Lab 4

Olivier Bergeron-Boutin

February 9th, 2021

Reviewing key concepts

What does randomization do?

Take a minute to think about what randomization does for us

What does randomization do?

Take a minute to think about what randomization does for us

True or false: randomization ensures that $Y_i \perp D_i$

What does randomization do?

Take a minute to think about what randomization does for us

True or false: randomization ensures that $Y_i \perp D_i$

- False: This would imply that $\tau_{ATE} = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = 0$

What does randomization do?

Take a minute to think about what randomization does for us

True or false: randomization ensures that $Y_i \perp D_i$

- False: This would imply that $\tau_{ATE} = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = 0$
- Randomization is about the *potential* outcomes, not the realized outcomes

What does randomization do?

Take a minute to think about what randomization does for us

True or false: randomization ensures that $Y_i \perp D_i$

- False: This would imply that $\tau_{ATE} = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = 0$
- Randomization is about the *potential* outcomes, not the realized outcomes

So what does randomization do then?

What does randomization do?

Take a minute to think about what randomization does for us

True or false: randomization ensures that $Y_i \perp D_i$

- False: This would imply that $\tau_{ATE} = \mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i | D_i = 0] = 0$
- Randomization is about the *potential* outcomes, not the realized outcomes

So what does randomization do then?

- It solves the selection problem: $\{Y_i(1), Y_i(0)\} \perp D_i$

What does randomization do?

Take a minute to think about what randomization does for us

True or false: randomization ensures that $Y_i \perp D_i$

- False: This would imply that $\tau_{ATE} = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = 0$
- Randomization is about the *potential* outcomes, not the realized outcomes

So what does randomization do then?

- It solves the selection problem: $\{Y_i(1), Y_i(0)\} \perp D_i$
- Assume you have access to the entire schedule of potential outcomes and show an example of the selection problem

What does randomization do?

Take a minute to think about what randomization does for us

True or false: randomization ensures that $Y_i \perp D_i$

- False: This would imply that $\tau_{ATE} = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = 0$
- Randomization is about the *potential* outcomes, not the realized outcomes

So what does randomization do then?

- It solves the selection problem: $\{Y_i(1), Y_i(0)\} \perp D_i$
- Assume you have access to the entire schedule of potential outcomes and show an example of the selection problem
- Selection problem: $\mathbb{E}[Y_i(0)|D_i = 1] \neq \mathbb{E}[Y_i(0)|D_i = 0]$

What does randomization do?

Take a minute to think about what randomization does for us

True or false: randomization ensures that $Y_i \perp D_i$

- False: This would imply that $\tau_{ATE} = \mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i | D_i = 0] = 0$
- Randomization is about the *potential* outcomes, not the realized outcomes

So what does randomization do then?

- It solves the selection problem: $\{Y_i(1), Y_i(0)\} \perp D_i$
- Assume you have access to the entire schedule of potential outcomes and show an example of the selection problem
- Selection problem: $\mathbb{E}[Y_i(0) | D_i = 1] \neq \mathbb{E}[Y_i(0) | D_i = 0]$
- If the treated units had not received treatment, they would still have systematic differences in outcome compared to the control units

SUTVA

Another one of those intuitively-named concepts...

“Gold standard”

What exactly do we mean when we say that randomized experiments are the gold standard?

“Gold standard”

What exactly do we mean when we say that randomized experiments are the gold standard?

- It’s about internal validity/identification without bias

“Gold standard”

What exactly do we mean when we say that randomized experiments are the gold standard?

- It’s about internal validity/identification without bias
- But ultimately, our standard is about learning about the substance of politics

“Gold standard”

What exactly do we mean when we say that randomized experiments are the gold standard?

- It’s about internal validity/identification without bias
- But ultimately, our standard is about learning about the substance of politics
- …which randomized experiments may or may not help with

“Gold standard”

What exactly do we mean when we say that randomized experiments are the gold standard?

- It’s about internal validity/identification without bias
- But ultimately, our standard is about learning about the substance of politics
- …which randomized experiments may or may not help with
- Recent debates: credibility revolution or narrow hypothesis testing?

“Gold standard”

What exactly do we mean when we say that randomized experiments are the gold standard?

- It’s about internal validity/identification without bias
- But ultimately, our standard is about learning about the substance of politics
- ...which randomized experiments may or may not help with
- Recent debates: credibility revolution or narrow hypothesis testing?

