Assignment 2b: Randomized Experiment Simulation

Yongchao Zhao

9/18/2018

Problem 1

Recall that in Assignment 1 we created a simulated dataset that could have manifested as a result of a completely randomized experiment. In that assignment, we asked about the difference between estimating ATE by using the difference in means versus using linear regression with pretest score as a covariate. In this exercise, we will explore the properties of these two different approaches to estimating our ATEs more deeply through simulation. We would like you to compare these approaches with respect to both bias and efficiency.

(a) Write a function to generate the data generating process (DGP) from Assignment 1 with arguments for sample size, the coefficient on the covariate, and the random seed. Then use this function to simulate a data set with sample size equal to 100, seed equal to 1234, and the coefficient on the covariate set to 1.1.

```
# function of data generating process (DGP)
dgp <- function(n, seed, coef){
      set.seed(seed)
      pre_score <- rnorm(n, mean=65, sd=3)

      error_0 <- rnorm(n, mean = 0, sd = 1)
      error_1 <- rnorm(n, mean = 0, sd = 1)

      post_0 <- 10 + coef*pre_score + 0 + error_0
      post_1 <- 10 + coef*pre_score + 5 + error_1

      return (data.frame(stuID = 1:n, pre_score, post_0, post_1))
}

#  simulated dataset with sample size of 100, seed equal to 1234, and the
coefficient on the covariate set to 1.1

mydata <- dgp(n=100, seed=1234, coef=1.1)

head(mydata)

##   stuID pre_score   post_0   post_1
## 1     1  61.37880 77.93121 83.00191
## 2     2  65.83229 81.94080 88.11229
## 3     3  68.25332 85.14465 90.26417
## 4     4  57.96291 73.25672 79.45993
## 5     5  66.28737 82.09011 88.22779
## 6     6  66.51817 83.33697 88.93045
```

```
summary(mydata[, -1])

##    pre_score         post_0          post_1     
##  Min.   :57.96   Min.   :73.26   Min.   :76.07  
##  1st Qu.:62.31   1st Qu.:78.69   1st Qu.:83.52  
##  Median :63.85   Median :80.44   Median :85.60  
##  Mean   :64.53   Mean   :81.02   Mean   :86.14  
##  3rd Qu.:66.41   3rd Qu.:83.19   3rd Qu.:88.33  
##  Max.   :72.65   Max.   :89.74   Max.   :95.38  

# calculate SATE

(SATE_t <- mean(mydata$post_1) - mean(mydata$post_0))

## [1] 5.11336
```

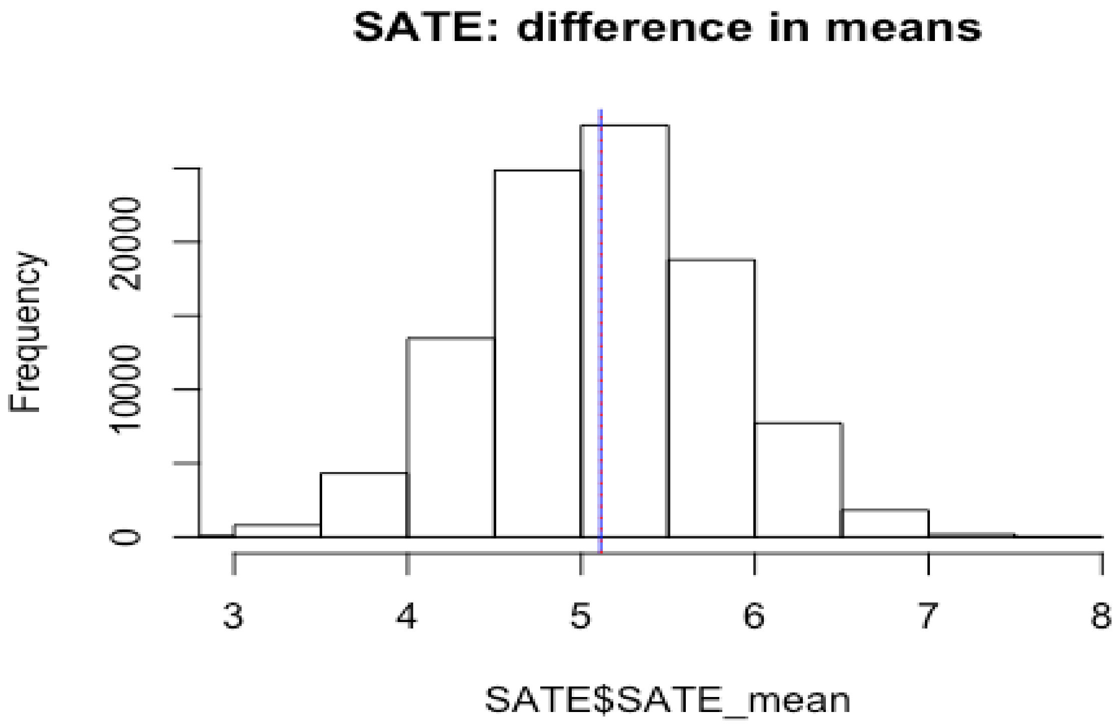
(b) We will now investigate the properties of two estimators of the SATE.

