Propensity score matching

Policy evaluation seeks to determine the effectiveness of a particular intervention. In economic policy analysis, we rarely can work with experimental data generated by purely random assignment of subjects to the treatment and control groups. Random assignment, analogous to the 'randomized clinical trial' in medicine, seeks to ensure that participation in the intervention, or treatment, is the only differentiating factor between treatment and control units.

In non-experimental economic data, we observe whether subjects were treated or not, but in the absence of random assignment, must be concerned with differences between the treated and non-treated. For instance, do those individuals with higher aptitude self-select into a job training program? If so, they are not similar to corresponding individuals along that dimension, even though they may be similar in other aspects.

The key concern is that of similarity. How can we find individuals who are similar on all observable characteristics in order to match treated and non-treated individuals (or plants, or firms...) With a single measure, we can readily compute a measure of distance between a treated unit and each candidate match. With multiple measures defining similarity, how are we to balance similarity along each of those dimensions?

The method of *propensity score matching* (PSM) allows this matching problem to be reduced to a single dimension: that of the propensity score. That score is defined as the probability that a unit in the full sample receives the treatment, given a set of observed variables. If all information relevant to participation and outcomes is observable to the researcher, the propensity score will produce valid matches for estimating the impact of an intervention. Thus, rather than matching on all values of the variables, individual units can be compared on the basis of their propensity scores alone.

An important attribute of PSM methods is that they do not require the functional form to be correctly specified. If we used OLS methods such as

$$y = X\beta + D\gamma + \epsilon$$

where y is the outcome, X are covariates and D is the treatment indicator, we would be assuming that the effects of treatment are constant across individuals. We need not make this assumption to employ PSM. As we will see, a crucial assumption is made on the contents of X, which should include all variables that can influence the probability of treatment.

Why use matching methods?

The greatest challenge in evaluating a policy intervention is obtaining a credible estimate of the *counterfactual:* what would have happened to participants (treated units) had they not participated? Without a credible answer, we cannot rule out that whatever successes have occurred among participants could have happened anyway. This relates to the *fundamental problem of causal inference*: it is impossible to observe the outcomes of the same unit in both treatment conditions at the same time.

The impact of a treatment on individual i, δ_i, is the difference between potential outcomes with and without treatment:

$$\delta_i = Y_{1i} - Y_{0i}$$

where states 0 and 1 corrrespond to non-treatment and treatment, respectively.

To evaluate the impact of a program over the population, we may compute the average treatment effect (ATE):

$$ATE = E[\delta_i] = E(Y_1 - Y_0)$$

Most often, we want to compute the average treatment effect on the treated (ATT):

$$ATT = E(Y_1 - Y_0|D = 1)$$

where $D = 1$ refers to the treatment.

The problem is that not all of these parameters are observable, as they rely on counterfactual outcomes. For instance, we can rewrite ATT as

$$ATT = E(Y_1|D = 1) - E(Y_0|D = 1)$$

The second term is the average outcome of treated individuals had they not received the treatment. We cannot observe that, but we do observe a corresponding quantity for the untreated, and can compute

$$\Delta = E(Y_1|D = 1) - E(Y_0|D = 0)$$

The difference between ATT and Δ can be defined as

$$\Delta = ATT + SB$$

where SB is the selection bias term: the difference between the counterfactual for treated units and observed outcomes for untreated units.

For the computable quantity Δ to be useful, the SB term must be zero. But selection bias in a non-experimental context is often sizable. For instance, those who voluntarily sign up for a teacher-training program may be the more motivated teachers, who might be more likely to do well (in terms of student test scores) even in the absence of treatment.

In other cases, the bias may not arise due to individuals self-selecting into treatment, but being selected for treatment on the basis of an interview or evaluation of their willingness to cooperate with the program. This gives rise to administrative selection bias or program placement bias.

Even in the case of a randomized experiment, participants selected for treatment may choose not to be treated, or may not comply with all aspects of the treatment regime. In this sense, even a randomized trial may involve bias in evaluating the effects of treatment, and nonexperimental methods may be required to adjust for that bias.