	Good identification	Poor identification
Substantively interesting	yep	?
Substantively trivial	?	nope

Matching

Review

The setup: observational data with self-selection intro treatment

Review

The setup: observational data with self-selection intro treatment

- Clearly, our assumption $\{Y_i(1), Y_i(0)\} \perp D_i$ does not hold

Review

The setup: observational data with self-selection intro treatment

- Clearly, our assumption $\{Y_i(1), Y_i(0)\} \perp D_i$ does not hold
- Gary King: experiment hidden in observational data

Review

The setup: observational data with self-selection intro treatment

- Clearly, our assumption $\{Y_i(1), Y_i(0)\} \perp D_i$ does not hold
- Gary King: experiment hidden in observational data
- What's the "hidden experiment"?

Review

The setup: observational data with self-selection intro treatment

- Clearly, our assumption $\{Y_i(1), Y_i(0)\} \perp D_i$ does not hold
- Gary King: experiment hidden in observational data
- What's the "hidden experiment"?
- If I can condition on a set of covariates X_i, treatment assignment D_i is independent of the potential outcomes

Review

The setup: observational data with self-selection intro treatment

- Clearly, our assumption $\{Y_i(1), Y_i(0)\} \perp D_i$ does not hold
- Gary King: experiment hidden in observational data
- What’s the “hidden experiment”?
- If I can condition on a set of covariates X_i, treatment assignment D_i is independent of the potential outcomes
- Conditional ignorability: $\{Y_i(1), Y_i(0)\} \perp D_i | X_i$

Review

The setup: observational data with self-selection intro treatment

- Clearly, our assumption $\{Y_i(1), Y_i(0)\} \perp D_i$ does not hold
- Gary King: experiment hidden in observational data
- What's the "hidden experiment"?
- If I can condition on a set of covariates X_i, treatment assignment D_i is independent of the potential outcomes
- Conditional ignorability: $\{Y_i(1), Y_i(0)\} \perp D_i | X_i$
- Selection on *observables*: to condition on X, I must observe it!

Review

The setup: observational data with self-selection intro treatment

- Clearly, our assumption $\{Y_i(1), Y_i(0)\} \perp D_i$ does not hold
- Gary King: experiment hidden in observational data
- What's the "hidden experiment"?
- If I can condition on a set of covariates X_i, treatment assignment D_i is independent of the potential outcomes
- Conditional ignorability: $\{Y_i(1), Y_i(0)\} \perp D_i | X_i$
- Selection on *observables*: to condition on X, I must observe it!

Causal inference

Given its limitations, is matching a causal inference method?

Causal inference

Given its limitations, is matching a causal inference method?

- A bit of a trick question: under the right assumption,

Example from Blattman and Annan (2010)

Setup

Children in Uganda are regularly abducted to serve as soldiers

Setup

Children in Uganda are regularly abducted to serve as soldiers

- Blattman and Annan are interested in the effect on education

Setup

Children in Uganda are regularly abducted to serve as soldiers

- Blattman and Annan are interested in the effect on education
- Problem: there's self-selection into treatment

Setup

Children in Uganda are regularly abducted to serve as soldiers

- Blattman and Annan are interested in the effect on education
- Problem: there's self-selection into treatment
- In this context, what does this mean?
  $\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \neq 0$

Setup

Children in Uganda are regularly abducted to serve as soldiers

- Blattman and Annan are interested in the effect on education
- Problem: there's self-selection into treatment
- In this context, what does this mean?
  $\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \neq 0$
- You should speculate (to yourself) about the nature and direction of selection bias

Setup

Children in Uganda are regularly abducted to serve as soldiers

- Blattman and Annan are interested in the effect on education
- Problem: there's self-selection into treatment
- In this context, what does this mean?
  $\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \neq 0$
- You should speculate (to yourself) about the nature and direction of selection bias
- What's a plausible story that could explain the inequality above?

Naive estimate

```r
child <- read.csv("child_soldiering.csv")
```

Y_i is *educ* and D_i is *abd*

Naive estimate

```r
child <- read.csv("child_soldiering.csv")
```

Y_i is ***educ*** and D_i is ***abd***

Produce a “naive” estimate of the average treatment effect of abduction on education

Naive estimate

```
child <- read.csv("child_soldiering.csv")
```

Y_i is *educ* and D_i is *abd*

Produce a “naive” estimate of the average treatment effect of abduction on education

```
mean(child$educ[child$abd==1]) - mean(child$educ[child$abd==0])
```

```
## [1] -0.5954243
```

```
sqrt(var(child$educ[child$abd==1])/length(child$educ[child$abd==1]) +
      var(child$educ[child$abd==0])/length(child$educ[child$abd==0]))
```

```
## [1] 0.222192
```

