Intuitive understanding of Propensity Score

Leyla Nunez
2019-10-15

GOAL

Hopefully give you some intuitional understanding of Propensity score (matching)

Our hypothetical example

- Suppose UN supports a program to build *High Schools* in poor cities in Nigeria, Africa
- The aim was to reduce *Teen Pregnancies*
- `T` - treatment variable
  - `T=1` - a city that got a new high school
  - `T=0` - a city that didn't
- The cities were not selected at random
- `BR` - Birth rate, our outcome of interest
  - The birth rate is measured in babies born to women aged 15–19 years per 1000 women, approximately two years after the high schools opened.

How did the program affect teen pregnancies?

city	T	BR
1	1	10
2	1	15
3	1	22
4	1	19
city	**T**	**BR**
5	0	25
6	0	19
7	0	4
8	0	8
9	0	6

- Compare the average birth rate in the 'treatment' city to that in the control city.

$$\frac{10+15+22+19}{4} - \frac{25+19+4+8+6}{5} = 4.1$$

- Birth rate did *increase*?!

What/why?

- Cities that got the high schools almost certainly had *high ⇑ BR before the program* was implemented
- If we had some information about the *pre-program BR*
  - We could have compared changes in the BR in the treatment group versus the changes in the control group
  - BUT suppose we don't have that information
- Our two samples may differ in terms of other factors, other than the fact that they were 'treated'

> **Selection bias (SB)**
> Individuals who experience a certain treatment often vary from individuals who don't recieved the treatment

- *4.1* reflects both the *effect of treatment* and some degree of *SB*
- A problem in any *observational studies*

In a perfect world

- We need to get the two groups to look as similar as possible
  - so that we can isolate the effect of 'treatment' (exposure)
- **Randomized controlled trial (RCT)**
  - Is the most powerful tool in order to achieve that
  - Provides groups comparable with respect of known and unknown confounders (variables)

- *Cities were not selected at random!*

Additional information available to us ...

- Two variables:
  - `povrate` - Poverty rate in the cities
  - `teachers` - The number of teachers per capita before the program was implemented

city	T	BR	povrate	teachers
1	1	10	0.5	1.5
2	1	15	0.6	2.5
3	1	22	0.7	1.5
4	1	19	0.6	2.5
city	**T**	**BR**	**povrate**	**teachers**
5	0	25	0.6	1.5
6	0	19	0.5	2.5
7	0	4	0.1	4.5
8	0	8	0.3	5.5
9	0	6	0.2	4.5

- *Take a moment to compare the treatment and control groups using those 2 variables*

Observed information ...

city	T	BR	povrate	teachers
1	1	10	0.5	1.5
2	1	15	0.6	2.5
3	1	22	0.7	1.5
4	1	19	0.6	2.5
city	**T**	**BR**	**povrate**	**teachers**
5	0	25	0.6	1.5
6	0	19	0.5	2.5
7	0	4	0.1	4.5
8	0	8	0.3	5.5
9	0	6	0.2	4.5

- On average the treatment group have
  - higher poverty rate 0.6 (0.34) and
  - fewer teachers per capita 2 (3.7)

We need to get the two groups to look as similar as possible

The basic idea

1. *Create a new control group*
    - For each observation in the treatment group,
    - select the control observation that looks most like it - based on the variables
2. *Compute the treatment effect*
    - Compare the average outcome in the treatment group with the average outcome in the new control group

Lets try with just one variable

- In a real study, you will have more cases and controls

city	T	BR	povrate
1	1	10	0.5
2	1	15	0.6
3	1	22	0.7
4	1	19	0.6
city	**T**	**BR**	**povrate**
5	0	25	0.6
6	0	19	0.5
7	0	4	0.1
8	0	8	0.3
9	0	6	0.2

Lets try with just one variable

- In a real study, you will have more cases and controls

city	T	BR	povrate
1	1	10	0.5
2	1	15	0.6
3	1	22	0.7
4	1	19	0.6
city	**T**	**BR**	**povrate**
5	0	25	0.6
6	0	19	0.5
7	0	4	0.1
8	0	8	0.3
9	0	6	0.2

Lets try with just one variable

- In a real study, you will have more cases and controls

city	T	BR	povrate
1	1	10	0.5
2	1	15	0.6
3	1	22	0.7
4	1	19	0.6
city	**T**	**BR**	**povrate**
5	0	25	0.6
6	0	19	0.5
7	0	4	0.1
8	0	8	0.3
9	0	6	0.2

- city *1* -> city *6*
- city *2* -> city *5*
- city *3* -> city *5*
- city *4* -> city *5*
  - Using the same controls *many* times - Not unusual!

