Blocking

Ryan T. Moore

American University

The Lab @ DC

2024-06-11

Table of contents I

What?

Motivation: A Causal Inference Question

"Would a canvassing policy increase enrollment in a health insurance program?"

Motivation: A Causal Inference Question

"Would a canvassing policy increase enrollment in a health insurance program?"

Precinct	Party	Canvass?	Enroll %
1	Dem		
2	Dem		
3	Rep		
4	Rep		

Motivation: A Causal Inference Question

Suppose we observationally measure

Precinct	Party	Canvass?	Enroll %
1	Dem	Yes	60
2	Dem	Yes	70
3	Rep	No	20
4	Rep	No	30
		Diff in Means: (Yes − No)	40

Motivation: A Causal Inference Question

Suppose we observationally measure

Precinct	Party	Canvass?	Enroll %
1	Dem	Yes	60
2	Dem	Yes	70
3	Rep	No	20
4	Rep	No	30
		Diff in Means: (Yes − No)	40

Causal claims? Concerns?

Motivation: A Causal Inference Question

Suppose we randomly assign 2 Tr, 2 Co, and measure

Precinct	Party	Canvass?	Enroll %
1	Dem	Yes	60
2	Dem	Yes	70
3	Rep	No	20
4	Rep	No	30
		Diff in Means: (Yes − No)	40

Motivation: A Causal Inference Question

Suppose we randomly assign 2 Tr, 2 Co, and measure

Precinct	Party	Canvass?	Enroll %
1	Dem	Yes	60
2	Dem	Yes	70
3	Rep	No	20
4	Rep	No	30
		Diff in Means: (Yes − No)	40

Causal claims? Concerns?

We knew: do better than draw

$$T_i \sim \text{Bern}(\pi)$$

We knew: do better than draw

$$T_i \sim \mathrm{Bern}(\pi)$$

- $\rightsquigarrow$ sample imbalance

We knew: do better than draw

$$T_i \sim \mathrm{Bern}(\pi)$$

- $\rightsquigarrow$ sample imbalance
- Might get no treateds!

We knew: do better than draw

$$T_i \sim \mathrm{Bern}(\pi)$$

- $\rightsquigarrow$ sample imbalance
- Might get no treateds!

We knew: do better than draw

$$T_i \sim \text{Bern}(\pi)$$

- $\rightsquigarrow$ sample imbalance
- Might get no treateds!

Seriously?

We knew: do better than draw

$$T_i \sim \mathrm{Bern}(\pi)$$

- $\rightsquigarrow$ sample imbalance
- Might get no treateds!

Seriously? Well, ...

We knew: do better than draw

$$T_i \sim \text{Bern}(\pi)$$

- $\rightsquigarrow$ sample imbalance
- Might get no treateds!

Seriously? Well, ...

```r
x <- sample(0:1, size = 10, replace = TRUE)
x
## [1] 1 1 1 1 1 1 1 1 1 1
```

Figure 1: Unlucky

We knew: do better than draw

$$T_i \sim \text{Bern}(\pi)$$

- $\rightsquigarrow$ sample imbalance
- Might get no treateds!
- CLT SE for diff in means:

We knew: do better than draw

$$T_i \sim \text{Bern}(\pi)$$

- $\rightsquigarrow$ sample imbalance
- Might get no treateds!
- CLT SE for diff in means:

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{N-1}\left[\frac{m\text{Var}(Y_i(0))}{N-m} + \frac{(N-m)\text{Var}(Y_i(1))}{m} + 2\text{Cov}(Y_i(0), Y_i(1))\right.}$$

- If $\text{Var}(Y_i(0)) = \text{Var}(Y_i(1))$, allocate units equally.

We knew: do better than draw

$$T_i \sim \text{Bern}(\pi)$$

- $\rightsquigarrow$ sample imbalance
- Might get no treateds!
- CLT SE for diff in means:

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{N-1}\left[\frac{m\text{Var}(Y_i(0))}{N-m} + \frac{(N-m)\text{Var}(Y_i(1))}{m} + 2\text{Cov}(Y_i(0), Y_i(1))\right.}$$

- If $\text{Var}(Y_i(0)) = \text{Var}(Y_i(1))$, allocate units equally. Say 5 treated, 4 control. What to do with 10th village?

