Gerber and Green (2012) Chapter 4 Problem 10

Margaret Moor and Alexander Coppock, Yale University

January 30, 2018

This script shows how to conduct the randomization inference procedure in Gerber and Green (2012) Chapter 4 Problem 10 two different ways: using the `ri2` package and by hand with a loop. The `ri` package does not support balance tests.

Chapter 4 Problem 10

The 2003 Kansas City voter mobilization experiment described in Chapter 3 is a cluster randomized design in which 28 precincts comprising 9,712 voters were randomly assigned to treatment and control. The study contains a wealth of covariates: the registrar recorded whether each voter participated in elections dating back to 1996.

(a) Assess the balance of the treatment and control groups by looking at whether past turnout predicts treatment assignment. Regress treatment assignment on the entire set of past votes, and calculate the sum of squared residuals. Use randomization inference to test the null hypothesis that none of the past turnout variables predict treatment assignment. Remember that to simulate the distribution of the F-statistic, you must generate a large number of random cluster assignments and calculate the F-statistic for each simulated assignment. Judging from the p-value of this test, what does the F-statistic seem to suggest about whether subjects in the treatment and control groups have comparable background characteristics?

SHOWN BELOW

(b) Regress turnout in 2003 (after the treatment was administered) on the experimental assignment and the full set of covariates. Interpret the estimated ATE. Use randomization inference to test the sharp null hypothesis that experimental assignment had no effect on any subject's decision to vote.

SHOWN BELOW

(c) When analyzing cluster randomized experiments with clusters of varying size, one concern is that difference-in-means estimation is prone to bias. This concern also applies to regression. In order to sidestep this problem, researchers may choose to use the difference-in-totals estimator in equation (3.24) to estimate the ATE. Estimate the ATE using this estimator.
(d) Use randomization inference to test the sharp null hypothesis that treatment assignment had no effect, using the difference-in-totals estimator.
(e) The difference-in-totals estimator can generate imprecise estimates, but its precision can be improved by incorporating information about covariates. Create a new outcome variable that is the difference between a subject's turnout (1 = vote, 0 = abstain) and the average rate of turnout in all past elections. Using this "differenced" outcome variable, estimate the ATE using the difference-in-totals estimator, and test the sharp null hypothesis of no effect.

NOT SHOWN

Setup

```
# Data from http://isps.yale.edu/FEDAI
library(haven)
data4.10 <- read_dta("datasets/4.10.dta")
```

```r
# Number of sims the same for both methods
sims <- 1000

# 14 treated clusted, 14 untreated clusters
with(data4.10, table(tapply(treatmen, INDEX = unit, FUN = unique)))
```

```
##
##  0  1
## 14 14
```

```r
covs <-
  "v_p2003 + v_m2003 + v_g2002 + v_p2002 + v_m2002 + v_s2001 +
  v_g2000 + v_p2000 + v_m2000 + v_s1999 + v_m1999 + v_g1998 +
  v_m1998 + v_s1998 + v_m1997 + v_s1997 + v_g1996 + v_p1996 +
  v_m1996 + v_s1996"
```

