Day 08: Heterogeneous treatment effects

Erin Rossiter

February 8, 2022

Plan

Moderators

- Motivate treatment effect heterogeneity
- Preliminary tests
- Structured tests
- Discuss caution needed
- A note on factorial experiments
- Labs
  - reinforce HTE tests
  - a note on power
- Discussion of your designs

Mediation – pushed back

Plan

Moderators

- Motivate treatment effect heterogeneity
- Preliminary tests
- Structured tests
- Discuss caution needed
- A note on factorial experiments
- Labs
  - reinforce HTE tests
  - a note on power
- Discussion of your designs

Mediation – pushed back

Plan

Moderators

- Motivate treatment effect heterogeneity
- Preliminary tests
- Structured tests
- Discuss caution needed
- A note on factorial experiments
- Labs
  - reinforce HTE tests
  - a note on power
- Discussion of your designs

Mediation – pushed back

Plan

Moderators

- Motivate treatment effect heterogeneity
- Preliminary tests
- Structured tests
- Discuss caution needed
- A note on factorial experiments
- Labs
  - reinforce HTE tests
  - a note on power
- Discussion of your designs

Mediation – pushed back

Plan

Moderators

- Motivate treatment effect heterogeneity
- Preliminary tests
- Structured tests
- Discuss caution needed
- A note on factorial experiments
- Labs
  - reinforce HTE tests
  - a note on power
- Discussion of your designs

Mediation – pushed back

Plan

Moderators

- Motivate treatment effect heterogeneity
- Preliminary tests
- Structured tests
- Discuss caution needed
- A note on factorial experiments
- Labs
  - reinforce HTE tests
  - a note on power
- Discussion of your designs

Mediation – pushed back

Plan

Moderators

- Motivate treatment effect heterogeneity
- Preliminary tests
- Structured tests
- Discuss caution needed
- A note on factorial experiments
- Labs
  - reinforce HTE tests
  - a note on power
- Discussion of your designs

Mediation – pushed back

Plan

Moderators

- Motivate treatment effect heterogeneity
- Preliminary tests
- Structured tests
- Discuss caution needed
- A note on factorial experiments
- Labs
  - » reinforce HTE tests
  - » a note on power
- Discussion of your designs

Mediation – pushed back

Plan

Moderators

- Motivate treatment effect heterogeneity
- Preliminary tests
- Structured tests
- Discuss caution needed
- A note on factorial experiments
- Labs
  - » reinforce HTE tests
  - » a note on power
- Discussion of your designs

Mediation – pushed back

Plan

Moderators

- Motivate treatment effect heterogeneity
- Preliminary tests
- Structured tests
- Discuss caution needed
- A note on factorial experiments
- Labs
  - » reinforce HTE tests
  - » a note on power
- Discussion of your designs

Mediation – pushed back

Plan

Moderators

- Motivate treatment effect heterogeneity
- Preliminary tests
- Structured tests
- Discuss caution needed
- A note on factorial experiments
- Labs
  - » reinforce HTE tests
  - » a note on power
- Discussion of your designs

Mediation – pushed back

Heterogeneous Treatment Effects (HTEs)

Heterogeneous Treatment Effects (HTEs)

- We know not every unit responds to treatment in the same way (board)
  - So far, we've talked about *average* treatment effects
  - Looking at the average effects ignores variability
    - (A feature not a bug of looking at the average)
- We might, however, be interested in this variability
  - For whom are there big effects?
  - For whom are there small effects?
  - For whom does treatment generate beneficial or adverse effects?
  - Examples?
- Before asking these specific questions, we first investigate if there's any evidence of heterogeneity or not.
  - Is $Var(\tau_i) > 0$?

Heterogeneous Treatment Effects (HTEs)

- We know not every unit responds to treatment in the same way (board)
  - So far, we've talked about *average* treatment effects
  - Looking at the average effects ignores variability
    - (A feature not a bug of looking at the average)
- We might, however, be interested in this variability
  - For whom are there big effects?
  - For whom are there small effects?
  - For whom does treatment generate beneficial or adverse effects?
  - Examples?
- Before asking these specific questions, we first investigate if there's any evidence of heterogeneity or not.
  - Is $Var(\tau_i) > 0$?

Heterogeneous Treatment Effects (HTEs)

- We know not every unit responds to treatment in the same way (board)
  - So far, we've talked about *average* treatment effects
  - Looking at the average effects ignores variability
    - (A feature not a bug of looking at the average)
- We might, however, be interested in this variability
  - For whom are there big effects?
  - For whom are there small effects?
  - For whom does treatment generate beneficial or adverse effects?
  - Examples?
- Before asking these specific questions, we first investigate if there's any evidence of heterogeneity or not.
  - Is $Var(\tau_i) > 0$?

Heterogeneous Treatment Effects (HTEs)

- We know not every unit responds to treatment in the same way (board)
  - So far, we've talked about *average* treatment effects
  - Looking at the average effects ignores variability
    - (A feature not a bug of looking at the average)
- We might, however, be interested in this variability
  - For whom are there big effects?
  - For whom are there small effects?
  - For whom does treatment generate beneficial or adverse effects?
  - Examples?
- Before asking these specific questions, we first investigate if there's any evidence of heterogeneity or not.
  - Is $Var(\tau_i) > 0$?

Heterogeneous Treatment Effects (HTEs)

- We know not every unit responds to treatment in the same way (board)
  - So far, we've talked about *average* treatment effects
  - Looking at the average effects ignores variability
    - (A feature not a bug of looking at the average)
- We might, however, be interested in this variability
  - For whom are there big effects?
  - For whom are there small effects?
  - For whom does treatment generate beneficial or adverse effects?
  - Examples?
- Before asking these specific questions, we first investigate if there's any evidence of heterogeneity or not.
  - Is $Var(\tau_i) > 0$?

Heterogeneous Treatment Effects (HTEs)

- We know not every unit responds to treatment in the same way (board)
  - So far, we've talked about *average* treatment effects
  - Looking at the average effects ignores variability
    - (A feature not a bug of looking at the average)
- We might, however, be interested in this variability
  - For whom are there big effects?
  - For whom are there small effects?
  - For whom does treatment generate beneficial or adverse effects?
  - Examples?
- Before asking these specific questions, we first investigate if there's any evidence of heterogeneity or not.
  - Is $Var(\tau_i) > 0$?

Heterogeneous Treatment Effects (HTEs)

- We know not every unit responds to treatment in the same way (board)
  - So far, we've talked about *average* treatment effects
  - Looking at the average effects ignores variability
    - (A feature not a bug of looking at the average)
- We might, however, be interested in this variability
  - For whom are there big effects?
  - For whom are there small effects?
  - For whom does treatment generate beneficial or adverse effects?
  - Examples?
- Before asking these specific questions, we first investigate if there's any evidence of heterogeneity or not.
  - Is $Var(\tau_i) > 0$?

Heterogeneous Treatment Effects (HTEs)

- We know not every unit responds to treatment in the same way (board)
  - So far, we've talked about *average* treatment effects
  - Looking at the average effects ignores variability
    - (A feature not a bug of looking at the average)
- We might, however, be interested in this variability
  - For whom are there big effects?
  - For whom are there small effects?
  - For whom does treatment generate beneficial or adverse effects?
  - Examples?
- Before asking these specific questions, we first investigate if there's any evidence of heterogeneity or not.
  - Is $Var(\tau_i) > 0$?

Heterogeneous Treatment Effects (HTEs)

- We know not every unit responds to treatment in the same way (board)
  - So far, we've talked about *average* treatment effects
  - Looking at the average effects ignores variability
    - (A feature not a bug of looking at the average)
- We might, however, be interested in this variability
  - For whom are there big effects?
  - For whom are there small effects?
  - For whom does treatment generate beneficial or adverse effects?
  - Examples?
- Before asking these specific questions, we first investigate if there's any evidence of heterogeneity or not.
  - Is $Var(\tau_i) > 0$?

Heterogeneous Treatment Effects (HTEs)

- We know not every unit responds to treatment in the same way (board)
  - So far, we've talked about *average* treatment effects
  - Looking at the average effects ignores variability
    - (A feature not a bug of looking at the average)
- We might, however, be interested in this variability
  - For whom are there big effects?
  - For whom are there small effects?
  - For whom does treatment generate beneficial or adverse effects?
  - Examples?
- Before asking these specific questions, we first investigate if there's any evidence of heterogeneity or not.
  - Is $Var(\tau_i) > 0$?

Heterogeneous Treatment Effects (HTEs)

- We know not every unit responds to treatment in the same way (board)
  - So far, we've talked about *average* treatment effects
  - Looking at the average effects ignores variability
    - (A feature not a bug of looking at the average)
- We might, however, be interested in this variability
  - For whom are there big effects?
  - For whom are there small effects?
  - For whom does treatment generate beneficial or adverse effects?
  - Examples?
- Before asking these specific questions, we first investigate if there's any evidence of heterogeneity or not.
  - Is $Var(\tau_i) > 0$?

