Identifying Causal Effects in Research and Evaluation: The Experimental and Quasi-Experimental Toolkit

Topic 2: Matching Approaches

Tim Maloney
Chief Economist
19-21 February 2019

MINISTRY OF SOCIAL DEVELOPMENT
TE MANATŪ WHAKAHIATO ORA

Introduction

- At the end of the previous topic we had a daunting task: control for all relevant determinants of some outcome to invoke the CIA. Problem is that many of these factors are unobservable.

- Put this in the context of an experiment. We want to know the effects of benefit status on wellbeing. Not everyone in the general population is likely to be eligible for a benefit. For CIA to work, those on and off the benefit need to look like random draws from some population. This gives us the causal interpretation we want.

- ***Matching*** is a strategy for 'trimming' or 'pruning' a sample. We want to compare treated individuals to a relevant group of controls (i.e., people who 'look like' beneficiaries).

CIA Revisited

- Recall that random assignment insured that selection bias term is zero:

$$E(Y_i|D_i=1) - E(Y_i|D_i=0) = \underbrace{E(Y_{1i}|D_i=1) - E(Y_{0i}|D_i=1)}_{\text{Average Treated Effect on the Treated}}$$

$$+ \underbrace{0}_{\text{Selection Bias}}$$

and the latent term can be replaced with its observable counterpart:

$$E(Y_i|D_i=1) - E(Y_i|D_i=0) = \underbrace{E(Y_{1i}|D_i=1) - E(Y_{0i}|D_i=0)}_{\text{Average Treated Effect on the Treated}}$$

- Random assignment gives us this unbiased treatment effect we want.

CIA Revisited

- Matching works if conditioning on some factor(s) (that partly determines selection into treatment) gives us essentially the same result as random assignment:

$$= \underbrace{E(Y_{1i}|\boldsymbol{X}_i, D_i = 1) - E(Y_{0i}|\boldsymbol{X}_i, D_i = 1)}_{\text{Average Treated Effect on the Treated}}$$

$$+ \underbrace{0}_{\text{Selection Bias}}$$

$$= \underbrace{E(Y_{1i}|\boldsymbol{X}_i, D_i = 1) - E(Y_{0i}|\boldsymbol{X}_i, D_i = 0)}_{\text{Average Treated Effect on the Treated}}$$

- Matching *could* give us this same unbiased treatment effect!

MATCHING EXAMPLE: One Covariate

- Suppose we have data on wellbeing measured with a Likert scale $Y_i = \{1,10\}$.

  For those on benefit ($D_i = 1$):

  $$\bar{Y}_1 = \frac{32}{5} = 6.4$$

  For those off benefit ($D_i = 0$):

  $$\bar{Y}_0 = \frac{56}{7} = 8.0$$

- As earlier, the benefit system appears to *reduce* wellbeing by 1.6 points.

Observation Number	Treatment D_i	Outcome Y_i
1	1	2
2	1	5
3	1	7
4	1	8
5	1	10
6	0	5
7	0	7
8	0	7
9	0	8
10	0	9
11	0	10
12	0	10

MATCHING EXAMPLE: One Covariate

- Suppose we also observe the *deprivation index* in the area of residence (1 = most deprived and 10 = least deprived).
- We want match treated with controls on deprivation scores.
- Do this mechanically:

#1 no counterpart – discard
#2 matches with #6
#3 matches with #8
#4 matches with #7 and #10
#5 matches with #9
Observations #11 and #12 are discarded

Observation Number	Treatment D_i	Outcome Y_i	Area Deprivation X_i
1	1	2	2
2	1	5	3
3	1	7	5
4	1	8	6
5	1	10	7
6	0	5	3
7	0	7	6
8	0	7	5
9	0	8	7
10	0	9	6
11	0	10	10
12	0	10	9

MATCHING EXAMPLE: One Covariate

- Using just the matched individuals, we get these mean wellbeing measures:

$$\bar{Y}_1 = \frac{5 + 7 + 8 + 10}{4} = \frac{30}{4} = 7.5$$

$$\bar{Y}_0 = \frac{5 + 7 + \frac{7+9}{2} + 8}{4} = \frac{28}{4} = 7.0$$

- Our matched estimator says that the benefit system *increases* wellbeing by 0.5 points!
- *Q*: Why did we get this change to the full sample? Is this enough?

