Lecture Notes on Propensity Score Matching

Jin-Lung Lin

This lecture note is intended solely for teaching. Some parts of the notes are taken from various sources listed below and no originality is claimed.

1 Introduction

A specific question: Is taking *math lessons after school* helpful in improving score? 補習數學有用嗎? (參考: 關秉寅、李敦義 (2008)。補習數學有用嗎? 一個「反事 實」的分析。台灣社會學刊,41,97-148)
A first attempt to answer this question would be computing the difference between the scores of those who took the after school lessons and those who don't.
By so doing, one assumes that all the students are similar and are randomly selected to take after school lessons.
In reality, these are two different groups with different characteristics that would affect the learning and scoring ability. In other words, there exist sample selection bias that seriously affects the validity of the analysis.

Some basic concepts:
Treatment = 補習數學 $(D = 1)$; 沒有補習數學 $(D = 0)$
$Y(1)$: 補習數學學生的數學成績; $Y(0)$: 沒有補習數學學生的數學成績

- ATE (Average Treatment Effect)
  如果所有的人國三都補數學的話, 數學成就會有差異嗎?
- ATT (Average Treatment Effect on the Treated):
  如果國三補數學的人, 沒補的話, 數學成就會有差異嗎?
- ATU (Average Treatment Effect on the Untreated):
  如果國三沒補數學的人, 補習的話, 數學成就會有差異嗎?

General questions: Is the treatment (for whatever) effective?
Fact: some people receive treatment.
Counterfactual question: *What would have happened to those who, in fact, did receive treatment, if they had not received treatment (or the converse)?*
In short, participants differ from nonparticipants and creates the *selection bias*. To minimize the

bias, we need to find a large group of nonparticipants those individuals who are similar to the participants in all relevant treatment characteristics.

Table 1: The counterfactual framework

	Potential outcomes	
Group	$Y(1)$	$Y(0)$
Treatment effect (D=1)	Observable $E(Y(1)\|D=1)$	Counterfactual $E(Y(0)\|D=1)$
Control group (D=0)	Counterfactual $E(Y(1)\|D=0)$	Observable $E(Y(0)\|D=0)$

2 Matching basics

Roy-Rubin model
main pillars: individual, treatment and potential outcome. For binary treatment, treatment indicator

$$D_i = \begin{cases} 1 & \text{if individual } i \text{ receives treatment} \\ 0 & \text{if individual } i \text{ does not receive treatment} \end{cases}$$

$Y_i(D_i)$ is the potential outcome for individual $i, i = 1, \cdots, N$. Treatment effect

$$\tau_i = Y_i(1) - Y_i(0)$$

Only one of $Y_i(1), Y_i(0)$ is observed and the other unobservable outcome is called *counterfactual outcome*. It is impossible to estimate τ_i for each i and we could only estimate the average treatment effect.

$$\begin{aligned} \tau_i &= Y_i(1) - Y_i(0) \\ \tau_{ATE} &= E(\tau) = E(Y(1) - Y(0)) \\ \tau_{ATT} &= E(\tau|D=1) = E(Y(1)|D=1) - E(Y(0)|D=1) \\ \tau_{ATU} &= E(\tau|D=0) = E(Y(1)|D=0) - E(Y(0)|D=0) \end{aligned}$$

Population average treatment effect (ATE), τ_{ATE}, answers the question *What is the expected effect of the outcome if individuals in the population were randomly assigned to treatment?* τ_{ATE} is not interesting because it includes the effects on persons not intended.
On the other hand, τ_{ATT}, average effect of the treated is defined as the difference between expected outcome values with and without treatment for those who actually participate in treatment. It determines the realized gross gain from the programme and can be compared with its cost. $E(Y(0)|D=1)$ is counterfactual (unobserved) and $E(Y(0)|D=0)$ is usually not a good proxy. There exists selection bias.

$$E(Y(1)|D=1) - E(Y(0)|D=0) = \tau_{ATT} + E(Y(0)|D=1) - E(Y(0)|D=0)$$

τ_{ATT} is only identified if the selection bias, $E(Y(0)|D=1) - E(Y(0)|D=0) = 0$

