Propensity Score [brief introduction]

✉ ehsan.karim@ubc.ca
Oct 7, 2020
SPPH 504/007

Reference

Austin, P. C. (2011).
A tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality.
Multivariate behavioral research, 46(1), 119-151.

Propensity score 🧪

1. **Definition**: the propensity score is
   - Probability of receiving treatment (exposure, A) given covariates (L).
   - P(L) = Pr(A=1|L)
2. **Properties**
   - Balancing score.
   - P(L)=0.5 in RCT.
3. **Assumes**
   - no unmeasured confounding $Y(1)$, $Y(0) \perp A \mid P(L)$.
   - positivity: 0<P(L)<1.
   - Sufficient overlap. If there is no overlap, can't compare $Y(0)$,$Y(1)$

Propensity score ⚗

- Modelling P(L) = Pr(A=1|L)
  - Any method that gives <u>good predictions</u> is useful.
    i. Logistic regression typically used
    ii. Machine learning methods also reasonable
  - only the <u>predictions matter</u>, the coefficients (in the PS model) don't
  - model can be <u>rich</u>

Propensity score 🧪

- Variables to include (requires subject area-expertise)
  - Include only <mark>pre-baseline</mark> measures
  - **Confounders**: important to include
  - **Risk factors** /**Predictors of Y**: include to reduce SE
  - **Instruments**/**Predictors of A only**: avoid
  - **Noise**: avoid (increases SE)
  - Don't look at outcome data while modelling PS

Various Propensity score analyses approaches

How can I use propensity scores?

- Matching ⊘
- Weighting ⊘
- Stratification (will not cover)
- Propensity score as a covariate (will not cover)

Propensity score Matching (ATT)

Propensity score matching

How to conduct propensity score matching? ⚙️⚙️

Step 1: Specify PS & fit model ⟶ ==Exposure model (RA)==

Step 2: Match subjects by PS

Step 3: Covariate balance in matched sample

Step 4: Estimate treatment effect ⟶ ==Outcome model (MI)==

For the purposes of illustration, we will first assume that our <u>data was collected via SRS</u>.

Propensity score matching: step 1

Step 1: Specify PS
Model and
fit that model

PS.model/formula specification: $A \sim L$

PS.fit = logistic($A \sim L$)

Predict from PS.fit
Get the predicted values from the fitted logistic regression

Propensity score matching: step 1

Step 1:

Plot the predicted values / propensity scores

Numerical summary of PS distribution

```
## $`Non-arthritis`
##     Min. 1st Qu.   Median     Mean 3rd Qu.     Max.
## 0.001809 0.013710 0.031450 0.064314 0.080949 0.733198
##
## $`Rheumatoid arthritis`
##     Min. 1st Qu.   Median     Mean 3rd Qu.     Max.
## 0.006047 0.087135 0.148875 0.190836 0.270072 0.792427
```


10
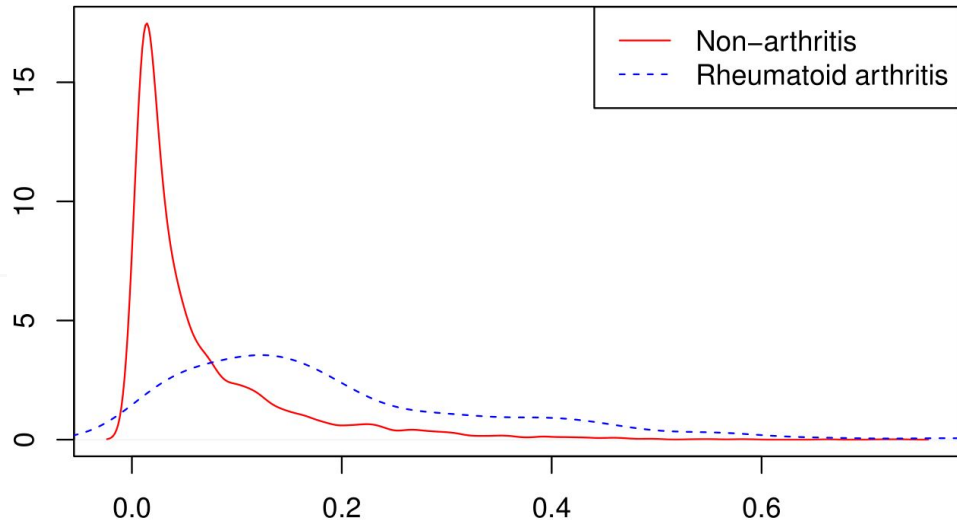
Propensity score matching: step 2

Step 2: Match subjects by PS

Different algorithms are available to match propensity scores

- Nearest Neighbor (NN) matching: <u>selects the closet PS in the control</u>
- NN & caliper matching: <u>pre-defined bound</u>
- Optimal matching
- Coarsened exact matching / CEM
- Full Matching