We knew: do better than draw

$$T_i \sim \text{Bern}(\pi)$$

- $\rightsquigarrow$ sample imbalance
- Might get no treateds!
- CLT SE for diff in means:

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{N-1}\left[\frac{m\text{Var}(Y_i(0))}{N-m} + \frac{(N-m)\text{Var}(Y_i(1))}{m} + 2\text{Cov}(Y_i(0), Y_i(1))\right.}$$

- If $\text{Var}(Y_i(0)) = \text{Var}(Y_i(1))$, allocate units equally. Say 5 treated, 4 control. What to do with 10th village?
  - $10^{th} \rightarrow$ control: $SE: \sqrt{\frac{5}{5} + \frac{5}{5}} = \sqrt{2}$

We knew: do better than draw

$$T_i \sim \text{Bern}(\pi)$$

- $\rightsquigarrow$ sample imbalance
- Might get no treateds!
- CLT SE for diff in means:

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{N-1}\left[\frac{m\text{Var}(Y_i(0))}{N-m} + \frac{(N-m)\text{Var}(Y_i(1))}{m} + 2\text{Cov}(Y_i(0), Y_i(1))\right.}$$

- If $\text{Var}(Y_i(0)) = \text{Var}(Y_i(1))$, allocate units equally. Say 5 treated, 4 control. What to do with 10th village?
  - $10^{th} \rightarrow$ control: $SE : \sqrt{\frac{5}{5} + \frac{5}{5}} = \sqrt{2}$
  - $10^{th} \rightarrow$ treated: $SE : \sqrt{\frac{6}{4} + \frac{4}{6}} = \sqrt{2.66}$

We knew: do better than draw

$$T_i \sim \text{Bern}(\pi)$$

- $\rightsquigarrow$ sample imbalance
- Might get no treateds!
- CLT SE for diff in means:

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{N-1}\left[\frac{m\text{Var}(Y_i(0))}{N-m} + \frac{(N-m)\text{Var}(Y_i(1))}{m} + 2\text{Cov}(Y_i(0), Y_i(1))\right.}$$

- If $\text{Var}(Y_i(0)) = \text{Var}(Y_i(1))$, allocate units equally. Say 5 treated, 4 control. What to do with 10th village?
  - $10^{th} \to$ control: $SE : \sqrt{\frac{5}{5} + \frac{5}{5}} = \sqrt{2}$
  - $10^{th} \to$ treated: $SE : \sqrt{\frac{6}{4} + \frac{4}{6}} = \sqrt{2.66}$
- If $\text{Var}(Y_i(0)) \neq \text{Var}(Y_i(1))$, allocate $\to$ higher-Variance condition

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{N-1}\left[\frac{m\text{Var}(Y_i(0))}{N-m} + \frac{(N-m)\text{Var}(Y_i(1))}{m} + 2\text{Cov}(Y_i(0), Y_i(1))\right.}$$

- If $\text{Cov}(Y_i(0), Y_i(1)) > 0$, larger SE, less precision
- If $\text{Cov}(Y_i(0), Y_i(1)) < 0$, smaller SE, *more* precision

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{N-1}\left[\frac{m\text{Var}(Y_i(0))}{N-m} + \frac{(N-m)\text{Var}(Y_i(1))}{m} + 2\text{Cov}(Y_i(0), Y_i(1))\right.}$$

- If $\text{Cov}(Y_i(0), Y_i(1)) > 0$, larger SE, less precision
- If $\text{Cov}(Y_i(0), Y_i(1)) < 0$, smaller SE, *more* precision

Demonstration:

Unit	Y_0	Y_1 (+ cov)	Y_1 (− cov)
1	0	0	10
2	5	5	5
3	10	10	0
Means	5	5	5

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{N-1}\left[\frac{m\text{Var}(Y_i(0))}{N-m} + \frac{(N-m)\text{Var}(Y_i(1))}{m} + 2\text{Cov}(Y_i(0), Y_i(1))\right.}$$

- If $\text{Cov}(Y_i(0), Y_i(1)) > 0$, larger SE, less precision
- If $\text{Cov}(Y_i(0), Y_i(1)) < 0$, smaller SE, *more* precision

Demonstration:

Unit	Y_0	Y_1 (+ cov)	Y_1 (− cov)
1	0	0	10
2	5	5	5
3	10	10	0
Means	5	5	5

Suppose assign 1 to Tr, 2 to Co.

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{N-1}\left[\frac{m\text{Var}(Y_i(0))}{N-m} + \frac{(N-m)\text{Var}(Y_i(1))}{m} + 2\text{Cov}(Y_i(0), Y_i(1))\right.}$$

- If $\text{Cov}(Y_i(0), Y_i(1)) > 0$, larger SE, less precision
- If $\text{Cov}(Y_i(0), Y_i(1)) < 0$, smaller SE, *more* precision

Demonstration:

Unit	Y_0	Y_1 (+ cov)	Y_1 (− cov)
1	0	0	10
2	5	5	5
3	10	10	0
Means	5	5	5

Suppose assign 1 to Tr, 2 to Co.