MATCHING EXAMPLE: Two Covariates

- Suppose treatment is also influenced by education.
- Matching now gets more difficult!
- Consider #2. Previously matched with #6 (same deprivation area). Years of education are quite different. No exact match. #2 looks like #8 on education.
- *Q*: How do we choose between these controls?

Observation Number	Treatment D_i	Outcome Y_i	Area Deprivation X_i	Years of Education S_i
1	1	2	2	7
2	1	5	3	11
3	1	7	5	9
4	1	8	6	12
5	1	10	7	13
6	0	5	3	8
7	0	7	6	10
8	0	7	5	11
9	0	8	7	14
10	0	9	6	9
11	0	10	10	16
12	0	10	9	13

Propensity Score Matching

- Imagine further complicating this example. Could observe multiple factors related to treatment. How do we choose the appropriate match from the control group? Could be further complicated by including covariates that are continuous (e.g., detailed benefit and other histories). Exact matches would be virtually impossible.
- How do we deal with this more complex situation?
- We want to estimate the probability of treatment conditional on *all* potential predictors:

$$Prob\left(D_i = 1 \middle| X_i, S_i, \dots\right)$$

- Often use either maximum likelihood logit or probit for this purpose. The fitted value is the predicted probability of treatment.

Propensity Score Matching

- Estimated probability of treatment is a 'summary measure' of all the predictive factors, weighted by their importance in predicting treatment.
- Let's try to match observations:

#1 closest to #6 – discard?
#2 matches with #8
#3 matches with #8
#4 matches with #10
#5 matches with #9
Observations #11 and #12 are discarded

Observation Number	Treatment D_i	Outcome Y_i	Propensity Score $\widehat{Prob}_i$
1	1	2	0.7941
2	1	5	0.5071
3	1	7	0.4348
4	1	8	0.3856
5	1	10	0.2561
6	0	5	0.5992
7	0	7	0.3657
8	0	7	0.4324
9	0	8	0.2116
10	0	9	0.3827
11	0	10	0.0564
12	0	10	0.0895

Propensity Score Matching

- Using these matches, compute these mean wellbeing measures:

$$\bar{Y}_1 = \frac{2+5+7+8+10}{5} = \frac{32}{5} = 6.4$$

$$\bar{Y}_0 = \frac{5+7+7+9+8}{5} = \frac{36}{5} = 7.2$$

- Our PSM estimator says that the benefit system *decreases* wellbeing by 0.8 points!
- If 1st observation was discarded, we'd get a 1.7-point *increase* in wellbeing. Suggests how fragile these results could be in a small sample with uncertainty over 'how close is close'.
- Less of a problem with a large sample, but 'common support' is critical.

Distance vs Propensity Score Matching

- Previously mentioned recent working paper by King and Nielsen who claim that the enormously popular PSM method is suboptimal and potentially harmful. Can lead to imbalance, inefficiency and bias.
- Propensity scores have a lot of valuable uses, but matching isn't one of them!
- I'll use some simple diagrams to illustrate their concerns. We'll use the current example of looking for the causal effect of the benefit system on wellbeing.
- Need to understand the difference between ***Classic Randomisation (CR)*** and ***Fully Blocked Randomisation (FBR)*** that's at the heart of their reasoning.

