Crossfitting-MPE

Lei Shi

2024-11-26

Setting 3: MPE setting

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.2.3
```

```
set.seed(2024)

K = 2e2
N = 2 * K

X1 = rnorm(N)
X2 = rnorm(N)

Y1 = rpois(N, lambda = exp(0.2 * X1 + 0.2 * X2))
Y0 = rpois(N, lambda = exp(0.1 * X1 + 0.1 * X2))
tau = mean(Y1 - Y0)

# Y = Y1 * Z + Y0 * (1-Z)

# initialization
MC = 1e3
record_MPE = data.frame(
  tau = rep(tau, MC),
  tau_hat = rep(0, MC),
  var_hat = rep(0, MC),
```

```
  tau_hat_cf = rep(0, MC),
  var_hat_cf = rep(0, MC)
)
```

```
pb = txtProgressBar(min = 0, max = MC, initial = 0, style = 3)
```

```
##   |                                                                      |
```

```
for (iter in 1:MC){
  # setup progress bar
  setTxtProgressBar(pb, iter)

  # simulate randomization
  Zo = rbinom(K, 1, 0.5) # odd units
  Ze = 1 - Zo # even units

  Z = rep(0, N)
  Z[(1:K)*2-1] = Zo
  Z[(1:K)*2] = Ze

  # Obtain treated and control outcomes & covariates for each pair
  Yt = Y1[Z == 1]
  Yc = Y0[Z == 0]
  X1t = X1[Z == 1]
  X1c = X1[Z == 0]
  X2t = X2[Z == 1]
  X2c = X2[Z == 0]

  # Construct point estimates and variance estimation without crossfitting
  tau_hat = mean(Yt - Yc)
  var_hat = var(Yt - Yc)/K

  record_MPE$tau_hat[iter] = tau_hat
  record_MPE$var_hat[iter] = var_hat

  # cross-fitting
  ## split the strata into two groups: a and b
  Ka = 100
  Kb = K - Ka
  S = sample(c(rep("a", Ka), rep("b", Kb)))

  df_t_a = data.frame(
    Y = Yt[S == "a"],
    X1 = X1t[S == "a"],
    X2 = X2t[S == "a"]
  )
  df_t_b = data.frame(
    Y = Yt[S == "b"],
    X1 = X1t[S == "b"],
    X2 = X2t[S == "b"]
  )
  df_c_a = data.frame(
    Y = Yc[S == "a"],
    X1 = X1c[S == "a"],
```

```
    X2 = X2c[S == "a"]
  )
  df_c_b = data.frame(
    Y = Yc[S == "b"],
    X1 = X1c[S == "b"],
    X2 = X2c[S == "b"]
  )

  ## Model fitting
  model_t_a = lm("Y ~ X1 + X2", data = df_t_a)
  model_t_b = lm("Y ~ X1 + X2", data = df_t_b)
  model_c_a = lm("Y ~ X1 + X2", data = df_c_a)
  model_c_b = lm("Y ~ X1 + X2", data = df_c_b)

  ## Cross fitting
  pred_t_b_with_t_a = predict(model_t_a, newdata = rbind(df_t_b, df_c_b))
  pred_t_a_with_t_b = predict(model_t_b, newdata = rbind(df_t_a, df_c_a))
  pred_c_b_with_c_a = predict(model_c_a, newdata = rbind(df_t_b, df_c_b))
  pred_c_a_with_c_b = predict(model_c_b, newdata = rbind(df_t_a, df_c_a))

  ## cf estimates
  tau_hat_cf_a = mean(df_t_a$Y - pred_t_a_with_t_b[1:Ka]) + mean(pred_t_a_with_t_b) -
    (mean(df_c_a$Y - pred_c_a_with_c_b[(Ka + 1):(2*Ka)]) + mean(pred_c_a_with_c_b))
  tau_hat_cf_b = mean(df_t_b$Y - pred_t_b_with_t_a[1:Kb]) + mean(pred_t_b_with_t_a) -
    (mean(df_c_b$Y - pred_c_b_with_c_a[(Kb + 1):(2*Kb)]) + mean(pred_c_b_with_c_a))

  var_hat_cf_a = var(df_t_a$Y - pred_t_a_with_t_b[1:Ka] - (df_c_a$Y - pred_c_a_with_c_b[(Ka + 1):(2*Ka)]
  var_hat_cf_b = var(df_t_b$Y - pred_t_b_with_t_a[1:Kb] - (df_c_b$Y - pred_c_b_with_c_a[(Kb + 1):(2*Kb)]

  tau_hat_cf = Ka/K * tau_hat_cf_a + Kb/K * tau_hat_cf_b
  var_hat_cf = (Ka/K)^2 * var_hat_cf_a + (Kb/K)^2 * var_hat_cf_b