Lets try with just one variable

- In a real study, you will have more cases and controls

city	T	BR	povrate
1	1	10	0.5
2	1	15	0.6
3	1	22	0.7
4	1	19	0.6
city	**T**	**BR**	**povrate**
5	0	25	0.6
6	0	19	0.5
7	0	4	0.1
8	0	8	0.3
9	0	6	0.2

- city *6* could have been matched to multiple treatment observations
  - city *2*
  - city *4*
  - No exact match but may still be 👌

Lets try with just one variable

city	T	BR	povrate
1	1	10	0.5
2	1	15	0.6
3	1	22	0.7
4	1	19	0.6
city	**T**	**BR**	**povrate**
5	0	25	0.6
6	0	19	0.5
7	0	4	0.1
8	0	8	0.3
9	0	6	0.2

- Many control observations (city *7-9*) might not match to any treatment observations
  - They have much lower poverty rates

Multiple background characteristics

- *Gets a lot more complicated!*
- city 1: Should we match it with city 5?

city	T	BR	povrate	teachers
1	1	10	0.5	1.5
2	1	15	0.6	2.5
3	1	22	0.7	1.5
4	1	19	0.6	2.5
city	**T**	**BR**	**povrate**	**teachers**
5	0	25	0.6	1.5
6	0	19	0.5	2.5
7	0	4	0.1	4.5
8	0	8	0.3	5.5
9	0	6	0.2	4.5

Multiple background characteristics

- *Gets a lot more complicated!*
- city 1: Should we match it with city 5? or city 6?

city	T	BR	povrate	teachers
1	1	10	0.5	1.5
2	1	15	0.6	2.5
3	1	22	0.7	1.5
4	1	19	0.6	2.5
city	**T**	**BR**	**povrate**	**teachers**
5	0	25	0.6	1.5
6	0	19	0.5	2.5
7	0	4	0.1	4.5
8	0	8	0.3	5.5
9	0	6	0.2	4.5

- Which variable was more important for WHO 🤷?

How do you actually match treatment observations to controls?

Propensity score

> **Propensity score**
>
> Conditional probability that an individual will be given the treatment just based on their background characteristics

$$P(T = 1 \mid X_1, X_2, \ldots X_p)$$

- $P(T = 1 \mid$
  - Probability of treatment given ...
- $X_1, X_2, \ldots X_p$
  - all our background characteristics that might influence selection of a city as a 'treatment' city
  - *povrate*
  - *teachers*

Propensity score calculation

- The likelihood that a city would have been in the 'treatment' group
- A value between *0* and *1*

city	T	BR	povrate	teachers	PScores
1	1	10	0.5	1.5	?
2	1	15	0.6	2.5	?
3	1	22	0.7	1.5	?
4	1	19	0.6	2.5	?
city	**T**	**BR**	**povrate**	**teachers**	**PScores**
5	0	25	0.6	1.5	?
6	0	19	0.5	2.5	?
7	0	4	0.1	4.5	?
8	0	8	0.3	5.5	?
9	0	6	0.2	4.5	?

Propensity score calculation

- Use **logistic regression** to estimate all the needed β coefficients

$$P(T = 1 \mid povrate \ \& \ teachers) = \beta_0 + \beta_1 \times povrate + \beta_2 \times teachers$$

```
pscores.model <- glm(T ~ povrate + teachers, # the model
                     family = binomial(link='logit'), #logistic regression
                     data = data) # our data set
```

The calculated coefficients

```
## (Intercept)     povrate    teachers
## -7.45368857 14.50004813 -0.08880023
```

$$P(T = 1 \mid povrate \ \& \ teachers) = -7.45 + 14.50 \times povrate - 0.09 \times teachers$$

Use the equation to compute the *predicted probability of treatment* using the background characteristics