Requirements for PSM validity

Two key assumptions underly the use of matching methods, and PSM in particular:

1. Conditional independence: there exists a set X of observable covariates such that after controlling for these covariates, the potential outcomes are independent of treatment status:

$$(Y_1, Y_0) \perp D|X$$

2. Common support: for each value fo X, there is a positive probability of being both treated and untreated:

$$0 < P(D = 1|X) < 1$$

The conditional independence assumption

$$(Y_1, Y_0) \perp D|X$$

implies that after controlling for X, the assignment of units to treatment is 'as good as random.' This assumption is also known as *selection on observables*, and it requires that all variables relevant to the probability of receiving treatment may be observed and included in X. This allows the untreated units to be used to construct an unbiased counterfactual for the treatment group.

The common support assumption

$$0 < P(D = 1|X) < 1$$

implies that the probability of receiving treatment for each possible value of the vector X is strictly within the unit interval: as is the probability of not receiving treatment. This assumption of common support ensures that there is sufficient overlap in the characteristics of treated and untreated units to find adequate matches.

When these assumptions are satisfied, the treatment assignment is said to be *strongly ignorable* in the terminology of Rosenbaum and Rubin (*Biometrika*, 1983).

Basic mechanics of matching

The procedure for estimating the impact of a program can be divided into three steps:

1. Estimate the propensity score
2. Choose a matching algorithm that will use the estimated propensity scores to match untreated units to treated units
3. Estimate the impact of the intervention with the matched sample and calculate standard errors

To estimate the propensity score, a logit or probit model is usually employed. It is essential that a flexible functional form be used to allow for possible nonlinearities in the participation model. This may involve the introduction of higher-order terms in the covariates as well as interaction terms.

There will usually be no comprehensive list of the clearly relevant variables that would assure that the matched comparison group will provide an unbiased estimate of program impact. Obviously explicit criteria that govern project or program eligibility should be included, as well as factors thought to influence self-selection and administrative selection.

In choosing a matching algorithm, you must consider whether matching is to be performed with or without replacement. Without replacement, a given untreated unit can only be matched with one treated unit. A criterion for assessing the quality of the match must also be defined. The number of untreated units to be matched with each treated unit must also be chosen.

Early matching estimators paired each treated unit with one unit from the control group, judged most similar. Researchers have found that estimators are more stable if a number of comparison cases are considered for each treated case, usually implying that the matching will be done with replacement.

The matching criterion could be as simple as the absolute difference in the propensity score for treated vs. non-treated units. However, when the sampling design oversamples treated units, it has been found that matching on the log odds of the propensity score ($p/(1-p)$) is a superior criterion.

The *nearest neighbor* matching algorithm merely evaluates absolute differences between propensity scores (or their log odds), where you may choose to use 1, 2, ... K nearest neighbors in the match. A variation, *radius matching*, specifies a 'caliper' or maximum propensity score difference. Larger differences will not result in matches, and all units whose differences lie within the caliper's radius will be chosen.

In many-to-one radius matching with replacement, the estimator of program impact may be written as

$$E(\Delta Y) = \frac{1}{N} \sum_{i=1}^{N} \left[Y_{1i} - \bar{Y}_{0j(i)} \right]$$

where $\bar{Y}_{0j(i)}$ is the average outcome for all comparison individuals matched with case i, Y_{1i} is the outcome for treated case i, and N is the number of treated cases.

As an alternative to radius matching, which rules out matches beyond the threshold of the caliper, the *kernel* and *local-linear* methods are nonparametric methods that compare each treated unit to a weighted average of the outcomes of all untreated units, with higher weights being placed on the untreated units with scores closer to that of the treated individual. These methods exhibit lower variance, but may suffer from the inclusion of information from poor matches. To use these methods, a kernel function must be chosen, and its bandwidth parameter must be specified.

The usual tradeoff between bias and efficiency arises in selecting a matching algorithm. By choosing only one nearest neighbor, we minimize bias by using the most similar observation. However, this ignores a great deal of information, and thus may yield less efficient estimates.

Evaluating the validity of matching assumptions

The conditional independence assumption cannot be directly tested, but several guidelines for model specification should be considered. The more transparent and well-controlled is the selection process, the more confidence you may have in arguing that all relevant variables have been included. Measures included in the PSM model should be stable over time, or deterministic (e.g., age), or measured before participation, so that they are not confounded with outcomes or the anticipation of treatment. The specification should allow for nonlinear covariate effects and potential interactions in order to avoid inappropriate constraints on the functional form.

