Propensity scores for the estimation of average treatment effects in observational studies

Leonardo Grilli and Carla Rampichini

Dipartimento di Statistica "Giuseppe Parenti"
Universit di Firenze

Training Sessions on Causal Inference
Bristol - June 28-29, 2011

Outline

1. Observational studies and Propensity score
2. Motivating example: effect of participation in a job training program on individuals earnings
3. Regression-based estimation under unconfoundedness
4. Matching
5. Propensity Scores
   - Propensity score matching
   - Propensity Score estimation
6. Matching strategy and ATT estimation
   - Propensity-score matching with STATA
   - Nearest Neighbor Matching
   - Example: PS matching
   - Example: balance checking
   - Caliper and radius matching
   - Overlap checking
   - pscore matching vs regression

Introduction

- In the evaluation problems, data often do not come from randomized trials but from (non-randomized) observational studies.
- Rosenbaum and Rubin (1983) proposed **propensity score matching** as a method to reduce the bias in the estimation of treatment effects with observational data sets.
- These methods have become increasingly popular in medical trials and in the evaluation of economic policy interventions.

Steps for designing observational studies

As suggested by Rubin (2008), we have to design observational studies to approximate randomized trial in order to obtain objective causal inference

- What was the hypotetical randomized experiment leading to the observed dataset?
- Are sample sizes in the data set adequate?
- Who are the decision makers for treatment assignment and what measurements where available to them?
  - what are the 'key' variables
  - how the treatment conditions were assigned
- are key covariates measured well?
- can balance be achieved on key covariates?

Regular Designs

For the analysis of observational data, we try to structure it so that we can conceptualize the data as having arisen from an underlying *regular assignment mechanism*.

- Regular designs are like completely randomized experiments except that the probabilities of treatment assignment are allowed to depend on covariates, and so can vary from unit to unit.
- Regular designs have two features:
  1. they are **unconfounded**, i.e.

$$Pr[W \mid X, Y(1), Y(0)] = Pr(W \mid X)$$

  2. the individual assignment possibilities as a function of unit i's value of the covariates, $p_i = Pr(W_i \mid X_i)$, are strictly between zero and one,

$$0 < p_i < 1$$

- The assignment probabilities, p_i, are called **propensity scores** (Rosenbaum and Rubin, 1983a).

Estimands I

The causal estimands of interest are usually average treatment effects on the whole population or on subpopulations.

- The parameter to estimate depends on the specific evaluation context and the specific question asked.
  - The ATE $= E[Y(1) - Y(0)]$ is useful to evaluate what is the expected effect on the outcome if individuals in the population were randomly assigned to treatment.
    - ⋆ Heckman (1997) notes that ATE might not be of relevance to policy makers because it includes the effect on persons for whom the programme was never intended.
  - the average treatment effect on the treated (ATT)

$$ATT = E[Y(1) - Y(0) \mid W = 1]$$

is useful to explicitly evaluate the effects on those for whom the programme is actually intended.

In the following we will consider ATT, the parameter of interest in most evaluation studies.

ATT identification, $ATT = E[Y(1) - Y(0) \mid W = 1]$

- Note that $E[Y(0) \mid W = 1]$, i.e. the *counterfactual mean* for those being treated is **not observed**

$\Rightarrow$ choose a proper substitute for it in order to estimate ATT.

- *Should we use the mean outcome of untreated individuals* $E[Y(0) \mid W = 0]$?
- in observational studies this is not a good idea because it could be that covariates which determine the treatment decision also determine the outcome variable of interest.

$$\Downarrow$$

The outcomes of individuals from the treatment and comparison groups would differ *even in the absence of treatment* leading to the so-called *selection bias*

ATT and selection bias

In general, if we compare the outcomes by treatment status, we obtain a biased estimate of the ATT. In fact:

$$E(Y^{obs} \mid W=1) - E(Y^{obs} \mid W=0) = E(Y(1) \mid W=1) - E(Y(0) \mid W=0)$$

leading to

$$\begin{aligned} E(Y(1) \mid W=1) - E(Y(0) \mid W=1) &+ [E(Y(0) \mid W=1) - (E(Y(0) \mid W=0)] \\ &= ATE + bias \end{aligned}$$

- the difference between treated and non treated outcomes in absence of treatment is the so-called *selection bias*
- ATT is identified only if $E(Y(0) \mid W=1) - (E(Y(0) \mid W=0) = 0$, i.e. if the outcomes of individuals from the treatment and comparison groups would NOT differ *in the absence of treatment*
- In experiments where assignment to treatment is random this is ensured and the treatment effect is identified.
- In observational studies, we must rely on some identifying assumptions to solve the selection problem.

Sources of selection bias

- non overlapping supports of X in the treated and comparison group (i.e., the presence of units in one group that cannot find suitable comparison in the other);
- unbalance in observed confounders between the groups of treated and control units (selection on observables)
- unbalance in unobserved confounders between the groups of treated and control units (selection on unobservables)

Matching approach

The underlying identifying assumption is **unconfoundedness** (selection on observables or conditional independence).

- *Intuition*: If the decision to take the treatment is *purely random* for individuals with similar values of the pre-treatment variables, then we could use the average outcome of some similar individuals who were not exposed to the treatment
  - for each i, matching estimators impute the missing outcome by finding other individuals in the data whose covariates are similar but who were exposed to the other treatment.
  - in this way, differences in outcomes of this well selected and thus adequate control group and of participants can be attributed to the treatment.

Matching approach (cont'd)

- Matching techniques have origins in experimental work from the first half of the twentieth century (see e.g. Rubin (1974) or Lechner (1998)) and were advanced and developed in a set of papers by Rosenbaum and Rubin (1983a, 1984, 1985a, 1985b)
- To ensure that the matching estimators identify and consistently estimate the treatment effects of interest, we assume
  - **unconfoundedness**: assignment to treatment is independent of the outcomes, conditional on the covariates

$$(Y(0); Y(1)) \perp\!\!\!\perp W \mid X$$

  - **overlap** or common support condition: the probability of assignment is bounded away from zero and one

$$0 < Pr(W = 1 \mid X) < 1$$

Overlap

$$0 < Pr(W = 1 \mid X) < 1$$

- The assignment mechanism can be interpreted as if, within subpopulations of units with the same value for the covariate, completely randomized experiment was carried out.
- We can analyze data from subsamples with the same value of the covariates, as if they came from a completely randomized experiment.

