

SOO AND MATCHING

Annie Chen

Jan. 29, 2020

REVIEW

$$ATT = \mathbb{E}[Y_i(1) - Y_i(0) | D_i = 1]$$

- Is this identified when D is not randomized?
- Fundamental Problem of Causal Inference

Under what assumptions can the ATE and ATT be non-parametrically identified?

- $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i \mid X_i = x$ for all $x \in \mathcal{X}$
- $0 < \Pr(D_i = 1 \mid X_i = x) < 1$ for all $x \in \mathcal{X}$

IDENTIFICATION OF ATT UNDER CI AND COMMON SUPPORT

$$\tau_{ATT} = \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]$$

- $= \int \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x, D_i = 1]f(x|D_i = 1)dx$ by law of iterated expectation¹
- $= \int \{\mathbb{E}[Y_i(1)|X_i = x, D_i = 1] - \mathbb{E}[Y_i(0)|X_i = x, D_i = 1]\}f(x|D_i = 1)dx$ linearity of expectation
- $= \int \mathbb{E}[Y_i|X_i = x, D_i = 1] - \mathbb{E}[Y_i|X_i = x, D_i = 0]f(x|D_i = 1)dx^2$

Similarly, $\tau_{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$

- $= \int \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]f(x)dx$
- $= \int (\mathbb{E}[Y_i|X_i = x, D_i = 1] - \mathbb{E}[Y_i|X_i = x, D_i = 0])f(x)dx$

¹Recall: $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$ and $\mathbb{E}[X] = \int xf(x)dx$ (continuous)

²Turns out, ATT is identifiable under a weaker set of assumptions. Compare overlap and ignorability assumptions required for identification in ATE v. ATT.

MATCHING PACKAGE

The workhorse of the package:

- `Match(Y, Tr, X, estimand = "ATT", M = 1, exact, Weight...)`
 - By default, the function estimates the ATT, `exact = NULL`, `replace = TRUE`, and uses one-to-one matching (`M = 1`).
 - Other arguments: (`ties`, `caliper`, `BiasAdjust`, `CommonSupport`)
 - `Weight = 2` (Mahalanobis distance), 3 (custom supplied by `Weight.matric`)
 - use `summary()` on matched object

MATCHING PACKAGE

Check the balance before and after matching using these functions to create tables:

- `MatchBalance(formula, data, match.out...)`
- `baltest.collect(matchbal.out, var.names)`

ALTERNATIVELY, THE MATCHIT PACKAGE

- `matchit(formula, data, method = "nearest", discard = "none", distance = "logit")`
- `method = ["genetic", "exact", "subclass", "nearest", ...]`
- I.e. nearest selects the r (default = 1) best control matches for each individual in the treatment group (excluding discarded).

JOB TRAINING EXAMPLE (LALONDE 1986)

```
data(lalonde)
```

This is a subsample of the original data consisting of the National Supported Work Demonstration (NSW) treated group and the comparison sample from the Population Survey of Income Dynamics (PSID).
(non-experimental study)

- `treat`: participation in the job training program
- `re78`: 1978 real earnings

JOB TRAINING EXAMPLE (LALONDE 1986)

A naive comparison of earnings for those who participated and those who did not.

```
lalonge %>% group_by(treat) %>%  
  summarise(Income1978 = mean(re78), n = n())
```

```
## # A tibble: 2 x 3  
##   treat Income1978      n  
##   <int>      <dbl> <int>  
## 1     0      6984.   429  
## 2     1      6349.   185
```


Specify pre-treatment covariates to match on.

```
bal_form <- formula(treat ~ age + educ + black + hispan +  
                    nodegree + married + re74 + re75 + re78)  
# or  
# as.formula(paste("treat~",paste(names(lalonde)[-1],collapse="+"))).
```

Let's check the pre-matching balance.

```
mb_unmatched <- MatchBalance(bal_form,  
                             data = lalonde, print.level=0)  
tab_unmatched <- baltest.collect(mb_unmatched,  
                                 var.names = colnames(lalonde)[-1],  
                                 after = FALSE)
```