Propensity score matching: step 2

Step 2: Match subjects by PS

Nearest Neighbor
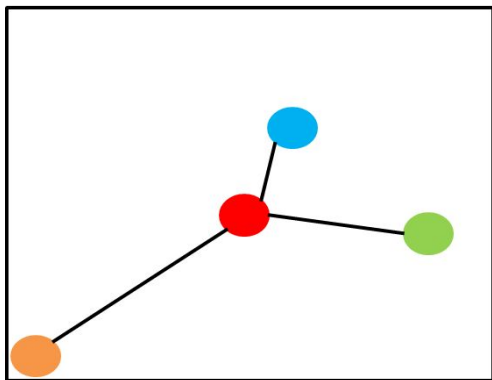
Nearest Neighbor + caliper

Treated

Control

Treated

Control

b
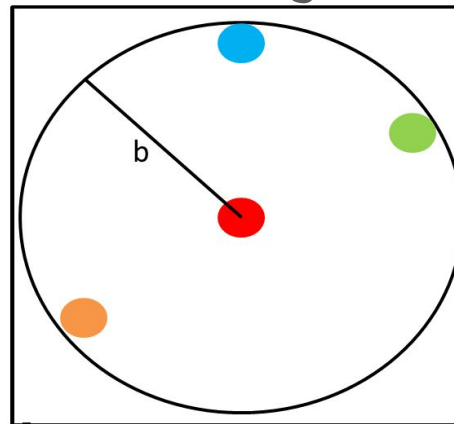
Randomness involved if tied

Propensity score matching: step 2

Step 2:

Match subjects

by PS

Good idea to set seed because some randomness is involved.

Match with the following criteria:
- First get PS from a <u>logistic regression</u> (logit link)
- Using those PS, perform <u>nearest-neighbor</u> matching
- Match <u>without replacement</u>
- Pair matching (ratio = <u>1:1</u> for RA vs. non-arthritis)
- <u>Caliper</u> = 0.2*sd(PS)

Summarize the PS

Propensity score matching: step 2

: Match subjects by PS

Distribution of Propensity Scores
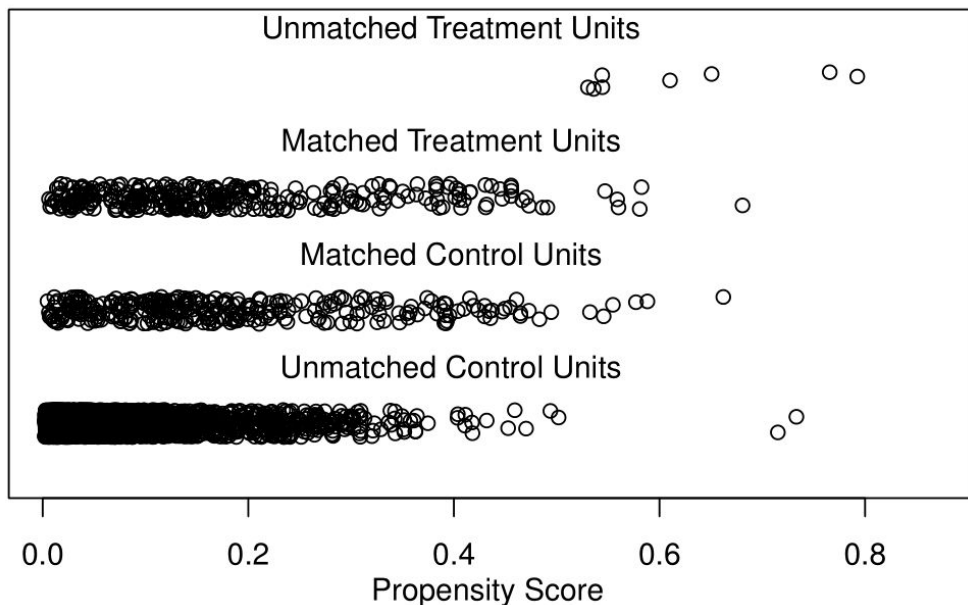

```
## Sample sizes:
##           Control Treated
## All          4089     325
## Matched       317     317
## Unmatched    3772       8
## Discarded       0       0
```