Furthermore, let $P(D=1)=\pi$, then

$$\begin{aligned}
\tau_{ATE} &= E(\tau)=E(Y(1)-Y(0)) \\
&= [\pi E(Y(1)|D=1)+(1-\pi)E(Y(1)|D=0)]-[\pi E(Y(0)|D=1)+(1-\pi)E(Y(0)|D=0)] \\
&= \pi[E(Y(1)|D=1)-E(Y(0)|D=1)]+(1-\pi)[E(Y(1)|D=0)-E(Y(0)|D=0)] \\
&= \pi E(\tau|D=1)+(1-\pi)E(\tau|D=0) \\
&= \pi ATT+(1-\pi)ATU
\end{aligned}$$

Regards to previous example,
不同組的人有同樣的因果效應嗎?
Yes, $\Rightarrow E[Y(1)|D=0]=E[Y(1)|D=1]$,
No, $E[Y(1)|D=0]-E[Y(1)|D=1]$ is the baseline bias.

不同組的人在未接受 treatment 前是一樣的嗎?
Yes, $\Rightarrow E[Y(0)|D=1]=E[Y(0)|D=0]$
No, $E[Y(0)|D=1]-E[Y(0)|D=0]$ is the differential effect bias

Then,

$$\begin{aligned}
E[Y(1)|D=1]-E[Y(0)|D=0] &= E(\tau)+\pi[E(Y(0)|D=1)-E(Y(0)|D=0)] \\
&+ (1-\pi)[E(\tau|D=1)-E(\tau|D=0)]
\end{aligned}$$

Naive Estimate = average causal effect + baseline bias + differential effect bias

Fundamental assumptions: Unconfoundedness and Common Support

Assumption 1: *Unconfoundedness*: $Y(0),Y(1)\perp D|X$
Given a set of observable covariates, X, which is not affected by treatment, potential outcomes are independent of treatment assignment. This implies that all variables that influence treatment assignment and potential outcomes simultaneously have to be observed by the researchers. Unconfoundedness is also called selection on observable or conditional independence.

Assumption 2: *Overlap*: $0<P(D=1|X)<1$
Persons with the same X values have a positive probability of being participants and nonparticipants.

Assumption 3: *Unconfoundedness for controls*: $Y(0)\perp D|X$

Assumption 4: *Weak overlap*: $P(D=1|X)<1$.

Troubles: as the dimension of X increases, the unconfoundedness is difficult to hold. Rosenbaum and Rubin (1983) suggested using balancing score $b(X)$. The propensity score, $P(D=1|X)=P(X)$, the probability for an individual to participate in a treatment given his observed covariates X, is one balancing score.

Corollary 1. *Unconcernedness given the propensity score*: $Y(0), Y(1) \perp D|P(X)$
Estimation strategy

$$\tau_{ATT}^{PSM} = E_{P(X)|D=1}(E(Y(1)|D=1, P(X)) - E(Y(0)|D=1, P(X))$$

PSM estimator is the mean difference in outcomes over the common support, appropriately weighted by the propensity score distribution of participants.

3 Implementation of Propensity Score Matching

3.1 Estimating the propensity score

Two choices:

1. Model to be used for the estimation
2. Variables to be included in this model

Model choice - Binary Treatment

- logit model
- probit model
- linear probability model

Model choice - Multiple treatments

- multinominal probit model
- multinominal logit model
- Series of binomial model
- linear probability model

variable choice

- Omitting important variables can seriously increase bias in the estimation.

- Only variables that influence simultaneously the participation decision and the outcome variable should be included.
- Only variables unaffected by participation should be included in the model.
- participants and nonparticipants should stem from the same source (dataset).
- Should avoid including too many variables as execrates the support problem and increases the variance.

3.2 Steps of Implementation PSM

Step 0: Decide between PSM and CVM (covariate matching)

Step 1: Propensity Score estimation

Step 2: Choose matching algorithm

Step 3: Check overlap/common support

Step 4: matching quality/effect estimation

Step 5: sensitivity analysis

3.3 Matching algorithm

Distance measures

1. Exact:
$$M_{ij} = \begin{cases} 0 & \text{if } X_i = X_j \\ \infty & \text{if } X_i \neq X_j \end{cases}$$
2. Mahalanobis:
$$M_{ij} = (X_i - X_j)'\Sigma^{-1}(X_i - X_j)$$
where Σ is the covariance matrix of X in the full control group.
3. Propensity score:
$$M_{ij} = |e_i - e_j|$$
4. Linear propensity score:
$$M_{ij} = |logit(e_i) - logit(e_j)|$$

5. Fine balance:

$$M_{ij} = \begin{cases} (Z_i - Z_j)'\Sigma^{-1}(Z_i - Z_j) & \text{if} \quad |logit(e_i) - logit(e_j)| \leq c \\ \infty & \text{if} \quad |logit(e_i) - logit(e_j)| > c \end{cases}$$

6. Prognosis score
   The predicted outcome each individual would have under the control condition. where c is the caliper.