- difference in means
- linear regression estimate of the treatment effect using the pretest score as a covariate

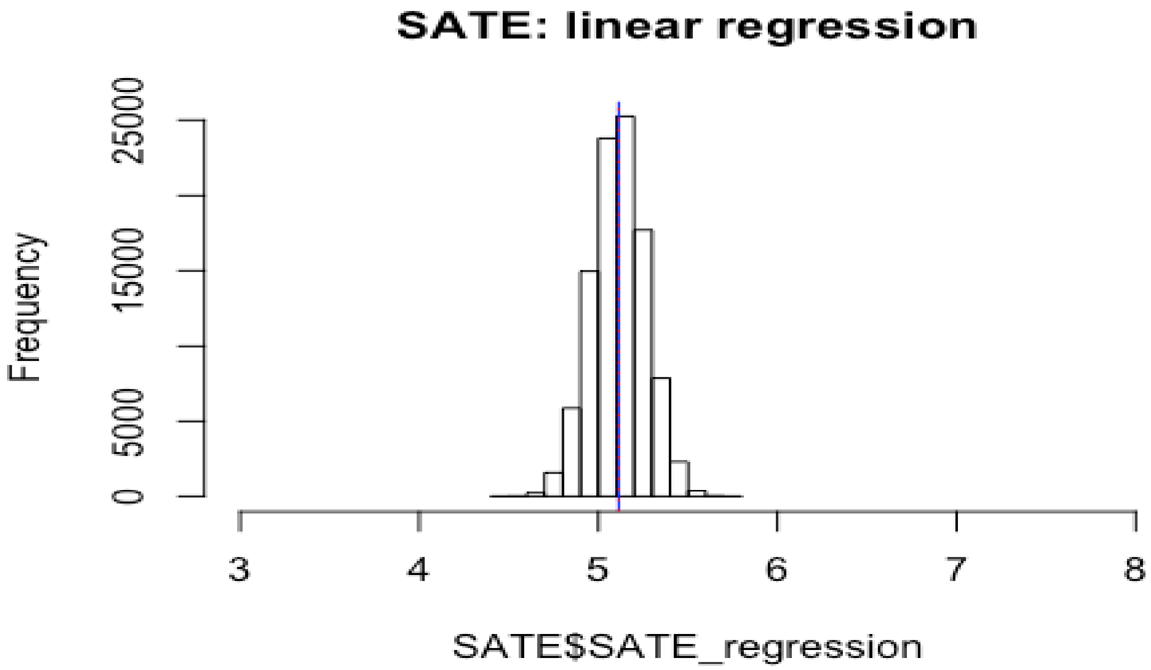
```
# Monte Carlo simulation (100,000 times)
repeats <- 100000
SATE <- data.frame(SATE_mean = rep(NA, repeats), SATE_regression =
rep(NA,repeats))

for (i in 1:repeats){
  newdata <- mydata
  selected <- sample(newdata$stuID, 50, replace = F)
  newdata$treatment <- ifelse(newdata$stuID %in% selected, 1, 0)
  newdata$outcome <- ifelse(newdata$treatment == 1, newdata$post_1,
                            newdata$post_0)

  avgs <- tapply(newdata$outcome, newdata$treatment, mean)
  SATE$SATE_mean[i] <- avgs[[2]] - avgs[[1]]

  mymodel<- lm(outcome ~ treatment + pre_score, data = newdata)
  SATE$SATE_regression[i] <- coef(mymodel)[[2]]
}

summary(SATE)
##    SATE_mean     SATE_regression
##  Min.   :2.439   Min.   :4.456  
##  1st Qu.:4.650   1st Qu.:5.011  
##  Median :5.112   Median :5.113  
##  Mean   :5.111   Mean   :5.113  
##  3rd Qu.:5.573   3rd Qu.:5.215  
##  Max.   :7.865   Max.   :5.721  
```