Propensity score matching: step 3

Step 3: Covariate balance in matched sample, check graphically
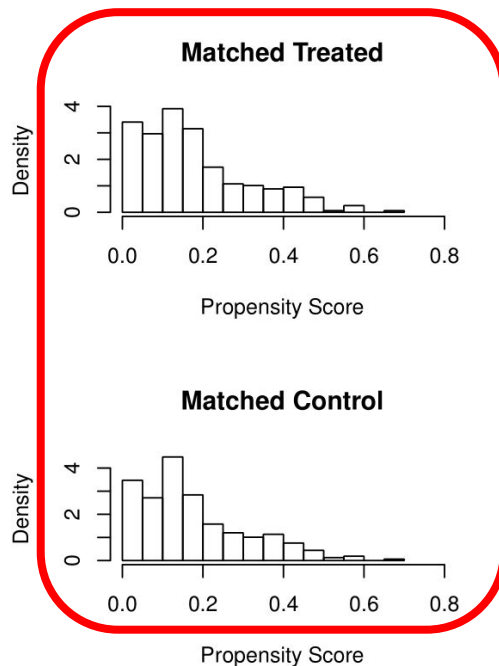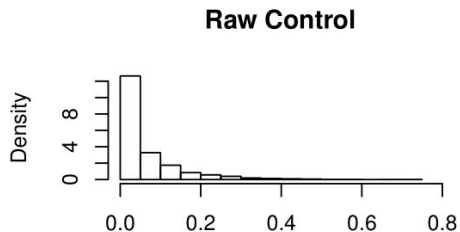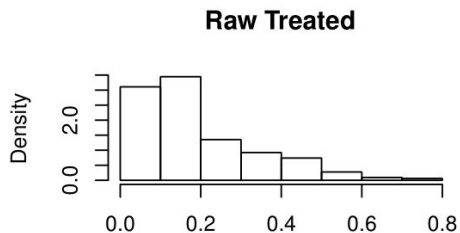
step 3

Step 3

Covariate balance checking using

SMD

(<0.2,

Or

<0.1)

Unmatched

```
##  Unmatched             Stratified by arthritis.type
##                        Non-arthritis Rheumatoid arthritis SMD
##   n                    4089          325
##   gender = Female (%)  1960 (47.9)    194 (59.7)          0.238
##   bmi = (25,80] (%)    2745 (67.1)    245 (75.4)          0.183
##   diabetes = Yes (%)    358 ( 8.8)     87 (26.8)          0.485
##   smoke = Yes (%)      1796 (43.9)    177 (54.5)          0.212
##   age (%)                                                 0.891
##     (0,50]             2577 (63.0)     74 (22.8)
##     (50,70]            1046 (25.6)    169 (52.0)
##     70+                 466 (11.4)     82 (25.2)
##   race (%)                                                0.347
##     White              1739 (42.5)    127 (39.1)
##     Black               843 (20.6)    114 (35.1)
##     Other              1507 (36.9)     84 (25.8)
##   born = USborn (%)    2912 (71.2)    262 (80.6)          0.221
##   education (%)                                           0.160
##     School              495 (12.1)     52 (16.0)
##     College            1892 (46.3)    127 (39.1)
##     High.School        1702 (41.6)    146 (44.9)
##   marriage = Married (%) 2468 (60.4)  152 (46.8)          0.275
##   annualincome (%)                                        0.531
##     <20k                820 (20.1)    135 (41.5)
##     20kto54k           1737 (42.5)    126 (38.8)
##     55k+               1532 (37.5)     64 (19.7)
##   physical.activity (%)                                   0.266
##     No                 2309 (56.5)    223 (68.6)
##     High                871 (21.3)     43 (13.2)
##     Moderate            909 (22.2)     59 (18.2)
##   medical.access = Yes (%) 3312 (81.0) 310 (95.4)         0.457
##   blood.pressure = Yes (%) 1057 (25.8) 204 (62.8)         0.801
##   healthy.diet (%)                                        0.213
##     Poor                210 ( 5.1)     34 (10.5)
##     Fair                951 (23.3)     81 (24.9)
##     Good               2928 (71.6)    210 (64.6)
##   covered.health = Yes (%) 2900 (70.9) 279 (85.8)         0.369
```