Propensity score

$$P(T = 1 \mid povrate \ \& \ teachers) = -7.45 + 14.50 \times povrate - 0.09 \times teachers$$

city	T	BR	povrate	teachers	PScores
1	1	10	0.5	1.5	0.4165712
2	1	15	0.6	2.5	0.7358171
3	1	22	0.7	1.5	0.9284516
4	1	19	0.6	2.5	0.7358171
city	**T**	**BR**	**povrate**	**teachers**	**PScores**
5	0	25	0.6	1.5	0.7527140
6	0	19	0.5	2.5	0.3951619
7	0	4	0.1	4.5	0.0016534
8	0	8	0.3	5.5	0.0268029
9	0	6	0.2	4.5	0.0070107

Lets match using the propensity score

city	T	BR	PScores
1	1	10	0.4165712
2	1	15	0.7358171
3	1	22	0.9284516
4	1	19	0.7358171
city	**T**	**BR**	**PScores**
5	0	25	0.7527140
6	0	19	0.3951619
7	0	4	0.0016534
8	0	8	0.0268029
9	0	6	0.0070107

Lets match using the propensity score

city	T	BR	PScores
1	1	10	0.4165712
2	1	15	0.7358171
3	1	22	0.9284516
4	1	19	0.7358171
city	**T**	**BR**	**PScores**
5	0	25	0.7527140
6	0	19	0.3951619
7	0	4	0.0016534
8	0	8	0.0268029
9	0	6	0.0070107

Balance in the covariates

city	T	BR	povrate	teachers	PScores
1	1	10	0.5	1.5	0.4165712
2	1	15	0.6	2.5	0.7358171
3	1	22	0.7	1.5	0.9284516
4	1	19	0.6	2.5	0.7358171
city	**T**	**BR**	**povrate**	**teachers**	**PScores**
5	0	25	0.6	1.5	0.7527140
6	0	19	0.5	2.5	0.3951619

- Similar poverty rate 0.6 (0.55) and
- Same number teachers per capita 2 (2)

What was the effect of treatment on BR?

$$\frac{10+15+22+19}{4} - \frac{25+19+19+19}{4} = -4$$

- The effect of constructing new high schools has led to an average birth rate reduction of 4 babies per 1000 women

That is how PS works!

Another example

lalonde

- The National Supported Work Demonstration (NSW) was a temporary employment program designed to help disadvantaged workers move into the labor market

```
## 'data.frame':    614 obs. of  10 variables:
##  $ treat   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ age     : int  37 22 30 27 33 22 23 32 22 33 ...
##  $ educ    : int  11 9 12 11 8 9 12 11 16 12 ...
##  $ black   : int  1 0 1 1 1 1 1 1 1 0 ...
##  $ hispan  : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ married : int  1 0 0 0 0 0 0 0 0 1 ...
##  $ nodegree: int  1 1 0 1 1 1 0 1 0 0 ...
##  $ re74    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ re75    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ re78    : num  9930 3596 24909 7506 290 ...
```

- 185 cases and 429 controls
- *We want to evaluate the effect of the NSW program on income*

Look at the data *BEFORE* matching

- **Standardized differences (smd)** to check covariate balance
- For continous variables:
  - e.g. *age*, *educ* - the number of years of schooling
  - the difference in means between groups, divided by the (pooled) standard deviation

$$smd = \frac{\bar{X}_{treatment} - \bar{X}_{control}}{\sqrt{\frac{s^2_{treatment} - s^2_{control}}{2}}}$$

> Rules of thumb:
>
> - values < *0.1* indicate adequate balance
> - value *0.1* - *0.2* are not too alarming
> - > *0.2* indicate serious imbalance

SMD before matching

- *re78* - is excluded since it is out outcome variable of interest
- *429* controls versus *185* treated patients

```
##                       Stratified by treat
##                        0                 1                 SMD
##   n                        429               185
##   age (mean (SD))        28.03 (10.79)     25.82 (7.16)     0.242
##   educ (mean (SD))       10.24 (2.86)      10.35 (2.01)     0.045
##   black (mean (SD))       0.20 (0.40)       0.84 (0.36)     1.668
##   hispan (mean (SD))      0.14 (0.35)       0.06 (0.24)     0.277
##   married (mean (SD))     0.51 (0.50)       0.19 (0.39)     0.719
##   nodegree (mean (SD))    0.60 (0.49)       0.71 (0.46)     0.235
##   re74 (mean (SD))     5619.24 (6788.75) 2095.57 (4886.62)  0.596
##   re75 (mean (SD))     2466.48 (3292.00) 1532.06 (3219.25)  0.287
```