Part 1 – is there heterogeneity?

Is there heterogeneity?

- Is $Var(\tau_i) > 0$?
- Imagine a test of null $Var(\tau_i) = 0$
  - In other words... $\tau_i = \tau \ \ \forall i$
  - Reject null $\rightarrow$ evidence of heterogeneity
- But, not possible to estimate $Var(\tau_i)$! Why?
  - $Var(\tau_i) = Var(Y_i(1) - Y_i(0)) =$
    $Var(Y_i(1)) + Var(Y_i(0)) - 2Cov(Y_i(1), Y_i(0))$

Is there heterogeneity?

- Is $Var(\tau_i) > 0$?
- Imagine a test of null $Var(\tau_i) = 0$
  - In other words... $\tau_i = \tau \ \ \forall i$
  - Reject null $\rightarrow$ evidence of heterogeneity
- But, not possible to estimate $Var(\tau_i)$! Why?
  - $Var(\tau_i) = Var(Y_i(1) - Y_i(0)) =$
    $Var(Y_i(1)) + Var(Y_i(0)) - 2Cov(Y_i(1), Y_i(0))$

Is there heterogeneity?

- Is $Var(\tau_i) > 0$?
- Imagine a test of null $Var(\tau_i) = 0$
  - In other words... $\tau_i = \tau \ \ \forall i$
  - Reject null $\rightarrow$ evidence of heterogeneity
- But, not possible to estimate $Var(\tau_i)$! Why?
  - $Var(\tau_i) = Var(Y_i(1) - Y_i(0)) =$
    $Var(Y_i(1)) + Var(Y_i(0)) - 2Cov(Y_i(1), Y_i(0))$

Is there heterogeneity?

- Is $Var(\tau_i) > 0$?
- Imagine a test of null $Var(\tau_i) = 0$
  - In other words... $\tau_i = \tau \ \ \forall i$
  - Reject null $\rightarrow$ evidence of heterogeneity
- But, not possible to estimate $Var(\tau_i)$! Why?
  - $Var(\tau_i) = Var(Y_i(1) - Y_i(0)) =$
    $Var(Y_i(1)) + Var(Y_i(0)) - 2Cov(Y_i(1), Y_i(0))$

Is there heterogeneity?

- Is $Var(\tau_i) > 0$?
- Imagine a test of null $Var(\tau_i) = 0$
  - In other words... $\tau_i = \tau \ \ \forall i$
  - Reject null $\rightarrow$ evidence of heterogeneity
- But, not possible to estimate $Var(\tau_i)$! Why?
  - $Var(\tau_i) = Var(Y_i(1) - Y_i(0)) =$
    $Var(Y_i(1)) + Var(Y_i(0)) - 2Cov(Y_i(1), Y_i(0))$

Is there heterogeneity?

- Is $Var(\tau_i) > 0$?
- Imagine a test of null $Var(\tau_i) = 0$
  - In other words... $\tau_i = \tau \ \ \forall i$
  - Reject null $\rightarrow$ evidence of heterogeneity
- But, not possible to estimate $Var(\tau_i)$! Why?
  - $Var(\tau_i) = Var(Y_i(1) - Y_i(0)) =$
    $Var(Y_i(1)) + Var(Y_i(0)) - 2Cov(Y_i(1), Y_i(0))$

Preliminary investigations

To answer is $Var(\tau_i) > 0$? we want to say something about null of $Var(\tau_i) = 0$.

1. Estimating bounds of $Var(\tau_i)$
2. Testing whether $Var(\tau_i) = 0$

1. Bounding $Var(\tau_i)$ (GG pg 292-293)

```r
Y <- c(1,2,3,4,5,6)
Z <- c(0,0,0,1,1,1)

# Pair in ascending order
sort(Y[Z==0])

## [1] 1 2 3

sort(Y[Z==1])

## [1] 4 5 6

# Lower bound -- ests of tau_i when Cov is as small as possible
var(sort(Y[Z==1]) - sort(Y[Z==0]))

## [1] 0

# Upper bound -- ests of tau_i when Cov is as large as possible
var(sort(Y[Z==1], decreasing = T) - sort(Y[Z==0]))

## [1] 4
```

$\rightarrow$ Lower bound $>>$ greater than 0 suggests TE heterogeneity

1. Bounding $Var(\tau_i)$ (GG pg 292-293)

```
Y <- c(1,2,3,4,5,6)
Z <- c(0,0,0,1,1,1)
```

```
# Pair in ascending order
sort(Y[Z==0])
```

```
## [1] 1 2 3
```

```
sort(Y[Z==1])
```

```
## [1] 4 5 6
```

```
# Lower bound -- ests of tau_i when Cov is as small as possible
var(sort(Y[Z==1]) - sort(Y[Z==0]))
```

```
## [1] 0
```

```
# Upper bound -- ests of tau_i when Cov is as large as possible
var(sort(Y[Z==1], decreasing = T) - sort(Y[Z==0]))
```

```
## [1] 4
```

$\rightarrow$ Lower bound $>>$ greater than 0 suggests TE heterogeneity

1. Bounding $Var(\tau_i)$ (GG pg 292-293)

```
Y <- c(1,2,3,4,5,6)
Z <- c(0,0,0,1,1,1)
```

```
# Pair in ascending order
sort(Y[Z==0])
```

```
## [1] 1 2 3
```

```
sort(Y[Z==1])
```

```
## [1] 4 5 6
```

```
# Lower bound -- ests of tau_i when Cov is as small as possible
var(sort(Y[Z==1]) - sort(Y[Z==0]))
```

```
## [1] 0
```

```
# Upper bound -- ests of tau_i when Cov is as large as possible
var(sort(Y[Z==1], decreasing = T) - sort(Y[Z==0]))
```

```
## [1] 4
```

$\rightarrow$ Lower bound $>>$ greater than 0 suggests TE heterogeneity

1. Bounding $Var(\tau_i)$ (GG pg 292-293)

```
Y <- c(1,2,3,4,5,6)
Z <- c(0,0,0,1,1,1)
```

```
# Pair in ascending order
sort(Y[Z==0])
```

```
## [1] 1 2 3
```

```
sort(Y[Z==1])
```

```
## [1] 4 5 6
```

```
# Lower bound -- ests of tau_i when Cov is as small as possible
var(sort(Y[Z==1]) - sort(Y[Z==0]))
```

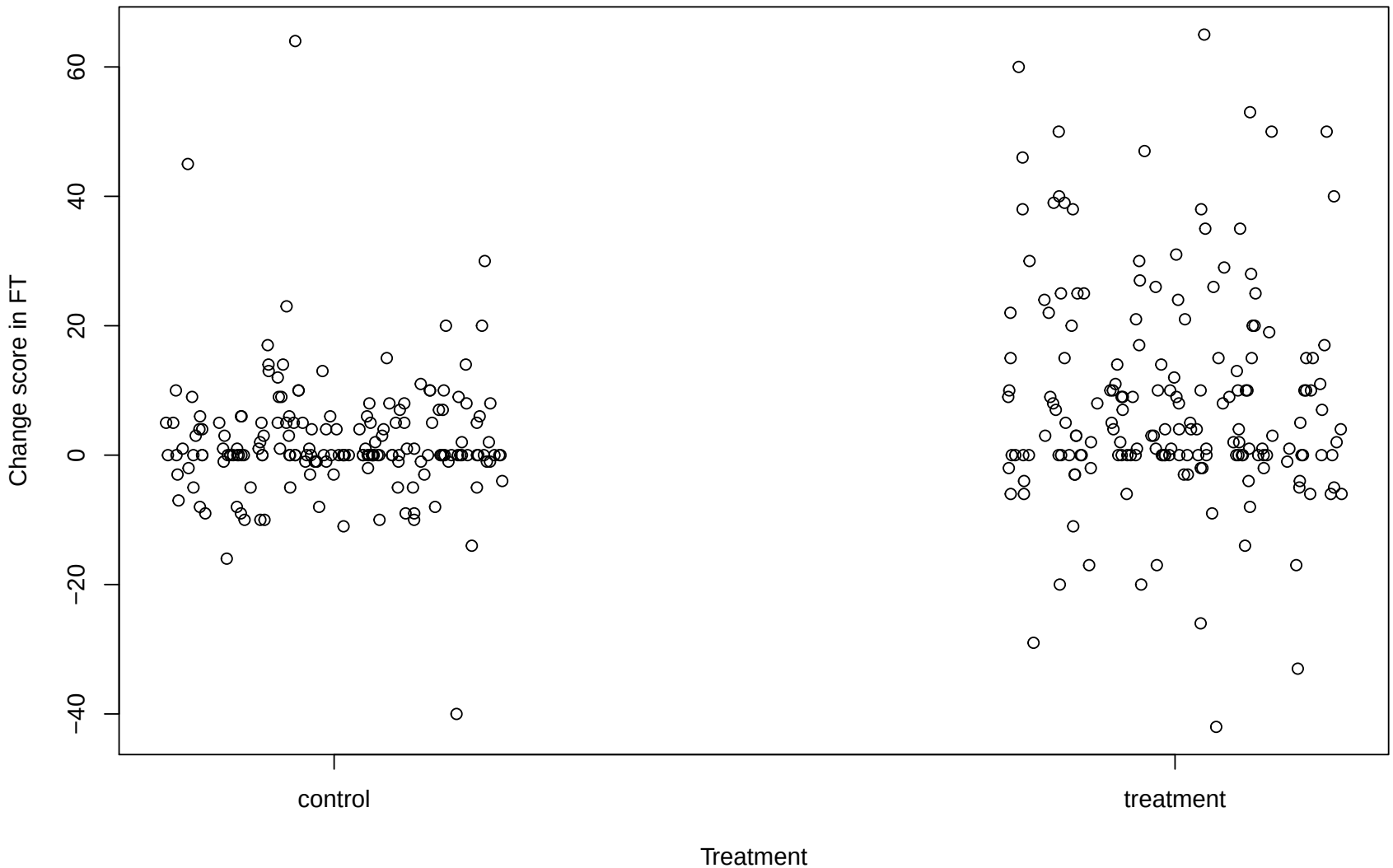
```
## [1] 0
```

```
# Upper bound -- ests of tau_i when Cov is as large as possible
var(sort(Y[Z==1], decreasing = T) - sort(Y[Z==0]))
```

```
## [1] 4
```