(c) Plot the (Monte Carlo estimate of the) randomization distribution for each of the two estimators: difference in means and regression. Either overlay the plots (with different colors for each) or make sure the xlim on both plots is the same. Also add vertical lines (using different colors) for the SATE and the mean of the randomizaton distribution.

```
# plots
hist(SATE$SATE_mean, xlim = c(3, 8), main = "SATE: difference in means")
abline(v = c(mean(SATE$SATE_mean), SATE_t), col = c("blue", "red"), lty =
c(1, 3))
```



```
hist(SATE$SATE_regression, xlim = c(3, 8), main = "SATE: linear regression")
abline(v = c(mean(SATE$SATE_regression), SATE_t), col = c("blue", "red"),lty
= c(1,3))
```



(d) What is the bias and efficiency of each of these two methods? What is the difference between them?

According to the output shown below, the linear regression method with pretest score as a covariate has smaller bias and efficiency values than the difference in means method. Thus, the linear regression approach yields a better estimate of the SATE.

```
# bias and efficiency for method 1: difference in means
(mean_bias <- mean(SATE$SATE_mean)- SATE_t)

## [1] -0.002845254

(mean_se <- sd(SATE$SATE_mean)/sqrt(repeats)) # standard error

## [1] 0.002154049

# bias and efficiency for method 2: linear regression using pretest score as
covariate
(regression_bias <- mean(SATE$SATE_regression)- SATE_t)

## [1] -0.0002056018

(regression_se <- sd(SATE$SATE_regression)/sqrt(repeats)) # standard error

## [1] 0.0004722776
```

(e) Re-run the simulation with a small coefficient (even 0) for the pretest covariate. Does the small coefficient lead to a different bias and efficiency estimate compared to when the coefficient for pretest was at 1.1 from before?

According to the output shown below, compared with pretest covariate coefficient equals to 1.1,

