Lecture 2

Propensity Score Matching

Policy Evaluation
Nova SBE - Universidade Nova de Lisboa / IGC
Pedro C. Vicente
http://www.pedrovicente.org/

Introducing PSM

- When a treatment cannot be randomized, the next best thing is to try to mimic randomization
- With matching methods, one tries to develop a counterfactual or control group that is as similar to the treatment group as possible in terms of observed characteristics
    - The idea is to find, from a large group of non-participants, individuals who are observationally similar to participants in terms of characteristics not affected by the program (e.g., pre-program characteristics not affected by subsequent participation in the program)
    - Each participant is matched with an observationally similar non-participant, and then the average difference of the outcome between the two groups is taken to get the program treatment effect

- Two fundamental assumptions:
    - Participation based solely on observed characteristics
    - Enough non-participants are available to match with participants
- Practical problem is: because many possible characteristics exist, the most common way of matching households is by using a single propensity score to match each participant to a non-participant, reflecting the probability of participating in the program conditional on their different observed characteristics
    - PSM avoids the curse of dimensionality associated with trying to match participants and non-participants on every possible characteristic when the set of observed characteristics X is large

• PSM constructs a statistical comparison group that is based on a model of the probability of participating in the treatment *T* conditional on observed characteristics *X*, or the propensity score:

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

• Rosenbaum and Rubin (Biometrika, 1983) show that, under the mentioned assumptions, matching on *P(X)* is as good as matching on *X*

Assumptions

• **Conditional independence** states that given a set of observable covariates *X* that are not affected by the treatment, potential outcomes Y^T, Y^C are independent of treatment assignment *T*:

$$(Y_i^T, Y_i^C) \perp T_i \mid X_i$$

• This assumption implies that uptake of the program is based entirely on observed characteristics

• Note that if unobserved characteristics determine program participation, conditional independence will be violated, and PSM is not an appropriate method

• **Common support** or **overlap condition** ensures that treatment observations have comparison observations nearby in the propensity score distribution:

$$0 < \Pr(T_i = 1 \mid X_i) < 1$$

• The effectiveness of PSM also depends on having a large number of participant and non-participant observations so that a substantial region of common support can be found, and inferences are made about causality

Figure 4.1 Example of Common Support

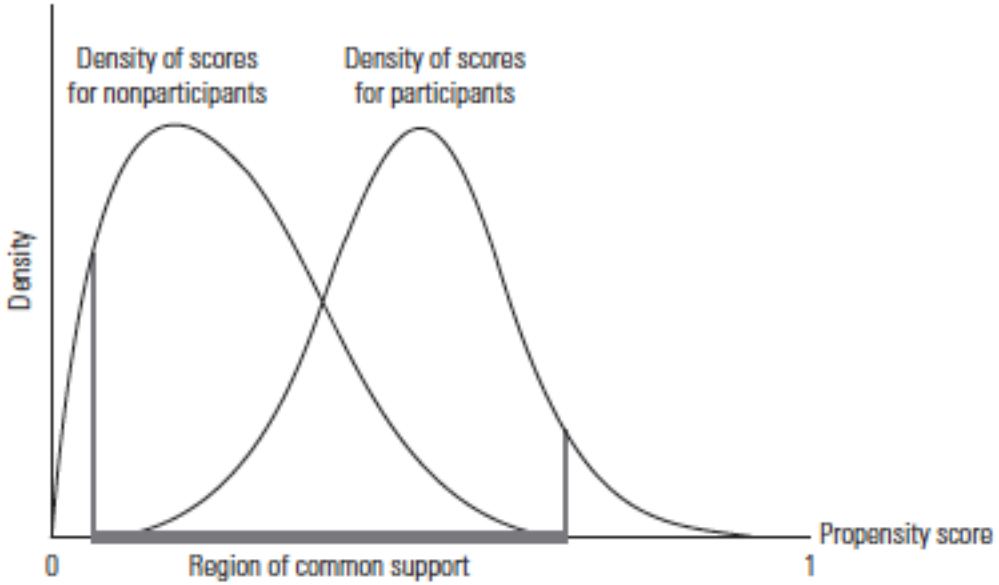
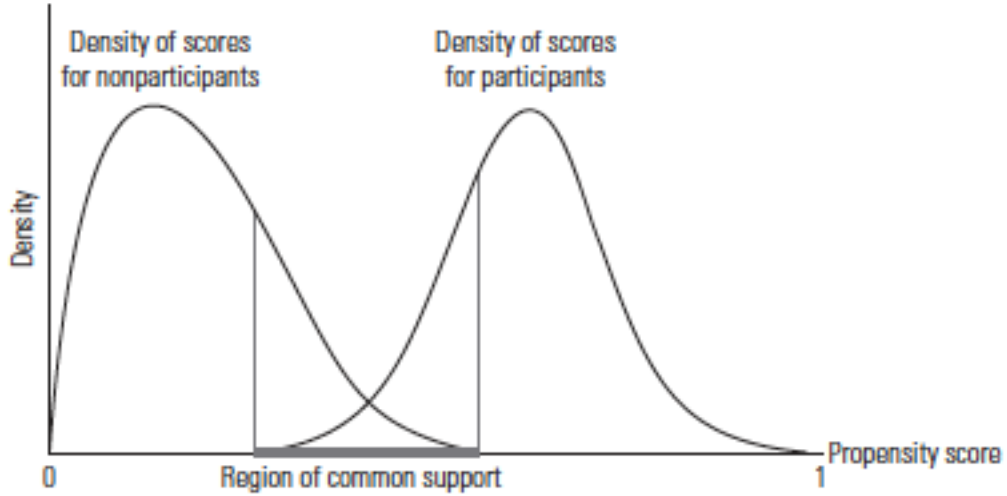