Types of Randomisation

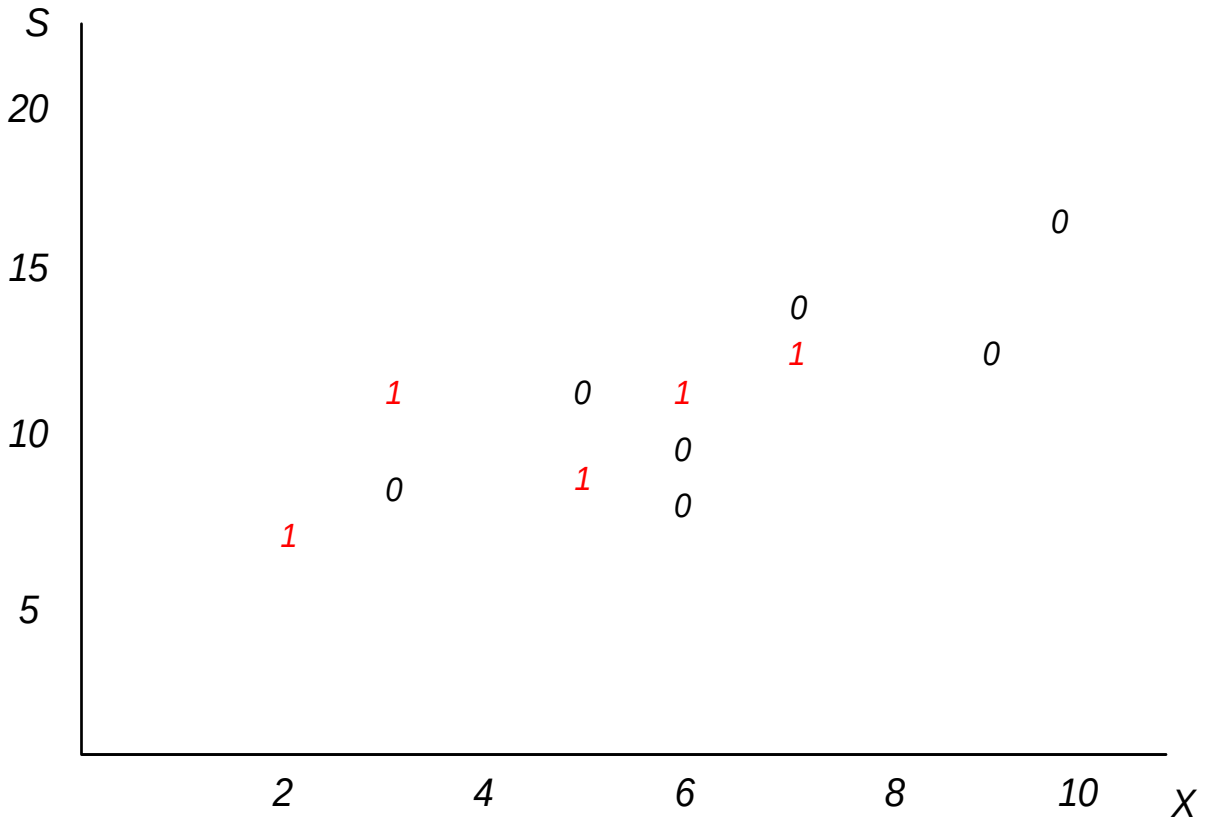
- ***Classic Randomisation*** is your typical RCT. Eligible individuals are randomly assigned to treatment and control groups. Balance occurs *on average* for both observable and unobservable factors.

- ***Fully Blocked Randomisation*** means that you find at least one control who matches a treated individual exactly. Random allocation happens within this 'block'. As a result, balance occurs *exactly* for observed factors (and still *on average* for unobservables).

- King and Nielsen show that ***FBR*** dominates ***CR*** in terms of balance, power, efficiency, bias and researcher costs. Any matching approach should try to emulate ***FBR***, not ***CR***. The former is the optimal target.

