Assignment_2a

Yongchao Zhao

9/11/2018

Question 1: Caliculating ATE (all seeing/omniscient)

(a) Please simulate a dataset consistent with the assumptions below while demonstrating an average treatment effect (ATE) of approximately 5.

```
# Data Generating Process (DGP)
set.seed(1234)

# population size is 1000 (sample size? since we are asked to compute SATE)
N <- 1000

# pretest quiz score ~ N(65, 3)
pre_score <- rnorm(N, mean = 65, sd = 3)

# error terms ~ N(0, 1)
error_0 <- rnorm(N, mean = 0, sd = 1)
error_1 <- rnorm(N, mean = 0, sd = 1)

# potential post-test score Y(0) and Y(1)
post_0 <- 10 + 1.1*pre_score + 0 + error_0
post_1 <- 10 + 1.1*pre_score + 5 + error_1

# generate the simulated dataset
my_data <- data.frame(stuID = 1:1000, pre_score, post_0, post_1)

head(my_data)
##   stuID pre_score   post_0   post_1
## 1     1  61.37880 76.31135 81.54286
## 2     2  65.83229 82.71698 87.31589
## 3     3  68.25332 83.53951 89.96792
## 4     4  57.96291 74.39457 79.95139
## 5     5  66.28737 83.61906 86.26023
## 6     6  66.51817 81.26410 87.12434

summary(my_data[, -1])
##    pre_score         post_0          post_1
##  Min.   :54.81   Min.   :69.21   Min.   :74.98
##  1st Qu.:62.98   1st Qu.:79.04   1st Qu.:84.11
##  Median :64.88   Median :81.41   Median :86.45
##  Mean   :64.92   Mean   :81.43   Mean   :86.44
##  3rd Qu.:66.85   3rd Qu.:83.78   3rd Qu.:88.65
##  Max.   :74.59   Max.   :92.45   Max.   :96.34
```

(b) What is your interpretation of tau?

Students who received the treatment effect (home tutors) have a score of 5 points higher on average than had the students NOT received the treatment.

(c) Please calculate SATE.

The SATE is 5.014443.

```
# individual treatement effect
TE <- my_data$post_1 - my_data$post_0

# compute SATE
(SATE_t <- mean(TE))
## [1] 5.014443
```

(d) Why is SATE different from tau?

The computed SATE from simulated dataset is different from tau = 5 because the two post-test scores Y(0) and Y(1) are generated based on two seperate linear functions. Althougn the error terms are both normal distributions with mean of 0 and standard deviation of 1, the actual data would be slightly different.

(e) How would you interpret the intercept in the DGP for Y(0) and Y(1)?

The intercept ($\beta_0 = 10$) in Y(0) indicates that students who scored 0 points in the pretest quiz would be expected to score 10 points on average in the post-test, if they don't receive extra home tutoring before the second attempt;

The intercept ($\beta_0 + \tau = 15$) in Y(1) indicates that students who scored 0 ponits in the pretest quiz would be expected to score 15 points on average in the post-test, if they received extra home tutoring before taking the quiz again.

(f) How would you interpret the β_1 coefficient?

Comparing students whose pretest quiz scores differ by 1 point, we would expect to see a difference of 1.1 points in their post-test scores.

Question 2: Estimating ATE (not all seeing/researchers' view)

(a) Using the same simulated dataset used in the previous case where $\tau = 5$, please randomly assign students to treatment and control groups. Then, create the observed data set which must include pretest scores, treatment assignment and observed y.
Hint: sample() is the command in R to draw out random samples.

```
# generate observed dataset
observed_data <- my_data
```

```r
# randomly assign students into treatment group (home tutors) and control
group (business-as-usual)
t_group <- sample(observed_data$stuID, 500, replace = F)
observed_data$treatment <- ifelse(observed_data$stuID %in% t_group, 1, 0)

# generate values for observed score (Y[1] for treatement group, Y[0] for
control group)
observed_data$observed_score <- ifelse(observed_data$treatment == 1,
observed_data$post_1, observed_data$post_0)

# check the dataset
head(observed_data)
##   stuID pre_score   post_0   post_1 treatment observed_score
## 1     1  61.37880 76.31135 81.54286         1       81.54286
## 2     2  65.83229 82.71698 87.31589         1       87.31589
## 3     3  68.25332 83.53951 89.96792         0       83.53951
## 4     4  57.96291 74.39457 79.95139         0       74.39457
## 5     5  66.28737 83.61906 86.26023         1       86.26023
## 6     6  66.51817 81.26410 87.12434         1       87.12434

observed_data$treatment <- as.factor(observed_data$treatment)
summary(observed_data[-1])
##    pre_score         post_0          post_1      treatment observed_score
##  Min.   :54.81   Min.   :69.21   Min.   :74.98   0:500     Min.   :69.21
##  1st Qu.:62.98   1st Qu.:79.04   1st Qu.:84.11   1:500     1st Qu.:80.99
##  Median :64.88   Median :81.41   Median :86.45             Median :83.97
##  Mean   :64.92   Mean   :81.43   Mean   :86.44             Mean   :83.97
##  3rd Qu.:66.85   3rd Qu.:83.78   3rd Qu.:88.65             3rd Qu.:87.06
##  Max.   :74.59   Max.   :92.45   Max.   :96.34             Max.   :96.34
```

(b) You can also use rbinom function to assign treatment. What's the difference between rbinom and sample function?

