Project

Junhao Yu

2023-01-11

Data Prepare

```
library("ggplot2")

data = load("./OK/OK.Rdata")
data <- eval(parse(text = data))
data$gender = rep(0, length(data$Y))
data$year = rep(0, length(data$Y))
data$delta = data$Y - data$s_hsgrade3
data$strata = rep(0, length(data$Y))
data$seq = 1: length(data$Y)
data[data$s_group == "M_0" | data$s_group == "M_1", ]$gender = 1
data[data$s_group == "F_1" | data$s_group == "M_1", ]$year = 1
data[data$s_group == "F_1", ]$strata = 1
data[data$s_group == "M_0", ]$strata = 2
data[data$s_group == "M_1", ]$strata = 3
summary(data)
```

```
##        Y               Z            s_hsgrade3    s_group   s_mtongue_english
##  Min.   :18.57   Min.   :0.0000   Min.   :54.00   F_0:339   0:693
##  1st Qu.:64.11   1st Qu.:0.0000   1st Qu.:78.08   F_1:441   1:510
##  Median :70.50   Median :0.0000   Median :82.33   M_0:181
##  Mean   :69.73   Mean   :0.3175   Mean   :82.46   M_1:242
##  3rd Qu.:76.59   3rd Qu.:1.0000   3rd Qu.:87.17
##  Max.   :94.80   Max.   :1.0000   Max.   :98.00
##      gender            year            delta              strata
##  Min.   :0.0000   Min.   :0.0000   Min.   :-53.595   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:-17.375   1st Qu.:0.000
##  Median :0.0000   Median :1.0000   Median :-11.905   Median :1.000
##  Mean   :0.3516   Mean   :0.5677   Mean   :-12.732   Mean   :1.271
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: -7.333   3rd Qu.:2.000
##  Max.   :1.0000   Max.   :1.0000   Max.   : 22.595   Max.   :3.000
##       seq
##  Min.   :   1.0
##  1st Qu.: 301.5
##  Median : 602.0
##  Mean   : 602.0
##  3rd Qu.: 902.5
##  Max.   :1203.0
```

CRE

Random Check

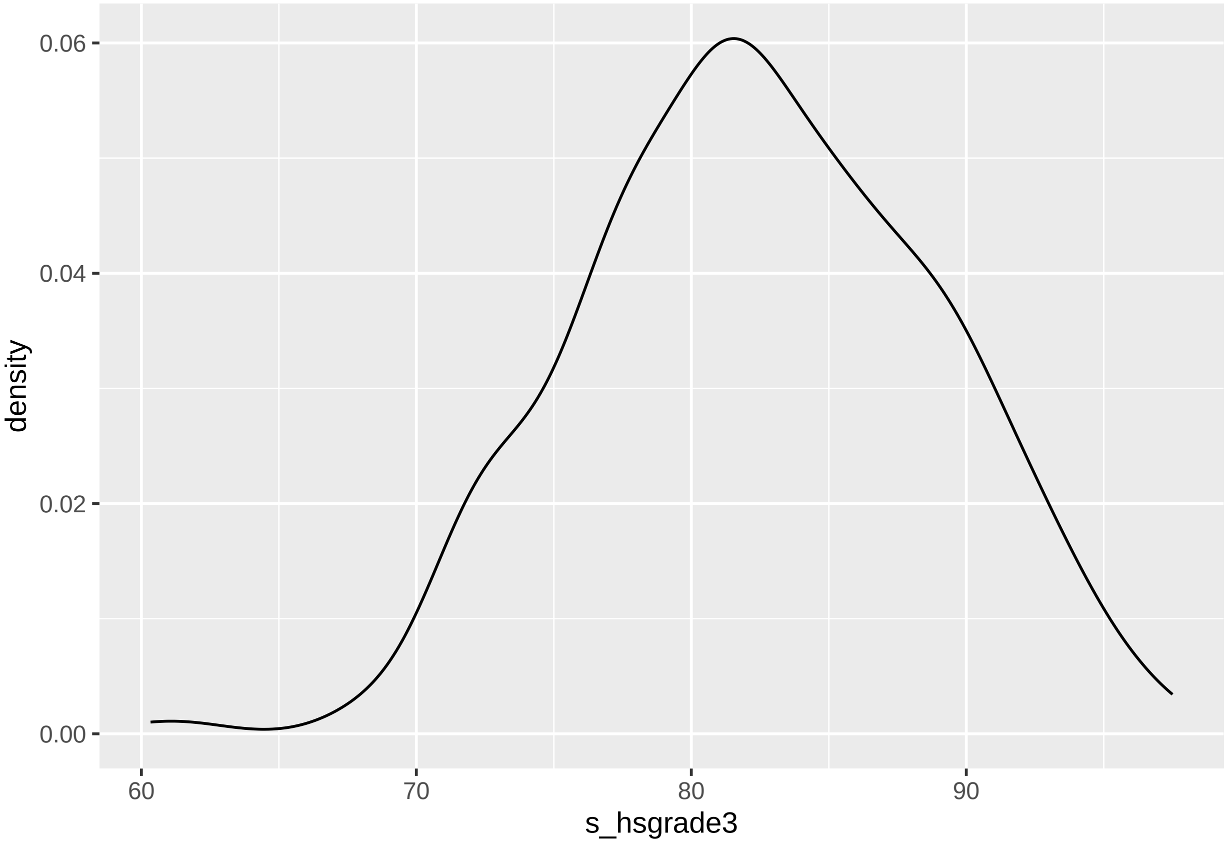
```r
t_data = data[data$Z == 1, ]
c_data = data[data$Z == 0, ]
summary(t_data)
```

```
##        Y               Z       s_hsgrade3     s_group   s_mtongue_english
##  Min.   :18.57   Min.   :1   Min.   :60.33   F_0: 92   0:220
##  1st Qu.:65.00   1st Qu.:1   1st Qu.:78.00   F_1: 99   1:162
##  Median :71.16   Median :1   Median :82.17   M_0: 91
##  Mean   :70.06   Mean   :1   Mean   :82.39   M_1:100
##  3rd Qu.:77.08   3rd Qu.:1   3rd Qu.:87.00
##  Max.   :93.40   Max.   :1   Max.   :97.50
##      gender         year             delta             strata
##  Min.   :0.0   Min.   :0.0000   Min.   :-53.595   Min.   :0.000
##  1st Qu.:0.0   1st Qu.:0.0000   1st Qu.:-17.000   1st Qu.:1.000
##  Median :0.5   Median :1.0000   Median :-11.150   Median :1.500
##  Mean   :0.5   Mean   :0.5209   Mean   :-12.335   Mean   :1.521
##  3rd Qu.:1.0   3rd Qu.:1.0000   3rd Qu.: -6.938   3rd Qu.:3.000
##  Max.   :1.0   Max.   :1.0000   Max.   : 22.595   Max.   :3.000
##       seq
##  Min.   :   1.0
##  1st Qu.: 295.2
##  Median : 568.0
##  Mean   : 591.4
##  3rd Qu.: 877.5
##  Max.   :1202.0
```