- **For the difference in means approach, the small coefficient (coef =0) leads to a different bias and efficiency estimate, both are smaller (bias: 0.00285 vs. 0.00033; efficiency: 0.00215 vs. 0.00047);**
- **For the linear regression approach, the small coefficient (coef=0) does not yield a different bias and efficiency estimate, both stay the same as coef =1.1 .**

```r
# new data with smaller coefficient for pretest coveriate (coef = 0)
mydata2 <- dgp(n=100, seed=1234, coef= 0)
head(mydata2)

##   stuID pre_score    post_0   post_1
## 1     1  61.37880 10.414524 15.48523
## 2     2  65.83229  9.525282 15.69677
## 3     3  68.25332 10.065993 15.18551
## 4     4  57.96291  9.497522 15.70073
## 5     5  66.28737  9.174001 15.31168
## 6     6  66.51817 10.166989 15.76046

summary(mydata2[, -1])

##    pre_score         post_0           post_1
##  Min.   :57.96   Min.   : 7.144   Min.   :11.77
##  1st Qu.:62.31   1st Qu.: 9.441   1st Qu.:14.62
##  Median :63.85   Median :10.033   Median :15.28
##  Mean   :64.53   Mean   :10.041   Mean   :15.15
##  3rd Qu.:66.41   3rd Qu.:10.628   3rd Qu.:15.68
##  Max.   :72.65   Max.   :13.044   Max.   :17.92

# calculate SATE

(SATE_t2 <- mean(mydata2$post_1) - mean(mydata2$post_0))

## [1] 5.11336

# Monte Carlo simulation (100,000 times)
SATE_2 <- data.frame(SATE_mean = rep(NA, repeats), SATE_regression =
rep(NA,repeats))
for (i in 1:repeats){
  newdata <- mydata2
  selected <- sample(newdata$stuID, 50, replace = F)
  newdata$treatment <- ifelse(newdata$stuID %in% selected, 1, 0)
  newdata$outcome <- ifelse(newdata$treatment == 1, newdata$post_1,
                            newdata$post_0)
  avgs <- tapply(newdata$outcome, newdata$treatment, mean)
  SATE_2$SATE_mean[i] <- avgs[[2]] - avgs[[1]]

  mymodel<- lm(outcome ~ treatment + pre_score, data = newdata)
  SATE_2$SATE_regression[i] <- coef(mymodel)[[2]]
}

summary(SATE_2)

##    SATE_mean      SATE_regression
##  Min.   :4.464   Min.   :4.456
##  1st Qu.:5.013   1st Qu.:5.011
##  Median :5.114   Median :5.113
##  Mean   :5.114   Mean   :5.113
##  3rd Qu.:5.214   3rd Qu.:5.215
##  Max.   :5.730   Max.   :5.721

# new bias and efficiency for method 1:difference in means
(mean_bias2 <- mean(SATE_2$SATE_mean)- SATE_t2)

## [1] 0.0003322121

(mean_se2 <- sd(SATE_2$SATE_mean)/sqrt(repeats))

## [1] 0.0004656902

# new bias and efficiency for method 2: linear regression using pretest score
as covariate
(regression_bias2 <- mean(SATE_2$SATE_regression)- SATE_t2)

## [1] -0.0002056018

(regression_se2 <- sd(SATE_2$SATE_regression)/sqrt(repeats))

## [1] 0.0004722776
```

Problem 2

In a randomized block design, randomization occurs separately within blocks. In many situations, the ratio of treatment to control observations is different across blocks. In addition, the treatment effect may vary across sites. For this problem, you will simulate data sets for a randomized block design that includes a binary indicator for female as a blocking variable. You will then estimate the ATE with two estimators: one that accounts for the blocking structure and one that does not. You will compare the bias and efficiency of these estimators. We will walk you through this in steps.

(a) First simulate the blocking variable and potential outcomes. In particular:
• Set the seed to by 1234
• Generate female as blocking variable (Female vs. Other Ratio (30:70))
• Generate Y(0) and Y(1) with the following features: – the intercept is 70 – the residual standard deviation is 1.

– treatment effect varies by block: observations with female=1 have treatment effect of 7 and those with female=0 have a treatment effect of 3. [Hint: Note that we are assuming that being female affects treatment effect but does not directly affect the average test score otherwise.]

```r
set.seed(1234)
# simulate dataset
female <- as.factor(c(rep(1, 30), rep(0, 70)))
y0 <- rnorm(100, mean=70, sd=1)
y1 <- c(rnorm(30, mean=77, sd=1),rnorm(70, mean=73, sd=1))

mydata <- data.frame(female, y0, y1)

# check the dataset
head(mydata)
```

```
##   female       y0       y1
## 1      1 68.79293 77.41452
## 2      1 70.27743 76.52528
## 3      1 71.08444 77.06599
## 4      1 67.65430 76.49752
## 5      1 70.42912 76.17400
## 6      1 70.50606 77.16699
```

```r
summary(mydata)
```

```
##  female       y0               y1
##  0:70   Min.   :67.65   Min.   :70.14
##  1:30   1st Qu.:69.10   1st Qu.:72.68
##         Median :69.62   Median :73.62
##         Mean   :69.84   Mean   :74.24
##         3rd Qu.:70.47   3rd Qu.:76.19
##         Max.   :72.55   Max.   :79.06
```

```r
tapply(mydata$y1, mydata$female, mean)
```

```
##        0        1
## 73.07429 76.96413
```

```r
tapply(mydata$y1, mydata$female, sd)
```

```
##         0         1
## 1.1155972 0.8167716
```

(b) Calculate the overall SATE and the SATE for each block.