Balancing tests consider whether the estimated propensity score adequately balances characteristics between the treatment and control group units. The assumption

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

is testable. If it is supported by the data, then after conditioning on the estimated propensity score $p(X)$, there should be no other variable that could be added to the conditioning set X that would improve the estimation, and after the application of matching, there should be no statistically significant differences between covariate means of the treated and comparison units. These mean comparisons can be contrasted with the unconditional means of the treatment and control groups, which are likely to be statistically significant in most applications.

Finally, the common support or overlap condition

$$0 < P(D = 1|X) < 1$$

should be tested. This can be done by visual inspection of the densities of propensity scores of treated and non-treated groups, or more formally via a comparison test such as the Kolmogorov–Smirnov nonparametric test. If there are sizable differences between the maxima and minima of the density distributions, it may be advisable to remove cases that lie outside the support of the other distribution. However, as with any trimming algorithm, this implies that results of the analysis are strictly valid only for the region of common support.

An empirical example

As an example of propensity score matching techniques, we follow Sianesi's 2010 presentation at the German Stata Users Group meetings (`http://ideas.repec.org/p/boc/dsug10/02.html`) and employ the `nsw_psid` dataset that has been used in several articles on PSM techniques. This dataset combines 297 treated individuals from a randomised evaluation of the NSW Demonstration job-training program with 2,490 non-experimental untreated individuals drawn from the Panel Study of Income Dynamics (PSID), all of whom are male. The outcome of interest is `re78`, 1978 earnings. Available covariates include age, ethnic status (black, Hispanic or white), marital status, years of education, an indicator for no high school degree and 1975 earnings (in 1978 dollars).

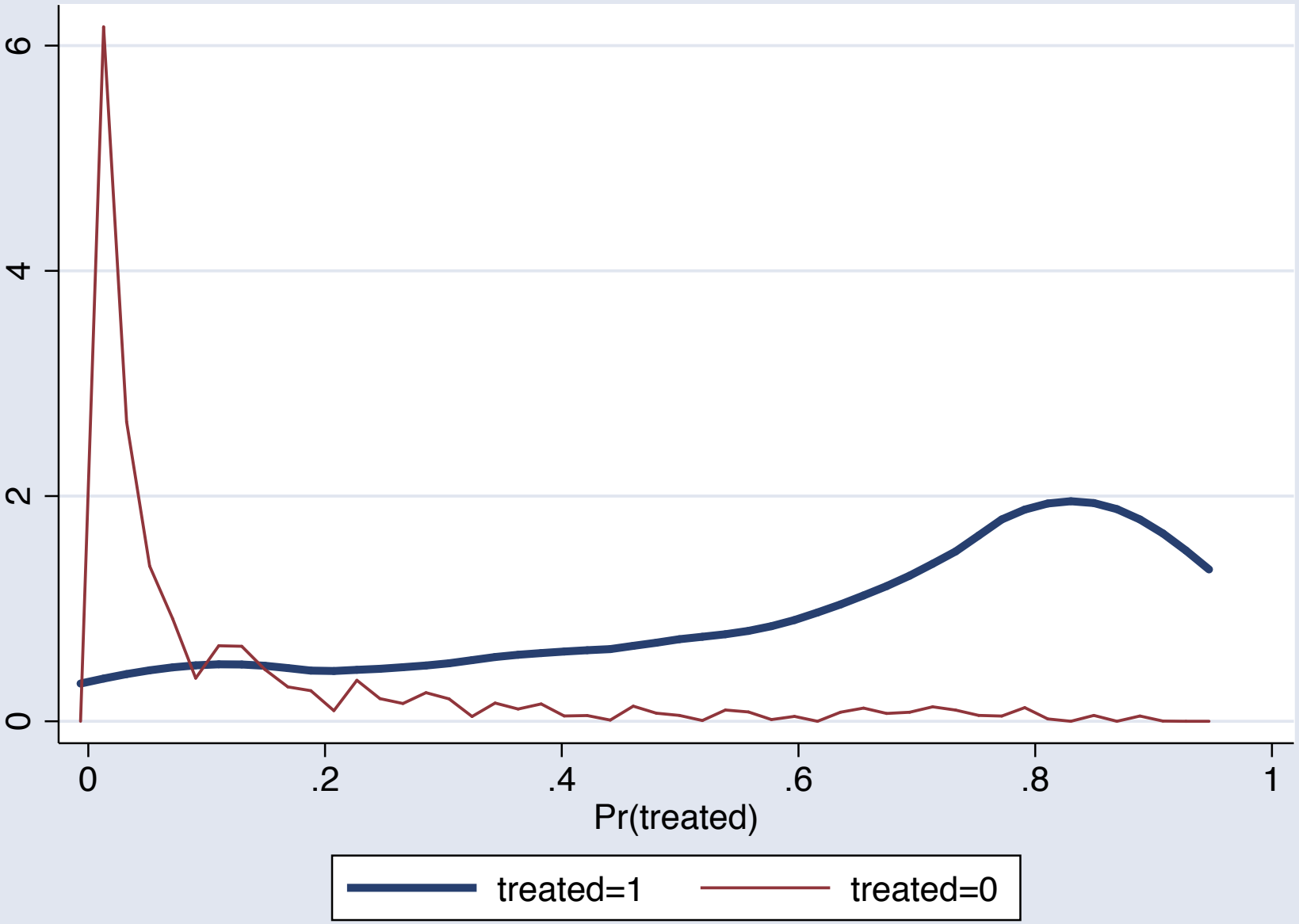
We use Leuven and Sianesi's `psmatch2` routine, available from SSC.