Strong ignorability

- The reduction to a paired-comparison should only be applied if unconfoundedness is a plausibly assumption based on the data and a detailed understanding of the institutional set-up by which selection into treatment takes place (see for example the discussion in Blundell et al., 2005).
- In their seminal article, Rosenbaum and Rubin (1983) define the treatment to be **strongly ignorable** when both unconfoudedness and overlap are valid.
- Given these two key assumptions of unconfoudddedness and overlap one can identify the average treatment effects (see next slide)

ATE identification under strong ignorability

- Given unconfoundedness, the following equality holds:

$$\begin{aligned} E[Y(w) \mid X = x] &= E[Y(w) \mid W = w, X = x] \\ &= E[Y^{obs} \mid W = w, X = x] \end{aligned}$$

- Thus one can estimate ATE by first estimating the average treatment effect for a subpopulation with covariates $X = x$:

$$E[Y(1) - Y(0) \mid X = x] = E[Y(1) \mid X = x] - E[Y(0) \mid X = x]$$

$$E[Y(1) \mid X = x, W = 1] - E[Y(0) \mid X = x, W = 0] =$$
$$= E[Y^{obs} \mid X, W = 1] - E[Y^{obs} \mid X, W = 0]$$

  - We need to estimate $E[Y(w) \mid X = x, W = w]$ for all values of w and x in the support of these variables.
  - If the overlap assumption is violated at $X = x$, it would be infeasible to estimate both $E[Y(1) \mid X = x, W = 1]$ and $E[Y(0) \mid X = x, W = 0]$.

Observational studies

How can we reduce the bias in estimating treatment effects?

- With an observational data set, we try to structure it so that we can conceptualize the data as having arisen from an underlying regular assignment mechanism.
- We need to adjust any difference in average outcomes for differences in pre-treatment characteristics (not being affected by the treatment)
  - Model-based imputation methods (e.g., regression models)
  - Matching methods
  - Methods based on propensity score
  - Stratification
  - Weighting
  - Mixed methods
- We will show some possible approaches following a well-known example, focusing on the propensity score matching approach

Effect of participation in a job training program on individuals earnings

Data used by Lalonde (1986)

- We are interested in the possible effect of participation in a job training program on individuals earnings in 1978
- This dataset has been used by many authors (Abadie *et al.* 2004, Becker and Ichino, 2002, Dehejia and Wahba, 1999).
- We use a subset of the data constructed by Dehejia and Wahba (1999, see their paper for details).
- Data available in STATA format at http:emlab.berkeley.eduusersimbens
- Variables:
  - treatment `t`: participation in the job training program
  - outcome `re78`: 1978 earnings of the individuals in the sample in terms of 1978 dollars.
  - the observable pre-treatment covariates that we use to identify similar individuals are given in the next slide.

Example: covariates

The data set includes information on pre-treatment (background; confounder) variables

Description	*Name*
age (in years)	age
years of education	educ
real yearly earnings in 1974 (in thousands of 1978)	re74
real yearlyearnings in 1975 (in thousands of 1978)	re75
afro-american (1 if African American, 0 otherwise)	ra
hispanic-american (1 if Hispanic, 0 otherwise)	rh
married (1 if married, 0 otherwise)	marr
more than grade school but less than high school education	nodegree
unemployed in 1974	u74
unemployed in 1975	u75

Regression-based estimation under unconfoundedness

We need to adjust any difference in average outcomes for differences in pre-treatment characteristics (not being affected by the treatment)

- We can adjust via specification of a conditional model for the potential outcome $\Rightarrow$ regression models
- In a standard regression approach, unconfoundedness is implicitly assumed together with other functional or distributional assumptions

$$\widehat{Y}_i^{obs} = \alpha + \tau W_i + \beta X_i + \varepsilon_i$$

with the usual exogeneity assumption that $\varepsilon_i \perp\!\!\!\perp W_i, X_i$.

- the regression of Y_i^{obs} on a constant, W and X implicitly assumes constant treatment effect
- the slope τ of the treatment indicator is an estimator of the average treatment effect
- Unconfoundedness is untestable.

Example: simple linear regression model for causal inference

- Consider the outcome $Y^{obs} = re78$, i.e. the observed individuals earnings in 1978

$$E(Y^{obs} \mid t) = \alpha + \tau t$$

- in this model, the treatment is the training status t, a dummy variable
- from the model we obtain:
  - average outcome for untreated units: $E(Y^{obs} \mid t = 0) = \alpha$
  - average outcome for treated units: $E(Y^{obs} \mid t = 1) = \alpha + \tau$

$\Rightarrow$ the difference in means estimator is given by the slope of t, i.e.

$$\tau = E(Y^{obs} \mid t = 1) - E(Y^{obs} \mid t = 0)$$

Example $E(Y_i^{obs} \mid t) = \alpha + \tau t_i$ STATA commands and results

```
tabulate treat, summarize(re78) means standard
```

	Summary of RE78	
TREAT	Mean	Std. Dev.
0	21594.38	15558.922
1	6349.15	7867.405
Total	20536.48	15638.517

```
reg re78 treat
```

re78	Coef.	Std. Err.	t	$P > \|t\|$
treat	**-15245.23**	1154.914	-13.20	0.000
cons	21594.38	304.233	70.98	0.000

- We should conclude that **the treatment is dangerous** because the expected average earning for treated is lower than for control! Is this a reliable result?
- Are the assumptions underlying the linear regression model plausible in this case?

