Causal Inference with R

Youjin Lee
Department of Biostatistics, Brown University

September 21, 2021
R-Ladies Philly

About me

- Assistant Professor in the Department of Biostatistics, Brown University (07/2021~)
- A postdoc fellow at the University of Pennsylvania
- Research interests: causal inference, social networks, biostatistics
  - e.g., the effect of Vaccine A compared to Vaccine B
- Github: https://github.com/youjin1207

Causal inference is hot topic now ..

- Public health policies
- Economics
- Education
- Medicine

Causal inference is important in different fields

- Does putting wearing a mask really work?
- What is the effect of marijuana legalization on opioid overdoses or crime rates?
- Is a Facebook advertisement effective?

Causal inference is important in different fields

- Does putting wearing a mask really work?
- What is the effect of marijuana legalization on opioid overdoses or crime rates?
- Is a Facebook advertisement effective?

Causal inference is about the effect of some **“treatment”** or **“intervention”** on an **outcome**.

- wearing a mask vs. not wearing a mask
- marijuana legalized vs. marijuana not legalized
- Facebook advertisement vs. no advertisement

Causal inference is important in different fields

- Does putting wearing a mask really work?
- What is the effect of marijuana legalization on opioid overdoses or crime rates?
- Is a Facebook advertisement effective?

Causal inference is about the effect of some “treatment” or “intervention” on an outcome.

- wearing a mask vs. not wearing a mask
- marijuana legalized vs. marijuana not legalized
- Facebook advertisement vs. no advertisement

Can we really compare the outcome under two different arms while controlling everything else?

Workshop objectives

1. To understand ‘causal’ problems and identify the outcome, intervention, and units.
2. To understand the fundamental problem of causal inference and address it using propensity score techniques available in R.
3. To advance the use of R in more complex settings, such as regression discontiunity and instrumental variables.

Preliminaries

- `R` packages: `Matching`, `tableone`, `WeighIt`, `MatchIt`, `survey`, `cobalt`
- Code and data are available at https://github.com/youjin1207/CausalTutorial.

Back to RStudio

data(lalonde) in library(Matching)

```
library(Matching)
data(lalonde)
head(lalonde)
```

```
##   age educ black hisp married nodegr re74 re75
## 1  37   11     1    0       1      1    0    0
## 2  22    9     0    1       0      1    0    0
## 3  30   12     1    0       0      0    0    0
## 4  27   11     1    0       0      1    0    0
## 5  33    8     1    0       0      1    0    0
## 6  22    9     1    0       0      1    0    0
##       re78 u74 u75 treat
## 1  9930.05   1   1     1
## 2  3595.89   1   1     1
## 3 24909.50   1   1     1
## 4  7506.15   1   1     1
## 5   289.79   1   1     1
## 6  4056.49   1   1     1
```

A causal question in data(lalonde)

- The treatment (treat) is an indicator of a labor training program; re78 is the outcome, real earnings in 1978[1].
- Units: 445 subjects (check with nrow(lalonde))
- Treatment group ($A = 1$): 185 subjects with treat==1
- Control group ($A = 0$): 260 subjects with treat==0
- Outcome (Y): real earnings in 1978 (re78)
- Covariates $\mathbf{X}$: age, educ, black, hisp, married, nodegr, re74, re75, u74, u75

[1] Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association, 94*(448), 1053-1062.

Potential outcomes

- $Y(A = 0) = Y(0)$: potential outcome under control, i.e., the outcome that would be observed if a unit gets the control
- $Y(A = 1) = Y(1)$: potential outcome under treatment, i.e.,the outcome that would be observed if a unit gets the treatment
- Causal effects are comparisons of “potential outcomes”, e.g., $Y(1) - Y(0)$

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology, 66*(5), 688.

The fundamental problem in causal inference

Unit i	treat_i	$Y_i = \text{re78}_i$	$Y_i(1)$	$Y_i(0)$	age_i	educ_i	$\cdots$
1	1	9930.05	9930.05	?	44	9	$\cdots$
2	1	3595.89	3595.89	?	22	9	$\cdots$
3	1	24909.50	24909.50	?	30	12	$\cdots$
444	0	7343.96	?	7343.96	25	9	$\cdots$
445	0	5448.80	?	5448.80	22	10	$\cdots$
Average	0.42	5300.77	?	?	25.37	10.20	$\cdots$

- Individual causal effect: $Y_i(1) - Y_i(0)$
- Average treatment effect, e.g., $E[Y_i(1) - Y_i(0)] = ?$
- The fundamental problem of causal inference:
  - We only observe either $Y_i(1)$ or $Y_i(0)$ for unit i
- We need identification assumptions to infer causal effects based on observational data.
  - e.g., if unit i and i''s age are the same, let $E(Y_i(1)) = E(Y_{i'}(1))$.

Two average causal effects

- Instead of individual causal effects, we aim for “average” causal effects.

Two average causal effects

- Instead of individual causal effects, we aim for “average" causal effects.

(1) Average treatment effect (**ATE**): $E[Y_i(1) - Y_i(0)]$

- What is the effect of the labor training program on everyone in population?
- Meaningful effect estimate when there is the potential to disseminate treatment to entire population.

Two average causal effects

- Instead of individual causal effects, we aim to “average" causal effects.

(1) Average treatment effect (**ATE**): $E[Y_i(1) - Y_i(0)]$

- What is the effect of the labor training program on everyone in population?
- Meaningful effect estimate when there is the potential to disseminate treatment to entire population.