$\rightarrow$ Lower bound $>>$ greater than 0 suggests TE heterogeneity

Example


Change score in FT
60
40
20
0
−20
−40
control
treatment
Treatment

Example

```
tau_ests_lwr <- (sort(results$outparty_change[results$Z == 1])
                 - sort(results$outparty_change[results$Z == 0]))
var(tau_ests_lwr)
```

```
## [1] 81.99062
```

```
tau_ests_upr <- (sort(results$outparty_change[results$Z == 1],
                      decreasing = T)
                 - sort(results$outparty_change[results$Z == 0]))
var(tau_ests_upr)
```

```
## [1] 630.2649
```

→ suggests there's heterogeneity!

- effect of outparty conversation varies from conversation to conversation
- not surprising, right?

Example

```
tau_ests_lwr <- (sort(results$outparty_change[results$Z == 1])
                 - sort(results$outparty_change[results$Z == 0]))
var(tau_ests_lwr)
```

```
## [1] 81.99062
```

```
tau_ests_upr <- (sort(results$outparty_change[results$Z == 1],
                      decreasing = T)
                 - sort(results$outparty_change[results$Z == 0]))
var(tau_ests_upr)
```

```
## [1] 630.2649
```

$\rightarrow$ suggests there's heterogeneity!

- effect of outparty conversation varies from conversation to conversation
- not surprising, right?

Example

```
tau_ests_lwr <- (sort(results$outparty_change[results$Z == 1])
                 - sort(results$outparty_change[results$Z == 0]))
var(tau_ests_lwr)
```

```
## [1] 81.99062
```

```
tau_ests_upr <- (sort(results$outparty_change[results$Z == 1],
                      decreasing = T)
                 - sort(results$outparty_change[results$Z == 0]))
var(tau_ests_upr)
```

```
## [1] 630.2649
```

$\rightarrow$ suggests there's heterogeneity!

- effect of outparty conversation varies from conversation to conversation
- not surprising, right?

Example

```
tau_ests_lwr <- (sort(results$outparty_change[results$Z == 1])
                 - sort(results$outparty_change[results$Z == 0]))
var(tau_ests_lwr)
```

```
## [1] 81.99062
```

```
tau_ests_upr <- (sort(results$outparty_change[results$Z == 1],
                      decreasing = T)
                 - sort(results$outparty_change[results$Z == 0]))
var(tau_ests_upr)
```

```
## [1] 630.2649
```

$\rightarrow$ suggests there's heterogeneity!

- effect of outparty conversation varies from conversation to conversation
- not surprising, right?

2. Testing for heterogeneity

Testing null that $Var(\tau_i) = 0$

- We *can* estimate $Var(Y_i(1))$ and $Var(Y_i(0))$
- So, test of null that distributions of potential outcomes are identical except for a constant shift τ
  - » Test null that $Var(Y_i(1)) = Var(Y_i(0))$
- Why do we get to do this instead? Board

How to do this:

1. RI
2. Regression

2. Testing for heterogeneity

Testing null that $Var(\tau_i) = 0$

- We *can* estimate $Var(Y_i(1))$ and $Var(Y_i(0))$
- So, test of null that distributions of potential outcomes are identical except for a constant shift τ
  - » Test null that $Var(Y_i(1)) = Var(Y_i(0))$
- Why do we get to do this instead? Board

How to do this:

1. RI
2. Regression

2. Testing for heterogeneity

Testing null that $Var(\tau_i) = 0$

- We *can* estimate $Var(Y_i(1))$ and $Var(Y_i(0))$
- So, test of null that distributions of potential outcomes are identical except for a constant shift τ
  - » Test null that $Var(Y_i(1)) = Var(Y_i(0))$
- Why do we get to do this instead? Board

How to do this:

1. RI
2. Regression

2. Testing for heterogeneity

Testing null that $Var(\tau_i) = 0$

- We *can* estimate $Var(Y_i(1))$ and $Var(Y_i(0))$
- So, test of null that distributions of potential outcomes are identical except for a constant shift τ
  - » Test null that $Var(Y_i(1)) = Var(Y_i(0))$
- Why do we get to do this instead? Board

How to do this:

1. RI
2. Regression

2. Testing for heterogeneity

Testing null that $Var(\tau_i) = 0$

- We *can* estimate $Var(Y_i(1))$ and $Var(Y_i(0))$
- So, test of null that distributions of potential outcomes are identical except for a constant shift τ
  - » Test null that $Var(Y_i(1)) = Var(Y_i(0))$
- Why do we get to do this instead? Board

How to do this:

1. RI
2. Regression

2. Testing for heterogeneity

Testing null that $Var(\tau_i) = 0$

- We *can* estimate $Var(Y_i(1))$ and $Var(Y_i(0))$
- So, test of null that distributions of potential outcomes are identical except for a constant shift τ
  - » Test null that $Var(Y_i(1)) = Var(Y_i(0))$
- Why do we get to do this instead? Board

How to do this:

1. RI
2. Regression

2. Testing for heterogeneity

Testing null that $Var(\tau_i) = 0$

- We *can* estimate $Var(Y_i(1))$ and $Var(Y_i(0))$
- So, test of null that distributions of potential outcomes are identical except for a constant shift τ
  - » Test null that $Var(Y_i(1)) = Var(Y_i(0))$
- Why do we get to do this instead? Board

How to do this:

1. RI
2. Regression

Example GG pg 295

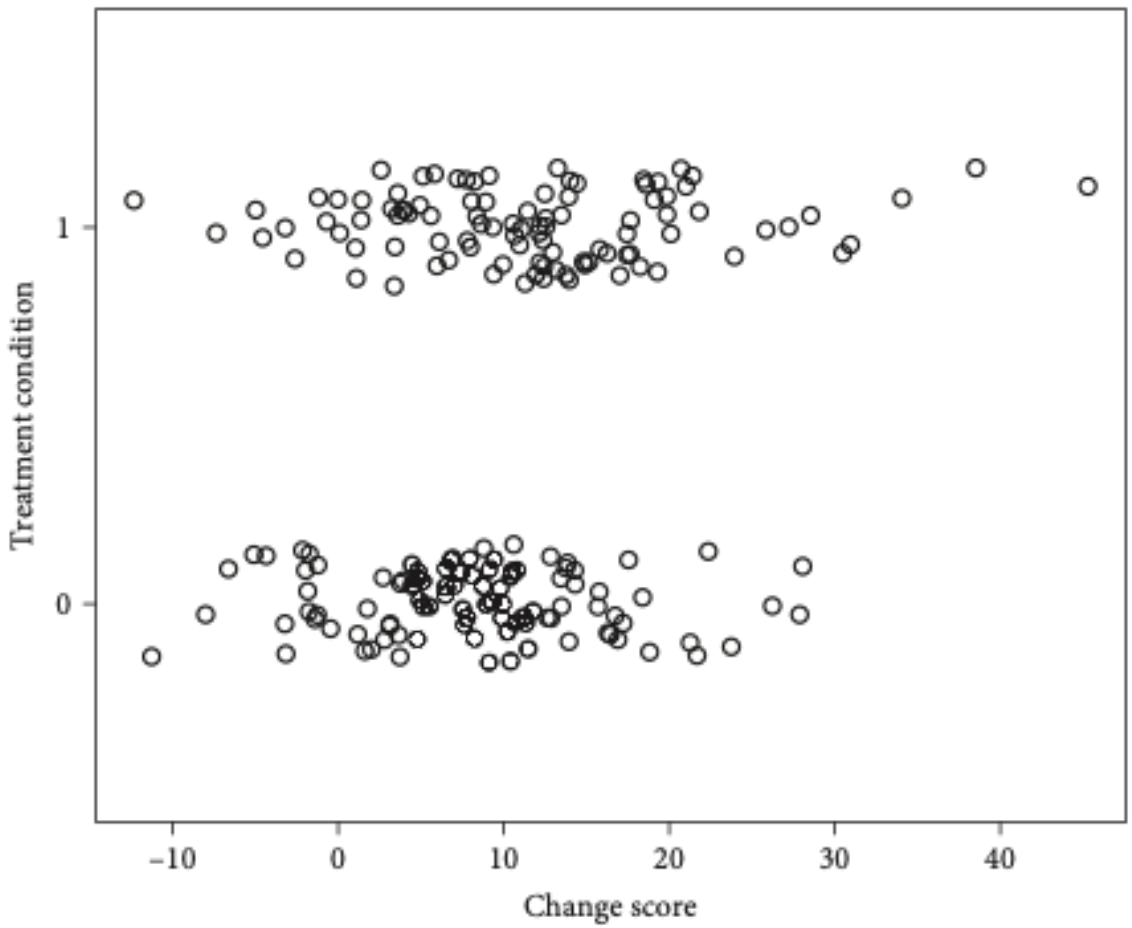
- Elementary schools
- Teachers in treated schools given bonuses (Z) according to standardized test scores
- Y = scores year 2 - scores year 1
- Find that $\hat{ATE}$ is 11.7 – big and significant