Example: multiple linear regression model for causal inference

- adjusting for confounding variables, we can estimate the Conditional Average Treatment Effect (CATE)

$$E[Y(1) - Y(0) \mid X = x]$$

- Imagine that the only confounder is EDUCATION,

$$E(Y_i^{obs} \mid t, educ) = \alpha + \tau t_i + \beta educ_i$$

- from this model we can estimate the average change in earnings due to training keeping education constant:
$E(Y^{obs} \mid t = 1, educ = c) - E(Y^{obs} \mid t = 0, educ = c) = \tau$

Multiple linear regression: STATA commands and results

$$\widehat{Y}_i^{obs} = -98.1288 - 12015.2 t_i + 1784.513 educ_i$$

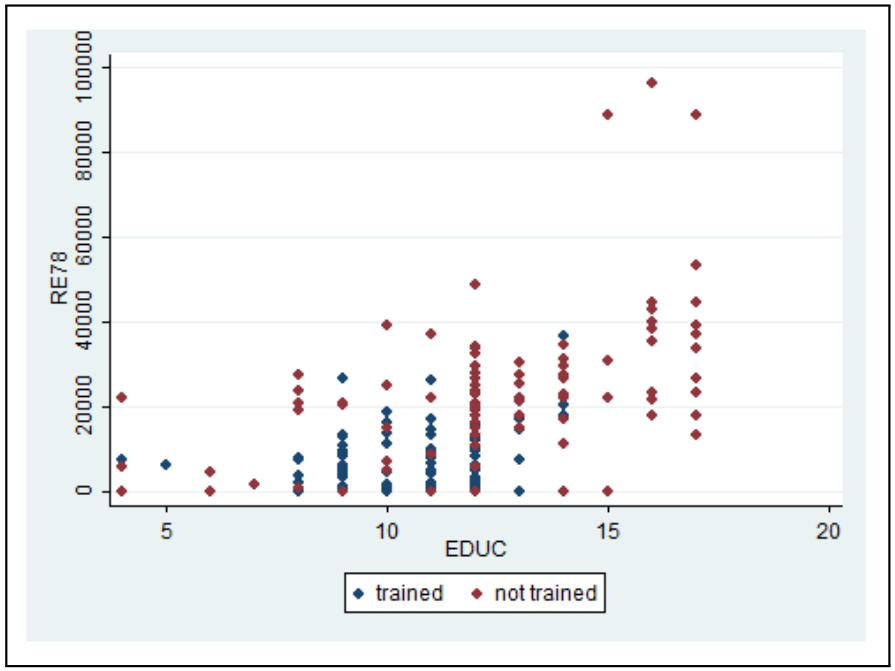
- the estimated average outcome for untreated units with zero years of education is given by $E(Y^{obs} \mid t = 0) = \widehat{\alpha} = -98.1288$ (not meaningfull)
- the estimated average change in earnings due to one additional year of education keeping training constant is positive $\widehat{\beta} = 1784.513$.
- the model assumes a homogeneous effect of training, i.e. the average change in earnings due to training keeping education constant is equal to

$$\widehat{\tau} = E(Y^{obs} \mid t = 1, educ = c) - E(Y^{obs} \mid t = 0, educ = c) = -12015.2$$

Multiple linear regression model for causal inference: identifying assumptions

The identifying assumption is unconfoundedness $(Y(1), Y(0) \perp\!\!\!\perp t | educ)$.

- At each value $educ = c$, we imagine an experiment took place and trained and not trained outcomes can be compared.
- e.g., for $educ = 10$ we compare the average re78 of trained and not trained and get the $ATE(educ = 10)$.
- Then we can average over the *educ* distribution to obtain the overall training effect.



Example: controlling for more covariates

Let us include all the pre-treatment variables available in the data set as independent variables $E(Y_i^{obs} \mid t, X) = \alpha + \tau t_i + \sum_{k=1}^{K} \beta_k X_{ki}$

```
reg re78 treat age educ ra rh marr re74 re75
```

re78	Coef.	Std. Err.	t	$P > \|t\|$
treat	864.3509	908.6754	0.95	0.342
age	-81.89111	20.7206	-3.95	0.000
educ	515.8077	76.85434	6.71	0.000
ra	-578.0219	496.0579	-1.17	0.244
rh	2415.323	1108.164	2.18	0.029
marr	1208.641	587.1255	2.06	0.040
re74	.2784607	.0279811	9.95	0.000
re75	.5692921	.0276004	20.63	0.000
cons	921.8091	1379.786	0.67	0.504

The estimated effect of training is positive (+864.3509 dollars) even tough it is not statistically significant (p-value $= 0.342$).

Regression: overlap problems

What could be wrong with the regression approach?

To identify causal effects, unconfoundedness is not enough, to achieve ignorability, we need also overlap, i.e. $0 < p_i(x) < 1$ for each value $x \epsilon X$

Let us consider the following example:

- we are interested in evaluate the effect of a binary treatment (trained or not) on a continuous outcome Y (e.g., earnings)
- we can assume unconfoundedness given an observed covariate X (e.g. education)
- it turns out that in the data at hand X assumes three values (1, 2, 3) for treated and only two (1 and 3) for control.
- This implies that treated with $X = 2$ cannot find good comparisons in the control group (no overlap).
- Unfortunately, the *regression analysis masks this fact* and assumes that the estimated equation is good for everybody (even for those never observed!!!).

Things are even worst with many confounders (we cannot easily see non overlap problems).

Pittfalls of the regression approach

- If the difference between the average values of the covariates in the two groups is large, the results are sensitive to the linearity assumption
- More generally, because we do not know the exact nature of dependence of the assignment on the covariates, this results in increased sensitivity to model and a priori assumptions
- Choice of covariates to be included in the model strongly affects results (cf. specification of propensity score)
- More recently, nonparametric regression estimators have been proposed (using a series approach or a local linear approach).
- in order to avoid model dependence we can consider **matching techniques**:
  - exact matching
  - propensity score matching

Matching

- Under unconfoundedness, the basic idea is to find in a large group of non-treated units, units similar to the treated subjects in all relevant pre-treatment characteristics X
- Since conditioning on all relevant covariates is limited in the case of a high dimensional vector X, we will use propensity score, i.e. the probability of being treated given observed characteristics X.
- Matching procedures based on this balancing score are known as propensity score matching.