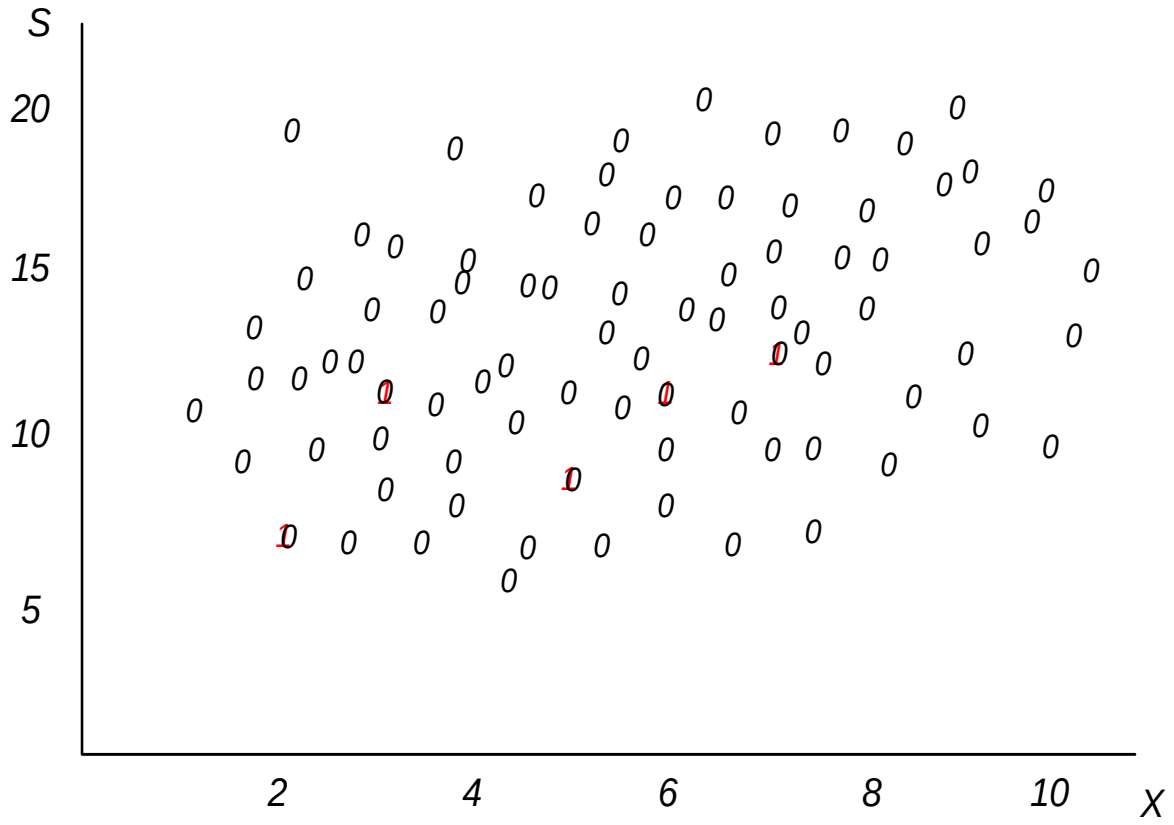
Graph of Two Covariates

- This is what our current data look like. Treated are *1*s and the control *0*s.