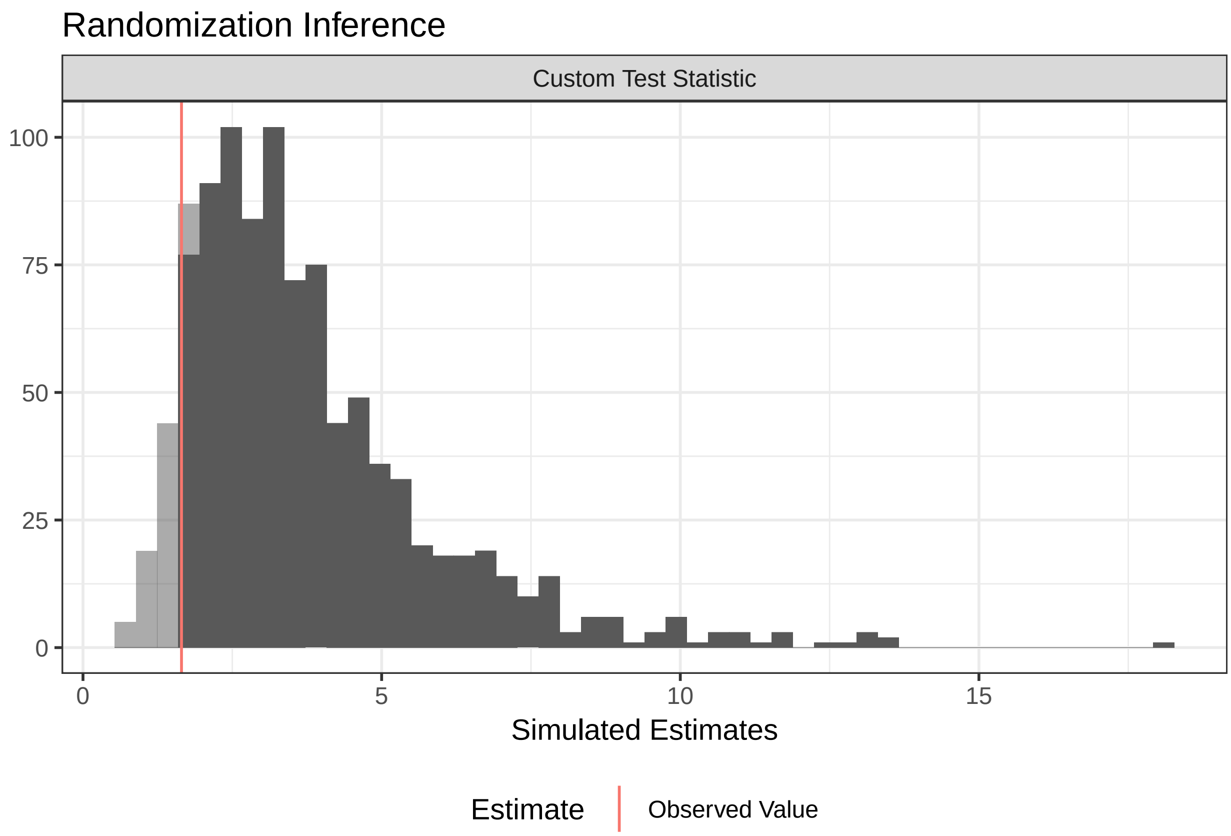
In ri2

Balance Test

```r
library(ri2)
# Declare randomization procedure
declaration <- declare_ra(clusters = data4.10$unit, m = 14)

# test function
balance_fun <- function(data) {
  fit <- lm(formula = paste0("treatmen ~ ", covs), data)
  f_stat <- summary(fit)$f[1]
  names(f_stat) <- NULL
  return(f_stat)
}

# Conduct Randomization Inference
ri2_out <- conduct_ri(declaration = declaration,
                      test_function = balance_fun,
                      assignment = "treatmen",
                      sims = sims,
                      data = data4.10)

summary(ri2_out)
```

```
##              coefficient estimate two_tailed_p_value null_ci_lower
## 1 Custom Test Statistic 1.649236              0.922      1.253666
##   null_ci_upper
## 1      9.672362
```

```
plot(ri2_out)
```