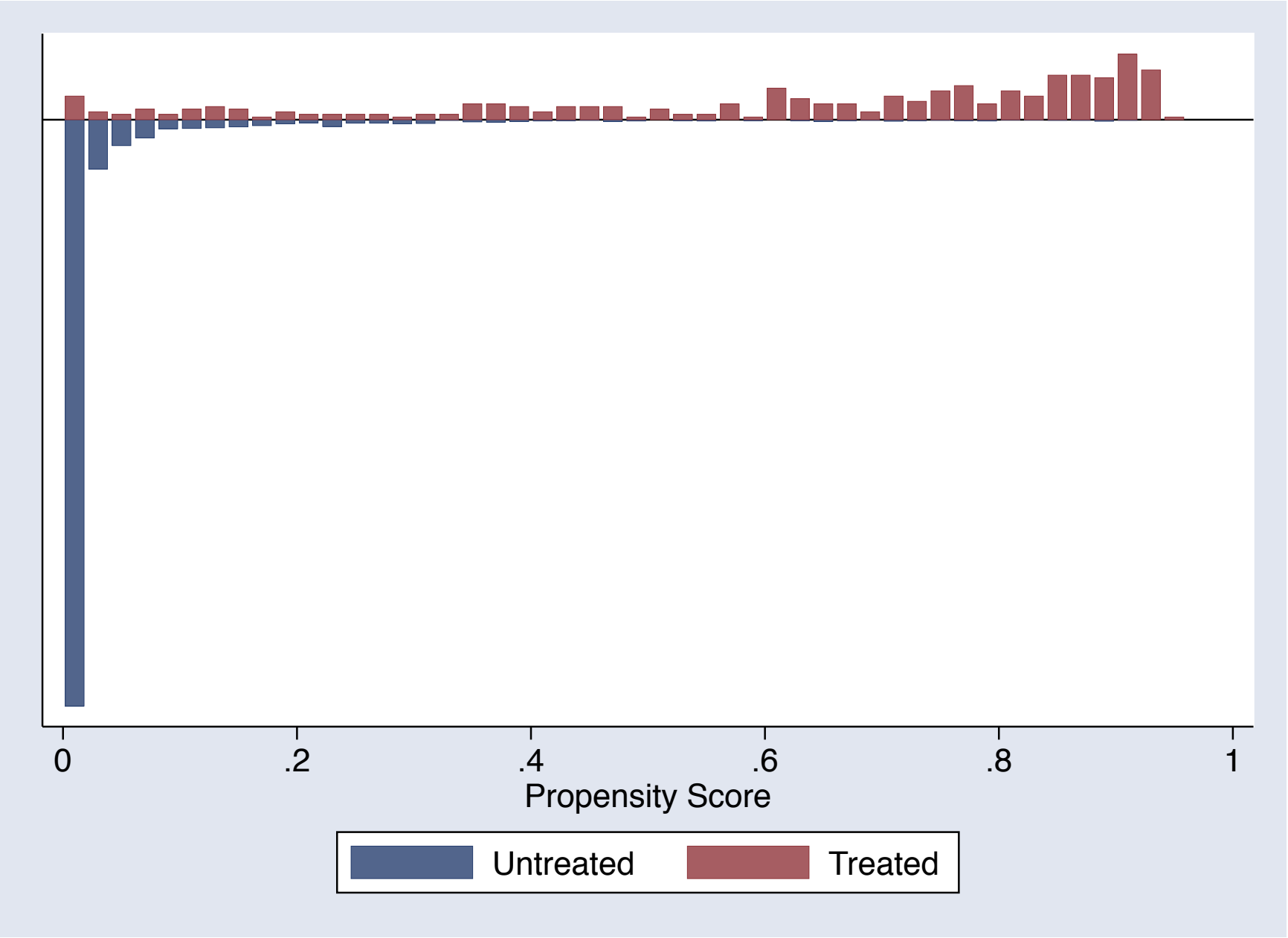
```
 . use nsw_psid, clear
(NSW treated and PSID non-treated)
. qui probit treated age black hispanic married educ nodegree re75
. margins, dydx(_all)
Average marginal effects                        Number of obs   =       2787
Model VCE    : OIM
Expression   : Pr(treated), predict()
dy/dx w.r.t. : age black hispanic married educ nodegree re75
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0035844	.000462	-7.76	0.000	-.0044899	-.002679
black	.0766501	.0088228	8.69	0.000	.0593577	.0939426
hispanic	.0831734	.0157648	5.28	0.000	.0522751	.1140718
married	-.0850743	.0070274	-12.11	0.000	-.0988478	-.0713009
educ	.0003458	.0023048	0.15	0.881	-.0041716	.0048633
nodegree	.0418875	.0108642	3.86	0.000	.0205942	.0631809
re75	-6.89e-06	5.89e-07	-11.71	0.000	-8.04e-06	-5.74e-06