$\widehat{ATE}_{+\text{cov}} = -7.5, 0, 7.5$

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{N-1}\left[\frac{m\text{Var}(Y_i(0))}{N-m} + \frac{(N-m)\text{Var}(Y_i(1))}{m} + 2\text{Cov}(Y_i(0), Y_i(1))\right.}$$

- If $\text{Cov}(Y_i(0), Y_i(1)) > 0$, larger SE, less precision
- If $\text{Cov}(Y_i(0), Y_i(1)) < 0$, smaller SE, *more* precision

Demonstration:

Unit	Y_0	Y_1 (+ cov)	Y_1 (− cov)
1	0	0	10
2	5	5	5
3	10	10	0
Means	5	5	5

Suppose assign 1 to Tr, 2 to Co.

$\widehat{ATE}_{+\text{cov}} = -7.5, 0, 7.5$

$\widehat{ATE}_{-\text{cov}} = 2.5, 0, -2.5$

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{N-1}\left[\frac{m\text{Var}(Y_i(0))}{N-m} + \frac{(N-m)\text{Var}(Y_i(1))}{m} + 2\text{Cov}(Y_i(0), Y_i(1))\right.}$$

- If $\text{Cov}(Y_i(0), Y_i(1)) > 0$, larger SE, less precision
- If $\text{Cov}(Y_i(0), Y_i(1)) < 0$, smaller SE, *more* precision

Demonstration:

Unit	Y_0	Y_1 (+ cov)	Y_1 (− cov)
1	0	0	10
2	5	5	5
3	10	10	0
Means	5	5	5

Suppose assign 1 to Tr, 2 to Co.

$\widehat{ATE}_{+\text{cov}} = -7.5, 0, 7.5$

$\widehat{ATE}_{-\text{cov}} = 2.5, 0, -2.5$ (less variance!)

Example: Canvassing and Enrollment

But, is obs difference causal?

Example: Canvassing and Enrollment

But, is obs difference causal?

What do we really want to know?

Example: Canvassing and Enrollment

But, is obs difference causal?

What do we really want to know?

Does canvassing actually *change* enrollment in precinct?
(Or, just Party $\rightarrow$ Enrollment?)

Example: Canvassing and Enrollment

But, is obs difference causal?

What do we really want to know?

Does canvassing actually *change* enrollment in precinct?
(Or, just Party $\rightarrow$ Enrollment?)

What *would have* happened to “No” precincts if “Yes”?

Example: Canvassing and Enrollment

But, is obs difference causal?

What do we really want to know?

Does canvassing actually *change* enrollment in precinct?
(Or, just Party $\rightarrow$ Enrollment?)

What *would have* happened to "No" precincts if "Yes"?

What would have happened under *other* conditions?

Example: Canvassing and Enrollment

Suppose we can know both *potential outcomes* ...

Precinct	Party	Canvass?	Enroll % if No Canvass	Enroll % if Canvass
1	Dem	_	20	60
2	Dem	_	30	70
3	Rep	_	20	30
4	Rep	_	30	40
		Means:	25	50

Example: Canvassing and Enrollment

Suppose we can know both *potential outcomes* ...

Precinct	Party	Canvass?	Enroll % if No Canvass	Enroll % if Canvass
1	Dem	_	20	60
2	Dem	_	30	70
3	Rep	_	20	30
4	Rep	_	30	40
		Means:	25	50

$\text{ATE} = 50 - 25 = 25$

Example: Canvassing and Enrollment

Suppose we can know both *potential outcomes* ...

Precinct	Party	Canvass?	Enroll % if No Canvass	Enroll % if Canvass
1	Dem	_	20	60
2	Dem	_	30	70
3	Rep	_	20	30
4	Rep	_	30	40
		Means:	25	50

$\text{ATE} = 50 - 25 = 25$

(True or an estimate?)

Example: Canvassing and Enrollment

Another way to think about same information:

Precinct	Party	Canvass?	Enroll % if No Canvass	Enroll % if Canvass	True Precinct Effect
1	Dem	_	20	60	40
2	Dem	_	30	70	40
3	Rep	_	20	30	10
4	Rep	_	30	40	10
		Means:	25	50	25

$$\text{ATE} = (40 + 40 + 10 + 10)/4 = 25$$

The "Fundamental Problem of Causal Inference' '

We can't observe both "Canvassed" and "Not Canvassed" for a precinct.