TABLE 1: Covariate Balance in Unmatched Data

	mean.Tr	mean.Co	sdiff	sdiff.pooled	var.ratio	T pval	KS pval
age	25.82	28.03	-30.94	-24.19	0.44	0.00	0.00
educ	10.35	10.24	5.50	4.48	0.50	0.58	0.02
black	0.84	0.20	175.68	166.77	0.82	0.00	
hispan	0.06	0.14	-34.89	-27.69	0.46	0.00	
married	0.71	0.60	24.43	23.50	0.86	0.01	
nodegree	0.19	0.51	-82.41	-71.95	0.62	0.00	
re74	2095.57	5619.24	-72.11	-59.58	0.52	0.00	0.00
re75	1532.06	2466.48	-29.03	-28.70	0.96	0.00	0.00
re78	6349.14	6984.17	-8.07	-8.37	1.16	0.35	0.14

Now, check the post-matching balance.

```
vars <- c("age", "educ", "black", "hispan", "nodegree", "married", "re74", "re75", "re78")

match_out <- Match(Y= lalonde$re78, Tr = lalonde$treat,
                  X = lalonde[, vars], exact = FALSE, Weight = 2)

mb_matched <- MatchBalance(bal_form, data= lalonde,
                          match.out = match_out, print.level=0)

tab_matched <- baltest.collect(mb_matched, var.names = colnames(lalonde)[-1],
                              after = TRUE)
```

TABLE 2: Covariate Balance in Matched Data

	mean.Tr	mean.Co	sdiff	sdiff.pooled	var.ratio	T pval	KS pval
age	25.82	24.35	20.47	20.47	0.69	0.00	0.00
educ	10.35	10.43	-4.30	-4.30	1.08	0.09	0.94
black	0.84	0.82	5.93	5.93	0.90	0.04	
hispan	0.06	0.06	0.00	0.00	1.00	1.00	
married	0.71	0.71	0.00	0.00	1.00	1.00	
nodegree	0.19	0.18	1.38	1.38	1.02	0.32	
re74	2095.57	1856.22	4.90	4.90	1.47	0.21	0.00
re75	1532.06	1112.68	13.03	13.03	2.05	0.00	0.41
re78	6349.14	5158.26	15.14	15.14	1.82	0.00	0.35

```
summary(match_out)
```

```
##  
## Estimate...    1190.9  
## AI SE.....    543.05  
## T-stat.....    2.1929  
## p.val.....    0.028312  
##  
## Original number of observations.....    614  
## Original number of treated obs.....    185  
## Matched number of observations.....    185  
## Matched number of observations (unweighted).    185  
##  
## Number of obs dropped by 'exact' or 'caliper'    0
```

PROPSENSITY SCORE

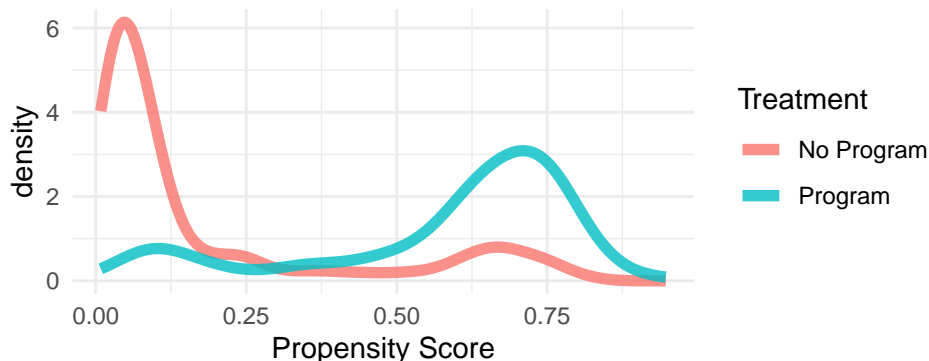
$$\pi(X_i) \equiv \Pr(D_i = 1|X_i)$$

What is the functional form of $\hat{\pi}(\cdot)$?