Naive ATE with regression

```
library(lmtest)
library(sandwich)
library(modelsummary)

model <- lm(educ ~ abd, data = child)
lmtest::coeftest(model, vcov = sandwich::vcovHC(model, type = "HC2"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  7.41577    0.18031 41.1271 < 2.2e-16 ***
## abd         -0.59542    0.22219 -2.6798  0.007531 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Balance test

```r
library(Matching)

# Storing formula: D_i on LHS, X_i on RHS
formul <- abd ~ C.ach + C.akw + C.ata + C.kma + C.oro + C.pad + C.paj + C.pal +
  age + fthr.ed + mthr.ed + orphan96 + fthr.frm + hh.size96 + educ

# nboots is nb. of iterations fo K-S test; print.level avoids output
unmatched_bal <- Matching::MatchBalance(formul = formul,
                                        data = child,
                                        nboots = 1000,
                                        print.level = 0)
# Function above creates ugly list object
# baltest.collect manipulates it to create a matrix
baltable_unmatched <- ebal::baltest.collect(matchbal.out = unmatched_bal,
                                            var.names = colnames(child)[-1],
                                            after = FALSE)
```

Balance table

```
baltable_unmatched[,-8:-10] %>% round(3) %>%  knitr::kable()
```

	mean.Tr	mean.Co	sdiff	sdiff.pooled	var.ratio	T pval	KS pval
C.ach	0.154	0.115	10.798	11.440	1.279	0.126	NA
C.akw	0.158	0.079	21.678	24.650	1.829	0.001	NA
C.ata	0.100	0.197	-32.549	-27.668	0.566	0.000	NA
C.kma	0.152	0.118	9.259	9.727	1.231	0.194	NA
C.oro	0.052	0.136	-37.924	-29.119	0.418	0.000	NA
C.pad	0.121	0.122	-0.199	-0.199	0.994	0.979	NA
C.paj	0.152	0.104	13.254	14.269	1.378	0.055	NA
C.pal	0.113	0.129	-5.208	-5.051	0.888	0.509	NA
age	21.366	20.151	24.242	24.495	1.043	0.001	0.010
fthr.ed	5.764	6.068	-8.590	-8.465	0.944	0.266	0.839
mthr.ed	2.093	2.495	-14.493	-13.311	0.729	0.086	0.362
orphan96	0.078	0.075	0.989	0.996	1.031	0.895	NA
fthr.frm	0.903	0.914	-3.834	-3.938	1.117	0.601	NA
hh.size96	8.090	8.695	-15.495	-14.624	0.803	0.058	0.034
educ	6.820	7.416	-21.337	-20.508	0.859	0.008	0.073

An aside on balance

What’s the null for this covariate balance test?

An aside on balance

What's the null for this covariate balance test?

- Given what we know about hypothesis tests, does that setup appear reasonable?

An aside on balance

What’s the null for this covariate balance test?

- Given what we know about hypothesis tests, does that setup appear reasonable?
- New-ish approach: equivalence tests

Matching procedure

```
vars <- c("age", "fthr.ed", "mthr.ed", "orphan96", "fthr.frm", "hh.size96")
match_out <- Matching::Match(Y = child$educ,
                             Tr = child$abd,
                             X = child[,vars],
                             M = 1,
                             Weight = 2,
                             exact = FALSE)
```

Matching procedure

```r
vars <- c("age", "fthr.ed", "mthr.ed", "orphan96", "fthr.frm", "hh.size96")
match_out <- Matching::Match(Y = child$educ,
                             Tr = child$abd,
                             X = child[,vars],
                             M = 1,
                             Weight = 2,
                             exact = FALSE)
```

- M is the number of matches for each treated unit

Matching procedure

```
vars <- c("age", "fthr.ed", "mthr.ed", "orphan96", "fthr.frm", "hh.size96")
match_out <- Matching::Match(Y = child$educ,
                             Tr = child$abd,
                             X = child[,vars],
                             M = 1,
                             Weight = 2,
                             exact = FALSE)
```

- `M` is the number of matches for each treated unit
- Here, `M = 1` means one-to-one matching

Matching procedure

```r
vars <- c("age", "fthr.ed", "mthr.ed", "orphan96", "fthr.frm", "hh.size96")
match_out <- Matching::Match(Y = child$educ,
                             Tr = child$abd,
                             X = child[,vars],
                             M = 1,
                             Weight = 2,
                             exact = FALSE)
```

- `M` is the number of matches for each treated unit
- Here, `M = 1` means one-to-one matching
- The **`Weight`** argument indicates the matching algorithm to use

Matching procedure