Table 1 in unmatched data and corresponding SMD

	Non-Ar	RA	**SMD**
Diabetes	8.8%	26.8%	**0.485**
Smoke	43.9%	54.5%	**0.212**

```
## Sample sizes:
##          Control Treated
## All         4089     325
## Matched      317     317
## Unmatched   3772       8
## Discarded      0       0
```

16

step 3

Step 3:

Unmatched

Table 1 in matched data and corresponding SMD

	Non-Ar	RA	**SMD**
Diabetes	21.8%	25.2%	**0.082**
Smoke	53.6%	53.6%	**<0.001**

SMD

(<0.2,

Or

<0.1)

```
## Sample sizes:
##             Control  Treated
## All            4089      325
## Matched         317      317
## Unmatched      3772        8
## Discarded         0        0
```

Matched

SMD		Stratified by arthritis.type		
			Non-arthritis	Rheumatoid arthritis	SMD
	##	n	317	317	
0.238	##	gender = Female (%)	173 (54.6)	186 (58.7)	0.083
0.183	##	bmi = (25,80] (%)	241 (76.0)	238 (75.1)	0.022
0.485	##	diabetes = Yes (%)	69 (21.8)	80 (25.2)	0.082
0.212	##	smoke = Yes (%)	170 (53.6)	170 (53.6)	<0.001
0.891	##	age (%)			0.052
	##	(0,50]	77 (24.3)	74 (23.3)	
	##	(50,70]	166 (52.4)	162 (51.1)	
	##	70+	74 (23.3)	81 (25.6)	
0.347	##	race (%)			0.110
	##	White	139 (43.8)	126 (39.7)	
	##	Black	108 (34.1)	107 (33.8)	
	##	Other	70 (22.1)	84 (26.5)	
0.221	##	born = USborn (%)	252 (79.5)	254 (80.1)	0.016
0.160	##	education (%)			0.066
	##	School	52 (16.4)	51 (16.1)	
	##	College	133 (42.0)	124 (39.1)	
	##	High.School	132 (41.6)	142 (44.8)	
0.275	##	marriage = Married (%)	150 (47.3)	149 (47.0)	0.006
0.531	##	annualincome (%)			0.034
	##	<20k	125 (39.4)	127 (40.1)	
	##	20kto54k	131 (41.3)	126 (39.7)	
	##	55k+	61 (19.2)	64 (20.2)	
0.266	##	physical.activity (%)			0.083
	##	No	205 (64.7)	215 (67.8)	
	##	High	52 (16.4)	43 (13.6)	
	##	Moderate	60 (18.9)	59 (18.6)	
0.457	##	medical.access = Yes (%)	300 (94.6)	302 (95.3)	0.029
0.801	##	blood.pressure = Yes (%)	202 (63.7)	196 (61.8)	0.039
0.213	##	healthy.diet (%)			0.049
	##	Poor	28 (8.8)	31 (9.8)	
	##	Fair	74 (23.3)	78 (24.6)	
	##	Good	215 (67.8)	208 (65.6)	
0.369	##	covered.health = Yes (%)	265 (83.6)	272 (85.8)	0.061

Propensity score matching: step 4

Step 4: Estimate treatment effect

Logistic regression

(Y ~ A)   OR = 1.55

fit in matched data

Propensity score matching vs. regression 🔍

Estimates of the <u>OR/CI</u> from <u>matching</u> are not very different than what we got from <u>regression</u>. Why would we do this then?

- ==Intuitive==: compare two similar groups
- ==Diagnostics== (balance checking) much easier compared to residual plot/influence
- Exposure and outcome models are ==seperate==
- Non-parametric (ML) approaches can be used to relax linearity assumption in estimating PS.

Propensity score matching directly gives you

ATT

ATE

Thanks!

ehsan.karim@ubc.ca

www.ehsankarim.com