1. Nearest neighbor matching

$$M_i = min_j |P_i - P_j|,\ j \in I_0$$

   nonparticipant with the value of M_j that is closet to P_i is selected as the match.

   - 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

2. 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

3. Caliper matching
   A match for person i is selected only if

$$|M_i - M_j| < \epsilon, j \in I_0$$

   where ϵ prespecified tolerance, usually $.25\sigma_m$.

   - 1-to-1 Nearest neighbor within caliper
   - 1-to-n Nearest neighbor within caliper

4. Radius matching
   1-NN only or more

5. Stratification and interval

   - Group sample into five categories based on propensity score (quintiles).
   - Within each quintile, calculate mean outcome for treated and nontreated groups.
   - Estimate the mean difference (average treatment effects) for the whole sample (i.e., all five groups) and variance using the following equations:

   $$\hat{\delta}\sum_{k=1}^{K}\frac{n_k}{N}[\bar{Y}_{0k}-\bar{Y}_{1k}], \quad Var(\hat{\delta})=\sum_{k=1}^{K}(\frac{n_k}{N})^2 Var[\bar{Y}_{0k}-\bar{Y}_{1k}]$$

   Number of strata intervals

6. Mahalanobis matching

   $$M_{ij}=(X_i-X_j)'\Sigma^{-1}(X_i-X_j)$$

   - Mahalanobis metric matching without p-score
   - Mahalanobis metric matching with p-score added (to X_i and X_j)

7. Local linear regression matching

8. Spline matching

3.4 So what does PSM do?

- Propensity score is the probability of taking treatment given a vector of observed variables.

  $$p(x)=Pr[D=1|X=x]$$

  If we take individuals with the same propensity score, and divide them into two groups- those who were and weren't treated-the groups will be approximately balanced on the variables predicting the propensity score.

- Among those with the same predicted probability of treatment $\hat{p}$, those who get treated and not treated differ only on their error term in the propensity score equation. But this error term is approximately independent of the X's. The treatment assignment Dis independent of Y, given the strata created by X's. This is why balancing should occur.

  $$Y\perp D|X$$

- Common support: the overlap condition for persons with the same x value in X are allowed to have a positive probability of being in treated and control groups. We only make inferences where we have sufficient data. Unlike ordinary regression, we dont extrapolate outside the range of the observed data points.
- Gives us weights for the control group to make them look as similar as possible in terms of X's as treatment group
- Nearest neighbor 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

3.5 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

3.6 Imposing Common Support

- In order for matching to be valid we need to observe participants and nonparticipants 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

4 A practical example

關秉寅、李敦義 (2008)。補習數學有用嗎? 一個「反事 實」的分析。台灣社會學刊,41,97-148。

研究使用的資料:TEPS 國中樣本 (公開使用版) 2001(N = 13,978) 2003(N = 13,247) 分析樣本 : 公立國中生;with common support (N = 10,013)

應變項: 國三數學能力 IRT, 轉換成 NCE (normal curve equivalence) 分數 (Range: 1 - 99; Mean: 50; S.D.: 21.06) 自變項 (Treatment): 國三補習數學 26個配對變項: 個人特性及學習特質: 性別、補習經驗、W1數學 IRT 等 (motivation, ability) 家庭背景: 父母教育程度、職業、教育期待等 班級/學校學習氣氛及環境

分析策略: 只研究國三數學補習的效果 國三補習主要是爲了準備基測 學校教育對數學能力的培養比較有影響力 以階層性模型探討國三補習的參與及補習效果, 以瞭解過往補習研究可能有的限制, 以及比較 OLS 及 PSM 兩者估計 ATE 可能有的差異 比較有及沒有 W1數學 IRT 做爲配對變項的差異

誰參加國三數學補習? 個人特性及學習特質: 先備能力較佳、過去沒補習經驗者、回家會複習功課、自己沒有補習的意願 家庭背景: 非原住民、與雙親同住、父母不是研究所學歷、白領職業、高收入、高父母教育期望、手足人數少 班級/學校情況: 位於在都市化程較高地區、班上讀書風氣盛、學業競爭激烈程度高