```
vars <- c("age", "fthr.ed", "mthr.ed", "orphan96", "fthr.frm", "hh.size96")
match_out <- Matching::Match(Y = child$educ,
                             Tr = child$abd,
                             X = child[,vars],
                             M = 1,
                             Weight = 2,
                             exact = FALSE)
```

- `M` is the number of matches for each treated unit
- Here, `M = 1` means one-to-one matching
- The **`Weight`** argument indicates the matching algorithm to use
- Here **`Weight`** `= 2` means the Mahalanobis distance

Matching procedure

```r
vars <- c("age", "fthr.ed", "mthr.ed", "orphan96", "fthr.frm", "hh.size96")
match_out <- Matching::Match(Y = child$educ,
                             Tr = child$abd,
                             X = child[,vars],
                             M = 1,
                             Weight = 2,
                             exact = FALSE)
```

- `M` is the number of matches for each treated unit
- Here, `M = 1` means one-to-one matching
- The **`Weight`** argument indicates the matching algorithm to use
- Here **`Weight = 2`** means the Mahalanobis distance
- `exact` to force exact matching and prune treated units without a perfect buddy

ATT estimate

```r
summary(match_out)
```

```
##
## Estimate...  -0.50152
## AI SE......  0.26428
## T-stat.....  -1.8976
## p.val......  0.057744
##
## Original number of observations..............  741
## Original number of treated obs...............  462
## Matched number of observations...............  462
## Matched number of observations  (unweighted).  521
##
## Number of obs dropped by 'exact' or 'caliper'  0
```

What exactly happened?

We can retrieve the underlying data using the Match object produced

What exactly happened?

We can retrieve the underlying data using the Match object produced

- the *`mdata`* data contains a list with three elements: a vector of Y values, a vector of *`Tr`* values, and a matrix of X values

What exactly happened?

We can retrieve the underlying data using the Match object produced

- the `mdata` data contains a list with three elements: a vector of Y values, a vector of `Tr` values, and a matrix of X values
- Using those, we can reconstruct the matched pairs

What exactly happened?

We can retrieve the underlying data using the Match object produced

- the `mdata` data contains a list with three elements: a vector of Y values, a vector of `Tr` values, and a matrix of X values
- Using those, we can reconstruct the matched pairs

```
Y <- match_out$mdata[[1]]
Tr <- match_out$mdata[[2]]
X <- match_out$mdata[[3]]

# The first treated observation (i=1) and its buddy (i=2)
rbind(X[1,], X[1+nrow(X)/2,]) %>% knitr::kable("latex")
```

age	fthr.ed	mthr.ed	orphan96	fthr.frm	hh.size96
21	7	4	0	1	9
20	7	4	0	1	8

What exactly happened?

We can retrieve the underlying data using the Match object produced

- the *mdata* data contains a list with three elements: a vector of *Y* values, a vector of *Tr* values, and a matrix of *X* values
- Using those, we can reconstruct the matched pairs

```
Y <- match_out$mdata[[1]]
Tr <- match_out$mdata[[2]]
X <- match_out$mdata[[3]]

# The first treated observation (i=1) and its buddy (i=2)
rbind(X[1,], X[1+nrow(X)/2,]) %>% knitr::kable("latex")
```

age	fthr.ed	mthr.ed	orphan96	fthr.frm	hh.size96
21	7	4	0	1	9
20	7	4	0	1	8

What do you think of this match? Good counterfactual?

What exactly happened?

We can retrieve the underlying data using the Match object produced

- the *`mdata`* data contains a list with three elements: a vector of *Y* values, a vector of *`Tr`* values, and a matrix of *X* values
- Using those, we can reconstruct the matched pairs

```
Y <- match_out$mdata[[1]]
Tr <- match_out$mdata[[2]]
X <- match_out$mdata[[3]]

# The first treated observation (i=1) and its buddy (i=2)
rbind(X[1,], X[1+nrow(X)/2,]) %>% knitr::kable("latex")
```

age	fthr.ed	mthr.ed	orphan96	fthr.frm	hh.size96
21	7	4	0	1	9
20	7	4	0	1	8

What do you think of this match? Good counterfactual? Seems like it…but the unobservables! Think about $\{Y_i(1), Y_i(0) \perp D_i | X_i\}$

Retrieving the Mahalanobis distance