(2) Average treatment effect on the treated (**ATT**): $E[Y_i(1) - Y_i(0) | A_i = 1]$

- What is the effect of the labor training program for those who were in the program?
- Meaningful effect when only a subset of entire population typically receives the treatment.

Can we compare two averaged outcomes?

```r
mean(lalonde$re78[lalonde$treat == 1])
```

```
## [1] 6349.145
```

```r
mean(lalonde$re78[lalonde$treat == 0])
```

```
## [1] 4554.802
```

```r
t.test(re78 ~ treat, data = lalonde)
```

```
##
##  Welch Two Sample t-test
##
## data:  re78 by treat
## t = -2.6741, df = 307.13, p-value = 0.007893
## alternative hypothesis: true difference in means between group 0 and
## 95 percent confidence interval:
##  -3114.6754  -474.0108
## sample estimates:
## mean in group 0 mean in group 1
##        4554.802        6349.145
```

Average treatment effect $E(Y_i(1)) - E(Y_i(0))$

```r
mean(lalonde$re78[lalonde$treat == 1])
mean(lalonde$re78[lalonde$treat == 0])
t.test(re78 ~ treat, data = lalonde)
```

- $E(Y_i(1)|A_i = 1) \neq E(Y_i(1))$ and $E(Y_i(0)|A_i = 0) \neq E(Y_i(0))$
- Unless the treatment A_i is randomized (i.e., $Y(0), Y(1) \perp\!\!\!\perp A$), the assignment may depend on the potential outcomes, e.g., those with higher income under treatment are more likely to receive the treatment.

Average treatment effect $E(Y_i(1)) - E(Y_i(0))$

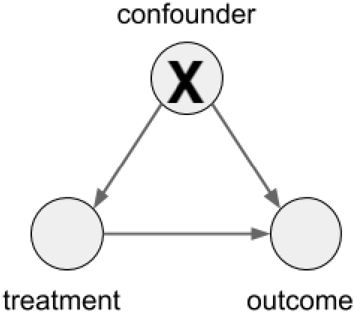
```
mean(lalonde$re78[lalonde$treat == 1])
mean(lalonde$re78[lalonde$treat == 0])
t.test(re78 ~ treat, data = lalonde)
```

- $E(Y_i(1)|A_i = 1) \neq E(Y_i(1))$ and $E(Y_i(0)|A_i = 0) \neq E(Y_i(0))$
- Unless the treatment A_i is randomized (i.e., $Y(0), Y(1) \perp\!\!\!\perp A$), the assignment may depend on the potential outcomes, e.g., those with higher income under treatment are more likely to receive the treatment.
- Instead, we will consider a more relaxed assumption: $Y(0), Y(1) \perp\!\!\!\perp A|\mathbf{X}$
  - Given observed covariates **X** (e.g., `age`, `educ`), the potential outcomes and the treatment assignment are conditionally independent.
  - In practice, for each of the treated and the control groups, we find groups of individuals whose distributions of **X** are similar each other.

Unit i	treat$_i$	Y_i = re78$_i$	$Y_i(1)$	$Y_i(0)$	age$_i$	educ$_i$	$\cdots$
1	1	9930.05	9930.05	$7343.96 + \epsilon_1$	25	9	$\cdots$
2	0	7343.96	$9930.05 + \epsilon_2$	7343.96	25	9	$\cdots$

Confounding

- Our main problem is that the treated and the control may be different on lots of factors (e.g., age). If these factors are associated both with the treatment assignment and the outcome, we call these “confounders”.



- If we can observe all of these confounders, then how can we solve confounding issue?

Recap..

- Causal inference is about the effect of the treatment/intervention on the outcome.
- Causal effects are comparisons of potential outcomes (i.e., the outcome that would be observed under each treamtent assignment).
- Due to the fundamental problem of causal inference, it is very challenging to estimate the individual causal effect; instead, we aim for the average causal effects (e.g., ATE, ATT).
- Still, due to confounding, we need several assumptions to identify the average causal effect, including $Y(0), Y(1) \perp\!\!\!\perp A|\mathbf{X}$.

Basic steps in performing causal analysis (Stuart, 2010)

(Roughly speaking, these are the basic steps to reduce the impact of measured confounders **X** on the average causal effect estimation)

1. Decide on covariates for which balance must be achieved;
2. Estimate the distance measure (e.g., propensity score);
3. Condition on the distance measure (e,g., using matching, weighting, or subclassification);
4. Assess balance on the covariates of interest; if poor, repeat steps 2-4;
5. Estimate the treatment effect in the conditional sample.

https://ngreifer.github.io/cobalt/articles/cobalt.html

The first step in performing causal analysis (Stuart, 2010)

(Roughly speaking, these are the basic steps to reduce the impact of measured confounders **X** on the average causal effect estimation)

1. **Decide on covariates for which balance must be achieved;**

- Distributions of **X** are significantly different between two treatment groups?
- Often, (standardized) mean and standard deviations are used to compare the distributions.