We can't observe both *potential outcomes* (*counterfactuals*).

So, how can we get a good causal estimate?

Example: Canvassing and Enrollment

Suppose we observe ...

Precinct	Party	Canvass?	Enroll % if No Canvass	Enroll % if Canvass
1	Dem	Yes		60
2	Dem	Yes		70
3	Rep	No	20	
4	Rep	No	30	
		Means:	25	65

Estimated ATE $= 65 - 25 = 40$

Example: Canvassing and Enrollment

Suppose we observe ...

Precinct	Party	Canvass?	Enroll % if No Canvass	Enroll % if Canvass
1	Dem	Yes		60
2	Dem	Yes		70
3	Rep	No	20	
4	Rep	No	30	
		Means:	25	65

Estimated ATE $= 65 - 25 = 40$ (too big)

Example: Canvassing and Enrollment

Or, we could have observed ...

Example: Canvassing and Enrollment

Or, we could have observed ...

Precinct	Party	Canvass?	Enroll % if No Canvass	Enroll % if Canvass
1	Dem	Yes		60
2	Dem	No	30	
3	Rep	Yes		30
4	Rep	No	30	
		Means:	30	45

Estimated ATE $= 45 - 30 = 15$

Example: Canvassing and Enrollment

Or, we could have observed ...

Precinct	Party	Canvass?	Enroll % if No Canvass	Enroll % if Canvass
1	Dem	Yes		60
2	Dem	No	30	
3	Rep	Yes		30
4	Rep	No	30	
		Means:	30	45

Estimated ATE $= 45 - 30 = 15$ (too small; closer)

Example: Canvassing and Enrollment

In our random allocation, possible data were

Assignments	Est ATE
YYNN	40

Example: Canvassing and Enrollment

In our random allocation, possible data were

Assignments	Est ATE
YYNN	40
NYNY	35
YNNY	25
NYYN	25
YNYN	15
NNYY	10

Example: Canvassing and Enrollment

In our random allocation, possible data were

Assignments	Est ATE
YYNN	40
NYNY	35
YNNY	25
NYYN	25
YNYN	15
NNYY	10

- Some closer to truth

Example: Canvassing and Enrollment

In our random allocation, possible data were

Assignments	Est ATE
YYNN	40
NYNY	35
YNNY	25
NYYN	25
YNYN	15
NNYY	10

- Some closer to truth
- $E(\hat{\bar{\tau}}) = \frac{1}{6} \cdot 40 + \frac{1}{6} \cdot 35 + \frac{1}{3} \cdot 25 + \frac{1}{6} \cdot 15 + \frac{1}{6} \cdot 10 = 25 = \bar{\tau}$

Example: Canvassing and Enrollment

In our random allocation, possible data were

Assignments	Est ATE
YYNN	40
NYNY	35
YNNY	25
NYYN	25
YNYN	15
NNYY	10

- Some closer to truth
- $E(\hat{\bar{\tau}}) = \frac{1}{6}\cdot 40 + \frac{1}{6}\cdot 35 + \frac{1}{3}\cdot 25 + \frac{1}{6}\cdot 15 + \frac{1}{6}\cdot 10 = 25 = \bar{\tau}$

Example: Canvassing and Enrollment

In our random allocation, possible data were

Assignments	Est ATE
YYNN	40
NYNY	35
YNNY	25
NYYN	25
YNYN	15
NNYY	10

- Some closer to truth
- $E(\hat{\bar{\tau}}) = \frac{1}{6}\cdot 40 + \frac{1}{6}\cdot 35 + \frac{1}{3}\cdot 25 + \frac{1}{6}\cdot 15 + \frac{1}{6}\cdot 10 = 25 = \bar{\tau}$ (unbiased)

A Problem

In practice, we don't know all potential outcomes.

A Problem

In practice, we don't know all potential outcomes.

Two assignments (YYNN and NNYY) leave treatment perfectly confounded with party.

A Problem

In practice, we don't know all potential outcomes.

Two assignments (YYNN and NNYY) leave treatment perfectly confounded with party.

We can never see all of Y_1, Y_0. But we can see all of X!

A Problem

In practice, we don't know all potential outcomes.

Two assignments (YYNN and NNYY) leave treatment perfectly confounded with party.

We can never see all of Y_1, Y_0. But we can see all of X!