```
. // compute the propensity score
. predict double score
(option pr assumed; Pr(treated))
```

```
 . // compare the densities of the estimated propensity score over groups
. density2 score, group(treated) saving(psm2a, replace)
(file psm2a.gph saved)
. graph export psm2a.pdf, replace
(file /Users/cfbaum/Documents/Stata/StataWorkshops/psm2a.pdf written in PDF for
> mat)
. psgraph, treated(treated) pscore(score) bin(50) saving(psm2b, replace)
(file psm2b.gph saved)
. graph export psm2b.pdf, replace
(file /Users/cfbaum/Documents/Stata/StataWorkshops/psm2b.pdf written in PDF for
> mat)
```


0
2
4
6
0
.2
.4
.6
.8
1
Pr(treated)
treated=1
treated=0


0
.2
.4
.6
.8
1
Propensity Score
Untreated
Treated

```
1 . // compute nearest-neighbor matching with caliper and replacement
2 . psmatch2 treated, pscore(score) outcome(re78) caliper(0.01)
  There are observations with identical propensity score values.
  The sort order of the data could affect your results.
  Make sure that the sort order is random before calling psmatch2.
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	5976.35202	21553.9209	-15577.5689	913.328457	-17.06
	ATT	6067.8117	5768.70099	299.110712	1078.28065	0.28

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support		
	Off suppo	On suppor	Total
Untreated	0	2,490	2,490
Treated	26	271	297
Total	26	2,761	2,787

```
3 . // evaluate common support
4 . summarize _support if treated
```