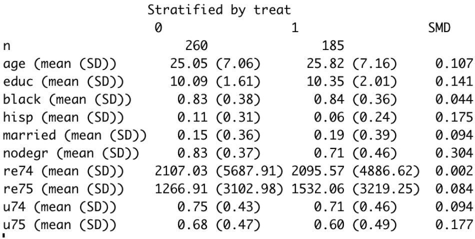
The first step in performing causal analysis

1. **Decide on covariates for which balance must be achieved;**

```
colnames(lalonde)
```

```
##  [1] "age"     "educ"    "black"   "hisp"
##  [5] "married" "nodegr"  "re74"    "re75"
##  [9] "re78"    "u74"     "u75"     "treat"
```

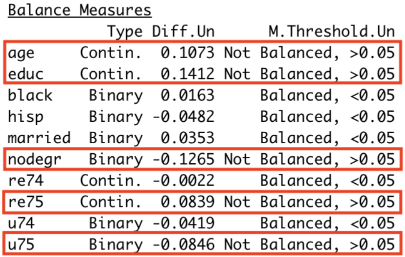
```
library(tableone)
xvars = colnames(lalonde)[!(colnames(lalonde) %in% c("treat", "re78"))]
table1 <- CreateTableOne(vars = xvars, strata = "treat",
                         data = lalonde, test = FALSE)
print(table1, smd = TRUE)
```



```
                      Stratified by treat
                       0                 1                 SMD
  n                        260               185
  age (mean (SD))        25.05 (7.06)      25.82 (7.16)     0.107
  educ (mean (SD))       10.09 (1.61)      10.35 (2.01)     0.141
  black (mean (SD))       0.83 (0.38)       0.84 (0.36)     0.044
  hisp (mean (SD))        0.11 (0.31)       0.06 (0.24)     0.175
  married (mean (SD))     0.15 (0.36)       0.19 (0.39)     0.094
  nodegr (mean (SD))      0.83 (0.37)       0.71 (0.46)     0.304
  re74 (mean (SD))     2107.03 (5687.91) 2095.57 (4886.62)  0.002
  re75 (mean (SD))     1266.91 (3102.98) 1532.06 (3219.25)  0.084
  u74 (mean (SD))         0.75 (0.43)       0.71 (0.46)     0.094
  u75 (mean (SD))         0.68 (0.47)       0.60 (0.49)     0.177
```

The first step in performing causal analysis

```r
library(cobalt) # I will introduce this package later
bal.tab(treat ~ age + educ + black + hisp +
            married + nodegr + re74 + re75 + u74 + u75,
        data = lalonde, estimand = "ATE", m.threshold = 0.05)
```



```
Balance Measures
           Type Diff.Un      M.Threshold.Un
age      Contin.  0.1073 Not Balanced, >0.05
educ     Contin.  0.1412 Not Balanced, >0.05
black     Binary  0.0163     Balanced, <0.05
hisp      Binary -0.0482     Balanced, <0.05
married   Binary  0.0353     Balanced, <0.05
nodegr    Binary -0.1265 Not Balanced, >0.05
re74     Contin. -0.0022     Balanced, <0.05
re75     Contin.  0.0839 Not Balanced, >0.05
u74       Binary -0.0419     Balanced, <0.05
u75       Binary -0.0846 Not Balanced, >0.05
```

- It looks like we must adjust for `age`, `educ`, `nodegr`, `re75`, `u75`.

Step 2-3 in performing causal analysis

1. Decide on covariates for which balance must be achieved: `age`, `educ`, `hisp`, `married`, `nodegr`, `re75`, `u74`, `u75` are selected.
2. **Estimate the distance measure (e.g., propensity score);**
3. **Condition on the distance measure**

Step 2-3 in performing causal analysis

1. Decide on covariates for which balance must be achieved: `age`, `educ`, `hisp`, `married`, `nodegr`, `re75`, `u74`, `u75` are selected.
2. **Estimate the distance measure (e.g., propensity score);**
3. **Condition on the distance measure**

Propensity score: probability of receiving the treatment A_i, given the covariates $\mathbf{X}_i$

$$e_i = Pr(A_i = 1 | \mathbf{X}_i)$$

- summary of all the covariates; a scalar value between 0 and 1.
- Within small range of propensity score values, treated and control individuals should look only randomly different on the observed covariates.
- Many packages implement step 2-3 together.

Propensity scores as summary of **X**

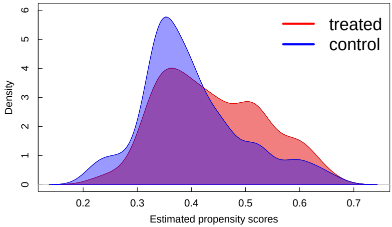
```r
fit <- glm(treat ~ age + educ + hisp +
             married + nodegr + re75 + u74 + u75,
           data = lalonde, family = binomial())

summary(fit$fitted.values[lalonde$treat == 1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2322  0.3581  0.4312  0.4381  0.5127  0.6513
```

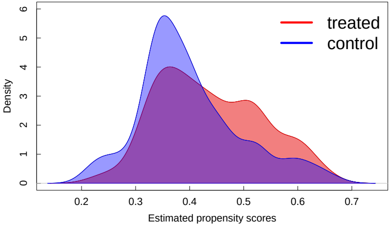
```r
summary(fit$fitted.values[lalonde$treat == 0])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2098  0.3391  0.3817  0.3998  0.4487  0.6711
```



The role of propensity scores in balancing

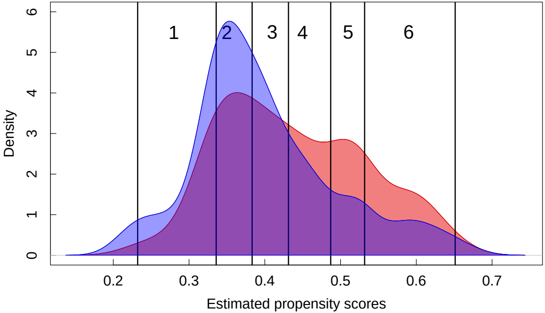
- Roughly speaking, we would like to balance the distribution of propensity scores between the treated and the control.