Let's ensure X does not predict T.

A Solution

Blocking:

Creating pre-treatmnt groups that look same on *predictors*.

A Solution

Blocking:

Creating pre-treatmnt groups that look same on *predictors*.

Precinct	Party	Canvass?	Enroll % if No Canvass	Enroll % if Canvass
1	Dem			
2	Dem			
3	Rep			
4	Rep			

A Solution

Blocking:

Creating pre-treatmnt groups that look same on *predictors*.

Precinct	Party	Canvass?	Enroll % if No Canvass	Enroll % if Canvass
1	Dem			
2	Dem			
3	Rep			
4	Rep			

(Then, randomize within groups.)

A Solution

Blocking:

Creating pre-treatmnt groups that look same on *predictors*.

Precinct	Party	Canvass?	Enroll % if No Canvass	Enroll % if Canvass
1	Dem	Y		60
2	Dem	N	30	
3	Rep	N	20	
4	Rep	Y		40

(Then, randomize within groups.)

Example: Canvassing and Enrollment

Blocking restricts possible data to

Assignments	Est TE
YYNN	40
NYNY	**35**
YNNY	**25**
NYYN	**25**
YNYN	**15**
NNYY	10

Example: Canvassing and Enrollment

Blocking restricts possible data to

Assignments	Est TE
YYNN	40
NYNY	**35**
YNNY	**25**
NYYN	**25**
YNYN	**15**
NNYY	10

Estimates have less variance, are closer to true ATE.

Example: Canvassing and Enrollment

Blocking restricts possible data to

Assignments	Est TE
YYNN	40
NYNY	**35**
YNNY	**25**
NYYN	**25**
YNYN	**15**
NNYY	10

Estimates have less variance, are closer to true ATE.

Under random allocation, 5 possible estimates: 40, 15, 25, 35, 10. Blocking restricts to 3 best: 15, 25, 35.

Example: Canvassing and Enrollment

Blocking restricts possible data to

Assignments	Est TE
YYNN	40
NYNY	**35**
YNNY	**25**
NYYN	**25**
YNYN	**15**
NNYY	10

Estimates have less variance, are closer to true ATE.

Under random allocation, 5 possible estimates: 40, 15, 25, 35, 10. Blocking restricts to 3 best: 15, 25, 35.

Why?

Why do we block?

- Covariate **balance**

Why do we block?

- Covariate **balance**
- Estimate **closer to truth**

Why do we block?

- Covariate **balance**
- Estimate **closer to truth**
- Increased **efficiency**

Why do we block?

- Covariate **balance**
- Estimate **closer to truth**
- Increased **efficiency**
- Triply-robust estimates: block, randomize, adjust

Why do we block?

- Covariate **balance**
- Estimate **closer to truth**
- Increased **efficiency**
- Triply-robust estimates: block, randomize, adjust
- Block-level effects

  $\rightsquigarrow$ different actors interested in different effects

Why do we block?

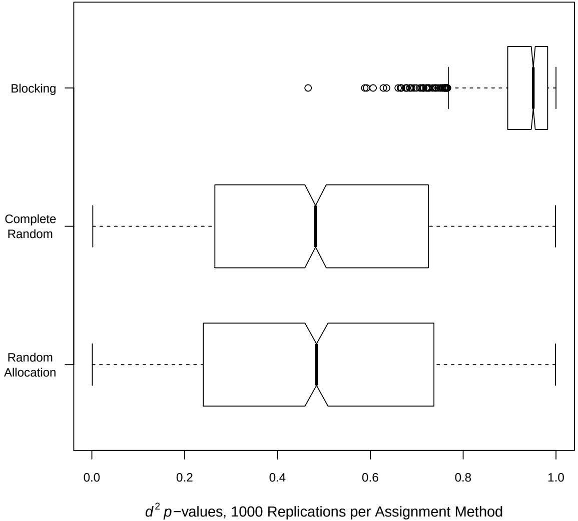
- Covariate **balance**
- Estimate **closer to truth**
- Increased **efficiency**
- Triply-robust estimates: block, randomize, adjust
- Block-level effects
  $\rightsquigarrow$ different actors interested in different effects
- Guidelines for limited/uncertain resources