國三補習數學有用嗎? Gross effect (OLS): 12.243(分析樣本 with common support) After controlling all matching variables (OLS): 3.017 - an estimate of ATE PSM results (all matching variables included): Total population (ATE): 2.956 Treated (ATT): 2.258 Untreated (ATU): 3.580

PSM的 ATE 估計大多比 OLS 的估計小 ATT比 ATU 小 都會受到未納入重要自變項的影響

針對國三補習數學者與從未補習者配對成功的4689對中進行敏感度分析的結果是, 如果未觀察到之變項對影響補習參與與否的影響力, 即 Gamma (γ), 是介於1.25到1.35時, 就可能會變成不顯著。取這些數値的自然對數, 則分別爲0.223及0.300 , 如與實際配對變項對是否補習之邏輯迴歸係數比較, 則此數値大約是完整家庭 (nuintact) 的係數 (.226)。也就是說, 這未觀察到變項對補習與否的影響力至少要像完整家庭一樣大, 才會影響 ATT 的變化

5 History

In 1983, Rosenbaum and Rubin published their seminal paper that first proposed this approach. · From the 1970s, Heckman and his colleagues focused on the problem of selection biases, and traditional approaches to program evaluation, including randomized experiments, classical matching, and statistical controls. Heckman later developed ”Difference-in-differences” method

6 Skeptics

Howard Bloom, MDRC · Sees PSM as a somewhat improved version of simple matching, but with many of the same limitations · Inclusion of propensity scores can help reduce large biases, but significant biases may remain · Local comparison groups are best- PSM is no miracle maker (it cannot match unmeasured contextual variables) · Short-term biases (2 years) are substantially less than medium term (3 to 5 year) biases- the value of comparison groups may deteriorate Michael Sosin, University of Chicago · Strong assumption that untreated cases were not treated at random · Argues for using multiple methods and not relying on PSM

7 Limitations

Limitations of Propensity Scores · Large samples are required · Group overlap must be substantial · Hidden bias may remain because matching only controls for observed variables (to the extent that they are perfectly measured) (Shadish, Cook, & Campbell, 2002)

8 Criteria

Identify treatment and comparison groups with substantial overlap · Match, as much as possible, on variables that are precisely measured and stable (to avoid extreme baseline scores that will regress toward the mean) · Use a composite variable- e.g., a propensity score- which minimizes group differences across many scores

9 Risk

They may undermine the argument for experimental designs- n argument that is hard enough to make, now · They may be used to act ”as if” a panel survey is an experimental design, overestimating the certainty of findings based on the PSM.

The crucial difference of PSM from conventional matching: match subjects on one score rather than multiple variables:” the propensity score is a monotone function of the discriminant score”(Rosenbaum & Rubin, 1984).
Run Logistic Regression: · Dependent variable: Y=1, if participate; Y = 0, otherwise. · Choose appropriate conditioning (instrumental) variables. · Obtain propensity score: predicted probability (p) or log[p/(1-p)].General Procedure Multivariate
Match Each Participant to One or More Nonparticipation Propensity Score..Nearest neighbor matching..Caliper matching..Mahalanobis metric matching in conjunction with PSM..Stratification matching..Difference-in-differences matching (kernel & local linear weights)

Sources

- 關秉寅 (2010),「因果推論新思維: 反事實分析架構, ”A New Paradigm for Causal Inference: The Counterfactual Framework,” available at

  `http://www3.nccu.edu.tw/~soci1005/Counterfactuals%20and%20PSM.ppt`

- Caliendo, Marco and Sabine Kopeinig, ”SOME PRACTICAL GUIDANCE FOR THE IMPLEMENTATION OF PROPENSITY SCORE MATCHING,” *Journal of Economic Surveys* 2008, 22, 31-2.

- Dehejia, Rajeev H. and Sadek Wahba ”PROPENSITY SCORE-MATCHING METHODS FOR NONEXPERIMENTAL CAUSAL STUDIES,” *The Review of Economics and Statistics*, 2002, 84, 151-161.

- Guo, S, R. Barth, and C. Gibbons (2004) , ”Introduction to Propensity Score Matching: A New Device for Program Evaluation,” available at

  `http://ssw.unc.edu/VRC/Lectures/PSM_SSWR_2004.pdf`

- Stuart, Elizabeth A., ”Matching methods for causal inference: A review and a look forward,” To appear in *Statistical Science*