- Here, we will introduce three methods to condition on the estimated propensity scores: (a) subclassification, (b) matching, and (c) weighting.

(a) Subclassification

Subclassification: to form subclasses, such that in each the distribution of the observed covariates (**or, equivalently, propensity scores**) for the treated and control groups are as similar as possible.



R package MatchIt: (a) Subclassification

```
m.out.subclass <- matchit(treat ~ age + educ + hisp +
             married + nodegr + re75 + u74 + u75,
                 data = lalonde, method = "subclass")
m.out.subclass
```

```
## A matchit object
##  - method: Subclassification (6 subclasses)
##  - distance: Propensity score
##              - estimated with logistic regression
##  - number of obs.: 445 (original), 445 (matched)
##  - target estimand: ATT
##  - covariates: age, educ, hisp, married, nodegr, re75, u74, u75
```

```
#print(summary(m.out.subclass, standardize = TRUE))
```

- You can use `summary()` to see how much balance is improved after subclassification.

(b) Propensity score matching

- Subclassification can be viewed as a form of coarsened exact matching. (`method = "subclass"`)

(b) Propensity score matching

- Subclassification can be viewed as a form of coarsened exact matching. (`method = "subclass"`)
- *k* to 1 nearest neighbor matching (`method = "nearest"`): for each treated unit, select *k* controls with closest propensity scores
- Optimal pair matching (`method = "optimal"`): attempts to choose matches that *collectively* optimize an overall criterion (e.g., minimizing the mean of the absolute pair distances)
- Optimal full matching (`method = "full"`): chooses number of subclasses and the assignment of units in an *optimal* way
- Exact matching (`method = "exact"`): creates subclasses based on unique combinations of covariates (e.g., same age and sex); most powerful matching

(b) Propensity score matching

- Subclassification can be viewed as a form of coarsened exact matching. (`method = "subclass"`)
- *k* **to 1 nearest neighbor matching** (`method = "nearest"`): for each treated unit, select *k* controls with closest propensity scores
- **Optimal pair matching** (`method = "optimal"`): attempts to choose matches that *collectively* optimize an overall criterion (e.g., minimizing the mean of the absolute pair distances)
- Optimal full matching (`method = "full"`): chooses number of subclasses and the assignment of units in an *optimal* way
- Exact matching (`method = "exact"`): creates subclasses based on unique combinations of covariates (e.g., same age and sex); most powerful matching
- **k:1 matching and optimal pair matching can only calculate ATT by design.**

(b) Propensity score matching

- **Subclassification** can be viewed as a form of coarsened exact matching. (`method = "subclass"`)
- *k* to 1 nearest neighbor matching (`method = "nearest"`): for each treated unit, select *k* controls with closest propensity scores
- Optimal pair matching (`method = "optimal"`): attempts to choose matches that *collectively* optimize an overall criterion (e.g., minimizing the mean of the absolute pair distances)
- **Optimal full matching** (`method = "full"`): chooses number of subclasses and the assignment of units in an *optimal* way
- Exact matching (`method = "exact"`): creates subclasses based on unique combinations of covariates (e.g., same age and sex); most powerful matching
- k:1 matching and optimal pair matching can only calculate ATT by design.
- **Optimal full matching and subclassification can calculate ATT or ATE.**

MatchIt: 1:1 nearest matching

```r
m.out <- matchit(treat ~ age + educ + hisp + married +
                   re75 + u74 + u75, data = lalonde,
                 method = "nearest", exact = c("nodegr"))
print(m.out)
```

```
## A matchit object
##  - method: 1:1 nearest neighbor matching without replacement
##  - distance: Propensity score
##              - estimated with logistic regression
##  - number of obs.: 445 (original), 348 (matched)
##  - target estimand: ATT
##  - covariates: age, educ, hisp, married, re75, u74, u75, nodegr
```

```r
#print(summary(m.out, standardize = TRUE))
```

MatchIt: 2:1 nearest matching

ratio: how many **control** units should be matched to each treated unit.

```
m.out2 <- matchit(treat ~ age + educ + hisp + married +
                     nodegr + re75 + u74 + u75, data = lalonde,
                  ratio = 2, method = "nearest")
print(m.out2)
```

```
## A matchit object
##  - method: 2:1 nearest neighbor matching without replacement
##  - distance: Propensity score
##              - estimated with logistic regression
##  - number of obs.: 445 (original), 445 (matched)
##  - target estimand: ATT
##  - covariates: age, educ, hisp, married, nodegr, re75, u74, u75
```

```
#print(summary(m.out, standardize = TRUE))
```

```
dat2 <- match.data(m.out2)
table(dat2$subclass)
```

```
## 
##   1   2   3   4   5   6   7   8   9  10  11  12
##   2   3   2   2   2   2   2   2   2   3   2   2
##  13  14  15  16  17  18  19  20  21  22  23  24
```

MatchIt: Optimal full matching

```
full.out <- matchit(treat ~ age + educ + hisp + married +
                     nodegr + re75 + u74 + u75, data = lalonde,
                 method = "full", estimand = "ATE")
## check estimand = "ATT"
print(full.out)
```

```
## A matchit object
##  - method: Optimal full matching
##  - distance: Propensity score
##              - estimated with logistic regression
##  - number of obs.: 445 (original), 445 (matched)
##  - target estimand: ATE
##  - covariates: age, educ, hisp, married, nodegr, re75, u74, u75
```

```
#print(summary(m.out, standardize = TRUE))
```