Visual inspection

Is treatment group variation in Y larger/smaller/same as control?

FIGURE 9.1

Distribution of outcomes for treatment and control groups in the teacher incentives experiment



Source: Muralidharan and Sundararaman 2011. The plotted circles have been jittered to make it easier to see each observation.

RI test

RI test

```
declaration <- randomizr::declare_ra(N = nrow(df)) #complete RA
```

```
var_fun <- function(data){
  abs(var(data$Y[data$Z == 1]) - var(data$Y[data$Z == 0]))
}
```

```
out <- ri2::conduct_ri(test_function = var_fun,
    declaration = declaration,
    assignment = "Z",
    sharp_hypothesis = ate_obs,
    data = df,
    p = "upper",
    sims = 10000)
out
```

```
##                    term estimate upper_p_value
## 1 Custom Test Statistic 57.44728        0.0476
```

RI test

```r
declaration <- randomizr::declare_ra(N = nrow(df)) #complete RA
```

```r
var_fun <- function(data){
  abs(var(data$Y[data$Z == 1]) - var(data$Y[data$Z == 0]))
}
```

```r
out <- ri2::conduct_ri(test_function = var_fun,
    declaration = declaration,
    assignment = "Z",
    sharp_hypothesis = ate_obs,
    data = df,
    p = "upper",
    sims = 10000)
out
```

```
##                    term estimate upper_p_value
## 1 Custom Test Statistic 57.44728        0.0476
```

RI test

```
declaration <- randomizr::declare_ra(N = nrow(df)) #complete RA
```

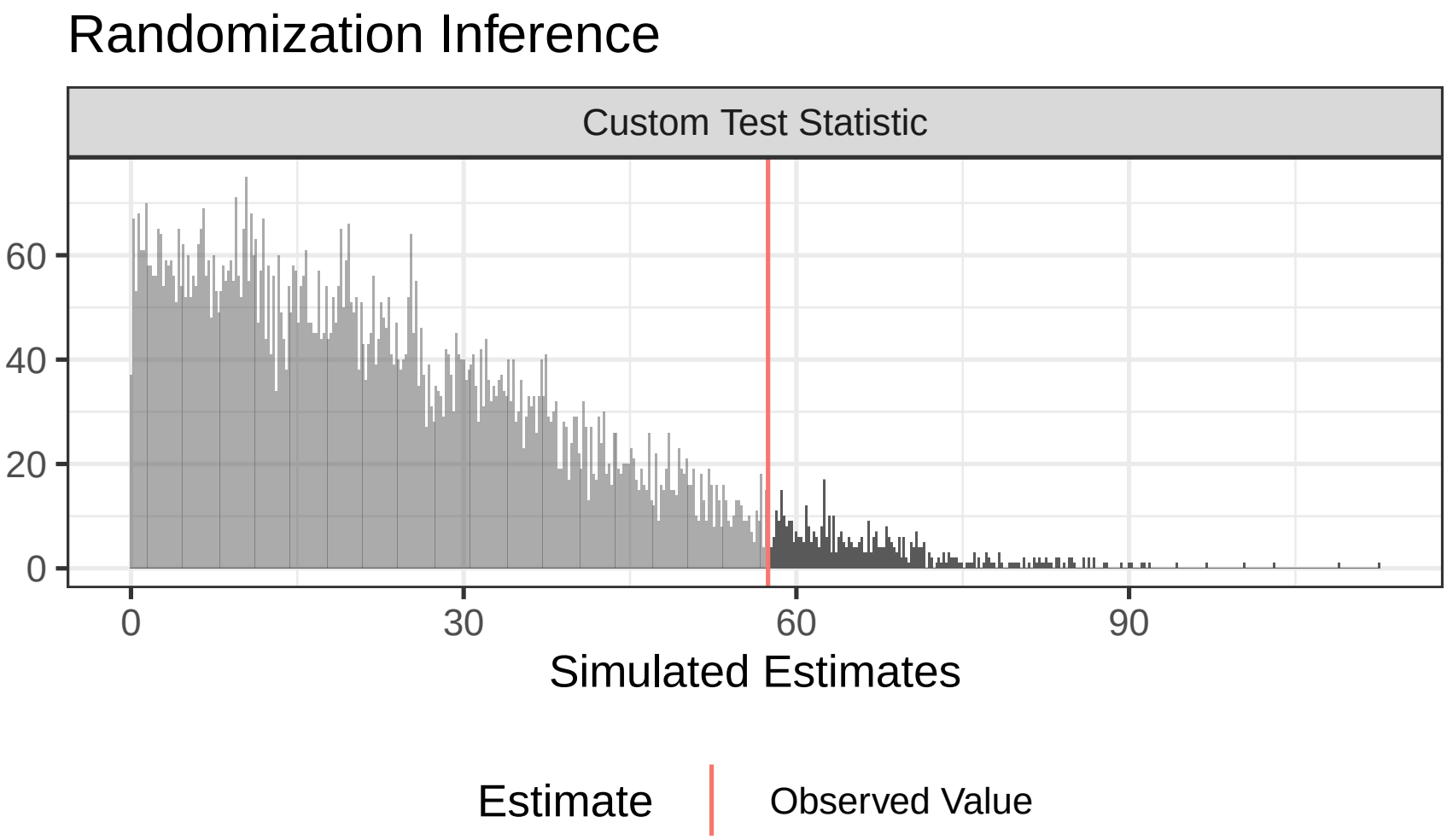
```
var_fun <- function(data){
  abs(var(data$Y[data$Z == 1]) - var(data$Y[data$Z == 0]))
}
```

```
out <- ri2::conduct_ri(test_function = var_fun,
    declaration = declaration,
    assignment = "Z",
    sharp_hypothesis = ate_obs,
    data = df,
    p = "upper",
    sims = 10000)
out
```

```
##                       term estimate upper_p_value
## 1 Custom Test Statistic 57.44728        0.0476
```

RI test

```
plot(out)
```



- Reject the null hypothesis that ATE $= 5.2$ for **all schools**
- Evidence to suggest effects vary from school to school

Beyond preliminary tests

What can and can't these preliminary tests do for us?

- Methods are for *continuous outcomes*
  - » Special methods needed for binary outcomes
- Good when we lack theoretical guidance about subgroups that may have different treatment effects
- Tests of equal variances tend to lack power
  - » First step in more structured assessment of HTE
    - Especially if evidence suggests heterogeneity

Next $\rightarrow$ what to do when we have theory or prelim tests that suggest heterogeneity

Beyond preliminary tests

What can and can't these preliminary tests do for us?

- Methods are for *continuous outcomes*
  - » Special methods needed for binary outcomes
- Good when we lack theoretical guidance about subgroups that may have different treatment effects
- Tests of equal variances tend to lack power
  - » First step in more structured assessment of HTE
    - Especially if evidence suggests heterogeneity

Next → what to do when we have theory or prelim tests that suggest heterogeneity

Beyond preliminary tests

What can and can't these preliminary tests do for us?

- Methods are for *continuous outcomes*
  - » Special methods needed for binary outcomes
- Good when we lack theoretical guidance about subgroups that may have different treatment effects
- Tests of equal variances tend to lack power
  - » First step in more structured assessment of HTE
    - Especially if evidence suggests heterogeneity

Next → what to do when we have theory or prelim tests that suggest heterogeneity

Beyond preliminary tests

What can and can't these preliminary tests do for us?