Matching vs OLS

The main assumption underlying the matching approaches (unconfoundedness) is the same as OLS.
$\Rightarrow$ as OLS, the matching is as good as its X are!

Why matching could be better than OLS?

- The additional **common support** condition focuses on comparison of comparable subjects
- matching is a **non-parametric** technique:
  - it avoids potential misspecification of $E(Y(0) \mid X)$
  - it allows for arbitrary heterogeneity in causal effects $E(Y(1) - Y(0) \mid X)$
- If OLS is correctly specified, it is more efficient and we can make OLS less parametric adding iteractions (see later on)

Balancing scores and propensity scores

Conditioning on all relevant covariates is limited in the case of a high dimensional vector X

- Rosenbaum and Rubin (1983) suggest the use of so-called *balancing scores* $b(X)$, i.e. functions of the relevant observed covariates X such that the conditional distribution of X given $b(X)$ is independent of assignment into treatment.

$$X_i \perp\!\!\!\perp W_i \mid b(X_i)$$

- Balancing scores are not unique
- One possible balancing score is the **propensity score**, i.e. the probability to be treated given observed characteristics X.

$$e(X) = Pr(W = 1 \mid X = x) = E[W \mid X = x]$$

Propensity Score properties

Rosenbaun and Rubin (1983) demonstrate five theorems whose conclusions may be summarized as follows

- The propensity score is a *balancing score*:

$$Pr(W_i = 1 \mid X_i, e(X_i)) = Pr(W_i = 1 \mid X_i) = e(X_i)$$

- any score that is 'finer' than the propensity score is a balancing score; moreover, X is the finest balancing score and the propensity score is the coarset.
- If treatment assignment is strongly ignorable given X, then it is strongly ignorable given any balancing score, i.e. W_i is independent of X_i given the propensity score
- At any value of a balancing score, the difference between the treatment and control means is an unbiased estimate of the average treatment effect at that value of the balancing score if treatment assignment is strongly ignorable.

Propensity Score properties (cont'd)

- Given these properties, with strongly ignorable treatment assignment:
  - pair matching on a balancing score
  - subclassification on a balancing score
  - and covariance adjustment on a balancing score

  can all produce unbiased estimates of treatment effects.
- Using sample estimates of balancing scores can produce sample balance on X.

Conditioning on the propensity score

If **uncounfoundedness** holds then

- all biases due to observable covariates can be removed by conditioning solely on the propensity score:

$$(Y(0), Y(1)) \perp\!\!\!\perp W \mid e(X)$$

- The proof consists in showing that

$$Pr(W = 1 \mid Y(0), Y(1), e(X)) = Pr(W = 1 \mid e(X)) = e(X)$$

  implying independence of $(Y(0), Y(1))$ and W conditional on $e(X)$ (see next slide).

Conditioning on the propensity score (cont'd)

$$P(W = 1 \mid Y(0), Y(1), e(X)) = E[W \mid Y(0), Y(1), e(X)]$$

$$\begin{aligned} &= E[E[W \mid Y(0), Y(1), e(X), X] \mid Y(0), Y(1), e(X)] = \\ &= E[E[W \mid Y(0), Y(1), X] \mid Y(0), Y(1), e(X)] = \\ &= E[E[W \mid X] \mid Y(0), Y(1), e(X)] = \\ &= E[e(X) \mid Y(0), Y(1), e(X)] \quad = \quad e(X) \end{aligned}$$

where the last equality follows from unconfoundedness.
The same argument shows that

$$\begin{aligned} Pr(W = 1 \mid e(X)) \quad &= \quad E[W = 1 \mid e(X)] \\ = E[E[W = 1 \mid X] \mid e(X)] \quad &= \quad E[e(X) \mid e(X)] = e(X) \end{aligned}$$

The role of propensity score

Many of the procedures for estimating and assessing causal effects under unconfoundedness involve the propensity score.

- If the balancing hypothesis

$$W \perp\!\!\!\perp X \mid e(x)$$

  is satisfied, observations with the same propensity score must have the same distribution of observable (and unobservable) characteristics independently of treatment status.

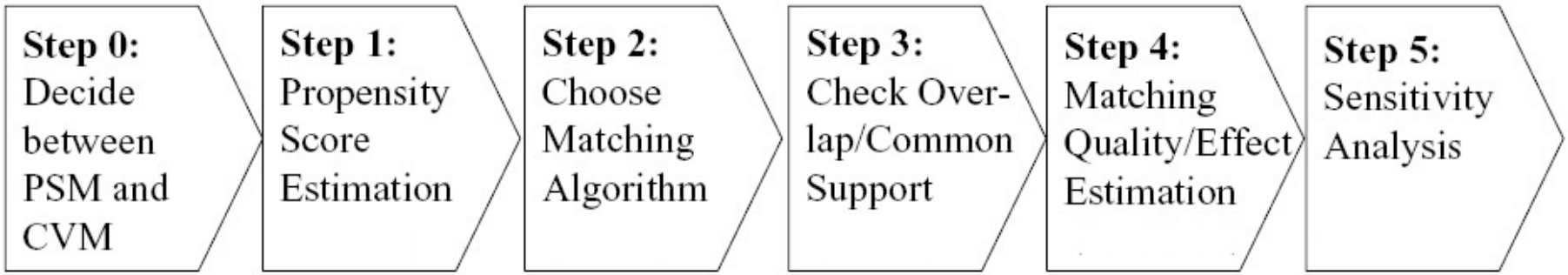
$\Rightarrow$ for a given propensity score, exposure to treatment is random and therefore treated and control units should be on average observationally identical.

The role of propensity score (cont'd)

- Usually, given a set of pre-treatment variables, unconfoundedness is viewed as a reasonable approximation to the actual assignment mechanism, with only vague a priori information about the dependence of the propensity score on the pre-treatment variables
- The true propensity score is generally unknown, so that the propensity score needs to be estimated.