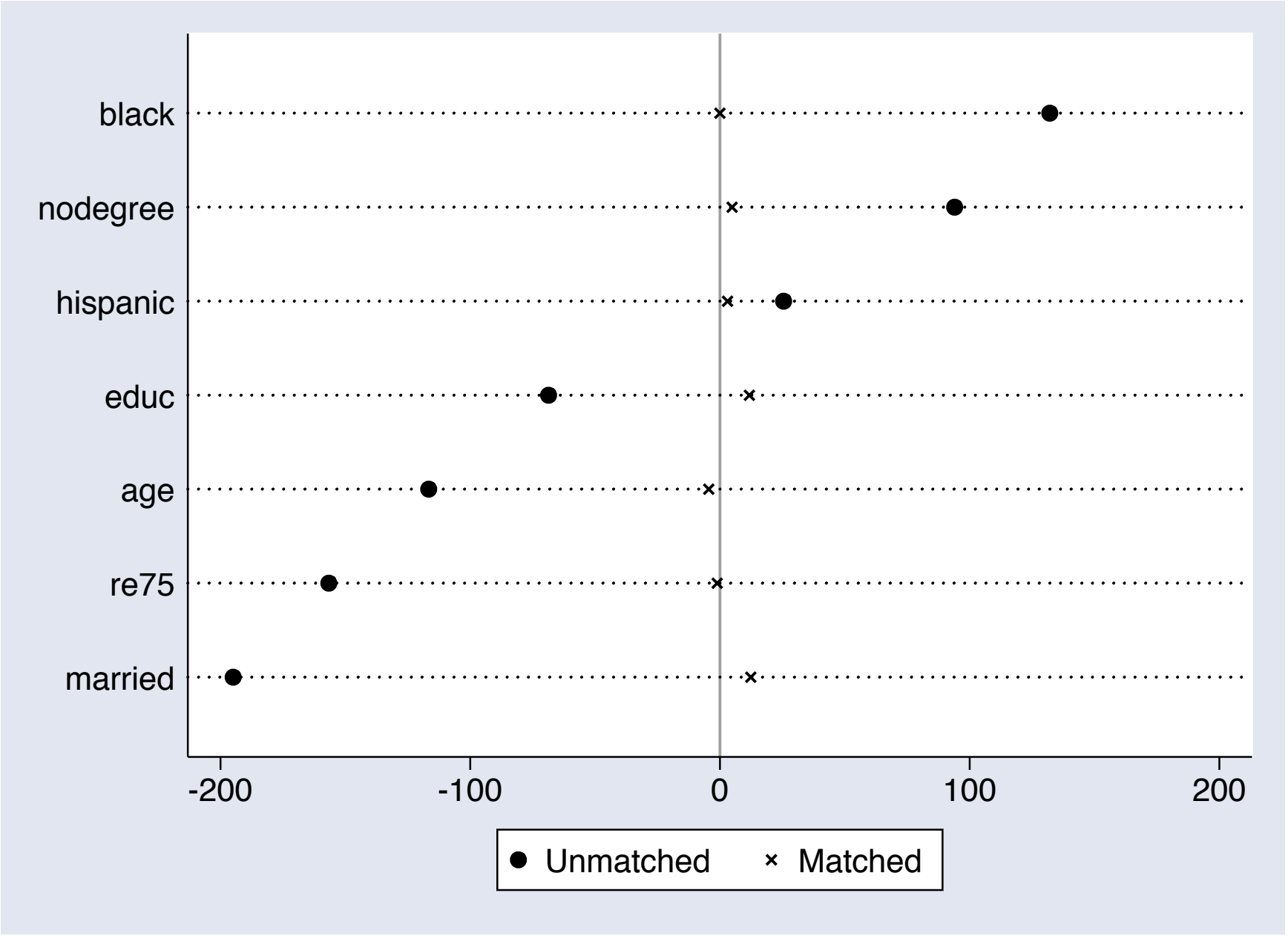
Variable	Obs	Mean	Std. Dev.	Min	Max
_support	297	.9124579	.2831048	0	1

```
5 . qui log close
```

```
1 . // evaluate quality of matching
2 . pstest2 age black hispanic married educ nodegree re75, sum graph
```

Variable	Sample	Mean Treated	Mean Control	%bias	%reduct |bias|	t-test t	t-test p>|t|
age	Unmatched	24.626	34.851	-116.6		-16.48	0.000
	Matched	25.052	25.443	-4.5	96.2	-0.61	0.545
black	Unmatched	.80135	.2506	132.1		20.86	0.000
	Matched	.78967	.78967	0.0	100.0	-0.00	1.000
hispanic	Unmatched	.09428	.03253	25.5		5.21	0.000
	Matched	.09594	.08856	3.0	88.0	0.30	0.767
married	Unmatched	.16835	.86627	-194.9		-33.02	0.000
	Matched	.1845	.14022	12.4	93.7	1.40	0.163
educ	Unmatched	10.38	12.117	-68.6		-9.51	0.000
	Matched	10.465	10.166	11.8	82.8	1.54	0.125
nodegree	Unmatched	.73064	.30522	94.0		15.10	0.000
	Matched	.71587	.69373	4.9	94.8	0.56	0.573
re75	Unmatched	3066.1	19063	-156.6		-20.12	0.000
	Matched	3197.4	3307.8	-1.1	99.3	-0.28	0.778