- 👎🚫 *Imbalance!!!* 🚫👎
  - A large difference in covariate distribution
  - We need balance! $\Rightarrow$ let's do *PS matching*

Matching methods

- *Exact match*?
  - We may not find cases/controls with the same PS
- *Greedy (nearest-neighbor) matching*
  - As soon as a match is made between two objects that match is kept even if a 'better' match could be found in the data
  - Tries to include as many cases as possible
  - Trade-offs: No perfect balance in covariates!
- *Caliper matching*
  - Matches two subjects within a given range (**caliper**)
- *Optimal matching*
  - The algorithm is able to reconsider a match
  - Is going to minimize the total distance
  - Involve a lot of computation

Matching methods

- *One-to-one matching*
- *One-to-many matching*
    - One case is matched with several controls
    - Useful if you have a large control group
- *without* replacement
    - ones a match has been made those subjects can not be used ones again
- *with* replacement
    - the same control subject can be matched to several cases

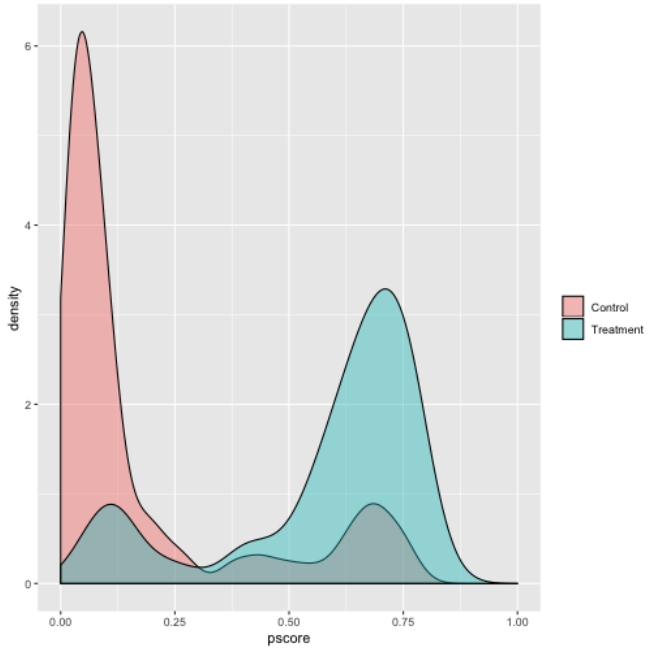
Step 1: Calculate PS

- Use a logistic regression model, where the outcome is *treatment*
- Include the 8 confounding variables in the model as predictors

```
# fit a PS model - logistic regression
psmodel <- glm(treat ~ age + educ + black + hispan + married
                           + nodegree + re74 + re75,
  family = binomial(),
  data = lalonde)
lalonde$pscore <- psmodel$fitted.values
head(lalonde[, c(1:9,11)])
```

```
##      treat age educ black hispan married nodegree re74 re75    pscore
## NSW1     1  37   11     1      0       1        1    0    0 0.6387699
## NSW2     1  22    9     0      1       0        1    0    0 0.2246342
## NSW3     1  30   12     1      0       0        0    0    0 0.6782439
## NSW4     1  27   11     1      0       0        1    0    0 0.7763241
## NSW5     1  33    8     1      0       0        1    0    0 0.7016387
## NSW6     1  22    9     1      0       0        1    0    0 0.6990699
```

Step 2: Examine PS values *BEFORE* matching



- Cases may be excluded!
  - The price we have to pay for matching
- Cases with extrem PS values
  - *very high PS*, close to *1*
  - *very low PS*, close to *0*

Step 3: Do the matching

- 1:1 matching
- without replacement
- no caliper

```
##                          Stratified by treat
##                           0                  1                  SMD
##   n                           185                185
##   age (mean (SD))           25.51 (10.69)      25.82 (7.16)      0.034
##   educ (mean (SD))          10.56 (2.65)       10.35 (2.01)      0.090
##   black (mean (SD))          0.47 (0.50)        0.84 (0.36)      0.852
##   hispan (mean (SD))         0.22 (0.42)        0.06 (0.24)      0.479
##   married (mean (SD))        0.22 (0.41)        0.19 (0.39)      0.067
##   nodegree (mean (SD))       0.65 (0.48)        0.71 (0.46)      0.116
##   re74 (mean (SD))        2554.11 (4508.53)  2095.57 (4886.62)  0.098
##   re75 (mean (SD))        1763.41 (2848.97)  1532.06 (3219.25)  0.076
```

- *185* cases and *185* controls
- 👎💩 *Still looks bad!!* 💩👎

Re-do the matching!