Figure 4.2 Example of Poor Balancing and Weak Common Support



• If conditional independence holds, and if there is a sizeable overlap in $P(X)$ across participants and non-participants, the PSM estimator for the TT can be specified as the mean difference in Y over the common support, weighting the comparison units by the propensity score distribution of participants:

$$TT_{PSM} = E_{P(X)|T=1}[E[Y^T \mid P(X), T=1] - E[Y^C \mid P(X), T=0]] =$$

$$= \frac{1}{N_T} \sum_{i \in T} \left[Y_i^T - \sum_{j \in C} \omega(i,j) Y_j^C \right]$$

where N_T is the number of participants i and $\omega(i,j)$ is the weight used to aggregate outcomes for the matched non-participants j

- Note that:
  - PSM vs. OLS: OLS controlling for X imposes a linear structure which is not needed in PSM; PSM focuses on region of common support unlike OLS; but OLS still assumes strict exogeneity which is analogous to the conditional independence condition in PSM
  - If S is not the full support of $P(X)$ for participants in the program, TT_{PSM} is going to be different from the TT from a randomized evaluation

Implementation

- **Step 1: estimating a model of program participation**
  - The samples of participants and non-participants should be pooled
  - *T* should be estimated on all observed covariates *X* in the data that are likely to determine participation
    - This can be implemented through probit or logit models
  - The predicted outcome represents the estimated probability of participation or propensity score
    - Every sampled participant and non-participant will have an estimated propensity score

$$\hat{P}r(T=1|X)=\hat{P}(X)$$

  - Possible bias due to omitted characteristics; same measurement conditions should be used for participants and non-participants
  - Including too many *X* variables should also be avoided, at risk of perfectly predicting participation (dropping observations out of common support)

- **Step 2: defining the region of common support and balancing tests**
  - The region of common support needs to be defined
  - Sampling bias may occur if dropped observations are systematically different in terms of observed characteristics – this requires close attention
  - Balancing tests can be conducted to check whether, within each quantile of the propensity score distribution, the average propensity score and mean of *X* are the same across comparison groups
    - Formally, one needs to check

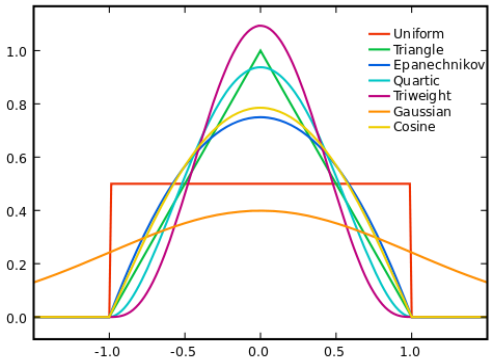
$$\hat{P}(X \mid T = 1) = \hat{P}(X \mid T = 0)$$