- Methods are for *continuous outcomes*
  - » Special methods needed for binary outcomes
- Good when we lack theoretical guidance about subgroups that may have different treatment effects
- Tests of equal variances tend to lack power
  - » First step in more structured assessment of HTE
    - Especially if evidence suggests heterogeneity

Next → what to do when we have theory or prelim tests that suggest heterogeneity

Beyond preliminary tests

What can and can't these preliminary tests do for us?

- Methods are for *continuous outcomes*
  - » Special methods needed for binary outcomes
- Good when we lack theoretical guidance about subgroups that may have different treatment effects
- Tests of equal variances tend to lack power
  - » First step in more structured assessment of HTE
    - Especially if evidence suggests heterogeneity

Next → what to do when we have theory or prelim tests that suggest heterogeneity

Beyond preliminary tests

What can and can't these preliminary tests do for us?

- Methods are for *continuous outcomes*
  - » Special methods needed for binary outcomes
- Good when we lack theoretical guidance about subgroups that may have different treatment effects
- Tests of equal variances tend to lack power
  - » First step in more structured assessment of HTE
    - Especially if evidence suggests heterogeneity

Next → what to do when we have theory or prelim tests that suggest heterogeneity

Beyond preliminary tests

What can and can't these preliminary tests do for us?

- Methods are for *continuous outcomes*
  - » Special methods needed for binary outcomes
- Good when we lack theoretical guidance about subgroups that may have different treatment effects
- Tests of equal variances tend to lack power
  - » First step in more structured assessment of HTE
    - Especially if evidence suggests heterogeneity

Next → what to do when we have theory or prelim tests that suggest heterogeneity

Beyond preliminary tests

What can and can't these preliminary tests do for us?

- Methods are for *continuous outcomes*
  - » Special methods needed for binary outcomes
- Good when we lack theoretical guidance about subgroups that may have different treatment effects
- Tests of equal variances tend to lack power
  - » First step in more structured assessment of HTE
    - Especially if evidence suggests heterogeneity

Next $\rightarrow$ what to do when we have theory or prelim tests that suggest heterogeneity

Part 2 – treatment by covariate interactions

Treatment by covariate interactions

- Partition units by some covariate (i.e., religion, age)
- Look at ATE *within* the subgroup
  - A new estimand!
  - CATE is **conditional average treatment effect**
  - e.g., $CATE_{catholic}$ is ATE among Catholics

Treatment by covariate interactions

- Partition units by some covariate (i.e., religion, age)
- Look at ATE *within* the subgroup
  - A new estimand!
  - CATE is **conditional average treatment effect**
  - e.g., $CATE_{catholic}$ is ATE among Catholics

Treatment by covariate interactions

- Partition units by some covariate (i.e., religion, age)
- Look at ATE *within* the subgroup
  - A new estimand!
  - CATE is **conditional average treatment effect**
  - e.g., $CATE_{catholic}$ is ATE among Catholics

Treatment by covariate interactions

- Partition units by some covariate (i.e., religion, age)
- Look at ATE *within* the subgroup
  - A new estimand!
  - CATE is **conditional average treatment effect**
  - e.g., $CATE_{catholic}$ is ATE among Catholics

Treatment by covariate interactions

- Partition units by some covariate (i.e., religion, age)
- Look at ATE *within* the subgroup
  - A new estimand!
  - CATE is **conditional average treatment effect**
  - e.g., $CATE_{catholic}$ is ATE among Catholics

Treatment by covariate interactions

Are the CATEs different for the two subgroups?

- Interact treatment with covariate $(X \cdot Z)$
  - Interaction is change/difference in treatment effect
  - $CATE_{catholic} - CATE_{noncatholic}$
    - Board
  - As long at treatment randomly assigned, we get an unbiased estimate of the difference between two CATEs
- Common examples?

Treatment by covariate interactions

Are the CATEs different for the two subgroups?

- Interact treatment with covariate $(X \cdot Z)$
  - » Interaction is change/difference in treatment effect
  - » $CATE_{catholic} - CATE_{noncatholic}$
    - Board
  - » As long at treatment randomly assigned, we get an unbiased estimate of the difference between two CATEs
- Common examples?

Treatment by covariate interactions

Are the CATEs different for the two subgroups?

- Interact treatment with covariate $(X \cdot Z)$
  - » Interaction is change/difference in treatment effect
  - » $CATE_{catholic} - CATE_{noncatholic}$
    - Board
  - » As long at treatment randomly assigned, we get an unbiased estimate of the difference between two CATEs
- Common examples?

Treatment by covariate interactions

Are the CATEs different for the two subgroups?

- Interact treatment with covariate $(X \cdot Z)$
  - » Interaction is change/difference in treatment effect
  - » $CATE_{catholic} - CATE_{noncatholic}$
    - Board
  - » As long at treatment randomly assigned, we get an unbiased estimate of the difference between two CATEs
- Common examples?

Treatment by covariate interactions

Are the CATEs different for the two subgroups?

- Interact treatment with covariate $(X \cdot Z)$
  - Interaction is change/difference in treatment effect
  - $CATE_{catholic} - CATE_{noncatholic}$
    - Board
  - As long at treatment randomly assigned, we get an unbiased estimate of the difference between two CATEs
- Common examples?

Treatment by covariate interactions

Are the CATEs different for the two subgroups?

- Interact treatment with covariate $(X \cdot Z)$
  - » Interaction is change/difference in treatment effect
  - » $CATE_{catholic} - CATE_{noncatholic}$
    - Board
  - » As long at treatment randomly assigned, we get an unbiased estimate of the difference between two CATEs
- Common examples?

Treatment by covariate interactions

Are the CATEs different for the two subgroups?

- Interact treatment with covariate $(X \cdot Z)$
  - Interaction is change/difference in treatment effect
  - $CATE_{catholic} - CATE_{noncatholic}$
    - Board
  - As long at treatment randomly assigned, we get an unbiased estimate of the difference between two CATEs
- Common examples?

Examples in print

Strength of partisanship, pages 31 and 32 (expressive partisanship)

PID, page 12 (opting out)

Strength of friendship, Figure 2c (61 million person experiment)

- "To measure a per-friend treatment effect, we compared behaviour in the friends connected to a user who received the social message to behaviour in the friends connected to a user in the control group"
- "For validated vote (Fig. 2c), the observed treatment effect is near zero for weak ties, but it spikes upwards and falls outside the null distribution for the top two deciles."

Hypothesis testing w/regression

Null hypothesis is that CATEs in both groups are **equal**

$$CATE_{catholic} = CATE_{noncatholic}$$

Null model has just one common ATE

$$Y_i = a + bZ_i + cX_i + u_i$$

Alternative model has two CATEs (board)

$$Y_i = a + bZ_i + cX_i + dZ_iX_i + u_i$$

Note these are nested models (null equals alternative when. . . ?)

- F stat!

Hypothesis testing w/regression

Null hypothesis is that CATEs in both groups are **equal**

$$CATE_{catholic} = CATE_{noncatholic}$$

Null model has just one common ATE

$$Y_i = a + bZ_i + cX_i + u_i$$

Alternative model has two CATEs (board)

$$Y_i = a + bZ_i + cX_i + dZ_iX_i + u_i$$

Note these are nested models (null equals alternative when. . . ?)

- F stat!

Hypothesis testing w/regression

Null hypothesis is that CATEs in both groups are **equal**

$$CATE_{catholic} = CATE_{noncatholic}$$

Null model has just one common ATE

$$Y_i = a + bZ_i + cX_i + u_i$$

Alternative model has two CATEs (board)

$$Y_i = a + bZ_i + cX_i + dZ_iX_i + u_i$$

Note these are nested models (null equals alternative when. . . ?)

- F stat!

Hypothesis testing w/regression

Null hypothesis is that CATEs in both groups are **equal**

$$CATE_{catholic} = CATE_{noncatholic}$$

Null model has just one common ATE

$$Y_i = a + bZ_i + cX_i + u_i$$

Alternative model has two CATEs (board)

$$Y_i = a + bZ_i + cX_i + dZ_iX_i + u_i$$

Note these are nested models (null equals alternative when. . . ?)

- F stat!

Hypothesis testing w/regression

Null hypothesis is that CATEs in both groups are **equal**

$$CATE_{catholic} = CATE_{noncatholic}$$

Null model has just one common ATE

$$Y_i = a + bZ_i + cX_i + u_i$$

Alternative model has two CATEs (board)

$$Y_i = a + bZ_i + cX_i + dZ_iX_i + u_i$$

Note these are nested models (null equals alternative when. . . ?)

- F stat!

Hypothesis testing w/regression

Null hypothesis is that CATEs in both groups are **equal**

$$CATE_{catholic} = CATE_{noncatholic}$$

Null model has just one common ATE

$$Y_i = a + bZ_i + cX_i + u_i$$

Alternative model has two CATEs (board)

$$Y_i = a + bZ_i + cX_i + dZ_iX_i + u_i$$

Note these are nested models (null equals alternative when. . . ?)