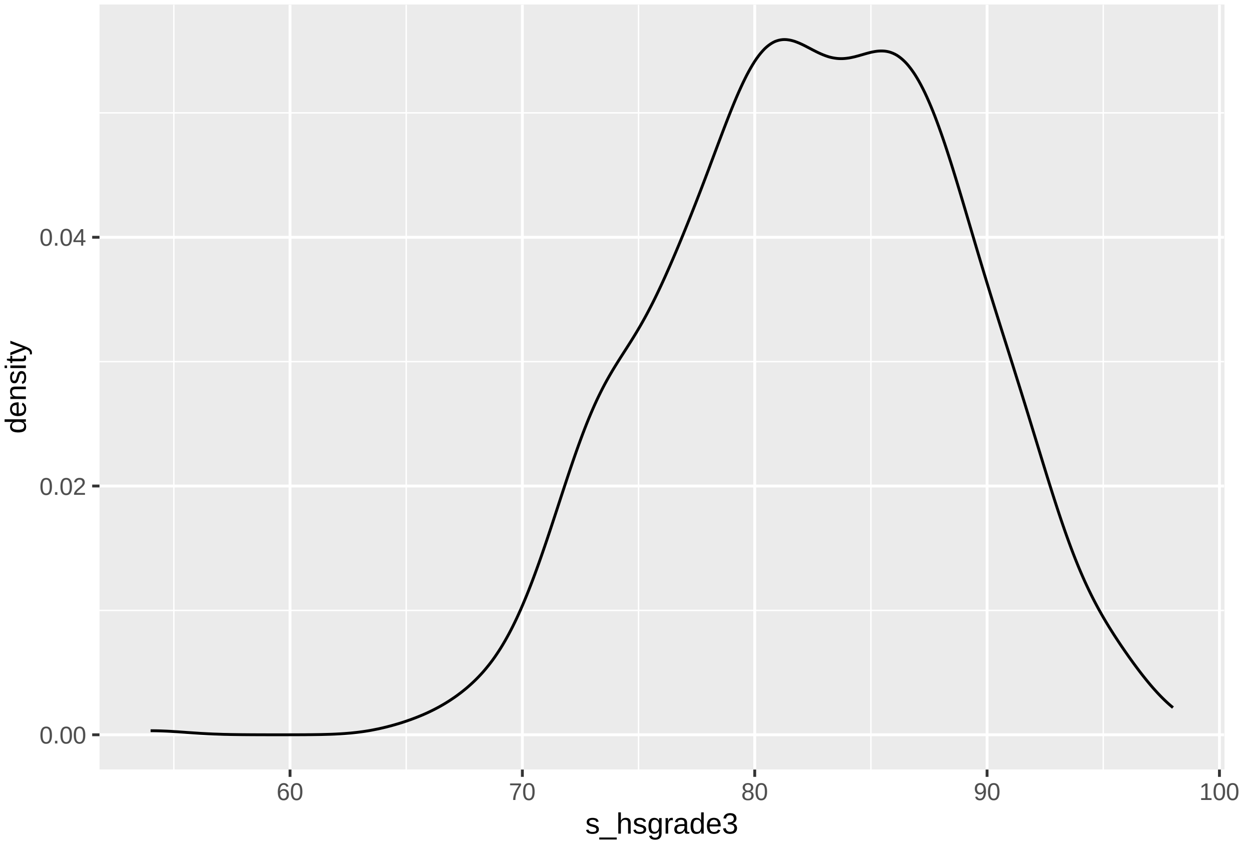
```r
summary(c_data)
```

```
##        Y               Z       s_hsgrade3     s_group   s_mtongue_english
##  Min.   :25.83   Min.   :0   Min.   :54.00   F_0:247   0:473
##  1st Qu.:63.88   1st Qu.:0   1st Qu.:78.17   F_1:342   1:348
##  Median :70.00   Median :0   Median :82.67   M_0: 90
##  Mean   :69.57   Mean   :0   Mean   :82.49   M_1:142
##  3rd Qu.:76.40   3rd Qu.:0   3rd Qu.:87.17
##  Max.   :94.80   Max.   :0   Max.   :98.00
##      gender            year             delta             strata
##  Min.   :0.0000   Min.   :0.0000   Min.   :-49.333   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:-17.778   1st Qu.:0.000
##  Median :0.0000   Median :1.0000   Median :-12.262   Median :1.000
##  Mean   :0.2826   Mean   :0.5895   Mean   :-12.917   Mean   :1.155
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: -7.625   3rd Qu.:2.000
##  Max.   :1.0000   Max.   :1.0000   Max.   : 15.167   Max.   :3.000
##       seq
##  Min.   :   2.0
##  1st Qu.: 306.0
##  Median : 616.0
##  Mean   : 606.9
##  3rd Qu.: 912.0
##  Max.   :1203.0
```

```
ggplot(data = t_data, aes(x = s_hsgrade3)) + geom_density()
```



```
ggplot(data = c_data, aes(x = s_hsgrade3)) + geom_density()
```