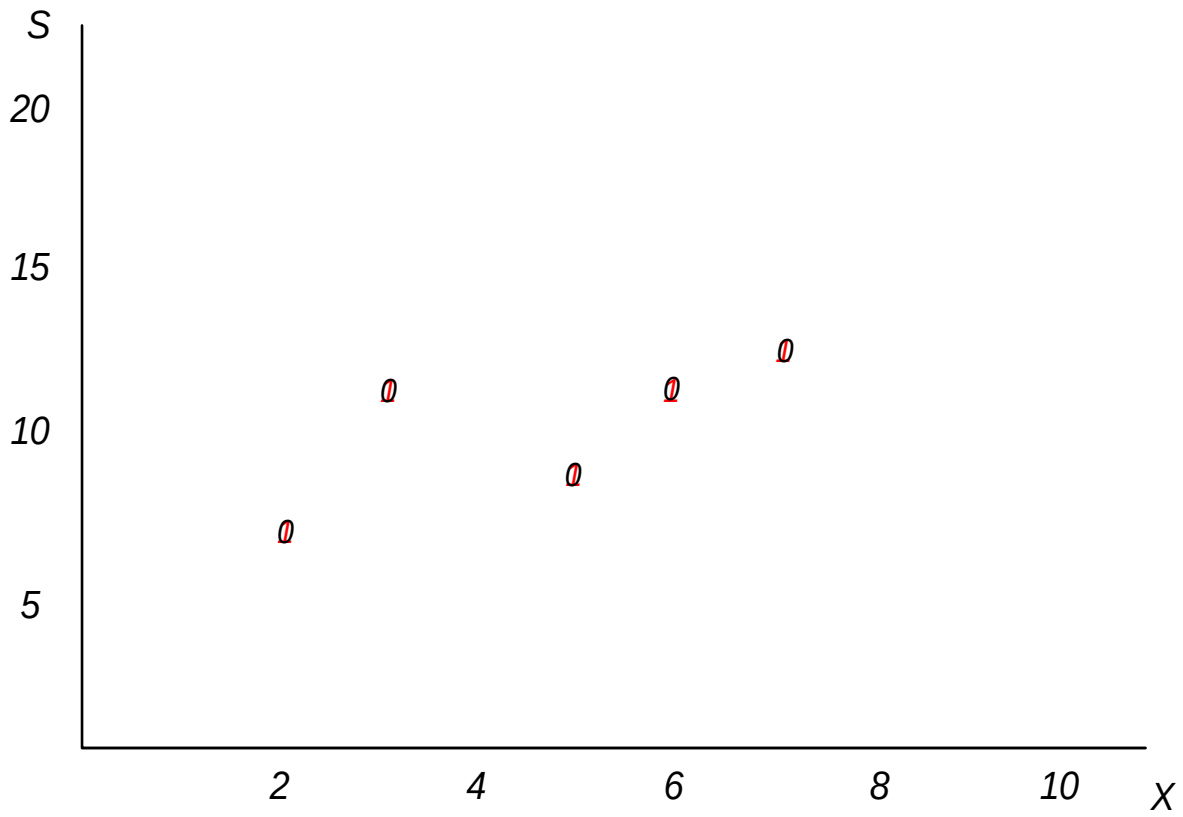
Best Case Scenario

- Ideally, space is saturated with controls. As sample increases, exact matches appear.



Pruning Gives Us the Counterpart of FBR

- This is known as ***Mahalanobis Distance Matching***.