(c) Propensity score weighting

- Recall $e_i = Pr(A_i = 1|\mathbf{X}_i)$

ATE weights

$$w_i = \begin{cases} 1/e_i & \text{if } A_i = 1 \\ 1/(1-e_i) & \text{if } A_i = 0 \end{cases} \tag{1}$$

The proportion of (Male):(Female) in the total population is 10:12

Treatment	**Control**	**Total**
8 Males 4 Females	2 Males 8 Females	10 Males 12 Females

(c) Propensity score weighting

- Recall $e_i = Pr(A_i = 1|\mathbf{X}_i)$

ATE weights

$$w_i = \begin{cases} 1/e_i & \text{if } A_i = 1 \\ 1/(1-e_i) & \text{if } A_i = 0 \end{cases}$$

The proportion of (Male):(Female) in the total population is 10:12

Treatment	**Control**	**Total**
8 Males	2 Males	10 Males
4 Females	8 Females	12 Females

- Goal: want to keep the ratio (10:12) in the treatment and the control group.

A	X	Propensity score	**ATE** weight	Observed population	Weighted population
Treatment	1:Male	$Pr(A = 1\|X = 1) =$???	1/??? =	8	
Treatment	0:Female	$Pr(A = 1\|X = 0) =$		4	
Control	1:Male	$Pr(A = 1\|X = 1) =$		2	
Control	0:Female	$Pr(A = 1\|X = 0) =$		8	

Table 2: After ATE weighting, (male):(female) = 10:12 both in the treatment and the control groups

(c) Propensity score weighting

- Recall $e_i = Pr(A_i = 1|\mathbf{X}_i)$

ATE weights

$$w_i = \begin{cases} 1/e_i & \text{if } A_i = 1 \\ 1/(1-e_i) & \text{if } A_i = 0 \end{cases}$$

The proportion of (Male):(Female) in the total population is 10:12

Treatment	Control	Total
8 Males	2 Males	10 Males
4 Females	8 Females	12 Females

- Goal: want to keep the ratio (10:12) in the treatment and the control group.

A	X	Propensity score	**ATE** weight	Observed population	Weighted population
Treatment	1:Male	$Pr(A=1\|X=1) = 0.8$	$1/0.8 = 1.25$	8	$1.25 \times 8 = 10$
Treatment	0:Female	$Pr(A=1\|X=0) =$		4	
Control	1:Male	$Pr(A=1\|X=1) =$		2	
Control	0:Female	$Pr(A=1\|X=0) =$		8	

Table 3: After ATE weighting, (male):(female) = 10:12 both in the treatment and the control groups

(c) Propensity score weighting

- Recall $e_i = Pr(A_i = 1|\mathbf{X}_i)$

ATE weights

$$w_i = \begin{cases} 1/e_i & \text{if } A_i = 1 \\ 1/(1-e_i) & \text{if } A_i = 0 \end{cases} \tag{2}$$

The proportion of (Male):(Female) in the total population is 10:12

Treatment	**Control**	**Total**
8 Males	2 Males	10 Males
4 Females	8 Females	12 Females

- Goal: want to keep the ratio (10:12) in the treatment and the control group.

A	X	Propensity score	**ATE** weight	Observed population	Weighted population
Treatment	1:Male	$Pr(A = 1\|X = 1) = 0.8$	1/0.8 = 1.25	8	10
Treatment	0:Female	$Pr(A = 1\|X = 0) = 0.33$	1/0.33 = 3	4	12
Control	1:Male	$Pr(A = 1\|X = 1) = 0.8$	1/(1-0.8) = 5	2	10
Control	0:Female	$Pr(A = 1\|X = 0) = 0.33$	1/(1-0.33) = 1.5	8	12

Table 4: After ATE weighting, (male):(female) = 10:12 both in the treatment and the control groups

(c) Propensity score weighting

- Recall $e_i = Pr(A_i = 1|\mathbf{X}_i)$

ATT weights

$$w_i = \begin{cases} 1 & \text{if } A_i = 1 \\ e_i/(1-e_i) & \text{if } A_i = 0 \end{cases} \quad (3)$$

The proportion of (Male):(Female) in the treated population is 8:4

Treatment	Control	Total
8 Males	2 Males	10 Males
4 Females	8 Females	12 Females

- Goal: want to keep the ratio (8:4) in the treatment and the control group.
  - We do not have to weight the treatment group.

A	X	Propensity score	**ATT** weight	Observed population	Weighted population
Treatment	1:Male	$Pr(A = 1\|X = 1) = 0.8$	1	8	8
Treatment	0:Female	$Pr(A = 1\|X = 0) = 0.33$	1	4	4
Control	1:Male	$Pr(A = 1\|X = 1) = 0.8$	0.8/(1-0.8) = 4	2	8
Control	0:Female	$Pr(A = 1\|X = 0) = 0.33$	0.33/(1-0.33) = 0.5	8	4

Table 5: After ATT weighting, (male):(female) = 8:4 both in the treatment and the control groups

R package WeightIt

```r
library(WeightIt)
w.out <- weightit(treat ~ age + educ + hisp + married +
                     nodegr + re75 + u74 + u75, data = lalonde,
                  estimand = "ATE")
summary(w.out)
```

```
##                   Summary of weights
##
## - Weight ranges:
##
##            Min
## treated 1.5353    |------------------------|
## control 1.2655 |---------------|
##            Max
## treated 4.3071
## control 3.0400
##
## - Units with 5 greatest weights by group:
##
##              44     87    100      2     28
##  treated 3.5286 3.9386 3.9733 4.1921 4.3071
##             372    365    422    428    382
##  control 2.7441 2.8178 2.8295 2.9837   3.04
##
```

