

# Propensity Score Matching - A beginners guide

Lorraine Dearden  
Institute for Fiscal Studies  
and Institute of Education  
Email: [ldearden@ifs.org.uk](mailto:ldearden@ifs.org.uk)

# The Evaluation Question

- Question which we want to answer is
  - What is the effect of some treatment ( $D_i=1$ ) on some outcome of interest ( $Y_{1i}$ ) compared to the outcome ( $Y_{0i}$ ) if the treatment had taken place ( $D_i=0$ )
- Problem is that it is impossible to observe both outcomes of interest to get the true causal effect
- PSM invokes certain assumptions to find this missing counterfactual (as do other methods)

# What are we trying to measure?

- Average treatment effect for the population (ATE)
- Average treatment effect on the treated (ATT)
- Average treatment effect on the non-treated (ATNT)
- Usually interested in ATT:
  - OLS -  $ATT=ATE=ATNT$
  - RCT – only ATT
  - PSM and more flexible regression methods- ATE, ATT & ATNT
- How can we find the missing counterfactual  $E(Y_0/D=1)$  ?

# Randomised Experiment

- Randomly assign people to treatment group and control group
- If groups large enough, the distribution of all pre-treatment characteristics in the two groups should be identical so any difference in outcome can be attributed to the treatment
  - Not generally available
  - Not always solution
  - Only recovers ATT

# Non-experimental Methods

- Non-experimental methods use a variety of statistical methods to identify the causal impact of a treatment on an outcome of interest
- Generally rely on having good quality individual level data
- But the methods differ in the assumptions they make in order to recover the missing counterfactual

# Propensity Score Matching

- Regression approaches are a form of matching approach
- Propensity score matching is another matching approach
- Shares a number of assumptions with regression based approaches
- A lot more flexible but also much more computationally expensive

# Assumptions

- Need to have a treatment group and some type of appropriate non-treated group from which you can select a control group
  - Finding an appropriate and convincing control group is often the most difficult evaluation task
- Assume ALL relevant differences between the groups pre-treatment can be captured by observable characteristics in your data (X)
  - Having high quality and extensive pre-treatment observables is crucial!
- Common support – return to this if time

# Matching

- Involves selecting from the non-treated pool a control group in which the distribution of observed variables is as similar as possible to the distribution in the treated group
- There are a number of ways of doing this but they almost always involve calculating the propensity score  $p_i(x) \equiv Pr\{D=1 | X=x\}$

# The propensity score

- The propensity score is the probability of being in the treatment group given you have characteristics  $X=x$
- How do you do this?
- Use parametric methods (e.g. logit or probit) and estimate the probability of a person being in the treatment group for all individuals in the treatment and non-treatment groups
- Rather than matching on the basis of ALL  $X$ 's can match on basis of this scalar propensity score (Rosenbaum and Rubin (1983))

# How do we match?

- Nearest neighbour matching
  - each person in the treatment group choose individual(s) with the closest propensity score to them
  - can do this with (most common) or without replacement
  - not very efficient as discarding a lot of information from the control group

## ■ Kernel based matching

- each person in the treatment group is matched to a weighted sum of individuals who have similar propensity scores with greatest weight being given to people with closer scores
- Some kernel based matching use ALL people in non-treated group (e.g. Gaussian kernel) whereas others only use people within a certain probability user-specified bandwidth (e.g. Epanechnikov )
- Choice of bandwidth involves a trade-off of bias with precision

# Other methods

- Radius matching
- Caliper matching
- Mahalanobis matching
- Local linear regression matching
- Spline matching.....

# So what does PSM do?

- Gives us weights for the control group to make them look as similar as possible in terms of X's as treatment group
- Nearest neighbour PSM these weights are integers
- Other methods non-integers
- Sum of weights for control group sums to number of observations in treatment group
- Use weighted difference in mean outcomes between treatment and control group to find effect
  - So only have to matching once to find impact of treatment on all outcomes of interest –always use same weights

# So how do we choose best method?

- If matching has worked, then none of the  $X$ 's should differ between control and treatment group
- So do another *weighted* probit/logit and check this is the case
  - If PSM has worked – none of the  $X$ 's should be significant in determining whether you are in the treatment group
- Check to see whether there are any significant differences in the weighted means of  $X$ 's between pilot and control areas (simple t-test)
- Usually find that one method works better than the rest
- But sometimes find that groups are just too different and no matching methods can come up with plausible weights
- Check to see if some flexible regression method gives you same answer as preferred matching method

# Imposing Common Support

- In order for matching to be valid we need to observe participants and non-participants with the same range of characteristics
  - i.e for all values of characteristics  $X$  there are treated and non-treated individuals
- If this cannot be achieved
  - treated units whose  $p$  is larger than the largest  $p$  in the non-treated pool are left unmatched

# Examples of Propensity Score matching

- EMA evaluation for DfES was first study to use propensity score matching
- Been extensively used since then in a number of studies
- Also used in returns to schooling literature e.g. Blundell et. al. (2005)
- Used to look at whether Ethnic Parity in Job Centre Plus programs and benefits
  - Very difficult to find comparable White customers to Ethnic Minority customers for a lot of programs and benefits – so couldn't estimate whether there was 'Ethnic Parity' in majority of cases
  - Other methods needed here and PSM not the answer

# Other Issues

- Intention to treat or impact on actual participants?
  - ITT usually what you want but not always possible
- How good is your data – administrative data may not have sufficiently rich  $X$ 's to do PSM
- May need to use other methods like difference in difference (DiD) methods or matched DiD

# Do you have a treatment and control group?

- Sometimes policy initiatives are just introduced – no pilot so finding a control group difficult
- Sometimes timing of introduction, e.g. phasing offers possibilities
- Sometimes just need to accept that quantitative evaluation is just not possible