Summary of the distribution of the abs(bias)


black
nodegree
hispanic
educ
age
re75
married
-200
-100
0
100
200
Unmatched
Matched

Alternatively, we can perform PSM with a kernel-based method. Notice that the estimate of ATT switches sign relative to that produced by the nearest-neighbor matching algorithm.

```
1 . // compute kernel-based matching with normal kernel
2 . psmatch2 treated, pscore(score) outcome(re78) kernel k(normal) bw(0.01)
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	**5976.35202**	**21553.9209**	**-15577.5689**	**913.328457**	**-17.06**
	ATT	**5976.35202**	**6882.18396**	**-905.831935**	**2151.26377**	**-0.42**

Note: S.E. does not take into account that the propensity score is estimated.

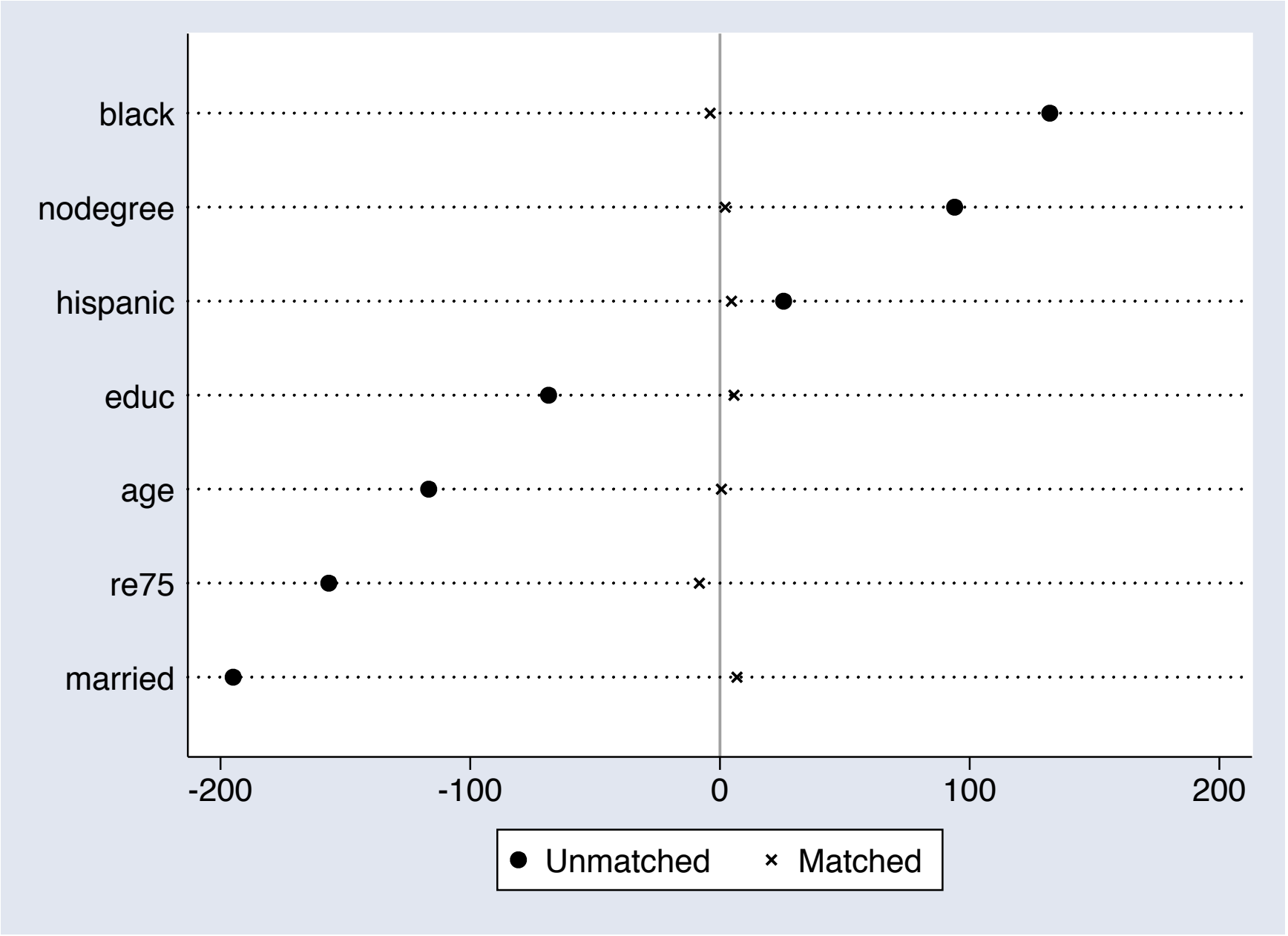
psmatch2: Treatment assignment	psmatch2: Common support On suppor	Total
Untreated	**2,490**	**2,490**
Treated	**297**	**297**
Total	**2,787**	**2,787**