Why Block: Balance

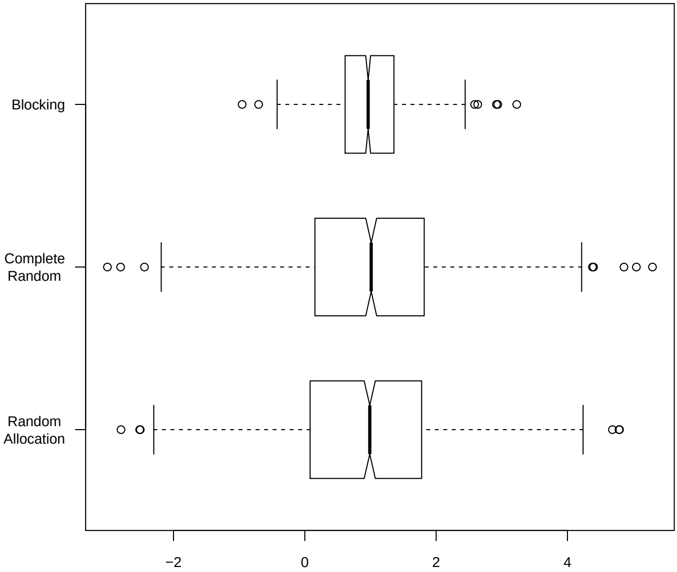
Simulation study: 100 units, $X_1 \sim N(0,1)$, $X_2 \sim \text{Unif}(0,1)$, $X_3 \sim \chi^2_2$; 1000 such experiments. Assg treatmnt in 3 ways.

Why Block: Balance

Simulation study: 100 units, $X_1 \sim N(0,1)$, $X_2 \sim \text{Unif}(0,1)$, $X_3 \sim \chi^2_2$; 1000 such experiments. Assg treatmnt in 3 ways.



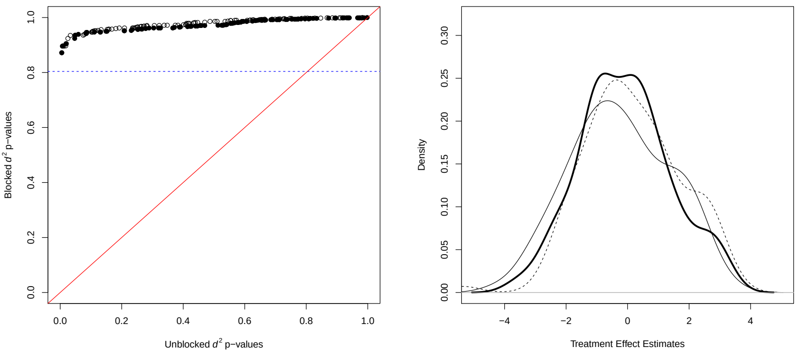
Why Block: Efficiency



Blocking in Applications: Balance and Efficiency

Moore (2012): Perry Preschool Experiment

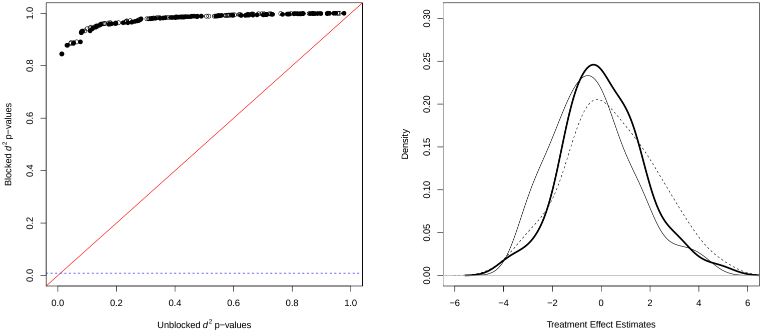
Left: QQ plot of balance (100 blocked vs. unblocked)
Right: Est TE under sharp null (100 blocked vs. unblocked)



(SES, sex, IQ)

Balance in Applications: Balance and Efficiency

Considering more variables ...



(+ siblings, AFDC, mom empl, educ, father, ...)

How: `blockTools`

Blocking with `blockTools`

Start with some sample data:

```
library(blockTools)
data(x100)

x100 |> head()
```

```
    id id2  b1  b2 g  ig
1 1001 101 156 795 b 729
2 1002 102 813 469 a 627
3 1003 103 950 978 a 959
4 1004 104 991 781 a 661
5 1005 105 613 759 a 819
6 1006 106 654 838 b 643
```

(Moore 2012; Moore and Schnakenberg 2023)

Blocking with blockTools

```r
b <- block(x100, id.vars = "id",
           block.vars = c("b1", "b2"))
```

```r
bl <- b$blocks$`1`

bl |> head()
```

```
  Unit 1 Unit 2   Distance
1   1043   1040 0.01240000
2   1100   1020 0.02259275
3   1065   1027 0.02912651
4   1085   1081 0.03498815
5   1088   1061 0.04789253
6   1064   1014 0.07985116
```

Blocking with blockTools

Why all this?

```
bl <- b$blocks$`1`
```

We are extracting just the blocked pairs themselves.