- **Step 3: matching participants to non-participants**
  - Different matching criteria can be used to assign participants to non-participants on the basis of the propensity score; doing so entails calculating a weight for each matched participant/non-participant set
    - **Nearest-neighbor matching (NNM)**: where each treatment unit is matched with the closest propensity score; one can also choose *n* nearest neighbors (e.g., 5); matching can be done with or without replacement (with replacement means that the same non-participant can be used as a match for different participants)
    - **Caliper or radius matching**: one problem with NNM is that the difference in propensity scores for a participant and its closest non-participant neighbor may still be high; this can be avoided by imposing a threshold on the maximum propensity score distance (caliper); this is matching with replacement within a certain radius; it is likely that more observations will be dropped

• **Stratification or interval matching**: it partitions the common support into different strata (or intervals) and calculates the program impact within each interval; a weighted average of these interval impact estimates yields the overall program impact, taking the share of participants in each interval as the weights

• **Kernel matching**: one risk with some of the other methods is that only a small subset of non-participants will ultimately be used; non-parametric matching estimators such as kernel matching use a weighted average of all non-participants to construct the counterfactual match for each participant

$$\omega(i,j)_{KM} = \frac{K\left(\frac{P_j - P_i}{a}\right)}{\sum_{k \in C} K\left(\frac{P_k - P_i}{a}\right)}$$



where *i* is participant and *j* is non-participant, K is a kernel function, and *a* is a bandwidth parameter

- **Step 4: estimating standard errors**
  - The estimated variance of the treatment effect in PSM should include the variance attributable to the derivation of the propensity score, the determination of the common support, and (if matching is done without replacement) the order in which treated individuals are matched
  - One solution is to use bootstrapping, where repeated samples are drawn from the original sample, and standard errors are estimated from those
    - Each bootstrap sample estimate includes the first steps of the estimation that derive the propensity score, common support, and so on

Example: Job Training Program

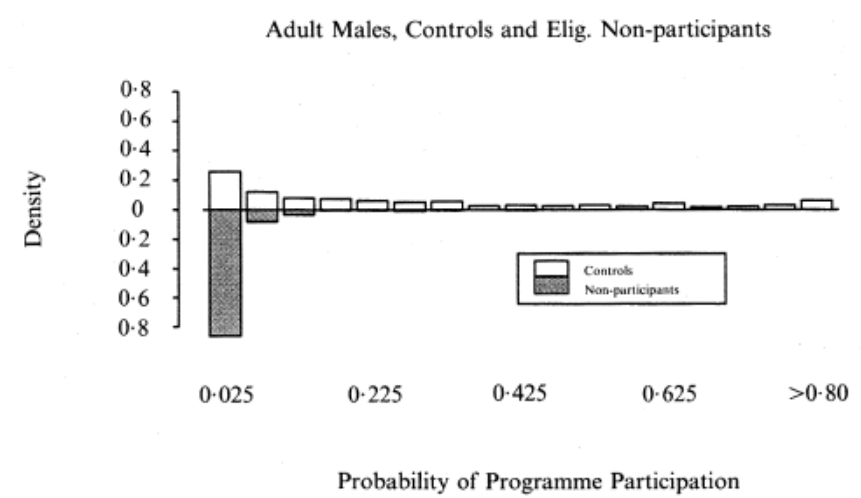
- Heckman, Ichimura, and Todd (RESTUD, 1997):
  - They combine non-experimental data on persons who chose not to participate in the program with data from a large-scale social experiment to examine the performance of various matching methods in estimating an averaged version of the effect of treatment on the treated
  - The program that is analyzed is the National Job Training Partnership Act (JTPA), a job training program for disadvantaged workers
    - Eligibility is based on having a family income near or below the poverty level for six months prior to application, or by participating in welfare and foodstamp programs
  - Authors collected longitudinal data from treatment, randomized-out control, and a comparison group of eligible non-participants (ENPs)