  ## record the results
  record_MPE$tau_hat_cf[iter] = tau_hat_cf
  record_MPE$var_hat_cf[iter] = var_hat_cf
}
```

```
##   |                                                                      |
```

```
# results checking
print(var(record_MPE$tau_hat))
```

```
## [1] 0.005761513
```

```
print(mean(record_MPE$var_hat))
```

```
## [1] 0.01073697
```

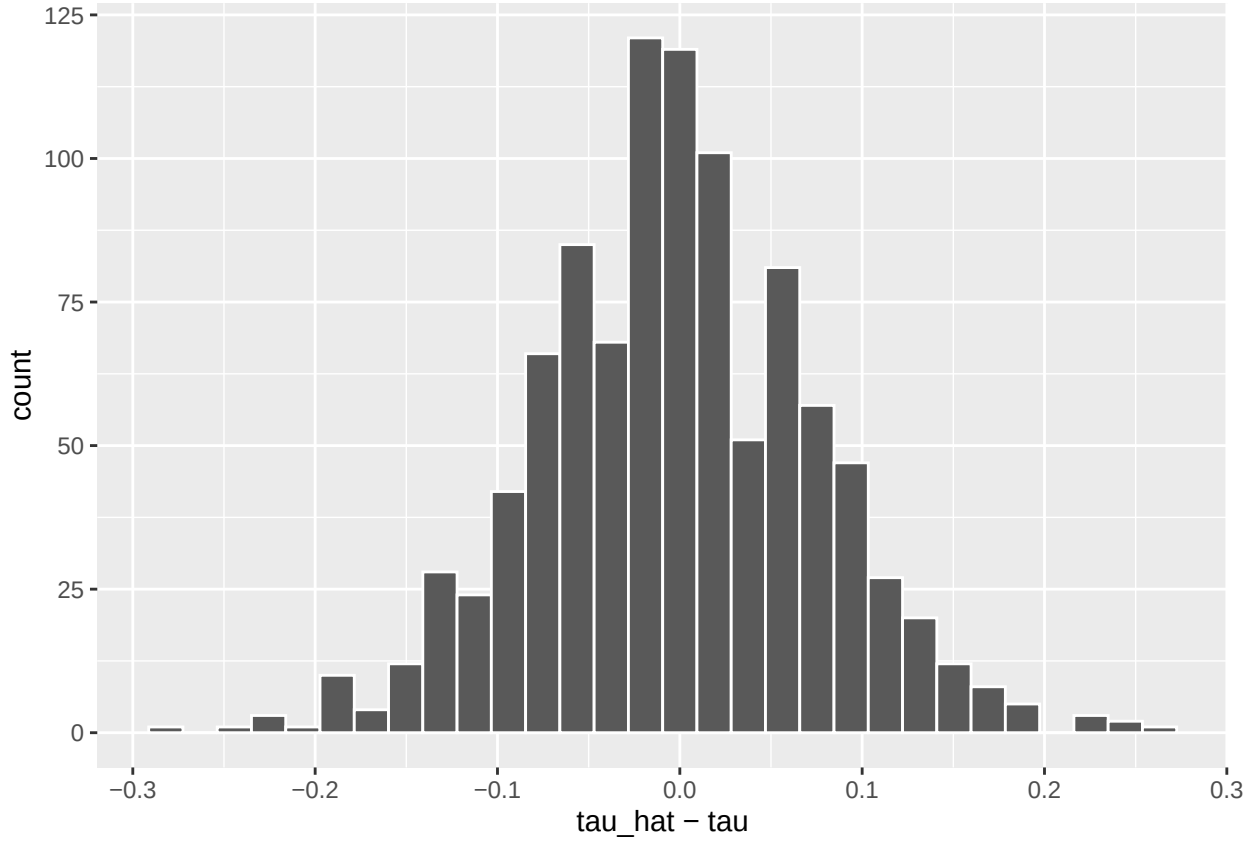
```
print(var(record_MPE$tau_hat_cf))
```