Fisher's Exact p-value

```
set.seed(42)
y_t = data[data$Z == 1, ]$Y
y_c = data[data$Z == 0, ]$Y
y_pool = data$Y
t_obs = abs(mean(y_t) - mean(y_c))
count = 0
for(i in 1:3000)
{
  y_t_sample = sample(y_pool, length(y_t))
  t_sample = abs(mean(y_t_sample) - (sum(y_pool) - sum(y_t_sample)) / length(y_c))
  if(t_sample > t_obs)
  {
    count = count + 1
  }
}
print(count/3000)
```

```
## [1] 0.4323333
```

Neyman

```
ate_hat = mean(y_t) - mean(y_c)
var_hat = var(y_t)/length(y_t) + var(y_c)/length(y_c)
print(ate_hat)
```

```
## [1] 0.487178
```

```
print(sqrt(var_hat))
```

```
## [1] 0.6230992
```

```
print(qnorm(0.975))
```

```
## [1] 1.959964
```

Regression

```
reg_cre1 = lm(Y ~ Z + s_hsgrade3 + s_mtongue_english + gender + year, data = data)
summary(reg_cre1)
```

```
##
## Call:
## lm(formula = Y ~ Z + s_hsgrade3 + s_mtongue_english + gender +
##     year, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.341  -4.296   0.737   5.272  30.062
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.58157    3.06224  -0.190   0.8494
## Z                   -0.02409    0.51042  -0.047   0.9624
## s_hsgrade3           0.87329    0.03673  23.777  < 2e-16 ***
## s_mtongue_english1 -0.65266    0.46890  -1.392   0.1642
## gender               1.62592    0.49713   3.271   0.0011 **
## year                -3.50463    0.46916  -7.470 1.54e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.036 on 1197 degrees of freedom
## Multiple R-squared:  0.3399, Adjusted R-squared:  0.3372
## F-statistic: 123.3 on 5 and 1197 DF,  p-value: < 2.2e-16
```