Implementation of PS matching

Implementation of PSM requires the answer to a lot of questions.
The following Figure (from Caliendo and Kopeinig, 2008) summarizes the necessary steps when implementing propensity score matching



CVM: Covariate Matching, PSM: Propensity Score Matching

Propensity Score estimation

When the propensity scores are unknown but the assignment mechanism is regular, an important first step from the assignment-based perspective is to estimate them.

- The goal is to obtain estimates of the propensity score that statistically balance the covariates between treated and control subpopulations, rather than one that estimates the true propensity score as accurately as possible.
- With sufficient overlap, the use of estimated rather than true propensity scores typically yields more precise, not less precise estimates (Rubin and Thomas, 1992b).
- If there is little or no overlap in the distributions of the estimated propensity scores in the treatment groups, there is no hope of drawing valid causal inferences from these data without making strong external assumptions involving model-based extrapolation, because the estimated propensities will all be essentially either 0 or 1.

Propensity Score estimation (cont'd)

Any standard probability model can be used to estimate the propensity score:

$$Pr(W_i = 1 \mid X_i) = F(h(X_i))$$

- $h(X_i)$ is a function of covariates with linear and higher order terms
- $F(.)$ is a cumulative distribution, e.g. the logistic distribution

$$P(W_i = 1 \mid X_1) = \frac{\exp(h(X_i))}{1 + \exp(h(X_i))}$$

Propensity Score estimation (cont'd)

The inclusion of higher order terms in $h(X_i)$ is determined only by the need to obtain an estimate of the propensity score that satisfies the **balancing property**.

- The specification of $h(X_i)$ that satisfies the balancing property is usually more parsimonious than the full set of interactions needed to match cases and controls on the basis of observables
- the propensity score reduces the dimensionality problem of matching treated and control units on the basis of the multidimensional vector X

Pre-treatment variables choice

What variables should be included in the model for the PS?

In general, the choice of covariates to insert in the propensity score model should be based on

- theory and previous empirical findings;
- formal (statistical) tests (e.g. Heckman et al. , 1998, Heckman and Smith, 1999 and Black and Smith, 2004)
- The model for the propensity scores does not need a behavioral interpretation.

Pre-treatment variables choice (cont'd)

In the literature, some advice is available regarding the inclusion (or exclusion) of covariates in the propensity score model.

- Only variables that influence simultaneously the treatment status and the outcome variable should be included (see e.g., Sianesi, 2004; Smith and Todd, 2005)
- given that unconfoudedness requires the outcome variable(s) to be independent of treatment conditional on the propensity score $\Rightarrow$ we must choose a set of X that *credibly* satisfy this condition.
- Only variables that are unaffected by treatment should be included in the model. To ensure this, variables should either be fixed over time or measured before participation.
- If $e(X) = 0$ or $e(X) = 1$ for some values 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 condition (overlap) fails and matches cannot be performed.

Pre-treatment variables choice (cont'd)

In cases of uncertainty of the proper specification, sometimes the question may arise whether it is better to include too many rather than too few variables.

- Omitting important variables can seriously increase bias in resulting estimates (Heckman et al., 1997; Dehejia and Wahba, 1999)
- However, there are two reasons why over-parameterized models should be avoided (see Bryson et al., 2002):
  - it may be the case that including extraneous variables in the propensity score model exacerbates the support problem.
  - although the inclusion of nonsignificant variables in the propensity score specification will not bias the propensity score estimates or make them inconsistent, it can increase their variance.
- On the other hand, a variable should only be excluded from analysis if there is consensus that the variable is either unrelated to the outcome or not a proper covariate. If there are doubts about these two points, Rubin and Thomas (1996) recommend to include the relevant variables in the propensity score estimation.

Pre-treatment variables choice (cont'd)

In finite samples there is a trade-off between the plausibility of the unconfoudedness assumption and the variance of the estimates (see Black and Smith, 2004).

- when using all the available covariates, bias arises from selecting a wide bandwidth in response to the weakness of the common support.
- when using a lower number of covariates, common support is not a problem but the plausibility of the unconfoundedness assumption is.
- Moreover, this trade-off also affects the estimated standard errors, which are smaller for the minimal specification where the common support condition poses no problem.
- Finally, checking the matching quality can also help to determine the propensity score specification.

Checking the balancing property of estimated PS

- We can investigate whether, approximately, $W_i \perp\!\!\!\perp X_i \mid \hat{e}(X_i)$, by stratifying the sample into subsamples (blocks) with *similar* value of $\hat{e}(X)$, and then testing independence of W_i and X_i within each resulting stratum.
- For each covariate, test whether the means for the treated and for the controls are statistically different in all blocks.
- If one covariate is not balanced in one block, split the block and test again within each finer block.
- If one covariate is not balanced in all blocks, modify the specification of the propensity score adding more interaction and higher order terms and then test again.
- Note that at this stage we do not use the outcome data ($\Rightarrow$ no way of biasing the final estimation for the treatment effects).

Matching strategy and ATT estimation

The standard matching strategy is the following:

- pair each treated subject i with one or more *comparable* non-treated subjects
- associate to the outcome Y_i^{obs} a matched outcome $\widehat{Y}_i(0)$ given by the (weighted) outcomes of its *neighbors* in the comparison group

$$\widehat{Y}_i(0) = \sum_{j \epsilon C(i)} w_{ij} Y_j^{obs}$$

where

- $C(i)$ is the set of neighbors with $W = 0$ of the treated subject i
- w_{ij} is the weight of non-treated j, with $\sum_{j \epsilon C(i)} w_{ij} = 1$

Matching strategy and ATT estimation (cont'd)

The ATT

$$E[Y_i(1) - Y_i(0) \mid W_i = 1]$$

can be estimated as follows:

$$\widehat{ATT} = \frac{1}{N^T} \sum_{i:W_i=1} \left[Y_i^{obs} - \widehat{Y}_i(0) \right]$$

where N^T is the number of matched treated in the sample.