- F stat!

Hypothesis testing w/regression

Null hypothesis is that CATEs in both groups are **equal**

$$CATE_{catholic} = CATE_{noncatholic}$$

Null model has just one common ATE

$$Y_i = a + bZ_i + cX_i + u_i$$

Alternative model has two CATEs (board)

$$Y_i = a + bZ_i + cX_i + dZ_iX_i + u_i$$

Note these are nested models (null equals alternative when. . . ?)

- F stat!

Hypothesis testing w/regression

Null hypothesis is that CATEs in both groups are **equal**

$$CATE_{catholic} = CATE_{noncatholic}$$

Null model has just one common ATE

$$Y_i = a + bZ_i + cX_i + u_i$$

Alternative model has two CATEs (board)

$$Y_i = a + bZ_i + cX_i + dZ_iX_i + u_i$$

Note these are nested models (null equals alternative when. . . ?)

- F stat!

F stat

- Recall: smaller SSR $\rightarrow$ better fitting model
- F stat: compares SSR from two nested models

$$F = \frac{\frac{\text{SSR under null hyp - SSR under alt hyp}}{\text{No. params in null mod - No. of params in alt mod}}}{\frac{\text{SSR under alt hyp}}{\text{N - No.of params in alt mod}}}$$

- Calculate p value of observing F-stat as large or larger given null hypothesis (of no interaction) is true
  - » if different ATEs for different subgroups reduces SSR, numerator is positive, F stat is bigger
    - in the tails? unlikely given null; evidence supportive of interaction
  - » what if allowing for different ATEs does nothing to fit?

F stat

- Recall: smaller SSR $\rightarrow$ better fitting model
- F stat: compares SSR from two nested models

$$F = \frac{\frac{\text{SSR under null hyp - SSR under alt hyp}}{\text{No. params in null mod - No. of params in alt mod}}}{\frac{\text{SSR under alt hyp}}{\text{N - No.of params in alt mod}}}$$

- Calculate p value of observing F-stat as large or larger given null hypothesis (of no interaction) is true
  - » if different ATEs for different subgroups reduces SSR, numerator is positive, F stat is bigger
    - in the tails? unlikely given null; evidence supportive of interaction
  - » what if allowing for different ATEs does nothing to fit?

F stat

- Recall: smaller SSR $\rightarrow$ better fitting model
- F stat: compares SSR from two nested models

$$F = \frac{\frac{\text{SSR under null hyp - SSR under alt hyp}}{\text{No. params in null mod - No. of params in alt mod}}}{\frac{\text{SSR under alt hyp}}{\text{N - No.of params in alt mod}}}$$

- Calculate p value of observing F-stat as large or larger given null hypothesis (of no interaction) is true
  - » if different ATEs for different subgroups reduces SSR, numerator is positive, F stat is bigger
    - in the tails? unlikely given null; evidence supportive of interaction
  - » what if allowing for different ATEs does nothing to fit?

F stat

- Recall: smaller SSR $\rightarrow$ better fitting model
- F stat: compares SSR from two nested models

$$F = \frac{\frac{\text{SSR under null hyp - SSR under alt hyp}}{\text{No. params in null mod - No. of params in alt mod}}}{\frac{\text{SSR under alt hyp}}{\text{N - No.of params in alt mod}}}$$

- Calculate p value of observing F-stat as large or larger given null hypothesis (of no interaction) is true
  - » if different ATEs for different subgroups reduces SSR, numerator is positive, F stat is bigger
    - in the tails? unlikely given null; evidence supportive of interaction
  - » what if allowing for different ATEs does nothing to fit?

F stat

- Recall: smaller SSR $\rightarrow$ better fitting model
- F stat: compares SSR from two nested models

$$F = \frac{\frac{\text{SSR under null hyp - SSR under alt hyp}}{\text{No. params in null mod - No. of params in alt mod}}}{\frac{\text{SSR under alt hyp}}{\text{N - No.of params in alt mod}}}$$

- Calculate p value of observing F-stat as large or larger given null hypothesis (of no interaction) is true
  - » if different ATEs for different subgroups reduces SSR, numerator is positive, F stat is bigger
    - in the tails? unlikely given null; evidence supportive of interaction
  - » what if allowing for different ATEs does nothing to fit?

F stat

- Recall: smaller SSR $\rightarrow$ better fitting model
- F stat: compares SSR from two nested models

$$F = \frac{\frac{\text{SSR under null hyp - SSR under alt hyp}}{\text{No. params in null mod - No. of params in alt mod}}}{\frac{\text{SSR under alt hyp}}{\text{N - No.of params in alt mod}}}$$

- Calculate p value of observing F-stat as large or larger given null hypothesis (of no interaction) is true
  - » if different ATEs for different subgroups reduces SSR, numerator is positive, F stat is bigger
    - in the tails? unlikely given null; evidence supportive of interaction
  - » what if allowing for different ATEs does nothing to fit?

F stat

- Recall: smaller SSR $\rightarrow$ better fitting model
- F stat: compares SSR from two nested models

$$F = \frac{\frac{\text{SSR under null hyp - SSR under alt hyp}}{\text{No. params in null mod - No. of params in alt mod}}}{\frac{\text{SSR under alt hyp}}{\text{N - No.of params in alt mod}}}$$

- Calculate p value of observing F-stat as large or larger given null hypothesis (of no interaction) is true
  - » if different ATEs for different subgroups reduces SSR, numerator is positive, F stat is bigger
    - in the tails? unlikely given null; evidence supportive of interaction
  - » what if allowing for different ATEs does nothing to fit?

Example

```
null_mod <- lm(outparty_change ~ Z + pid2, data = results[results$full_cluster,])
summary(null_mod)
```

```
##
## Call:
## lm(formula = outparty_change ~ Z + pid2, data = results[results$full_cluster,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.604  -7.600  -1.255   3.748  62.748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2553     0.9763   2.310   0.0212 *
## Z             7.3485     1.1209   6.556 1.23e-10 ***
## pid2R        -1.0035     1.1207  -0.895   0.3710
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.47 on 575 degrees of freedom
## Multiple R-squared:  0.07075,    Adjusted R-squared:  0.06752
## F-statistic: 21.89 on 2 and 575 DF,  p-value: 6.88e-10
```

Example

```r
alt_mod <- lm(outparty_change ~ Z*pid2, data = results[results$full_cluster,])
summary(alt_mod)
```

```
##
## Call:
## lm(formula = outparty_change ~ Z * pid2, data = results[results$full_cluster,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.238  -7.908  -1.599   3.966  62.092
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.5986     1.1302   1.414    0.158
## Z             8.6395     1.5847   5.452 7.42e-08 ***
## pid2R         0.3099     1.5984   0.194    0.846
## Z:pid2R      -2.5820     2.2412  -1.152    0.250
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.47 on 574 degrees of freedom
## Multiple R-squared:  0.0729, Adjusted R-squared:  0.06805
## F-statistic: 15.04 on 3 and 574 DF,  p-value: 1.942e-09
```

Example

```
anova(null_mod, alt_mod)
```

```
## Analysis of Variance Table
##
## Model 1: outparty_change ~ Z + pid2
## Model 2: outparty_change ~ Z * pid2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    575 104362
## 2    574 104121  1    240.76 1.3272 0.2498
```

Side note

P-values are the same because of the relationship between the t and F distributions

Final intuition

$$Y_i = a + bZ_i + cX_i + dZ_iX_i + u_i$$

- b is $CATE_D$
- $b + d$ is $CATE_R$
- d is interaction effect, or difference in $CATE$s

Final intuition

$$Y_i = a + bZ_i + cX_i + dZ_iX_i + u_i$$

- b is $CATE_D$
- $b + d$ is $CATE_R$
- d is interaction effect, or difference in $CATE$s

Final intuition

$$Y_i = a + bZ_i + cX_i + dZ_iX_i + u_i$$

- b is $CATE_D$
- $b + d$ is $CATE_R$
- d is interaction effect, or difference in $CATEs$

Final intuition

$$Y_i = a + bZ_i + cX_i + dZ_iX_i + u_i$$

- b is $CATE_D$
- $b + d$ is $CATE_R$
- d is interaction effect, or difference in $CATE$s

Why do we need caution?

Discipline is skeptical, so **be very intentional, transparent, and careful about your HTE expectations and interpretations**

1. Multiple comparisons problem

- more on this later in the semester
- evaluate enough hypothesis, and one is likely to be significant at .05 level even though no effect
- reject a true null; false positive!

2. Pre-register a specific set of subgroup analyses

- without PAP, book advises to "regard hypothesis tests with skepticism pending replication by another study"

3. Subgroup analysis is fundamentally non-experimental!

- Moderator not randomly assigned
- Republican/Democrat is a marker of many things...
  - » predictive vs. causal interpretation
- Therefore, think of this as either **descriptive or exploratory**

Why do we need caution?