Bayesian

Check the Balance of Variable delta

```
delta_c = data[data$Z == 0, ]$delta
delta_t = data[data$Z == 1, ]$delta
length(y_c)
```

```
## [1] 821
```

```
length(y_t)
```

```
## [1] 382
```

```
mean(delta_c)
```

```
## [1] -12.91667
```

```r
mean(delta_t)
```

```
## [1] -12.33508
```

```r
var(delta_c)
```

```
## [1] 62.69825
```

```r
var(delta_t)
```

```
## [1] 81.93903
```

```r
set.seed(42)
mu_std = 15
y_t_std = 9
y_c_std = 8
t_bay = rep(0, 1000)
for(i in 1:1000){
  mu_t = rnorm(1, 0, mu_std)
  mu_c = rnorm(1, 0, mu_std)
  sample_bay = data[sample(nrow(data), 20), ]
  for(j in 1:20){
    if(as.numeric(sample_bay[j, "Z"]) == 1){
      t_bay[i] = t_bay[i] + (as.numeric(sample_bay[j, "Y"]) - rnorm(1, mu_c, y_c_std))
    }else{
      t_bay[i] = t_bay[i] + (rnorm(1, mu_t, y_t_std) - as.numeric(sample_bay[j, "Y"]))
    }
  }
  t_bay[i] = t_bay[i] / 20
}
ate_bay = mean(t_bay)
var_bay = var(t_bay)
ate_bay
```

```
## [1] -25.09619
```

```r
sqrt(var_bay)
```

```
## [1] 18.4618
```

SRE

Neyman

```r
data_00 = data[data$gender == 0 & data$year == 0, ]
data_01 = data[data$gender == 0 & data$year == 1, ]
data_10 = data[data$gender == 1 & data$year == 0, ]
data_11 = data[data$gender == 1 & data$year == 1, ]

y_00_t = data_00[data_00$Z == 1, ]$Y
y_00_c = data_00[data_00$Z == 0, ]$Y
ate_00_hat = mean(y_00_t) - mean(y_00_c)
var_00_hat = var(y_00_t)/length(y_00_t) + var(y_00_c)/length(y_00_c)
print("00")
```

```
## [1] "00"
```

```
print(ate_00_hat)
```

```
## [1] 0.9699417
```

```
print(sqrt(var_00_hat))
```

```
## [1] 1.092539
```

```
print(qnorm(0.975))
```

```
## [1] 1.959964
```

```
y_01_t = data_01[data_01$Z == 1, ]$Y
y_01_c = data_01[data_01$Z == 0, ]$Y
ate_01_hat = mean(y_01_t) - mean(y_01_c)
var_01_hat = var(y_01_t)/length(y_01_t) + var(y_01_c)/length(y_01_c)
print("01")
```

```
## [1] "01"
```

```
print(ate_01_hat)
```

```
## [1] -0.2127408
```

```
print(sqrt(var_01_hat))
```

```
## [1] 1.192675
```

```
print(qnorm(0.975))
```

```
## [1] 1.959964
```

```
y_10_t = data_10[data_10$Z == 1, ]$Y
y_10_c = data_10[data_10$Z == 0, ]$Y
ate_10_hat = mean(y_10_t) - mean(y_10_c)
var_10_hat = var(y_10_t)/length(y_10_t) + var(y_10_c)/length(y_10_c)
print("10")
```

```
## [1] "10"
```

```
print(ate_10_hat)
```

```
## [1] 0.4703871
```

```
print(sqrt(var_10_hat))
```

```
## [1] 1.331536
```

```
print(qnorm(0.975))
```

```
## [1] 1.959964
```

```
y_11_t = data_11[data_11$Z == 1, ]$Y
y_11_c = data_11[data_11$Z == 0, ]$Y
ate_11_hat = mean(y_11_t) - mean(y_11_c)
var_11_hat = var(y_11_t)/length(y_11_t) + var(y_11_c)/length(y_11_c)
print("10")
```

```
## [1] "10"
```

```
print(ate_11_hat)
```

```
## [1] -0.9562276
```

```
print(sqrt(var_11_hat))
```

```
## [1] 1.404488
```

```
print(qnorm(0.975))
```

```
## [1] 1.959964
```

```
ate_hat = (length(y_00_c) + length(y_00_t))/(length(data$Y)) *ate_00_hat + (length(y_01_c)  + length(y_
var_ate = ((length(y_00_c) + length(y_00_t))/(length(data$Y)))^2 *var_00_hat + ((length(y_01_c) + lengt

print(ate_hat)
```

```
## [1] 0.07375274
```

```
print(sqrt(var_ate))
```

```
## [1] 0.637105
```