Exact matching

- Unconfoundedness suggests the following strategy for the estimation of the ATT:
  - stratify the data into cells defined by each particular value of X;
  - within each cell (i.e. conditioning on X) compute the difference between the average outcomes of the treated and the controls;
  - average these differences with respect to the distribution of X in the population of treated units.
- Exact matching may not be feasible if the sample is small, the set of covariates is large and many of them are multivalued, or, worse, continue.
  - With K binary variables the number of cells is 2^K. The number increases further if some variables take more than two values.
  - If the number of cells is very large with respect to the size of the sample it is possible that cells contain only treated or only control subjects.

Propensity-score matching

- Propensity score matching has the advantage of reducing the dimensionality of matching to a single dimension.
- Matching on the *true* propensity score leads to a $\sqrt{N}$-consistent, asymptotically normally distributed estimator.
- the first step in PSM is the estimation of the propensity score: this affects the large sample distribution of propensity score matching estimators.
- Abadie and Imbens (2009) derive the large sample distribution of PSM estimators and propose an adjustment to the large sample variance of propensity score matching estimators that corrects for first step estimation of the propensity score.

Matching methods

- An estimate of the propensity score is not enough to estimate the ATT
- In fact, the probability of observing two units with exactly the same value of the propensity score is in principle zero since $e(X)$ is a continuous variable.
- Several matching methods have been proposed in the literature. The most widely used are
  - Nearest-Neighbor Matching (with or without caliper)
  - Radius Matching
  - Kernel Matching
  - Stratification Matching.
- Typically, one treatment case is matched to several control cases, but one-to-one matching is also common and may be preferred (Glazerman, Levy, and Myers 2003).

Propensity-score matching with STATA

The Stata command `psmatch2` (Leuven and Sianesi 2003) will perform PSM

- many matching methods are available: nearest neighbor (with or without within caliper, with or without replacement), k-nearest neighbors, radius, kernel, local linear regression, and Mahalanobis matching;
- it includes routines for common support graphing (`psgraph`) and covariate imbalance testing (`pstest`);
- Standard errors are obtained using bootstrapping methods or variance approximation;
- It has a useful help file;
- Type `ssc describe psmatch2` to see a description and `ssc install psmatch2` to install

Propensity-score matching with STATA (cont'd)

Another useful Stata command is `pscore` (Becker and Ichino 2002)

- `pscore` estimates propensity scores and then ATT using various matching techniques.
- To obtain standard errors one can choose between bootstrapping and the variance approximation.
- Additionally, the program offer balancing tests based on stratification.
- Type `net search pscore` to find this command, or see http://www.lrz.de/ sobecker/pscore.html

Propensity-score matching with STATA (cont'd)

- The STATA command `nnmatch` (Abadie *et al.* 2004) implements covariate matching, where the user can choose between several different distance metrics. It also allows for
  - exact matching (or as close as possible) on a subset of variables
  - bias correction of the treatment effect, and estimation of either the sample or population variance, with or without assuming a constant treatment effect (homoskedasticity).
  - using the observations as a match more than once.
- A list of software for matching available in other packages (R, SAS, SPSS) is provided by Elisabeth Stuart: http://www.biostat.jhsph.edu/ estuart/propensityscoresoftware.html

Nearest Neighbor Matching

- NN match treated and control units taking each treated unit and searching for the control unit with the closest propensity score; i.e., the Nearest Neighbor.
- Although it is not necessary, the method is usually applied with replacement, in the sense that a control unit can be a best match for more than one treated unit.
- Once each treated unit is matched with a control unit, the difference between the outcome of the treated units and the outcome of the matched control units is computed.
- The ATT of interest is then obtained by averaging these differences.
- All treated units find a match. However, it is obvious that some of these matches are fairly poor because for some treated units the nearest neighbor may have a very different propensity score, and, nevertheless, he would contribute to the estimation of the treatment effect independently of this difference.

Nearest Neighbor Matching (cont'd)

- Let $e_i(x_i) = p_i$ the propensity score of the i-th unit
- Given a treated unit i, let $l_{m(i)}$ denote the index of the non-treated unit that is the m-th closest to unit i in terms of the distance measure based on the norm $\|.\|$

$$\sum_{j:W_j \neq W_i} \mathbb{I}\{\|p_j - p_i\| \leq \|p_l - p_i\|\} = m$$

- Let $C(i)_M$ denote the set of indices for the first M matches for unit i: $C(i)_M = \{l_1(i), \ldots, l_M(i)\}$

$$\widehat{Y}_i(0) = \frac{1}{M} \sum_{j \epsilon C(i)_M} Y_j^{obs}$$

Nearest Neighbor Matching (cont'd)

The formula for of the NN matching estimator is:

$$
\begin{aligned}
ATT^{NN} &= \frac{1}{N^T} \sum_{i:W_i=1} \left[Y_i^{obs} - \sum_{j \epsilon C(i)_M} w_{ij} Y_j^{obs} \right] \\
&= \frac{1}{N^T} \sum_{i:W_i=1} Y_i^{obs} - \frac{1}{N^T} \sum_{j \epsilon C(i)_M} w_j Y_j^{obs}
\end{aligned}
$$

- N^T is the number of observations in the treated group
- N_i^C is the number of controls matched with treated observation i
- w_{ij} is equal to $\frac{1}{N_i^C}$ if j is a control units of i, and zero otherwise
- $w_j = \sum_i w_{ij}$
- See Becher and Ichino (2002) for the variance formula of this estimator.

NN Matching: Trade-off between bias and variance

- How many nearest neighbors should we use?
  - Matching just one nearest neighbor minimizes bias at the cost of larger variance.
  - Matching using additional nearest neighbors increase the bias but decreases the variance.
- Matching with or without replacement?
  - Matching with replacement keeps bias low at the cost of larger variance.
  - Matching without replacement keeps variance low at the cost of potential bias.