PSM is Suboptimal

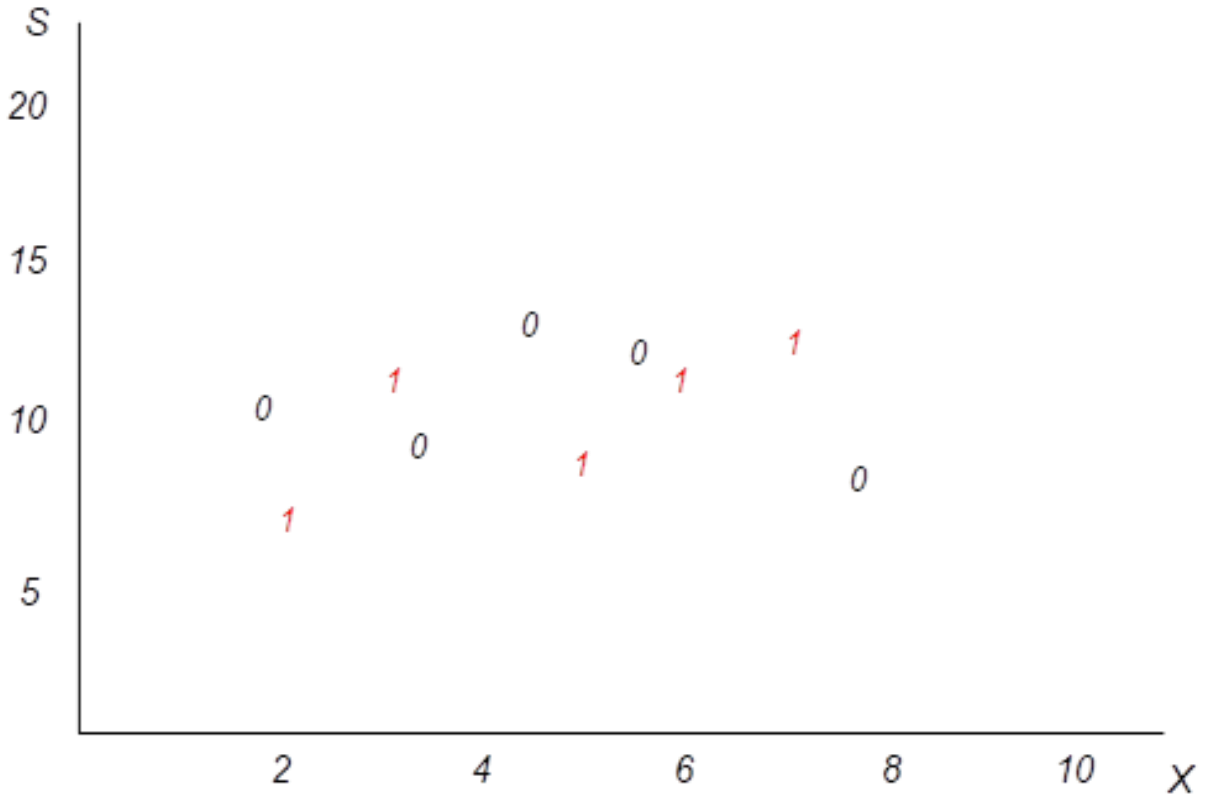
- King and Nielsen show that ***PSM*** is generally inferior. Reason is that ***PSM*** emulates ***CR***, while ***MDM*** emulates ***FBR*** (*and* ***FBR*** dominates ***CR***). So ***PSM*** is suboptimal because its target is 'second-best'.

- They motivate this by going back to the saturated data. ***PSM*** estimates the probability of treatment and matches individuals by minimising the probability distances. Suppose (and this is key) that everyone has the same predicted probability (all equally likely to be treated – the ideal situation). Matches will be chosen and controls discarded at random.

- The basic problem is $X_T = X_C \Rightarrow \hat{P}_T = \hat{P}_C$,
  but $\hat{P}_T = \hat{P}_C \not\Rightarrow X_T = X_C$.

PSM is Suboptimal

- Let's show what this would like in this extreme example.



MINISTRY OF SOCIAL DEVELOPMENT
TE MANATŪ WHAKAHIATO ORA

The PSM Paradox

- King and Nielsen summarise this with the ***PSM Paradox***: '*When you do better, you do worse*.'
- Pruning at random means that ***PSM*** can lead to imbalance, inefficiency and bias.
- My concern is that this is true for this extreme scenario where we have identical predicted probabilities of treatment. Not sure if this holds for the more typical situations where probabilities vary over the sample. [K&N say it doesn't matter because this says that the matches are poor anyway!]
- ***Conclusion***: Recent concerns over the appropriateness of PSM. Other matching methods may be far better from a conceptual and practical point of view.

Matching vs Regression Analysis

- Motivations behind matching and regression analysis are similar: Isolate treatment effect while holding other factors constant.

Relative Advantages of Matching

1. Less sensitive to functional form assumptions
2. Easier to assess when it's working well
3. Matching discards observations that aren't comparable
4. Easier to explain

Matching vs Regression Analysis

Relative Advantages of Regression Analysis

1. Matching works well for simple binary treatment, but not for more complex situations

2. Regressions estimate partial effects of all covariates on the outcome of interest

3. Allows interactions of treatment with other covariates

4. Replication is easier (in some sense)

In the end, the biggest threat to both matching and regression analysis is the same: Unobserved factors that influence treatment and the outcome of interest.