According to the output:

- **The overall SATE is 4.40;**
- **For Female =0 block, the SATE = 3.17;**
- **For Female =1 block, the SATE = 7.26.**

```
# overall SATE
(SATE_all <- mean(mydata$y1 -mydata$y0))

## [1] 4.398005

# SATE for each block
library(dplyr)

mydata %>% group_by(female) %>% summarise(mean(y1-y0))

## # A tibble: 2 x 2
##   female   `mean(y1 - y0)`
##   <fct>              <dbl>
## 1 0                   3.17
## 2 1                   7.26
```

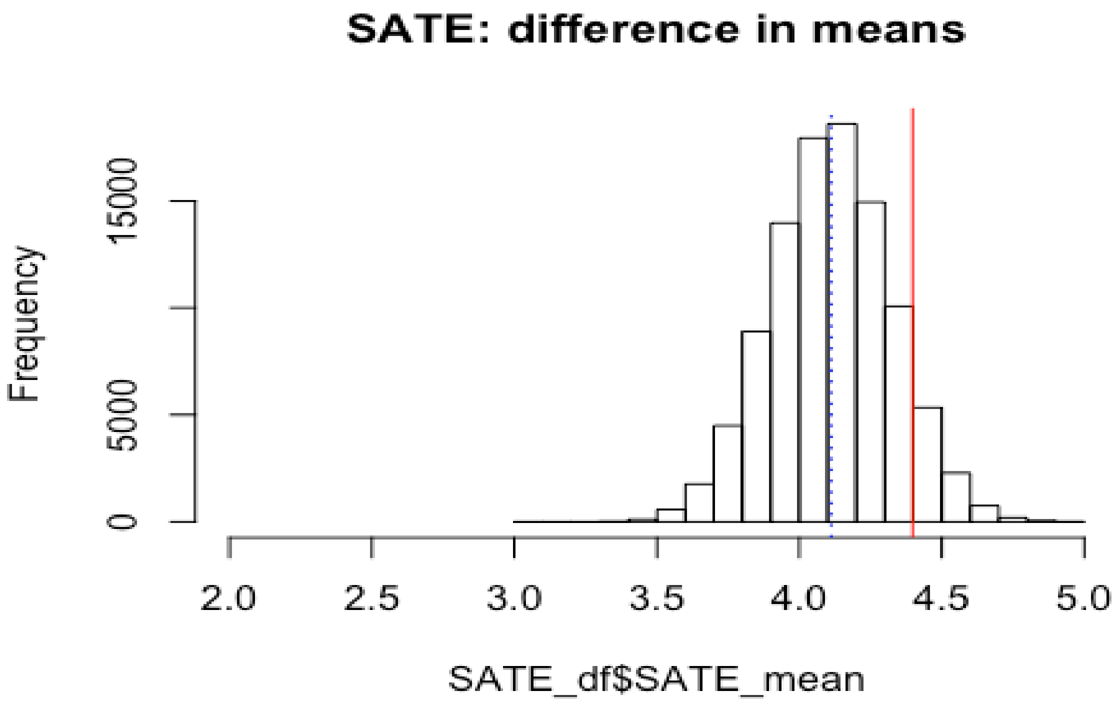
Now create a function for assigning the treatment. In particular, within each block create different assignment probabilities:
* Pr(Z=1 | female=0) = .6
* Pr(Z=1 | female=1) = .4
Generate the treatment and create a vector for the observed outcomes implied by that treatment. We will use this to create a randomization distribution for two different estimators for the SATE. Obtain 100,000 draws from that distribution.

```
assign.treatment <- function(dataset){
   newdata <- dataset
   # assign treatment with specific probabilities
   newdata$treatment <- NA
   newdata[newdata$female==0,]$treatment <- sample(c(1,0),
             nrow(newdata[newdata$female==0,]), replace=T, prob=c(0.6, 0.4))

   newdata[newdata$female==1,]$treatment <- sample(c(1,0),
             nrow(newdata[newdata$female==1,]),replace=T, prob=c(0.4, 0.6))

   # create observed outcome variable
   newdata$outcome <- ifelse(newdata$treatment == 1, newdata$y1, newdata$y0)
   return(newdata)
}
# test the function
head(assign.treatment(mydata))

##   female       y0       y1 treatment  outcome
## 1      1 68.79293 77.41452         0 68.79293
## 2      1 70.27743 76.52528         0 70.27743
## 3      1 71.08444 77.06599         0 71.08444
## 4      1 67.65430 76.49752         1 76.49752
## 5      1 70.42912 76.17400         0 70.42912
## 6      1 70.50606 77.16699         1 77.16699
```