- Why `b$blocks`? Since `b` has 3 components:

```
names(b)
```

```
[1] "blocks"    "level.two" "call"
```

- Why `blocks$1`'? Since this is (default-named) first (and only) "group":

```
names(b$blocks)
```

```
[1] "1"
```

Blocking with blockTools

What else could we do?

```r
b_3groups_3conditions <- block(
  x100,
  groups = "g",                    # (Factor variable in data)
  n.tr = 3,
  id.vars = "id",
  block.vars = c("b1", "b2"),
  distance = "mve"
)
```

```r
b_3groups_3conditions$blocks
```

```
$a
   Unit 1 Unit 2 Unit 3 Max Distance
1    1076   1039   1056    0.2755751
2    1084   1058   1017    0.3991725
3    1073   1029   1098    0.4242714
4    1065   1061   1002    0.4469259
```

Blocking with blockTools

Some rows from each "group":

```
rows_a <- b_3groups_3conditions$blocks$a |> slice(1:2) |> mutate
rows_b <- b_3groups_3conditions$blocks$b |> slice(1:2) |> mutate
rows_c <- b_3groups_3conditions$blocks$c |> slice(1:2) |> mutate
bind_rows(rows_a, rows_b, rows_c)
```

```
  Unit 1 Unit 2 Unit 3 Max Distance group
1   1076   1039   1056    0.2755751     a
2   1084   1058   1017    0.3991725     a
3   1043   1040   1009    0.1721362     b
4   1031   1025   1048    0.2110529     b
5   1095   1092   1049    0.3439317     c
6   1088   1027   1066    0.3552557     c
```

Blocking with `blockTools`

Other arguments to `block()`

- `vcov.data`
- `groups`: for exact-blocks
- `n.tr`
- `id.vars`
- `block.vars`
- `algorithm`: `optGreedy`, `optimal`, `naiveGreedy`, `randGreedy`, `sortGreedy`
- `distance`: `mahalanobis`, `mcd`, `mve`, `euclidean`, $k \times k$ matx
- `weight`
- `level.two`: block states by most similar cities
- `valid.var`, `valid.range`: Goldilocks
- `seed.dist`: (for `mcd` and `mve`)

Assign

```r
a <- assignment(b, seed = 71573706)
a
```

```
Assignments:

    Treatment 1  Treatment 2  Distance
1   1040         1043         0.01240000
2   1100         1020         0.02259275
3   1065         1027         0.02912651
4   1081         1085         0.03498815
5   1088         1061         0.04789253
6   1014         1064         0.07985116
7   1032         1070         0.08279625
8   1097         1098         0.08882421
9   1038         1018         0.09316331
10  1031         1048         0.10391953
11  1084         1058         0.10835825
```

Get Assignments

```
a |> extract_conditions(x100, id.var = "id")
```

```
  [1] 2 1 2 2 2 2 1 2 2 1 1 1 1 1 1 1 2 2 1 2 2 2 1 2 1 1 2
 [38] 1 1 1 2 2 2 2 2 2 1 2 1 2 2 2 2 1 1 2 1 2 1 1 2 2 1 2
 [75] 1 2 2 2 1 2 1 1 1 1 2 1 2 1 1 2 1 2 1 1 2 2 1 2 1 1
```

```
x100 |> mutate(
  condition = extract_conditions(a, x100, id.var = "id"))
```

```
      id id2  b1   b2 g  ig condition
1   1001 101 156  795 b 729         2
2   1002 102 813  469 a 627         1
3   1003 103 950  978 a 959         2
4   1004 104 991  781 a 661         2
5   1005 105 613  759 a 819         2
6   1006 106 654  838 b 643         2
7   1007 107 640  645 c  12         1
8   1008 108 681  404 a 221         2
```

Assign 3 Conditions, within Groups

```
a3 <- assignment(b_3groups_3conditions, seed = 979677744)
a3
```

```
Assignments:

Group: a
    Treatment 1  Treatment 2  Treatment 3  Max Distance
1   1056         1076         1039         0.2755751
2   1017         1058         1084         0.3991725
3   1029         1073         1098         0.4242714
4   1061         1002         1065         0.4469259
5   1059         1081         1046         0.4973664
6   1060         1067         1004         0.6482068
7   1089         1024         1032         0.6880429
8   1052         1054         1030         0.8059432
9   1093         1026         1068         0.8745853
10  1036         1008         1091         1.4250668
```