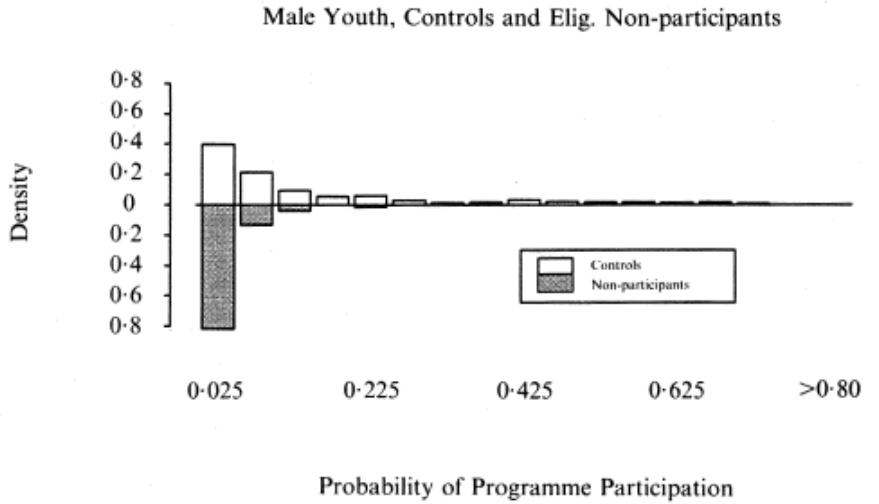
• Members of the ENP comparison group reside in the same narrowly-defined geographic regions as program applicants, are eligible for the program, but do not apply to it; they were administered the same survey as the experimental individuals
• Data available from a baseline, and two follow-up surveys over 36 months
• Study design enables computation of bias by comparing outcomes of control and ENP comparison groups:

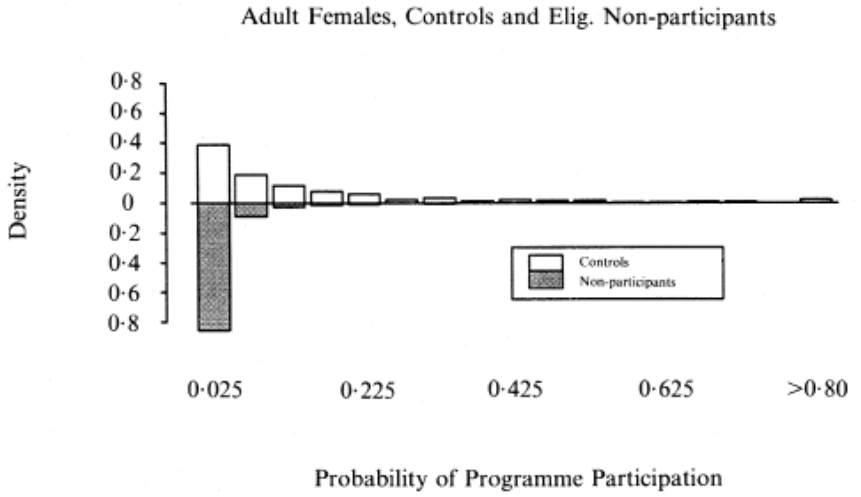
$$B = E(Y^C \mid T = 1) - E(Y^C \mid T = 0)$$

• This bias could arise from (i) non-overlapping support, (ii) different distributions of *X*, and (iii) selection on unobservables

- Results in Heckman et al (1997):
  - Region of common support for PSM can be very small
  - Good news (for this specific application): selection on unobservables accounts for a small share of the treatment effect bias
  - Experimental *TT* varies substantially depending on region of support
  - Bias arising from TT_{PSM} is substantial even though generally lower than simple difference in means (OLS)


Adult Males, Controls and Elig. Non-participants
Density
0·8
0·6
0·4
0·2
0
0·2
0·4
0·6
0·8
Controls
Non-participants
0·025
0·225
0·425
0·625
>0·80
Probability of Programme Participation