Discipline is skeptical, so **be very intentional, transparent, and careful about your HTE expectations and interpretations**

1. Multiple comparisons problem

- more on this later in the semester
- evaluate enough hypothesis, and one is likely to be significant at .05 level even though no effect
- reject a true null; false positive!

2. Pre-register a specific set of subgroup analyses

- without PAP, book advises to "regard hypothesis tests with skepticism pending replication by another study"

3. Subgroup analysis is fundamentally non-experimental!

- Moderator not randomly assigned
- Republican/Democrat is a marker of many things...
  - » predictive vs. causal interpretation
- Therefore, think of this as either **descriptive or exploratory**

Why do we need caution?

Discipline is skeptical, so **be very intentional, transparent, and careful about your HTE expectations and interpretations**

1. Multiple comparisons problem

- more on this later in the semester
- evaluate enough hypothesis, and one is likely to be significant at .05 level even though no effect
- reject a true null; false positive!

2. Pre-register a specific set of subgroup analyses

- without PAP, book advises to "regard hypothesis tests with skepticism pending replication by another study"

3. Subgroup analysis is fundamentally non-experimental!

- Moderator not randomly assigned
- Republican/Democrat is a marker of many things...
  - » predictive vs. causal interpretation
- Therefore, think of this as either **descriptive or exploratory**

Why do we need caution?

Discipline is skeptical, so **be very intentional, transparent, and careful about your HTE expectations and interpretations**

1. Multiple comparisons problem

- more on this later in the semester
- evaluate enough hypothesis, and one is likely to be significant at .05 level even though no effect
- reject a true null; false positive!

2. Pre-register a specific set of subgroup analyses

- without PAP, book advises to "regard hypothesis tests with skepticism pending replication by another study"

3. Subgroup analysis is fundamentally non-experimental!

- Moderator not randomly assigned
- Republican/Democrat is a marker of many things...
  - » predictive vs. causal interpretation
- Therefore, think of this as either **descriptive or exploratory**

Why do we need caution?

Discipline is skeptical, so **be very intentional, transparent, and careful about your HTE expectations and interpretations**

1. Multiple comparisons problem

- more on this later in the semester
- evaluate enough hypothesis, and one is likely to be significant at .05 level even though no effect
- reject a true null; false positive!

2. Pre-register a specific set of subgroup analyses

- without PAP, book advises to "regard hypothesis tests with skepticism pending replication by another study"

3. Subgroup analysis is fundamentally non-experimental!

- Moderator not randomly assigned
- Republican/Democrat is a marker of many things...
  - » predictive vs. causal interpretation
- Therefore, think of this as either **descriptive or exploratory**

Why do we need caution?

Discipline is skeptical, so **be very intentional, transparent, and careful about your HTE expectations and interpretations**

1. Multiple comparisons problem

- more on this later in the semester
- evaluate enough hypothesis, and one is likely to be significant at .05 level even though no effect
- reject a true null; false positive!

2. Pre-register a specific set of subgroup analyses

- without PAP, book advises to "regard hypothesis tests with skepticism pending replication by another study"

3. Subgroup analysis is fundamentally non-experimental!

- Moderator not randomly assigned
- Republican/Democrat is a marker of many things...
  - » predictive vs. causal interpretation
- Therefore, think of this as either **descriptive or exploratory**

Why do we need caution?

Discipline is skeptical, so **be very intentional, transparent, and careful about your HTE expectations and interpretations**

1. Multiple comparisons problem

- more on this later in the semester
- evaluate enough hypothesis, and one is likely to be significant at .05 level even though no effect
- reject a true null; false positive!

2. Pre-register a specific set of subgroup analyses

- without PAP, book advises to "regard hypothesis tests with skepticism pending replication by another study"

3. Subgroup analysis is fundamentally non-experimental!

- Moderator not randomly assigned
- Republican/Democrat is a marker of many things...
  - » predictive vs. causal interpretation
- Therefore, think of this as either **descriptive or exploratory**

Why do we need caution?

Discipline is skeptical, so **be very intentional, transparent, and careful about your HTE expectations and interpretations**

1. Multiple comparisons problem

- more on this later in the semester
- evaluate enough hypothesis, and one is likely to be significant at .05 level even though no effect
- reject a true null; false positive!

2. Pre-register a specific set of subgroup analyses

- without PAP, book advises to "regard hypothesis tests with skepticism pending replication by another study"

3. Subgroup analysis is fundamentally non-experimental!

- Moderator not randomly assigned
- Republican/Democrat is a marker of many things...
  - » predictive vs. causal interpretation
- Therefore, think of this as either **descriptive or exploratory**

Why do we need caution?

Discipline is skeptical, so **be very intentional, transparent, and careful about your HTE expectations and interpretations**

1. Multiple comparisons problem

- more on this later in the semester
- evaluate enough hypothesis, and one is likely to be significant at .05 level even though no effect
- reject a true null; false positive!

2. Pre-register a specific set of subgroup analyses

- without PAP, book advises to "regard hypothesis tests with skepticism pending replication by another study"

3. Subgroup analysis is fundamentally non-experimental!

- Moderator not randomly assigned
- Republican/Democrat is a marker of many things...
  - » predictive vs. causal interpretation
- Therefore, think of this as either **descriptive or exploratory**

Why do we need caution?

Discipline is skeptical, so **be very intentional, transparent, and careful about your HTE expectations and interpretations**

1. Multiple comparisons problem

- more on this later in the semester
- evaluate enough hypothesis, and one is likely to be significant at .05 level even though no effect
- reject a true null; false positive!

2. Pre-register a specific set of subgroup analyses

- without PAP, book advises to "regard hypothesis tests with skepticism pending replication by another study"

3. Subgroup analysis is fundamentally non-experimental!

- Moderator not randomly assigned
- Republican/Democrat is a marker of many things...
  - predictive vs. causal interpretation
- Therefore, think of this as either **descriptive or exploratory**

Why do we need caution?

Discipline is skeptical, so **be very intentional, transparent, and careful about your HTE expectations and interpretations**

1. Multiple comparisons problem

- more on this later in the semester
- evaluate enough hypothesis, and one is likely to be significant at .05 level even though no effect
- reject a true null; false positive!

2. Pre-register a specific set of subgroup analyses

- without PAP, book advises to "regard hypothesis tests with skepticism pending replication by another study"

3. Subgroup analysis is fundamentally non-experimental!

- Moderator not randomly assigned
- Republican/Democrat is a marker of many things...
  - » predictive vs. causal interpretation
- Therefore, think of this as either **descriptive or exploratory**

Why do we need caution?

Discipline is skeptical, so **be very intentional, transparent, and careful about your HTE expectations and interpretations**

1. Multiple comparisons problem

- more on this later in the semester
- evaluate enough hypothesis, and one is likely to be significant at .05 level even though no effect
- reject a true null; false positive!

2. Pre-register a specific set of subgroup analyses

- without PAP, book advises to "regard hypothesis tests with skepticism pending replication by another study"

3. Subgroup analysis is fundamentally non-experimental!

- Moderator not randomly assigned
- Republican/Democrat is a marker of many things...
  - » predictive vs. causal interpretation
- Therefore, think of this as either **descriptive or exploratory**

Part 3 – Treatment by Treatment interactions

Treatment by Treatment interactions

Factorial experiment

- Overcomes causal limitations of subgroup analyses
- Experiments with two (or more) factors
- Each factor has two (or more) experimental conditions
- Example
  - FactorA: Talk (1), Don't talk (0)
  - FactorB: Political topic (1), Non-political topic (0)
  - Factor1 randomly assigned, then Factor2 randomly assigned
    - "Fully crossed"
  - Board
- Analysis with regression follows as before, but with causal interpretation

Treatment by Treatment interactions

Factorial experiment

- Overcomes causal limitations of subgroup analyses
- Experiments with two (or more) factors
- Each factor has two (or more) experimental conditions
- Example
  - FactorA: Talk (1), Don't talk (0)
  - FactorB: Political topic (1), Non-political topic (0)
  - Factor1 randomly assigned, then Factor2 randomly assigned
    - "Fully crossed"
  - Board
- Analysis with regression follows as before, but with causal interpretation

Treatment by Treatment interactions

Factorial experiment

- Overcomes causal limitations of subgroup analyses
- Experiments with two (or more) factors
- Each factor has two (or more) experimental conditions
- Example
  - FactorA: Talk (1), Don't talk (0)
  - FactorB: Political topic (1), Non-political topic (0)
  - Factor1 randomly assigned, then Factor2 randomly assigned
    - "Fully crossed"
  - Board
- Analysis with regression follows as before, but with causal interpretation

Treatment by Treatment interactions

Factorial experiment

- Overcomes causal limitations of subgroup analyses
- Experiments with two (or more) factors
- Each factor has two (or more) experimental conditions
- Example
  - FactorA: Talk (1), Don't talk (0)
  - FactorB: Political topic (1), Non-political topic (0)
  - Factor1 randomly assigned, then Factor2 randomly assigned
    - "Fully crossed"
  - Board
- Analysis with regression follows as before, but with causal interpretation