R package WeightIt

```r
library(WeightIt)
w.out2 <- weightit(treat ~ age + educ + hisp + married +
                     nodegr + re75 + u74 + u75, data = lalonde,
                   estimand = "ATT")
summary(w.out2)
```

```
##                   Summary of weights
##
## - Weight ranges:
##
##             Min                                 Max
## treated 1.0000             ||                 1.00
## control 0.2655 |---------------------------| 2.04
##
## - Units with 5 greatest weights by group:
##
##                6      5      4      3    1
##  treated       1      1      1      1    1
##              372    365    422    428  382
##  control 1.7441 1.8178 1.8295 1.9837 2.04
##
## - Weight statistics:
##
```

Recap

- We estimate the propensity score as a summary of all the observed covariates and use the score to balance two treatment groups.
- As a way to adjust for potential confounders using propensity scores, we can consider (a) stratification, (b) matching, and (c) weighting.
- Before implementing each method, we need to specify the target estimand (e.g., ATE, ATT)

The next step of causal analysis after propensity score adjustment

4. **Assess balance on the covariates of interest; if poor, repeat steps 2-4**

- Covariate balance is typically assessed and reported by using statistical measures, including **standardized mean differences**, variance ratios, and *t*-test. (similar to Step 1)
- **Standardized mean difference (SMD)**: the difference in the proportions/means across the treatment group divided by the standard deviation within the treatment group.
- For SMD, a threshold of 0.1 can be used to determine whether balance of each covariate is satisfactory.

R package cobalt: Assessing balance

4. **Assess balance on the covariates of interest; if poor, repeat steps 2-4**

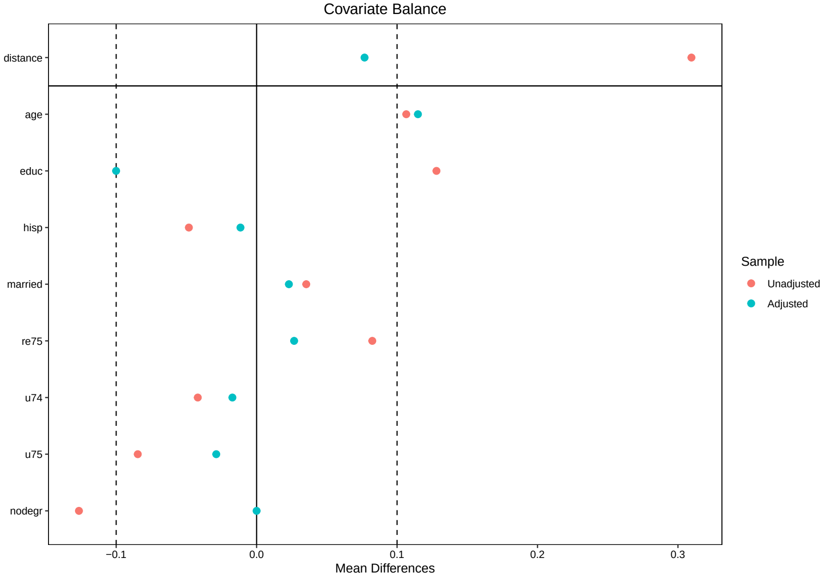
cobalt: provides the balance assessment tools and allows researchers to report balance on observed covariates before and after conditioning.

```
library(cobalt)
bal.tab(m.out, stats = "m", thresholds = c(m = 0.1))
```

```
## Call
##  matchit(formula = treat ~ age + educ + hisp + married + re75 +
##     u74 + u75, data = lalonde, method = "nearest", exact = c("nodegr
##
## Balance Measures
##              Type Diff.Adj        M.Threshold
## distance Distance   0.0769     Balanced, <0.1
## age        Contin.   0.1149 Not Balanced, >0.1
## educ       Contin.  -0.1000 Not Balanced, >0.1
## hisp        Binary  -0.0115     Balanced, <0.1
## married     Binary   0.0230     Balanced, <0.1
## re75       Contin.   0.0267     Balanced, <0.1
## u74         Binary  -0.0172     Balanced, <0.1
## u75         Binary  -0.0287     Balanced, <0.1
## nodegr      Binary   0.0000     Balanced, <0.1
##
## Balance tally for mean differences
```

R package cobalt: Assessing balance

```
love.plot(m.out,  thresholds = c(m = 0.1))
```