Male Youth, Controls and Elig. Non-participants
Density
0·8
0·6
0·4
0·2
0
0·2
0·4
0·6
0·8
Controls
Non-participants
0·025
0·225
0·425
0·625
>0·80
Probability of Programme Participation


Adult Females, Controls and Elig. Non-participants
Density
0·8
0·6
0·4
0·2
0
0·2
0·4
0·6
0·8
Controls
Non-participants
0·025
0·225
0·425
0·625
>0·80
Probability of Programme Participation

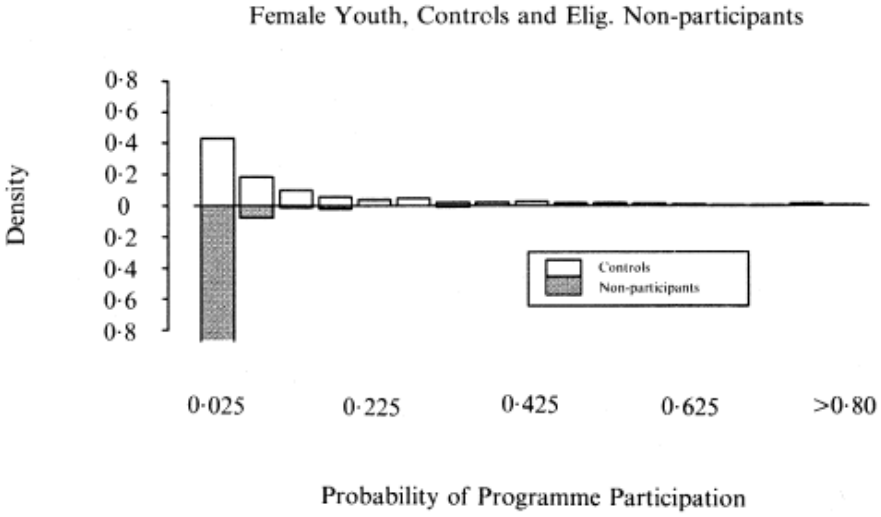

Female Youth, Controls and Elig. Non-participants
Density
0·8
0·6
0·4
0·2
0
0·2
0·4
0·6
0·8
Controls
Non-participants
0·025
0·225
0·425
0·625
>0·80
Probability of Programme Participation

TABLE 2

Decomposition of difference in post-programme mean earnings
Bootstrapped standard errors shown in parentheses‡
Percentage of mean difference attributable to components in square brackets
Earnings measured in average monthly dollars

	Mean difference $\hat{B}$	Non-overlap* $\hat{B}_1$	Density weighting $\hat{B}_2$	Selection bias $\hat{B}_3$	Average bias $(\hat{B}_{S_r})$	Selection bias** $(\hat{B}_{S_r})$ as a % of treatment impact
Experimental Controls and eligible nonparticipants (ENPs)†						
Adult males	−342	218	−584	23	38	87%
(std. err.) [%]	(47)	(38) [−64%]	(41) [170%]	(33) [−7%]	(63)	
Adult females	33	80	−78	31	38	129%
(std. err.) [%]	(26)	(13) [242%]	(18) [−235%]	(26) [94%]	(33)	
Male youth	20	142	−131	9	14	23%
(std. err.) [%]	(57)	(28) [704%]	(35) [−650%]	(42) [46%]	(64)	
Female youth	42	74	−67	35	49	7239%
(std. err.) [%]	(36)	(17) [177%]	(26) [−161%]	(28) [84%]	(42)	

TABLE 4

Estimated programme impacts
*Monthly average impacts in dollars over 18 months after random assignment**

	Experimental impact $M(S_E)$	Experimental impact for persons in overlap support region $M(S_P)$	Bias from non-overlap	% bias
Adult males	44	61	17	39
	(17)	(17)		
Adult females	29	35	6	21
	(9)	(12)		
Male youth	−58	−36	22	38
	(14)	(18)		
Female youth	−1	25	26	2500
	(11)	(18)		

* In our data, the experimental control group was administered a long-baseline survey that gathered five years of retrospective data while the experimental treatment group was not. Since information on recent labour force status and on recent earnings is missing for treatments, we are only able to obtain coarse estimates of P for the treated group. We use the coarse II model described in the notes to Table 6(a). The support region in the nonexperimental analysis is determined using the best predictor P model, so it is necessary to estimate which treatment group members would be excluded by imposing a common support to obtain impact estimates using nonexperimental methods. The impact estimates in the support region were obtained as follows. For controls and treatment, we first divide the coarse P distribution into 20 equal-size bins, then within-bin treatment estimates are estimated. The impact estimate in the overlap region is obtained as the weighted average of the within-bin estimates, with the weights given by the proportion of controls within each bin after deleting controls whose values of P lie outside the overlap region.