Treatment by Treatment interactions

Factorial experiment

- Overcomes causal limitations of subgroup analyses
- Experiments with two (or more) factors
- Each factor has two (or more) experimental conditions
- Example
  - FactorA: Talk (1), Don't talk (0)
  - FactorB: Political topic (1), Non-political topic (0)
  - Factor1 randomly assigned, then Factor2 randomly assigned
    - "Fully crossed"
  - Board
- Analysis with regression follows as before, but with causal interpretation

Treatment by Treatment interactions

Factorial experiment

- Overcomes causal limitations of subgroup analyses
- Experiments with two (or more) factors
- Each factor has two (or more) experimental conditions
- Example
  - FactorA: Talk (1), Don't talk (0)
  - FactorB: Political topic (1), Non-political topic (0)
  - Factor1 randomly assigned, then Factor2 randomly assigned
    - "Fully crossed"
  - Board
- Analysis with regression follows as before, but with causal interpretation

Treatment by Treatment interactions

Factorial experiment

- Overcomes causal limitations of subgroup analyses
- Experiments with two (or more) factors
- Each factor has two (or more) experimental conditions
- Example
  - FactorA: Talk (1), Don't talk (0)
  - FactorB: Political topic (1), Non-political topic (0)
  - Factor1 randomly assigned, then Factor2 randomly assigned
    - "Fully crossed"
  - Board
- Analysis with regression follows as before, but with causal interpretation

Treatment by Treatment interactions

Factorial experiment

- Overcomes causal limitations of subgroup analyses
- Experiments with two (or more) factors
- Each factor has two (or more) experimental conditions
- Example
  - FactorA: Talk (1), Don't talk (0)
  - FactorB: Political topic (1), Non-political topic (0)
  - Factor1 randomly assigned, then Factor2 randomly assigned
    - "Fully crossed"
  - Board
- Analysis with regression follows as before, but with causal interpretation

Treatment by Treatment interactions

Factorial experiment

- Overcomes causal limitations of subgroup analyses
- Experiments with two (or more) factors
- Each factor has two (or more) experimental conditions
- Example
  - FactorA: Talk (1), Don't talk (0)
  - FactorB: Political topic (1), Non-political topic (0)
  - Factor1 randomly assigned, then Factor2 randomly assigned
    - "Fully crossed"
  - Board
- Analysis with regression follows as before, but with causal interpretation

Treatment by Treatment interactions

Factorial experiment

- Overcomes causal limitations of subgroup analyses
- Experiments with two (or more) factors
- Each factor has two (or more) experimental conditions
- Example
  - FactorA: Talk (1), Don't talk (0)
  - FactorB: Political topic (1), Non-political topic (0)
  - Factor1 randomly assigned, then Factor2 randomly assigned
    - "Fully crossed"
  - Board
- Analysis with regression follows as before, but with causal interpretation

Treatment by Treatment interactions

Factorial experiment

- Overcomes causal limitations of subgroup analyses
- Experiments with two (or more) factors
- Each factor has two (or more) experimental conditions
- Example
  - FactorA: Talk (1), Don't talk (0)
  - FactorB: Political topic (1), Non-political topic (0)
  - Factor1 randomly assigned, then Factor2 randomly assigned
    - "Fully crossed"
  - Board
- Analysis with regression follows as before, but with causal interpretation

Treatment by Treatment interactions

Factorial experiment

- Overcomes causal limitations of subgroup analyses
- Experiments with two (or more) factors
- Each factor has two (or more) experimental conditions
- Example
  - FactorA: Talk (1), Don't talk (0)
  - FactorB: Political topic (1), Non-political topic (0)
  - Factor1 randomly assigned, then Factor2 randomly assigned
    - "Fully crossed"
  - Board
- Analysis with regression follows as before, but with causal interpretation

In sum

- Investigating treatment effect heterogeneity is challenging
  - » be methodical and cautious
  - » plan in the PAP!
- Without (or with) a PAP
  - » start with prelim tests to shed light on the question
  - » if you find a tentative yes, use theory to guide testable interaction effects
- Ideally, build additional factors into the design
  - » random assignment allows for causal interpretation!
  - » if you can manipulate it, try to, but often we can't
  - » treatment-by-covariate interactions are less informative in this sense, but informative and important (gender, religion, years of education)

In sum

- Investigating treatment effect heterogeneity is challenging
  - » be methodical and cautious
  - » plan in the PAP!
- Without (or with) a PAP
  - » start with prelim tests to shed light on the question
  - » if you find a tentative yes, use theory to guide testable interaction effects
- Ideally, build additional factors into the design
  - » random assignment allows for causal interpretation!
  - » if you can manipulate it, try to, but often we can't
  - » treatment-by-covariate interactions are less informative in this sense, but informative and important (gender, religion, years of education)

In sum

- Investigating treatment effect heterogeneity is challenging
  - » be methodical and cautious
  - » plan in the PAP!
- Without (or with) a PAP
  - » start with prelim tests to shed light on the question
  - » if you find a tentative yes, use theory to guide testable interaction effects
- Ideally, build additional factors into the design
  - » random assignment allows for causal interpretation!
  - » if you can manipulate it, try to, but often we can't
  - » treatment-by-covariate interactions are less informative in this sense, but informative and important (gender, religion, years of education)

In sum

- Investigating treatment effect heterogeneity is challenging
  - » be methodical and cautious
  - » plan in the PAP!
- Without (or with) a PAP
  - » start with prelim tests to shed light on the question
  - » if you find a tentative yes, use theory to guide testable interaction effects
- Ideally, build additional factors into the design
  - » random assignment allows for causal interpretation!
  - » if you can manipulate it, try to, but often we can't
  - » treatment-by-covariate interactions are less informative in this sense, but informative and important (gender, religion, years of education)

In sum

- Investigating treatment effect heterogeneity is challenging
  - » be methodical and cautious
  - » plan in the PAP!
- Without (or with) a PAP
  - » start with prelim tests to shed light on the question
  - » if you find a tentative yes, use theory to guide testable interaction effects
- Ideally, build additional factors into the design
  - » random assignment allows for causal interpretation!
  - » if you can manipulate it, try to, but often we can't
  - » treatment-by-covariate interactions are less informative in this sense, but informative and important (gender, religion, years of education)

In sum

- Investigating treatment effect heterogeneity is challenging
  - » be methodical and cautious
  - » plan in the PAP!
- Without (or with) a PAP
  - » start with prelim tests to shed light on the question
  - » if you find a tentative yes, use theory to guide testable interaction effects
- Ideally, build additional factors into the design
  - » random assignment allows for causal interpretation!
  - » if you can manipulate it, try to, but often we can't
  - » treatment-by-covariate interactions are less informative in this sense, but informative and important (gender, religion, years of education)

In sum

- Investigating treatment effect heterogeneity is challenging
  - » be methodical and cautious
  - » plan in the PAP!
- Without (or with) a PAP
  - » start with prelim tests to shed light on the question
  - » if you find a tentative yes, use theory to guide testable interaction effects
- Ideally, build additional factors into the design
  - » random assignment allows for causal interpretation!
  - » if you can manipulate it, try to, but often we can't
  - » treatment-by-covariate interactions are less informative in this sense, but informative and important (gender, religion, years of education)

In sum

- Investigating treatment effect heterogeneity is challenging
  - » be methodical and cautious
  - » plan in the PAP!
- Without (or with) a PAP
  - » start with prelim tests to shed light on the question
  - » if you find a tentative yes, use theory to guide testable interaction effects
- Ideally, build additional factors into the design
  - » random assignment allows for causal interpretation!
  - » if you can manipulate it, try to, but often we can't
  - » treatment-by-covariate interactions are less informative in this sense, but informative and important (gender, religion, years of education)

In sum

- Investigating treatment effect heterogeneity is challenging
  - » be methodical and cautious
  - » plan in the PAP!
- Without (or with) a PAP
  - » start with prelim tests to shed light on the question
  - » if you find a tentative yes, use theory to guide testable interaction effects
- Ideally, build additional factors into the design
  - » random assignment allows for causal interpretation!
  - » if you can manipulate it, try to, but often we can't
  - » treatment-by-covariate interactions are less informative in this sense, but informative and important (gender, religion, years of education)

In sum

- Investigating treatment effect heterogeneity is challenging
  - » be methodical and cautious
  - » plan in the PAP!
- Without (or with) a PAP
  - » start with prelim tests to shed light on the question
  - » if you find a tentative yes, use theory to guide testable interaction effects
- Ideally, build additional factors into the design
  - » random assignment allows for causal interpretation!
  - » if you can manipulate it, try to, but often we can't
  - » treatment-by-covariate interactions are less informative in this sense, but informative and important (gender, religion, years of education)

Labs