- I.e. estimate propensity scores using logistic regression.
- Exercise: Try matching on a PS! Create a balance table before and after matching.
- Workflow: $\hat{\pi}(X_i) \rightarrow \text{Match} \rightarrow \text{MatchBalance} \rightarrow \text{baltest.collect}$

PROPENSITY SCORE

```
ps_model <- glm(bal_form, data = lalonde,  
               family = binomial(link = logit))  
#ps_model$fitted.values  
#lalonde$pscore <- predict(ps_model, type = "response")
```



Matching on the PS.

```
match_pscore <- Match(Y = lalonde$re78,  
                      Tr = lalonde$treat,  
                      X = ps_model$fitted.values)  
  
#summary(match_pscore)
```

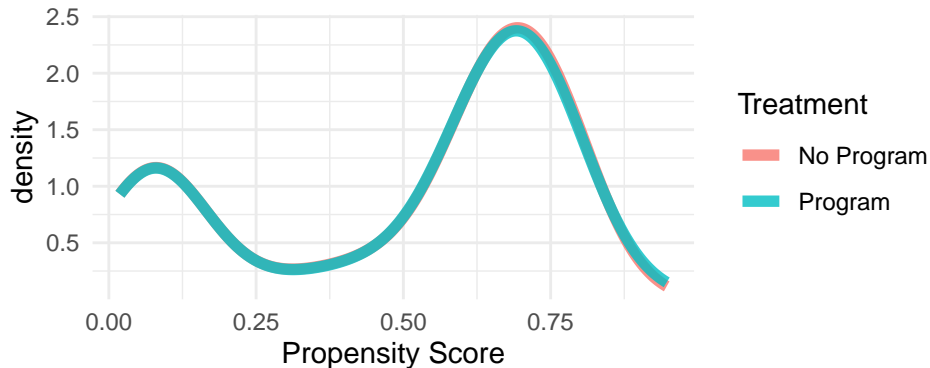






TABLE 3: Covariate Balance in Propensity Score Matched Data

	mean.Tr	mean.Co	sdiff	sdiff.pooled	var.ratio	T pval	KS pval
age	25.82	24.51	18.28	18.28	0.44	0.18	0.00
educ	10.35	10.48	-6.79	-6.79	0.63	0.55	0.09
black	0.84	0.85	-2.97	-2.97	1.06	0.32	
hispan	0.06	0.03	10.42	10.42	1.67	0.16	
married	0.71	0.65	12.74	12.74	0.91	0.23	
nodegree	0.19	0.22	-6.88	-6.88	0.91	0.42	
re74	2095.57	2100.16	-0.09	-0.09	1.58	0.99	0.00
re75	1532.06	1920.09	-12.05	-12.05	1.00	0.24	0.00
re78	6349.14	7035.08	-8.72	-8.72	1.19	0.33	0.00

- How does this compare to our previous approach to matching?

WEIGHTING

- Idea: weight each observation in the control group such that it looks like the treatment group (i.e., good covariate balance)³
- Suppose, there are two types of cats, each with a different probability of receiving treatment D . In this example, we can assign the lone cats in the off-diagonal cells a weight of 3.

$P(D = 1) = 0.75$ $P(D = 1) = 0.25$	
$D = 1$	 
$D = 0$	 

³Matching is a special case of weighting!

INVERSE PROBABILITY WEIGHTING (IPW)

Weighting on the Propensity Score $\pi(X_i)$

$$\tau_{ATE} = \mathbb{E} \left[Y_i \cdot \frac{D_i - \hat{\pi}(X_i)}{\hat{\pi}(X_i) \cdot [1 - \hat{\pi}(X_i)]} \right]$$

The sample analog is $(\hat{\tau}_{ATE})$:

$$\frac{1}{N} \sum_{i=1}^N \left\{ Y_i \cdot \frac{D_i - \hat{\pi}(X_i)}{\hat{\pi}(X_i) \cdot [1 - \hat{\pi}(X_i)]} \right\} = \frac{1}{N} \sum_{i=1}^N \left\{ \frac{D_i \cdot Y_i}{\hat{\pi}(X_i)} \right\} - \frac{1}{N} \sum_{i=1}^N \left\{ \frac{(1 - D_i) \cdot Y_i}{[1 - \hat{\pi}(X_i)]} \right\}$$

INVERSE PROBABILITY WEIGHTING (IPW)

$$\tau_{ATT} = \mathbb{E} \left[Y_i \cdot \frac{D_i - \hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} \right] \cdot \mathbb{P}(D_i = 1)^{-1}$$

With sample analog:

$$\hat{\tau}_{ATT} = \frac{1}{N_1} \sum_{i=1}^N \left\{ Y_i \cdot \frac{D_i - \hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} \right\}$$

ADDITIONAL NOTES