TABLE 5(a)

Estimated bias for alternative nonparametric matching methods Experimental controls and eligible nonparticipants (ENPs)†‡*

Quarter	Difference in means ($\hat{\beta}$)	Nearest neighbour without common support	Nearest neighbour with common support	Local linear P score matching	Regression-adjusted local linear matching	Difference-in-differences from local linear P score matching	Difference-in-differences from regression-adjusted local linear matching
			Adult males				
$t=1$	−418 (38)	221 (56)	123 (67)	33 (59)	39 (60)	97 (62)	104 (63)
$t=2$	−349 (47)	−166 (151)	77 (83)	37 (61)	39 (64)	77 (89)	77 (92)
$t=3$	−337 (55)	−58 (206)	53 (96)	29 (78)	21 (80)	90 (114)	74 (114)
$t=4$	−286 (57)	161 (178)	86 (96)	80 (77)	65 (82)	112 (90)	98 (91)
$t=5$	−305 (57)	167 (196)	87 (100)	64 (77)	50 (83)	19 (95)	−5 (99)
$t=6$	−328 (63)	45 (191)	34 (113)	37 (82)	17 (90)	4 (105)	−35 (111)
Ave. 1 to 6	−337 (47)	62 (127)	77 (80)	47 (60)	38 (64)	67 (71)	52 (74)
As a % of impact**	775%	142%	177%	108%	87%	153%	120%
As a % of adjusted impact	552%	102%	126%	77%	62%	109%	85%
			Adult females				
$t=1$	−26 (24)	115 (30)	67 (36)	45 (33)	55 (36)	65 (31)	74 (30)
$t=2$	29 (25)	113 (53)	47 (46)	48 (37)	55 (39)	53 (40)	60 (39)
$t=3$	38 (26)	124 (107)	63 (59)	26 (48)	31 (52)	10 (56)	14 (59)
$t=4$	55 (30)	106 (102)	58 (52)	36 (39)	35 (45)	12 (53)	7 (56)
$t=5$	62 (34)	92 (111)	47 (51)	48 (40)	48 (45)	29 (51)	23 (53)
$t=6$	40 (36)	79 (84)	−6 (54)	23 (40)	16 (42)	−5 (51)	−18 (51)
Ave. 1 to 6	33 (26)	105 (69)	46 (43)	38 (33)	40 (38)	27 (38)	27 (39)
As a % of impact**	113%	358%	157%	130%	137%	93%	91%
As a % of adjusted impact	94%	300%	131%	109%	114%	78%	76%

* The table reports the bias for alternative matching methods. The bias in the first column is $\hat{\beta}$. The estimator in the second column does not restrict matches to a common support region. The estimators in the third through seventh columns restrict matches to a common support region and the bias estimates correspond to $\hat{B}_{S_P}$.

† The best predictor model given in the second footnote to Table 2, is used to estimate the probability of programme participation. The conditioning variables in the regression adjusted local linear models are site, race, age, education, previous training, work experience in months, the local unemployment rate, indicator variables for marital status and for the presence of a child aged less than 6 in the household, and indicators for the quarter of the year and the year.

‡ A 2% trimming rule is used to determine the region of overlapping support (see Appendix C). A fixed bandwidth equal to 0·06 and a biweight kernel, defined in Appendix A, are used for the nonparametric estimates.

** The impacts in the table are mean monthly impacts for the six post-programme quarters, estimated using the experimental treatment and control data for the four JTPA training sites in our study. See the experimental impacts and the adjusted impacts in Table 4.