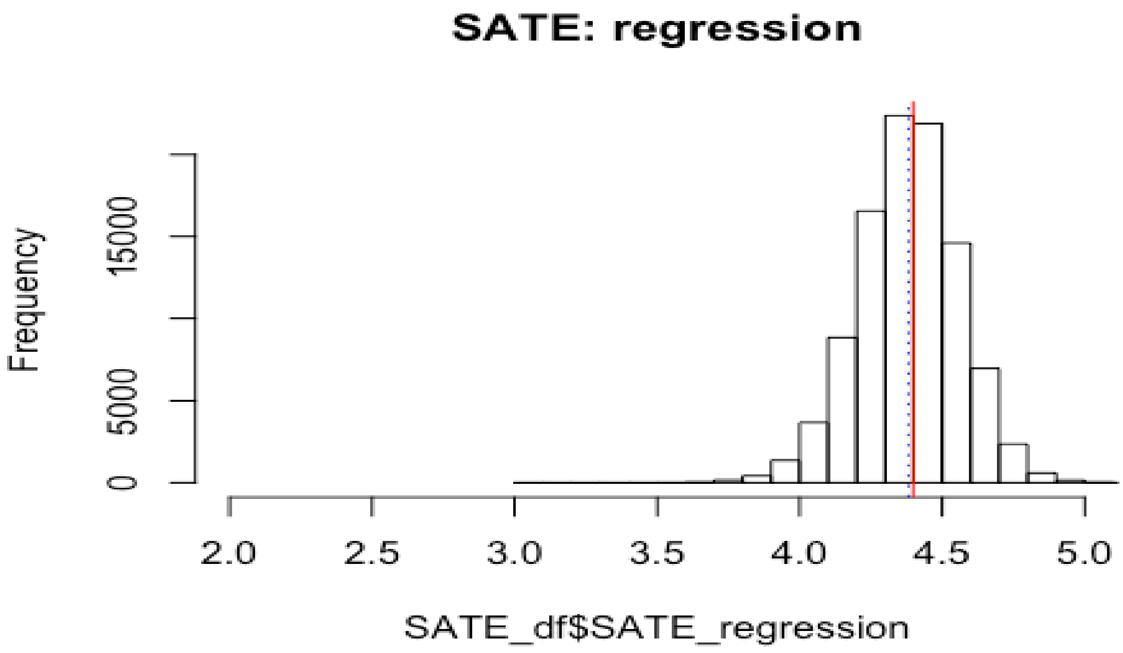
```
table(treatment = assign.treatment(mydata)$treatment, female =
assign.treatment(mydata)$female)

##            female
## treatment    0  1
##         0   41 16
##         1   29 14
```

(c) Plot the (Monte Carlo estimate of the) randomization distribution for each of the two estimators: difference in means and regression. Either overlay the plots (with different colors for each) or make sure the xlim on both plots is the same.

```
times <- 100000
SATE_df <- data.frame(SATE_mean=rep(NA, times), SATE_regression = rep(NA,
times))

for (i in 1:times){
  mydata_new <- assign.treatment(mydata)

  # calculate SATE: difference in means
  avgs <- tapply(mydata_new$outcome, mydata_new$treatment, mean)
  SATE_df$SATE_mean[i] <- avgs[[2]] - avgs[[1]]

  # calculate SATE: regression
  SATE_df$SATE_regression[i] <- coef(lm(outcome ~ treatment + female,
                                              data= mydata_new))[[2]]
}

summary(SATE_df)
##    SATE_mean     SATE_regression
##  Min.   :3.081   Min.   :3.082
##  1st Qu.:3.969   1st Qu.:4.268
##  Median :4.112   Median :4.385
##  Mean   :4.112   Mean   :4.381
##  3rd Qu.:4.254   3rd Qu.:4.498
##  Max.   :4.981   Max.   :5.123

# plots
hist(SATE_df$SATE_mean, xlim=c(2,5), main = "SATE: difference in means")
abline(v=c(mean(SATE_df$SATE_mean), SATE_all), col= c("blue", "red"), lty =
c(3,1))
```



```
hist(SATE_df$SATE_regression, xlim=c(2,5), main = "SATE: regression")
abline(v= c(mean(SATE_df$SATE_regression), SATE_all), col= c("blue", "red"),
lty=c(3,1))
```