- 1:1 matching
- without replacement
- *caliper = 0.1*

```
##                         Stratified by treat
##                          0                   1                   SMD
##    n                         111                 111
##    age (mean (SD))         26.27 (11.10)       26.22 (7.18)       0.006
##    educ (mean (SD))        10.37 (2.66)        10.25 (2.31)       0.047
##    black (mean (SD))        0.72 (0.45)         0.74 (0.44)       0.040
##    hispan (mean (SD))       0.11 (0.31)         0.10 (0.30)       0.029
##    married (mean (SD))      0.24 (0.43)         0.24 (0.43)      <0.001
##    nodegree (mean (SD))     0.66 (0.48)         0.65 (0.48)       0.019
##    re74 (mean (SD))      2704.56 (4759.89)   2250.49 (5746.14)    0.086
##    re75 (mean (SD))      1969.10 (3169.08)   1222.25 (3081.19)    0.239
```

- *111* cases and *111* controls
- 👍 *Not perfect but much better!* 👍
- *Overt (measured) bias* would occur if you had imbalance and you carried out your outcome analysis anyway

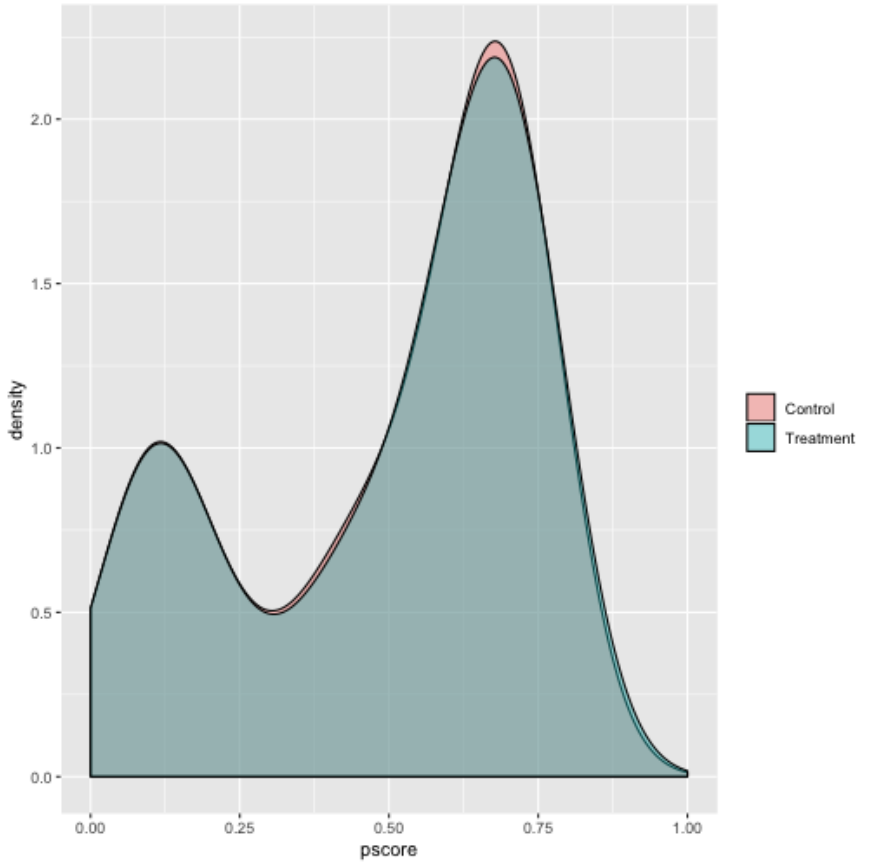
2nd matching attempt

- 1:1 matching
- without replacement
- *caliper = 0.1*

```
##                         Stratified by treat
##                          0                 1                 SMD
##   n                          111               111
##   age (mean (SD))          26.77 (11.64)     26.22 (7.18)     0.057
##   educ (mean (SD))         10.32 (2.68)      10.25 (2.31)     0.025
##   black (mean (SD))         0.72 (0.45)       0.74 (0.44)     0.040
##   hispan (mean (SD))        0.10 (0.30)       0.10 (0.30)    <0.001
##   married (mean (SD))       0.25 (0.44)       0.24 (0.43)     0.021
##   nodegree (mean (SD))      0.67 (0.47)       0.65 (0.48)     0.038
##   re74 (mean (SD))       2606.37 (4589.94) 2250.49 (5746.14)  0.068
##   re75 (mean (SD))       2030.48 (3184.86) 1222.25 (3081.19)  0.258
```

```
warning("Do NOT play with the data!")
```