TABLE 5(b)

Estimated bias for alternative nonparametric matching methods Experimental controls and eligible nonparticipants (ENPs)†‡*

Quarter	Difference in means ($\hat{\beta}$)	Nearest neighbour without common support	Nearest neighbour with common support	Local linear *P* score matching	Regression-adjusted local linear matching	Difference-in-differences from local linear *P* score matching	Difference-in-differences from regression-adjusted local linear matching
			Male youth				
$t=1$	−51 (58)	146 (92)	49 (75)	3 (64)	8 (61)	43 (72)	80 (77)
$t=2$	2 (60)	197 (92)	98 (82)	40 (64)	28 (55)	43 (60)	61 (60)
$t=3$	5 (73)	202 (105)	83 (119)	33 (81)	−8 (77)	92 (80)	70 (86)
$t=4$	17 (69)	246 (105)	98 (94)	44 (81)	4 (71)	9 (74)	−5 (77)
$t=5$	82 (73)	283 (118)	138 (89)	84 (93)	42 (76)	18 (88)	−11 (81)
$t=6$	65 (77)	258 (145)	129 (121)	28 (93)	−31 (92)	−23 (89)	−64 (84)
Ave. 1 to 6	20 (57)	222 (88)	99 (78)	39 (66)	7 (53)	30 (49)	22 (48)
As a % of impact**	34%	382%	170%	67%	12%	52%	38%
As a % of adjusted impact	56%	617%	275%	108%	19%	84%	61%
			Female youth				
$t=1$	6 (31)	67 (54)	−7 (60)	31 (42)	−8 (46)	−7 (38)	−14 (41)
$t=2$	54 (40)	85 (57)	23 (60)	79 (53)	27 (49)	60 (49)	27 (47)
$t=3$	89 (44)	142 (62)	97 (78)	121 (60)	49 (52)	135 (59)	83 (58)
$t=4$	42 (50)	89 (56)	24 (72)	37 (59)	−28 (59)	45 (57)	4 (59)
$t=5$	64 (41)	121 (57)	51 (63)	65 (54)	8 (54)	45 (61)	−7 (63)
$t=6$	31 (46)	107 (82)	34 (70)	34 (65)	1 (62)	31 (70)	6 (69)
Ave. 1 to 6	48 (36)	102 (49)	37 (56)	61 (45)	8 (42)	52 (39)	17 (39)
As a % of impact**	7059%	15000%	5441%	8971%	1176%	7574%	2426%
As a % of adjusted impact	195%	415%	150%	248%	33%	209%	67%

* The table reports the bias for alternative matching methods. The bias in the first column is $\hat{\beta}$. The estimator in the second column does not restrict matches to a common support region. The estimators in the third through seventh columns restrict matches to a common support region and the bias estimates correspond to $\hat{B}_{S_P}$.

† The best predictor model given in the second footnote to Table 2, is used to estimate the probability of programme participation. The conditioning variables in the regression adjusted local linear models are site, race, age, education, previous training, the local unemployment rate, indicator variables for marital status and the presence of a child aged less than 6 in the household, and indicators for the quarter of the year and the year.

‡ A 5% trimming rule is used to determine the region of overlapping support (see Appendix C), and a fixed bandwidth equal to 0·06 and a biweight kernel, defined in Appendix A, are used for nonparametric estimates.

** The impacts in the table are mean monthly impacts for the six post-programme quarters, estimated using the experimental treatment and control data for the four JTPA training sites in our study. See the experimental impacts and the adjusted impacts in Table 4.

What have we learnt?

- PSM is a useful method to evaluate policies when selection is on observables
- Main advantage: few assumptions relative to OLS controlling for observables
- However, as a way to identify *TT*, it relies on finding a large region of common support for $P(X)$; also, many different ways to implement, and these may yield different results
- Several improvements were proposed in the literature on PSM (e.g., Heckman et al, RESTUD, 1997, 1998)

Problem Set 2

Exercise: Follow the instructions in chapter 13 of Khandker et al (2009). You should produce a do file and a log file, which should be commented to show that you understood the results. These should be emailed to the grader (Matilde Grácio): only one email per group, please.

Note: We should be able to run the do file given the original datafile and produce the raw log file.

Due date: Friday, February 25.