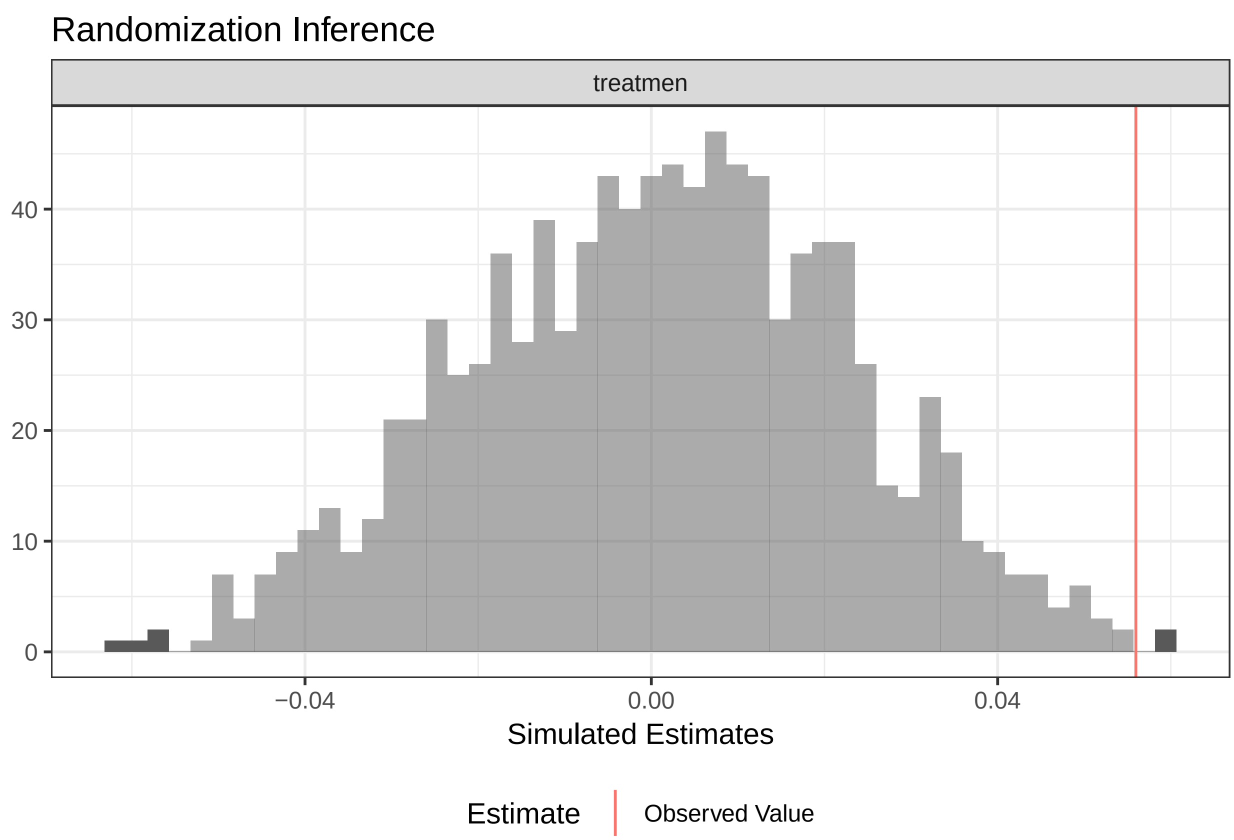


Hypothesis Test

```
# Conduct Randomization Inference
ri2_out <- conduct_ri(formula = paste0("vote03 ~ treatmen + ", covs),
                      declaration = declaration,
                      assignment = "treatmen",
                      sims = sims,
                      data = data4.10)

summary(ri2_out)
```

```
##   coefficient   estimate two_tailed_p_value null_ci_lower null_ci_upper
## 1    treatmen 0.05596196              0.006   -0.04157238    0.04258617
```

```
plot(ri2_out)
```