Common support condition

- We can consider only the observations whose propensity score belongs to the intersection of the supports of the propensity score of treated and controls.
- The quality of the matches may be improved by imposing the common support restriction.
- However, matches may be lost at the boundaries of the common support and the sample may be considerably reduced $\Rightarrow$ imposing the common support restriction is not necessarily better (see Lechner 2001).

PS matching algorithm

1. Assume unconfoundedness: is it plausible on the basis of theory knowledge/common sense? If so, go ahead.
2. Estimate the probability of getting the treatment as a function of observable pre-treatment covariates (e.g., using a logit model).
3. Use predicted values of step 2 to generate propensity score $p_i(x)$ for all treatment and control unitis
4. Restrict samples to ensure common support.
5. Match treated units: for each unit find a sample of controls with similar $p_i(x)$.
6. Check the balancing: test that the means of each covariate do not differ between treated and control units.
7. If the means of one or more covariate differ, the balancing property is not satisfied and a less parsimonious specification of $h(X_i)$ is needed.
8. If balancing is not achieved, repeat steps 2-3-4-5.

PS matching algorithm (cont'd)

- The algorithm is recursive: if in step 6 balancing is not satisfactory then repeat steps 2-5
  - modify the matching algorithm (step 5)
  - and/OR modify the propensity score model (step 2).
- Note that the outcome plays no role in the algorithm for the estimation of the propensity score (similar to controlled experiments in which the design of the experiment has to be specified independently of the outcome).

Example: PS matching

Consider again the job training example and re-estimate the causal effect of training on `re78` using propensity score matching

- Assume unconfoundedness holds
- Estimate a logit model for the PS

  ```
  .logit treat age educ ra rh marr re74 re75 un74 un75
  ```

- Predict the $p_i(x)$ for each i

  ```
  .predict pscore, pr
  ```

- use `psmatch2` for matching: a simple NN matching without replacement; conditioning on the common support.
  - Since there are observations with identical propensity score values, the sort order of the data could affect matching results.
  - it is advisible to sort randomly the data before calling psmatch2.
- use `pstest` fot test the balancing

Example: Real earning and unemployed subjects

- In the distributions of real earnings before the treatment (`re74` and `re75`) there are some 0.

```
.sum re74 re75
```

Variable	Obs	Mean	Std. Dev.	Min	Max
re74	2666	18235.12	13719.03	0	137149
re75	2666	17861	13882.53	0	156653

- subjects with zero values were unemployed.
- The unemployed are likely to be the most interested in receiving the training
- In order to balance the proportion of unemployed in the treatment and control groups, we created two dummy indicators for unemployment and use these new variables together with real earnings in the propensity score model

```
.gen un74 =(re74==0)
.gen un75= (re75==0)
```

Example: psmatch2 output

```
. psmatch2 treat, pscore(pscore) outcome(re78) common noreplacement
```

- the `common` option imposes a common support by dropping treatment observations whose pscore is higher than the maximum or less than the minimum pscore of the controls.
- Default matching method is single nearest-neighbour (without caliper).
- the `noreplacement` option perform 1-to-1 matching without replacement (available for NN PS matching only).

Example: psmatch2 output (cont'd)

Summary of units off and on support (here we discard 3 treated units).

psmatch2: Treatment assignment	psmatch2: Common support Off suppo	On suppor	Total
Untreated	0	2,481	2,481
Treated	**3**	182	185
Total	3	2,663	2,666

Example: psmatch2 output (cont'd)

Estimated ATT

Var	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	6349.14537	21594.3797	-15245.2343	1154.91439	-13.20
	ATT	6258.48804	7311.88121	**-1053.39316**	1006.83285	-1.05

- We need to check balancing before trusting the ATT estimation!

Example: balance checking

`pstest` calculates several measures of the balancing of the variables before and after matching,

- `pstest` only considers balancing for the treated.
- the balance is checked considering (type `help ptest` for details) :
  - t-tests for equality of means in the treated and non-treated groups, both before and after matching: for good balancing, these should be non significant after matching.
  - the standardized bias before and after matching (formulae from Rosenbaum and Rubin, 1985): this should be less than 5% after matching.
  - the `summary` option outputs some diagnostics of covariate balancing before and after matching

Example: balance checking (cont'd)

- Almost none of the covariates is well balanced (requires %bias after matching $< 5\%$).
- The matching was not effective in building a good control group!
- Try other methods or model specifications.

```
.pstest age educ ra rh marr re74 re75 un74 un75 , sum
```

		Mean			%red	t-test	
var	Sample	Treated	Control	%bias	|bias|	t	$p>\|t\|$
age	Unmatched	25.816	34.801	-100.5		-11.53	0.000
	Matched	25.934	29.467	-39.5	60.7	-3.73	0.000
educ	Unmatched	10.346	12.156	-70.6		-8.02	0.000
	Matched	10.357	10.5	-5.6	92.1	-0.56	0.578
ra	Unmatched	.84324	.2503	148.1		18.14	0.000
	Matched	.84066	.71978	30.2	79.6	2.81	0.005

Example: balance checking (cont'd)

- The `sum` option of `ptest` gives a summary of the distribution of the abs(bias) before and after matching.
- The average % absolute bias before matching was 128.14%. After matching it becomes 39.06.
  $\Rightarrow$ matching reduces starting unbalancing but not satisfactorily (it should be $< 5\%$).

Is there something that we can do to improve the matching?

- We can change the matching method
  - in the NN method, all treated units find a match. However, some of these matches are fairly poor because for some treated units the nearest neighbor may have a very different propensity score
  - caliper matching and radius matching (among others) offer a solution to this prolem
- we can change the propensity score model and re-do the matching

Caliper matching

- NN matching (consider M=1): treated unit i is matched to the non-treated unit j such that

$$\|p_i - p_j\| = min_{k \epsilon W=0} \|p_i - p_k\|$$

- Caliper matching (Cochran and Rubin, 1973) is a variation of NN matching that attempts to avoid *bad* matches (i.e. p_j far from p_i) by imposing a tolerance on the maximum distance $\|p_i.p_j\|$ allowed.
- That is, for a a prespecified $\delta > 0$ treated unit i is matched to the non-treated unit j if

$$\delta > \|p_i - p_j\| = min_{k \epsilon W=0} \|p_i - p_k\|$$

- If none of the non-treated units is within δ from treated unit i, i is excluded from the analysis (which is one way of imposing a common support condition).
- A drawback of Caliper matching is that it is difficult to know a priori what choice for the tolerance level is reasonable.