```
X_control <- child[child$abd==0,vars]
X_treat <- child[child$abd==1,vars]

library(StatMatch)
X_control$dist_i1 <-
    mahalanobis.dist(data.x = X_treat[1,vars],
                     data.y = X_control[,vars]) %>% as.vector()

knitr::kable(plyr::rbind.fill(X_treat[1,], X_control[1:7,]))
```

age	fthr.ed	mthr.ed	orphan96	fthr.frm	hh.size96	dist_i1
21	7	4	0	1	9	NA
29	7	0	0	1	11	2.066150
14	0	0	0	1	17	3.303101
19	4	4	0	1	6	1.106657
16	0	0	0	1	8	2.270651
16	0	0	0	1	5	2.351817
15	0	0	0	1	4	2.513462
29	7	0	0	1	8	2.030444

To prune or not to prune?

Without a caliper, the algorithm will find a match for **every** treated unit, whether the match is a plausible counterfactual or not

To prune or not to prune?

Without a caliper, the algorithm will find a match for **every** treated unit, whether the match is a plausible counterfactual or not

- Low caliper: more plausible counterfactual, but more pruning

To prune or not to prune?

Without a caliper, the algorithm will find a match for **every** treated unit, whether the match is a plausible counterfactual or not

- Low caliper: more plausible counterfactual, but more pruning
- We lose our ability to speak about the entire sample of treated units, but make conditional ignorability more plausible

To prune or not to prune?

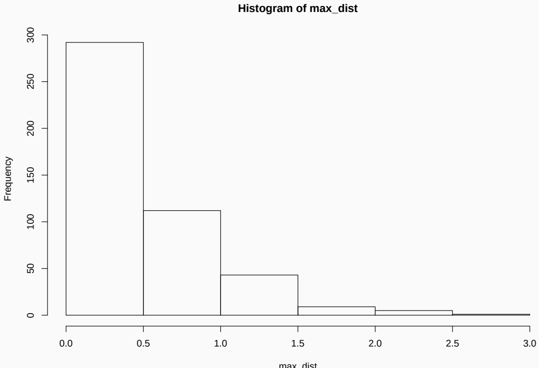
Without a caliper, the algorithm will find a match for **every** treated unit, whether the match is a plausible counterfactual or not

- Low caliper: more plausible counterfactual, but more pruning
- We lose our ability to speak about the entire sample of treated units, but make conditional ignorability more plausible
- King, Lucas, and Nielsen 2017: balance-sample size frontier

Let's find the treated unit with the most dissimilar buddy...

To prune or not to prune?

```
dist <- list()
for(i in 1:nrow(X_treat)){
  dist[[i]] <- mahalanobis.dist(data.x = X_treat[i,vars],
                                data.y = X_control[,vars]) %>% as.vector()
}
max_dist <- sapply(dist, min)
hist(max_dist)
```



To prune or not to prune

```
X_buddies <- X[(nrow(X)/2+1):(nrow(X)),]
treat_max <- X_treat[which.max(max_dist),]
control_max <- matrix(X_buddies[which.max(max_dist),], nrow=1) %>%
  as.data.frame() %>%
  setNames(vars)

knitr::kable(rbind(treat_max, control_max))
```

	age	fthr.ed	mthr.ed	orphan96	fthr.frm	hh.size96
609	21	12	4	1	0	6
1	16	7	0	0	1	13

Balance atfter matching

```
matched_bal <- Matching::MatchBalance(formul = formul,
                                      data = child,
                                      match.out = match_out,
                                      ks = TRUE,
                                      nboots = 1000,
                                      print.level = 0)

baltable_matched <- ebal::baltest.collect(matchbal.out = matched_bal,
                                          var.names = colnames(child)[-1],
                                          after = TRUE) %>% as.data.frame()

baltable_matched[,-8:-10] %>% round(3) %>%  knitr::kable()
```

Balance after matching

	mean.Tr	mean.Co	sdiff	sdiff.pooled	var.ratio	T pval	KS
C.ach	0.154	0.100	14.888	14.888	1.446	0.015	
C.akw	0.158	0.098	16.301	16.301	1.498	0.005	
C.ata	0.100	0.222	-40.727	-40.727	0.520	0.000	
C.kma	0.152	0.116	10.010	10.010	1.258	0.106	
C.oro	0.052	0.137	-38.452	-38.452	0.416	0.000	
C.pad	0.121	0.089	9.937	9.937	1.317	0.120	
C.paj	0.152	0.116	9.910	9.910	1.254	0.102	
C.pal	0.113	0.122	-3.079	-3.079	0.931	0.654	
age	21.366	21.133	4.640	4.640	1.068	0.009	
fthr.ed	5.764	5.682	2.324	2.324	1.086	0.170	
mthr.ed	2.093	2.069	0.859	0.859	0.999	0.489	
orphan96	0.078	0.078	0.000	0.000	1.000	1.000	
fthr.frm	0.903	0.903	0.000	0.000	1.000	1.000	

Propensity score

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

Generate PS for each unit and match on PS