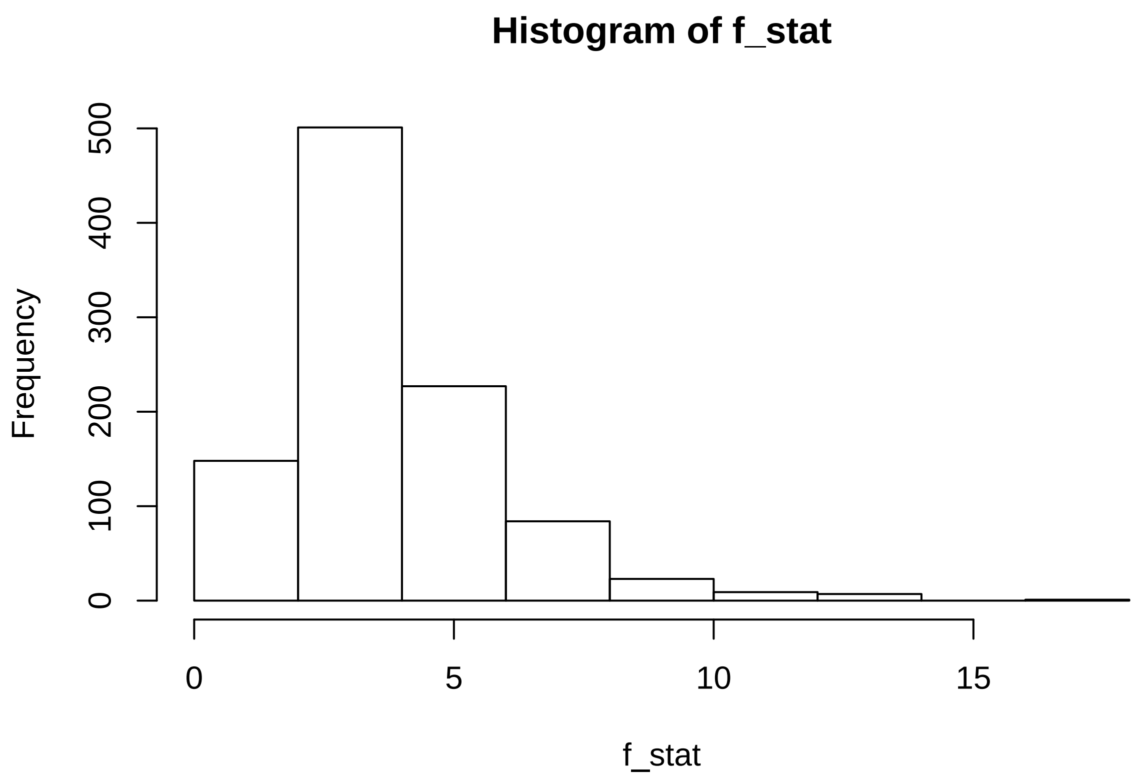
By hand

Balance Test

```
fit <- lm(formula = paste0("treatmen ~ ", covs), data4.10)
obs_fstat <- summary(fit)$fstatistic[1]

f_stat <- rep(NA, sims)

for(i in 1:sims) {
  data4.10$Z_sim <- cluster_ra(data4.10$unit, m = 14)
  fit_sim <- lm(formula = paste0("Z_sim ~ ", covs), data = data4.10)
  f_stat[i] <- summary(fit_sim)$fstatistic[1]
}

p_two_tailed <- mean(abs(f_stat) >= abs(obs_fstat))
p_upper <- mean(f_stat >= obs_fstat)
p_lower <- mean(f_stat <= obs_fstat)

c(observed_ate, p_two_tailed, p_upper, p_lower)
```

```
## [1] -110.1765    0.9360    0.9360    0.0640
```

```
hist(f_stat, breaks = 10)
```

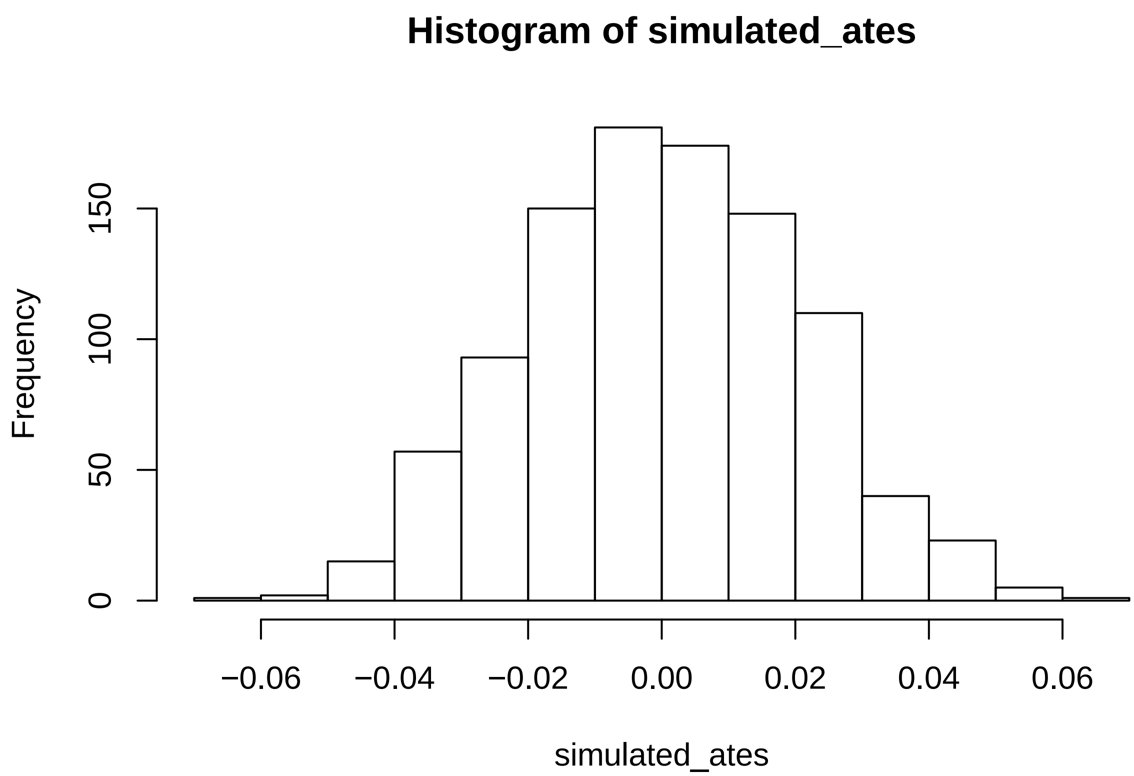


Hypothesis Test

```
fit_2 <- lm(formula = paste0("vote03 ~ treatmen + ", covs), data4.10)
observed_ate <- coef(fit_2)[2]

simulated_ates <- rep(NA, sims)

for (i in 1:sims){
  data4.10$Z_sim <- cluster_ra(data4.10$unit, m = 14)
  fit_sim <- lm(formula = paste0("vote03 ~ Z_sim + ", covs), data4.10)
  simulated_ates[i] <- coef(fit_sim)[2]
}

p_two_tailed2 <- mean(abs(simulated_ates) >= abs(observed_ate))
p_upper2 <- mean(simulated_ates >= observed_ate)
p_lower2 <- mean(simulated_ates <= observed_ate)

c(observed_ate, p_two_tailed, p_upper, p_lower)
```

```
##   treatmen
## 0.05596196 0.93600000 0.93600000 0.06400000
```

```
hist(simulated_ates, breaks = 10)
```


Histogram of simulated_ates
Frequency
0
50
100
150
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
simulated_ates