How: `randomizr`

Blocking with `randomizr::block_ra()`

```r
library(randomizr)

tr <- block_ra(x100$g)

# Better:

x100 |> mutate(tr = block_ra(x100$g))
```

```
       id id2  b1   b2 g  ig tr
1    1001 101 156  795 b 729  1
2    1002 102 813  469 a 627  1
3    1003 103 950  978 a 959  1
4    1004 104 991  781 a 661  1
5    1005 105 613  759 a 819  0
6    1006 106 654  838 b 643  1
7    1007 107 640  645 c  12  1
8    1008 108 681  404 a 221  0
9    1009 109 530  823 b 321  1
```

Then what?

blockTools: diagnose, get block IDs, check balance

Diagnose:

```r
diagnose(a, data = x100, id.vars = "id",
         suspect.var = "b1", suspect.range = c(0, 5))
```

blockTools: diagnose, get block IDs, check balance

Diagnose:

```
diagnose(a, data = x100, id.vars = "id",
         suspect.var = "b1", suspect.range = c(0, 5))
```

Get block IDs

```
createBlockIDs(a, data = x100, id.var = "id")
```

`blockTools`: diagnose, get block IDs, check balance

Diagnose:

```
diagnose(a, data = x100, id.vars = "id",
         suspect.var = "b1", suspect.range = c(0, 5))
```

Get block IDs

```
createBlockIDs(a, data = x100, id.var = "id")
```

Get balance:

```
assg2xBalance(a, x100, id.var = "id",
              bal.vars = c("b1", "b2"))
```

Analysis

- Generally, use Lin or Blocked Diff-in-Means

Analysis

- Generally, use Lin or Blocked Diff-in-Means
- LSDV (block indicators) weights w/in block TE's by

$$p_j(1 - p_j)n_j$$

Analysis

- Generally, use Lin or Blocked Diff-in-Means
- LSDV (block indicators) weights w/in block TE's by

$$p_j(1-p_j)n_j$$

Analysis

- Generally, use Lin or Blocked Diff-in-Means
- LSDV (block indicators) weights w/in block TE's by

$$p_j(1 - p_j)n_j$$

where

- p_j = share of block j treated
- n_j = size of block j
- (I.e., $p_j(1 - p_j) = \text{var}(TE)$ in block j)

Analysis

- Generally, use Lin or Blocked Diff-in-Means
- LSDV (block indicators) weights w/in block TE's by

$$p_j(1-p_j)n_j$$

where

- p_j = share of block j treated
- n_j = size of block j
- (I.e., $p_j(1-p_j) = \text{var}(TE)$ in block j)

Safer when

- p_j constant across blocks j
- n_j constant across blocks j

Analysis

```r
library(estimatr)

df <- x100 |> mutate(tr = block_ra(x100$g))

lm_lin(b1 ~ tr, covariates = ~ g, data = df)
```

```
              Estimate Std. Error     t value      Pr(>|t|)
(Intercept)  492.800000   39.08723 12.60769882 6.405020e-22
tr           -50.760000   51.70890 -0.98164922 3.287925e-01
gb_c          -4.673611  100.16089 -0.04666104 9.628824e-01
gc_c          -4.111111   92.84835 -0.04427770 9.647770e-01
tr:gb_c     -190.763889  129.80730 -1.46959290 1.450111e-01
tr:gc_c     -112.826389  128.65776 -0.87694974 3.827497e-01
            DF
(Intercept) 94
tr          94
gb_c        94
gc_c        94
```

Analysis

Can I just ignore blocks and pool?

Analysis

Can I just ignore blocks and pool?

- If p_j varies, no

Thanks!

rtm@american.edu
www.ryantmoore.org

References I

Coppock, Alexander. 2023. *randomizr: Easy-to-Use Tools for Common Forms of Random Assignment and Sampling.* https://CRAN.R-project.org/package=randomizr.

Moore, Ryan T. 2012. "Multivariate Continuous Blocking to Improve Political Science Experiments." *Political Analysis* 20 (4): 460–79. https://doi.org/10.1093/pan/mps025.

Moore, Ryan T., and Keith Schnakenberg. 2023. *blockTools: Blocking, Assignment, and Diagnosing Interference in Randomized Experiments.* http://www.ryantmoore.org/html/software.blockTools.html.