Radius matching

Each treated unit is matched **only** with the control units whose propensity score falls into a predefined neighborhood of the propensity score of the treated unit.

- all the control units with p_j falling within a radius r from p_i

$$\|p_i - p_j\| < r,$$

  are matched to the treated unit i.
  How to choose the radius?
- The smaller the radius ...
  - ... the better the quality of the matches.
  - ... the higher the possibility that some treated units are not matched because the neighborhood does not contain control units.

Example: changing the propensity score model to improve balancing

- for not well balanced variables: include higher order terms (e.g., squared values) and/or interactions (guidelines in Caliendo and Kopeining, 2006 and Dehejia and Wahba, 1999)
- after many trials (see the do file) we found a good overall matching quality with the following specification:
  `logit treat age age2 educ educ2 educ_ra age ra rh marr re74 re75 re74_2 re75_2 un74 un75 ra_un74`
  - higher order terms (e.g. squared value of age, $age2 = age^2$)
  - interaction terms (e.g., $educ\_ra = educ \times ra$).
- for this specification of the pscore model
  - we observe the smallest average $\%abs(bias)$ after matching
  - only 1 variable had $\%abs(bias) > 5\%$ after matching and this was not extremely high (about 11%).
- with alternative specifications the mean $\%abs(bias)$ is worst and many covariates show $\%abs(bias) > 5\%$

Example: ATT estimation via pscore matching

```
. psmatch2 treat, pscore(pscore21) outcome(re78) common caliper(0.01)
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-st
re78	Unmatched	6349.14537	21594.3797	-15245.2343	1154.91439	-13.2
ATT	6208.06449	4688.79355	**1519.27095**	1946.17307	0.78	

- The **estimated ATT** is now positive (as expected, this data are very famous!)
- We must check the balancing with `pstest` to validate this result (see the do file)
  - The balancing is good for all covariates: $abs(bias) < 5\%$ and t-test not significant for all covariates with the only exception of un74 (showing a small unbalance of 11%).
  - the overall matching performance is good: after matching the average $abs(bias)$ is 3.39 (with the first pscore model it was 29.77).

Example: overlap checking

- we can summarize the pscores in the treatment and control group and count how many units are off-support (see the do-file)
  - the common support is $[0.0003456, 0.9870998]$
  - ... there are 0 treated and 1305 not-treated subjects out of the common support (due to caliper)!
  - an histogram of pscores by treatment group highlight overlap problems (to avoid the problem of controls with extremely low pscores we discard units with $p_i < 0.1$)

Example: pscore matching and regression

- Instead of matching, we can estimate a flexible (and similar to a matching) regression model by including interactions (allowing for heterogeneous effects) and higher order terms.
  - The `film` command of STATA allows to perform various steps to increase the flexibility of the regression model until it resembles a matching estimator
  - we can impose of the common support whit the common option
    `film re78 treat age age2 educ educ2 educ_ra ra rh marr re74 re75 re74_2 re75_2 un74 un75 ra_un74, common`
- The estimated ATT (1,708) is quite similar to the matching result.
- However, this specification was suggested by the matching procedure
- we probably would not have used the same model without being driven to it by the balancing checking!!!

Is matching better than regression?

- Both methods are appropriate ONLY when **unconfoundedness** (selection on observables) is plausible!
- The PSM forces the researcher to design the evaluation framework and check the data **before looking at the outcomes** (this should avoid cheating from the evaluator).
- PSM makes more explicit the comparison of treated and control units.
- Matching techniques are **nonparametric** (or semi-parametric if the pscores are estimated using a parametric model like the logit) and tend to focus attention on the common support condition.
- Matching do not impose any restriction on the heterogeneity of treatment effects (regression with interactions allow for heterogeneity but this is still limited by the functional form we impose)
- If treatment effects are homogeneous (rarely) or you know the correct functional form (rarely), then regression-based estimators are more efficient (lower variance).

When to not use matching ...

- if unconfoudedness is not a plausible assumption: when the selection into treatment (also) depend on unobservables that are correlated to the outcomes of interest matching estimators are biased
- In small samples: an acceptable balance on important covariates is rarely achieved

▶ the relevance of matching methods depends on the data availability for the specific policy evaluation problem.

What to do whit selection on unobservables ...

- Instrumental variable techniques.
- Bounding. The drawback is that they often give (if we are not willing to impose strong assumptions) quite wide an uninformative bounds.
- Sensitivity analysis: to assess the bias of causal effect estimates when the unconfoundedness assumption is assumed to fail in some specific and meaningful ways
  - e.g. Ichino, Mealli and Nannicini, 2007 proposed a strategy implemented in the `sensatt` command of STATA.
  - for our example it turns out that the baseline estimate is rather robust to deviations from the unconfoundedness assumption

References I

Caliendo M. and Kopeinig S. (2008). SOME PRACTICAL GUIDANCE FOR THE IMPLEMENTATION OF PROPENSITY SCORE MATCHING *Journal of Economic Surveys* , Vol. 22, No. 1, pp. 3172.

Cochran, W. and Rubin, D.B. (1973), Controlling Bias in Observational Studies, *Sankyha*, 35, 417-446.

Lechner, M. (2001). Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption, in: Lechner, M., Pfeiffer, F. (eds), *Econometric Evaluation of Labour Market Policies*, Heidelber ca/Springer, p. 43-58.

Guo S. and Fraser W.M. (2009). *Propensity Score Analysis: Statistical Methods and Applications*. Thousand Oaks, CA: Sage Publications.

Leuven E. and Sianesi B. (2003). "psmatch2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing". http://ideas.repec.org/c/boc/bocode/s432001.html.