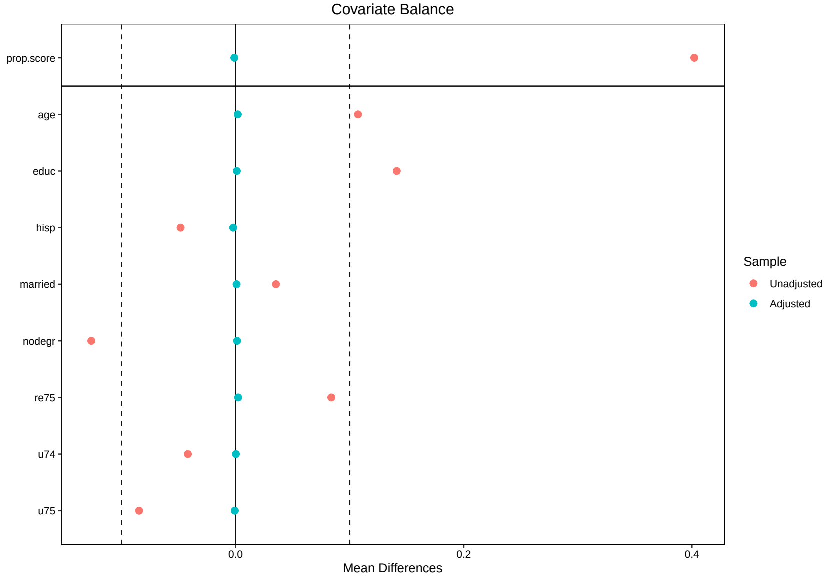
R package cobalt: Assessing balance

```
bal.tab(w.out) ## w.out <- weightit(treat ~ ...)
```

```
## Call
##  weightit(formula = treat ~ age + educ + hisp + married + nodegr +
##     re75 + u74 + u75, data = lalonde, estimand = "ATE")
##
## Balance Measures
##                 Type Diff.Adj
## prop.score Distance  -0.0011
## age        Contin.   0.0019
## educ       Contin.   0.0010
## hisp        Binary  -0.0022
## married     Binary   0.0008
## nodegr      Binary   0.0013
## re75       Contin.   0.0022
## u74         Binary   0.0002
## u75         Binary  -0.0007
##
## Effective sample sizes
##            Control Treated
## Unadjusted  260.     185.
## Adjusted    250.78   176.1
```

R package cobalt: Assessing balance

```
love.plot(w.out,  thresholds = c(m = 0.1))
```



Last step in performing causal analysis

5. **Estimate the treatment effect in the conditional sample.**

- So, we have adjusted our dataset using propensity scores. How can we estimate the ATE or ATT using this stratified/matched/weighted data?
- **Subclassification**: either (1) estimates effects within subclass and then combines, or (2) includes subclass terms in the outcome model (e.g., `subclass*treat`).
- **Matching**: runs regression on matched samples (i.e., `data = match.data(m.out)`)
- **Weighting**: runs regression with weights.

Last step in performing causal analysis

Matching + Regression

- In the outcome model, should we include other observed covariates (e.g., `age`)?
  - can include covariates in both models (propensity score and outcome) if not interested in coefficients of that covariate in the outcome model.
  - A coefficient associated with `treat` can be interpreted as an average causal effect.

```
fit.m <- lm(re78 ~ treat + age + educ + hisp +
              married + nodegr + re75 + u74 + u75,
            data = match.data(m.out))
## the target estimand of m.out was ATT
confint(fit.m)[2,]
```

```
##     2.5 %    97.5 %
##  193.4309 3091.0300
```

Last step in performing causal analysis

Weighted population + Regression

- R package survey: fit a generalized linear model to data from a complex survey design, with inverse-probability weighting and design-based standard errors.

```
library(survey)
design.w <- svydesign(~1, weights = w.out$weights, data = lalonde)
fit.w <- svyglm(re78 ~ treat + age + educ + hisp +
          married + nodegr + re75 + u74 + u75,
          design = design.w)
## the target estimand w.out was ATE
confint(fit.w)[2,]
```

```
##     2.5 %    97.5 %
##  314.9374 2891.0605
```

- implement the same code with w.out2 (with estimand = "ATT")

Exercise: Right heart catheterization data

- Units: ICU patients in 5 hospitals
- Outcome : binary indicator for death (yes/no)
- Treatment: right heart cathetization (rhc) vs. not
- Confounders: demographics, insurance, disease diagnoses, etc.
- (A subset of) Data is available in the data folder.

```
dat = read.csv("../data/rhc.csv", sep = ",", header = TRUE)
```

Exercise: Right heart catheterization data

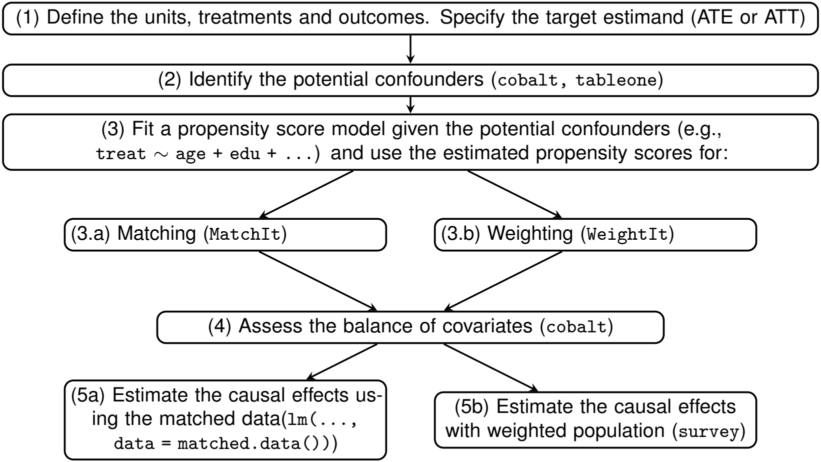
```r
head(dat)
```

```
##   ARF CHF Cirr colcan Coma lungcan MOSF sepsis
## 1   0   0    0      0    0       0    0      0
## 2   0   0    0      0    0       0    0      1
## 3   0   0    0      0    0       0    1      0
## 4   1   0    0      0    0       0    0      0
## 5   0   0    0      0    0       0    0      1
## 6   0   0    0      0    0       0    0      0
##          age female treatment died
## 1 70.25098      0         0    0
## 2 78.17896      1         1    1
## 3 46.09198      1         1    0
## 4 75.33197      1         0    1
## 5 67.90997      0         1    1
## 6 86.07794      1         0    0
```

```r
xvars <- c("ARF", "CHF", "Cirr", "colcan", "Coma", "lungcan",
           "MOSF", "sepsis", "age", "female", "meanbp1")
```

