrically around the treated observation, e.g. at boundary points, or when there are gaps in the propensity score distribution.

How to choose matching algorithm

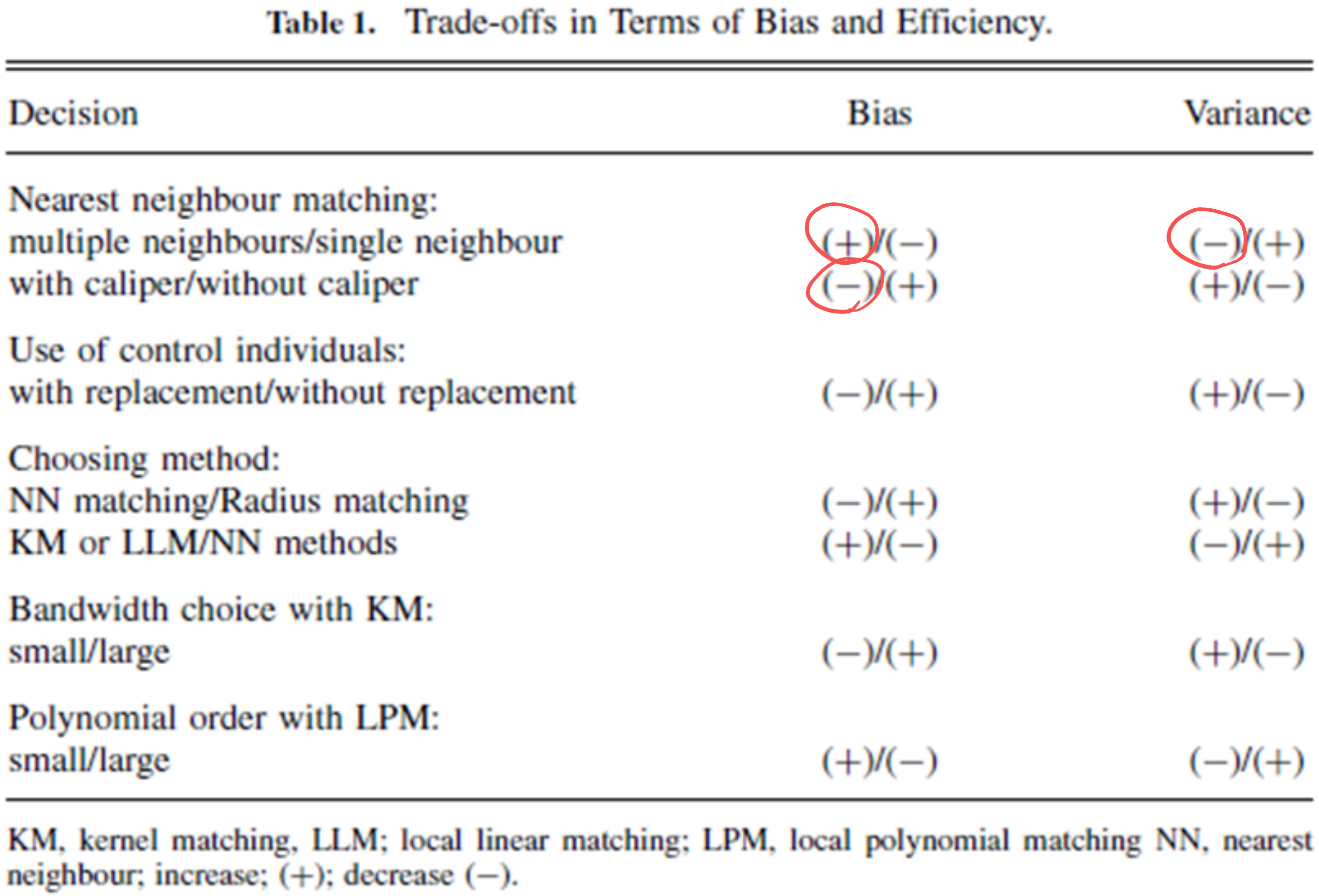


Table 1. Trade-offs in Terms of Bias and Efficiency.

Decision	Bias	Variance
Nearest neighbour matching:		
multiple neighbours/single neighbour	(+)/(−)	(−)/(+)
with caliper/without caliper	(−)/(+)	(+)/(−)
Use of control individuals:		
with replacement/without replacement	(−)/(+)	(+)/(−)
Choosing method:		
NN matching/Radius matching	(−)/(+)	(+)/(−)
KM or LLM/NN methods	(+)/(−)	(−)/(+)
Bandwidth choice with KM:		
small/large	(−)/(+)	(+)/(−)
Polynomial order with LPM:		
small/large	(+)/(−)	(−)/(+)

KM, kernel matching, LLM; local linear matching; LPM, local polynomial matching NN, nearest neighbour; increase; (+); decrease (−).

Rules of Thumb

- If comparison observations are few, single nearest neighbor matching without replacement is a bad idea.
- If comparison observations are many and are evenly distributed, multiple nearest neighbor matching
- If comparison observations are many but asymmetrically distributed, kernel matching
- If many observations have P(X) near zero or one, local linear matching.
- Choice of matching algorithm always means a trade-off between bias and variance.
- Sensitivity analysis

Specification Test

- *balancing property* of the propensity score

$$\Pr\left(X \mid |D = 1, p(X)\right) = \Pr\left(X \mid D = 0, p(X)\right)$$

- This is a testable assumption and should be tested

PSM + DID 还是存在 OLS 内生性问题.

DID + PSM ✓

DID 有效，PSM能改进其估计效果.

Propensity score matching with difference-in-differences (PSMDD)

- If we have panel data, we can combine PSM with difference-in-differences to get a PSMDD estimator of ATT.

先match，再作差.　被解释变量换为两期差异.

$$ATT^{PSMDD}=E_{P(X_i)|D_i\underline{=1}}\{E(Y_{\underline{1}it_1}-Y_{\underline{1}it_0}\mid \underline{D_i=1,}P(X_i))-E(Y_{\underline{0}it_1}-Y_{\underline{0}it_0}\mid \underline{D_i=0,}P(X_i))\}$$

- Similarly, we can estimate ATU and ATE.

Stata syntax

- *Type the following command in the Stata command window, then psmatch2 package will be download and installed into your computer.*

```
ssc install psmatch2
```

- Almost all the matching algorithm and specification test can be realized through psmatch2 and pstest

Illustrate with an example

Thanks for your Attention!

Any questions or comments please write to:

chenglingguo@nju.edu.cn