Regression

```
reg_sre1 = lm(Y ~ Z + s_hsgrade3, data = data_00)
summary(reg_sre1)
```

```
##
## Call:
## lm(formula = Y ~ Z + s_hsgrade3, data = data_00)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.429  -4.088   0.109   5.162  18.683
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.73819    5.62926   3.151  0.00177 **
## Z            0.65503    0.89379   0.733  0.46415
## s_hsgrade3   0.64872    0.06821   9.511  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.313 on 336 degrees of freedom
## Multiple R-squared:  0.2143, Adjusted R-squared:  0.2096
## F-statistic: 45.82 on 2 and 336 DF,  p-value: < 2.2e-16
```

```
reg_sre2 = lm(Y ~ Z + s_hsgrade3, data = data_01)
summary(reg_sre2)
```

```
##
## Call:
## lm(formula = Y ~ Z + s_hsgrade3, data = data_01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.271  -4.137   0.999   5.544  21.763
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -21.73394    4.86707  -4.466 1.02e-05 ***
## Z            -0.47052    0.90886  -0.518    0.605    
## s_hsgrade3    1.08148    0.05852  18.480  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.963 on 438 degrees of freedom
## Multiple R-squared:  0.4381, Adjusted R-squared:  0.4356 
## F-statistic: 170.8 on 2 and 438 DF,  p-value: < 2.2e-16
```

```r
reg_sre3 = lm(Y ~ Z + s_hsgrade3, data = data_10)
summary(reg_sre3)
```

```
## 
## Call:
## lm(formula = Y ~ Z + s_hsgrade3, data = data_10)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -32.558  -4.170   0.543   4.475  26.465 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 13.81808    7.26702   1.901   0.0589 .  
## Z            0.65169    1.14662   0.568   0.5705    
## s_hsgrade3   0.70340    0.08813   7.981 1.72e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.711 on 178 degrees of freedom
## Multiple R-squared:  0.2641, Adjusted R-squared:  0.2558 
## F-statistic: 31.93 on 2 and 178 DF,  p-value: 1.409e-12
```

```r
reg_sre4 = lm(Y ~ Z + s_hsgrade3, data = data_11)
summary(reg_sre4)
```

```
## 
## Call:
## lm(formula = Y ~ Z + s_hsgrade3, data = data_11)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -37.676  -4.781   1.656   5.734  16.831 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -2.46737    7.37512  -0.335    0.738    
## Z           -0.64411    1.17301  -0.549    0.583    
## s_hsgrade3   0.87773    0.08902   9.859   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.982 on 239 degrees of freedom
## Multiple R-squared:  0.2905, Adjusted R-squared:  0.2846
## F-statistic: 48.94 on 2 and 239 DF,  p-value: < 2.2e-16
```

```
(length(y_00_c) + length(y_00_t))/(length(data$Y)) *as.numeric(reg_sre1$coefficients["Z"]) + (length(y_(
```

```
## [1] -0.01942201
```

```
var_ate = ((length(y_00_c) + length(y_00_t))/(length(data$Y)))^2 *0.89379^2 + ((length(y_01_c) + length

print(sqrt(var_ate))
```

```
## [1] 0.5097877
```

Pairwise

```
set.seed(42)
pair1 = rep(0, length(data$Y))
t_pair = rep(0, length(data$Y))
j = 1

for(i in 1: length(data$Y)){
  pair_data = data[as.numeric(data[i, "s_hsgrade3"]) - 0.1 < data$s_hsgrade3 & data$s_hsgrade3 < as.nume
  if(length(pair_data$Y) > 0){
    randomPair = round(runif(1, 1, length(pair_data$Y)))
    pair1[i] = as.numeric(pair_data[randomPair, "seq"])
    if(as.numeric(data[i, "Z"]) == 1){
      t_pair[j] = as.numeric(data[i, "Y"]) - as.numeric(pair_data[randomPair, "Y"])
    }
    else{
      t_pair[j] = as.numeric(pair_data[randomPair, "Y"]) - as.numeric(data[i, "Y"])
    }
    j = j + 1
  }
}

t_pair = t_pair[1: j - 1]
ate_pair = mean(t_pair)
var_pair = 1/((j - 1)) * var(t_pair)
```

```
ate_pair
```

```
## [1] 0.305334
```

```
sqrt(var_pair)
```

```
## [1] 0.4014235
```