10 Minute Countdown

Flowchart



(1) Define the units, treatments and outcomes. Specify the target estimand (ATE or ATT)

(2) Identify the potential confounders (`cobalt`, `tableone`)

(3) Fit a propensity score model given the potential confounders (e.g., `treat ~ age + edu + ...`) and use the estimated propensity scores for:

(3.a) Matching (`MatchIt`)

(3.b) Weighting (`WeightIt`)

(4) Assess the balance of covariates (`cobalt`)

(5a) Estimate the causal effects using the matched data(`lm(..., data = matched.data())`)

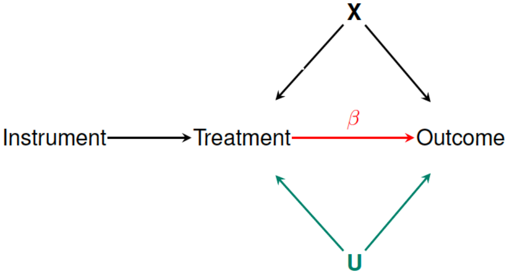
(5b) Estimate the causal effects with weighted population (`survey`)

Back to R Studio

Other useful `R` packages for causal inference

Instrumental variable method

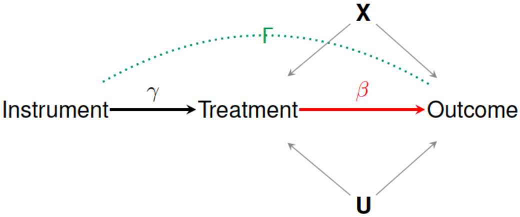
- The instrumental variables method is a popular method to estimate the causal effect of a treatment, exposure or policy on an outcome when **unmeasured** confounding is present.



- We can use the randomization of the “instrumental variable" (IV) to help us estimate the effect we are interested in (β).
- Useful for dealing with noncompliance in randomized trials.

Instrumental variable method

- The instrumental variable methods extract variation in the treatment that is free of the unmeasured confounders and use this confounder-free variation in the treatment to estimate the causal effect of the treatment[5].



- Two stage least squares method to estimate β
- Useful Coursera course: https://www.coursera.org/lecture/crash-course-in-causality/iv-analysis-in-r-D19Ae
- R packages: `ivreg`, `ivmodel`, ...

[5] Baiocchi, M., Cheng, J., & Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in medicine*, 33(13), 2297-2340.

Regression discontinuity

- Use a cutoff or threshold to assign an intervention.

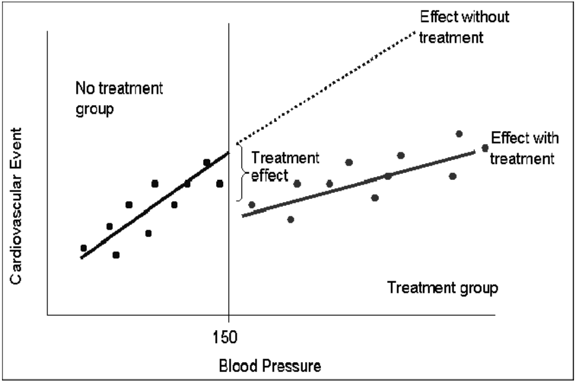
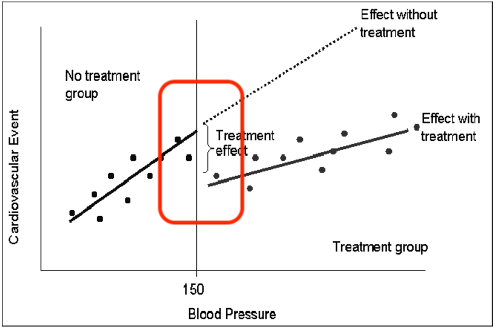


Figure 1: The treatment is assigned when blood pressure exceed 150, and we want to compare the treatment effect on the outcomes (e.g., cardiovascular events)

Regression discontinuity

- A key assumption is that people just before and just above the cutoff (e.g., BP 150) are only randomly different.



- R package: `rdd`, `rddtools`
- https://www.econometrics-with-r.org/13-4-quasi-experiments.html

Figure: Tsai, J. H. C., Tu, S. P., Perrin, N. A., & Breslau, E. S. (2016). Implementation Research and Asian American/Pacific Islander Health. Asian/Pacific Island Nursing Journal, 1(2), 24-34.

References

- Dr. Stuart webpage: Software for implementing matching methods and propensity scores
  - https://www.elizabethstuart.org/psoftware/rcode/
- Noah Greifer's cobalt package vignettes
  - https://cran.microsoft.com/snapshot/2017-08-01/web/packages/cobalt/vignettes/cobalt_basic_use.html
- UseR! 2020: Causal inference in R (Lucy D'Agostino McGowan, Malcom Barrett)
  - Video: https://youtu.be/n8c-UK19hbA
- Greifer, N. (2021). cobalt: Covariate Balance Tables and Plots. R package version 4.3.1.9000.

Thank you!

Email: youjin_lee@brown.edu