(d) Calculate the bias and efficiency of each estimator. Also calculate the root mean squared error.

```
# bias, efficiency and RMSE of SATE estimator using differences in means
(mean_bias <- mean(SATE_df$SATE_mean - SATE_all)) #bias
## [1] -0.2860087

(mean_se <- sd(SATE_df$SATE_mean)/sqrt(times)) #efficiency
## [1] 0.0006656124

(mean_rmse <- sqrt(mean((SATE_df$SATE_mean - SATE_all)^2))) #rmse
## [1] 0.355112

# bias, efficiency and RMSE of SATE estimator using regression
(regression_bias <- mean(SATE_df$SATE_regression - SATE_all))#bias
## [1] -0.01671058

(regression_se <- sd(SATE_df$SATE_regression)/sqrt(times)) #efficiency
## [1] 0.0005552788

(regression_rmse <- sqrt(mean((SATE_df$SATE_regression - SATE_all)^2))) #rmse
## [1] 0.176387
```

(e) Why is the estimator that ignores blocks biased? Is the efficiency meaningful here? Why did I have you calculate the RMSE?

According to the results from (d) and the plots from (c), the SATE estimator using differences in means which ignores blocks is biased, because the ratio of treated to control groups is different across blocks (i.e., treatment assignment probability for female and others is 0.4 vs. 0.6), which leads to imbalanced potential outcomes between treatment and control groups.

The efficiency is not meaningful here because both approaches yield almost the same standard error values.

RMSE measures the difference between values predicted by a model or an estimator and the values observed, a smaller value indicates better model performance or better estimation. In our case, the RMSE value of SATE estimator using regression is smaller (0.176 vs. 0.355) than the RSME value of the difference-in-means method, thus, the regression is a better approach for estimating SATE.

(f) Describe one possible real-life scenario where treatment assignment probabilities and/or treatment effects vary across levels of a covariate.

For example, we want to design a study to examine the effect of a math training program on college students' math performance. The treatment is receiving the math training program, while the control condition stays business-as-usual. The treatment effect is then the difference between a student's post-training math score and pre-test score. According to previous results, we know boys and girls perform differently in math. Moreover, the accessible population we are going to use for

drawing random samples has more girls than boys. Thus, gender is a covariate in this study. Therefore, to make a good design for this study, the treatment assignment probabilities and treatment effects might vary across the two gender groups.

(g) How could you use a regression to estimate the treatment effects separately by group? Calculate estimates for our original sample and treatment assignment (with seed 1234).

```
# regression by separate groups
set.seed(1234)
mydata2 <- assign.treatment(mydata) #original sample and treatment assignment

# regression for females
summary(lm(outcome ~ treatment, data = mydata2, subset = (female == 1)))
## Call:
## lm(formula = outcome ~ treatment, data = mydata2, subset = (female == 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1602 -0.4676 -0.1878  0.4047  1.3724
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.7120     0.1433  486.53   <2e-16 ***
## treatment     7.1660     0.2775   25.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6721 on 28 degrees of freedom
## Multiple R-squared:  0.9597, Adjusted R-squared:  0.9583
## F-statistic:   667 on 1 and 28 DF,  p-value: < 2.2e-16

# regression for others
summary(lm(outcome ~ treatment, data = mydata2, subset = (female == 0)))
## Call:
## lm(formula = outcome ~ treatment, data = mydata2, subset = (female == 0))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.03143 -0.58402 -0.09766  0.67693  2.86810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.8367     0.2330  299.77   <2e-16 ***
## treatment     3.3390     0.2874   11.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.141 on 68 degrees of freedom
## Multiple R-squared:  0.665,  Adjusted R-squared:  0.6601
## F-statistic:   135 on 1 and 68 DF,  p-value: < 2.2e-16
```