```
## [1] 0.005298129
```

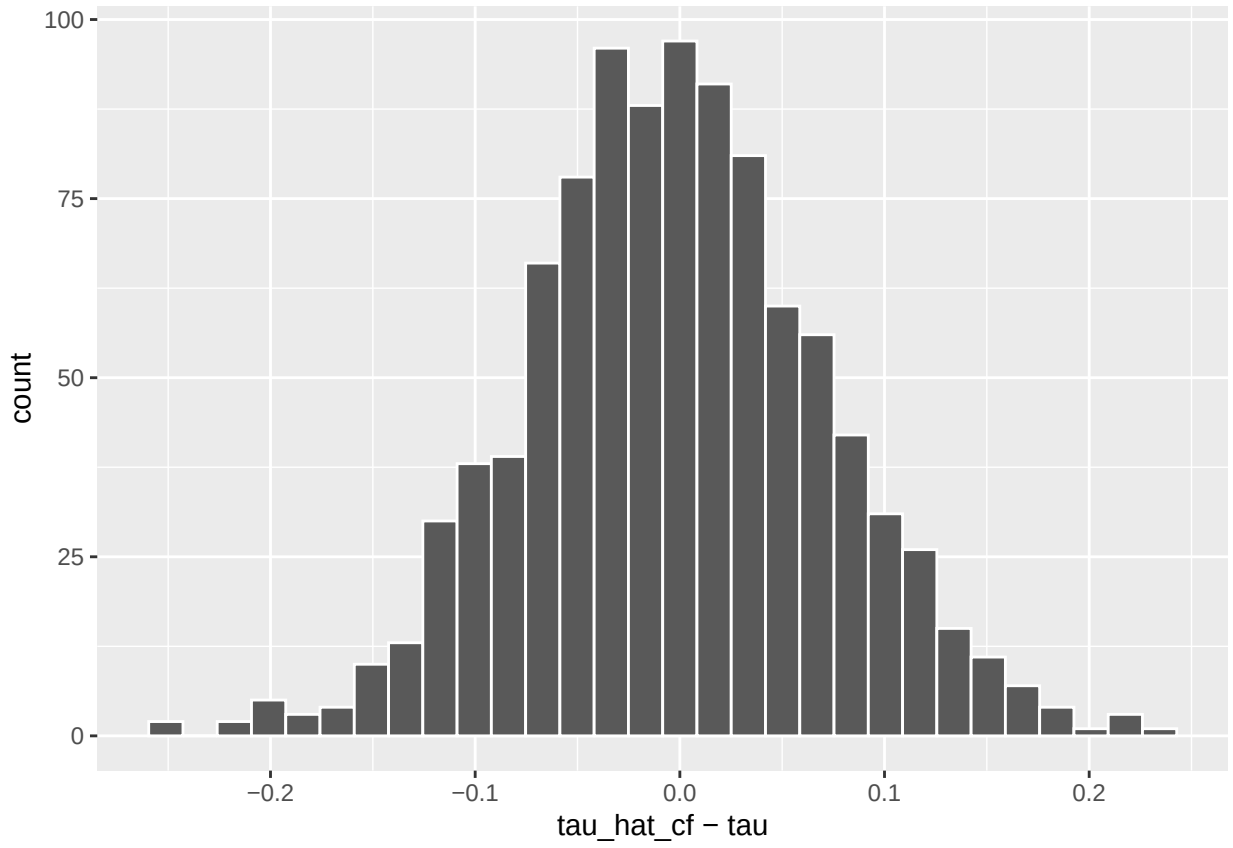
```
print(mean(record_MPE$var_hat_cf))
```

```
## [1] 0.01047374
```

```
record_MPE %>%
  ggplot() +
  geom_histogram(aes(x = tau_hat - tau), col = "white", bins = 30)
```




```
record_MPE %>%
  ggplot() +
  geom_histogram(aes(x = tau_hat_cf - tau), col = "white", bins = 30)
```


100
75
50
25
0
count
−0.2
−0.1
0.0
0.1
0.2
tau_hat_cf − tau