Step 4: Compare PS after matching



Step 5: Carry out the outcome analysis

- If out matching algorithm has done a good job!
  - We have *balance* in our covariates
  - Compare *re78* - real earnings in 1978 - in the treatment group versus the control group

```
lm_model1 <- lm(re78 ~ treat,  data = matched2)
tidy(lm_model1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    4904.      597.      8.22 1.77e-14
## 2 treat          1247.      844.      1.48 1.41e- 1
```

- *p-values = 0.141*, we have no evidence that **NSW program had any effect on income**

Traditional covariate adjustment

- *Add all covaraies into the model*
  - Using the original data

```
lm_model2 <- lm(re78 ~ treat + age + educ + black + hispan +
                       married + nodegree + re74 + re75,  data = lalonde)
tidy(lm_model2)
```

```
## # A tibble: 10 x 5
##    term         estimate std.error statistic     p.value
##    <chr>           <dbl>     <dbl>     <dbl>       <dbl>
##  1 (Intercept)    66.5   2437.         0.0273 0.978
##  2 treat        1548.     781.         1.98   0.0480
##  3 age            13.0     32.5        0.399  0.690
##  4 educ          404.     159.         2.54   0.0113
##  5 black       -1241.     769.        -1.61   0.107
##  6 hispan        499.     942.         0.530  0.597
##  7 married       407.     695.         0.585  0.559
##  8 nodegree      260.     847.         0.307  0.759
##  9 re74            0.296    0.0583     5.09   0.000000489
## 10 re75            0.232    0.105      2.21   0.0273
```

- **POS** - Shows you the effect of *all* covariates not just the *treatment*
- **NEG** - May not be suitable with many covariates in small studies

PS covariate adjustment

- Use of PS as the only covariate in the model
- Alternative methods are available - adding all covariates + PS

```
lm_model3 <- lm(re78 ~ treat + pscore, data = matched2)

tidy(lm_model3)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic      p.value
##   <chr>          <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)    5950.     1040.      5.72 0.0000000344
## 2 treat          1252.      843.      1.49 0.139
## 3 pscore        -2120.     1728.     -1.23 0.221
```

- We have no evidence of treatment effect!

Doubly robust methods

- *Controll for unbalanced covariates*
  - *re75* - real earnings in 1975 in dollars

```
lm_model4 <- lm(re78 ~ treat + re75, data = matched2)

tidy(lm_model4)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic       p.value
##   <chr>          <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept) 4027.      639.         6.30 0.00000000159
## 2 treat       1579.      831.         1.90 0.0586
## 3 re75           0.445     0.133      3.36 0.000919
```

- We have no evidence of treatment effect!

What we estimated is NOT the average treatment effect but the average treatment effect on the treated (ATT).

Our conclusions can only be generalised to cases that look like the treatment group

Limitations: Hidden bias

We can only expect to achieve balance on *observed* variables

> **Hidden bias** would occur if there is *imbalance* on *unobserved* variables and these unobserved variables are actually confounders

- We still have to worry about the possibility that our observations differ on other *unobserved* ways that we haven't accounted for by the PS

Advantages and disadvantages of PS matching

POS

- Provides excellent covariate balance in most circumstances
- Simple to analyse, present and interpret

NEG

- Some patients are unmatched $\Rightarrow$ important information being excluded from the analysis
- Matching tended to give less precise estimates in some cases

Sensitivity analysis

- We typically believe that there is some degree of *unmeasured confounding* 🙈

Are the conclusions we're making *sensitive* to just *minor violations* of our key assumption or is it *very sensitive* to *violations*

For example:

- *Change from statistically significant to not*
    - How much hidden bias would they have to be before we would *not* have a significant result?
- *Change in direction of effect*
    - How much bias would there have to be before the sign actually changed?

Propensity score methods are NOT necessarily superior to conventional covariate adjustment, and care should be taken to select the most suitable method.

Slides created via the R package xaringan