```
3 . qui log close
```

```
1 . // evaluate quality of matching
2 . pstest2 age black hispanic married educ nodegree re75, sum graph
```

Variable	Sample	Mean Treated	Mean Control	%bias	%reduct |bias|	t-test t	t-test p>|t|
age	Unmatched	24.626	34.851	-116.6		-16.48	0.000
	Matched	24.626	24.572	0.6	99.5	0.09	0.926
black	Unmatched	.80135	.2506	132.1		20.86	0.000
	Matched	.80135	.81763	-3.9	97.0	-0.50	0.614
hispanic	Unmatched	.09428	.03253	25.5		5.21	0.000
	Matched	.09428	.08306	4.6	81.8	0.48	0.631
married	Unmatched	.16835	.86627	-194.9		-33.02	0.000
	Matched	.16835	.1439	6.8	96.5	0.82	0.413
educ	Unmatched	10.38	12.117	-68.6		-9.51	0.000
	Matched	10.38	10.238	5.6	91.8	0.81	0.415
nodegree	Unmatched	.73064	.30522	94.0		15.10	0.000
	Matched	.73064	.72101	2.1	97.7	0.26	0.793
re75	Unmatched	3066.1	19063	-156.6		-20.12	0.000
	Matched	3066.1	3905.8	-8.2	94.8	-1.99	0.047

Summary of the distribution of the abs(bias)


black
nodegree
hispanic
educ
age
re75
married
-200
-100
0
100
200
Unmatched
Matched

We could also employ Mahalanobis matching, which matches on the whole vector of X values (and possibly the propensity score as well), using a different distance metric.

An additional important issue: how might we address unobserved heterogeneity, as we do in a panel data context with fixed effects models? A *differences-in-differences matching estimator* (DID) has been proposed, in which rather than evaluating the effect on the outcome variable, you evaluate the effect on the change in the outcome variable, before and after the intervention. Akin to DID estimators in standard policy evaluation, this allows us to control for the notion that there may be substantial unobserved differences between treated and untreated units, relaxing the 'selection on observables' assumption.

Regression discontinuity models

The idea of Regression Discontinuity (RD) design, due to Thistlewaite and Campbell (*J. Educ. Psych.*, 1960) and Hahn et al. (*Econometrica*, 2001) is to use a discontinuity in the level of treatment related to some observable to get a consistent estimate of the LATE: the local average treatment effect. This compares those just eligible for the treatment (above the threshold) to those just ineligible (below the threshold).

Among non-experimental or quasi-experimental methods, RD techniques are considered to have the highest internal validity (the ability to identify causal relationships in this research setting). Their external validity (ability to generalize findings to similar contexts) may be less impressive, as the estimated treatment effect is local to the discontinuity.

What could give rise to a RD design? In 1996, a number of US states adopted a policy that while immigrants were generally ineligible for food stamps, a form of welfare assistance, those who had been in the country legally for at least five years would qualify. At a later date, one could compare self-reported measures of dietary adequacy, or measures of obesity, between those immigrants who did and did not qualify for this assistance. The sharp discontinuity in this example relates to those on either side of the five-year boundary line.

Currently, US states are eligible for additional Federal funding if their unemployment rate is above 8% (as most currently are). This funding permits UI recipients to receive a number of additional weeks of benefit. Two months ago, the Massachusetts unemployment rate dropped below the threshold: good news for those who are employed, but bad news for those still seeking a job, as the additional weeks of benefit are now not available to current recipients.