The sample function takes a random sample of specified size from a population, with or without replacement, like choosing a number of cards from a pack of shuffled cards. For our case, it is a non-replacement sampling. We would get a balanced group size of treatment and control (500 in both groups).

The rbinom function considers the probability of success on a trial, for example, flip a coin according to a given probability of 0.5. For our case, we simulate 1000 flips to assign students into either treatment or control group. The group size could be unbalanced, as shown in the example below (505 vs 495).

```r
observed_data2 <- my_data
observed_data2$treatment <- rbinom(n = 1000, size = 1, prob = 0.5)
table(observed_data2$treatment, useNA = "ifany")
##
##   0   1
## 505 495
```

(c) Now please estimate SATE using a difference in means.

```
# compute the means of treatment and control group
(avgs <- tapply(observed_data$observed_score, observed_data$treatment, mean))
##        0        1
## 81.39460 86.53686

# estimated SATE via the diffenrence in group means
(SATE_e <- avgs[[2]] -avgs[[1]])
## [1] 5.142255
```

(d) Is this estimate close to the true SATE? Divide the difference between SATE and estimated SATE by the standard deviation of the observed outcome, Y.

The estimated SATE (SATE_e = 5.142255) is relatively close to the true SATE (SATE_t = 5.014443).

```
(SATE_t-SATE_e)/sd(observed_data$observed_score)

## [1] -0.02952664
```

(e) Why is estimated SATE different from SATE and tau?

From researchers' view, we now only have factual post-test scores available, the observed SATE is an estimated difference of treatment and control goup means. Thus, the observed SATE is different from the true SATE and tau.

Moreover, because the treatment is randomly assigned in our example, the observed SATE can be used to get an unbiased estimate of the SATE.

Question 3: Use Linear Regression to estimate the treatment effect

(a) Now we will use linear regression to estimate SATE for the observed data set created by Question 2. With this set up, we will begin to better understand some fundamental assumptions crucial for the later R homework assignments.

Using the observed dataset from Question 2, the simple linear regression using only observed post-test scores as response varible yields an estimate of SATE as 5.1423 (with a standard error of 0.2203).

```
# simple linear regression
summary(lm(observed_score ~ treatment, data = observed_data))
##
## Call:
## lm(formula = observed_score ~ treatment, data = observed_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.184  -2.342   0.072   2.281  10.377
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  81.3946     0.1558  522.52   <2e-16 ***
## treatment1    5.1423     0.2203   23.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.483 on 998 degrees of freedom
## Multiple R-squared:  0.3532, Adjusted R-squared:  0.3525
## F-statistic: 544.9 on 1 and 998 DF,  p-value: < 2.2e-16
```

(b) What is gained by using linear regression to estimate ATE instead of the mean difference estimation from above?

Using gained scores as transformed response variable, the linear regression yields an estimate of AMT as 5.04036 (with a standard error of 0.06726).

```
# linear regression using gained scores as transformed response variable
summary(lm(observed_score - pre_score ~ treatment, data = observed_data))
##
## Call:
## lm(formula = observed_score - pre_score ~ treatment, data = observed_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3144 -0.7157  0.0310  0.7469  2.8906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.52534    0.04756  347.47   <2e-16 ***
## treatment1   5.04036    0.06726   74.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.063 on 998 degrees of freedom
## Multiple R-squared:  0.8491, Adjusted R-squared:  0.849
## F-statistic:  5616 on 1 and 998 DF,  p-value: < 2.2e-16
```

(c) What assumptions do we need to make in order to believe this estimate?

In order to believe the estimate using gained scores from (b), an extra assumption needs to be make, namely, the coefficient for pretest score ("pre_score" shown in the linear regression below) is 1.

Plus, as the following regression output shows, the effciency of the estimate is improved by controlling for pretest scores (estimate of ATE as 5.02874 with standard error of 0.06374).

```
# controling for pre-treatment predictors
summary(lm(observed_score ~ treatment + pre_score, data = observed_data))
## Call:
## lm(formula = observed_score ~ treatment + pre_score, data = observed_data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1834 -0.6886  0.0358  0.7007  3.0074
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.12808    0.69279   13.18   <2e-16 ***
## treatment1   5.02874    0.06374   78.89   <2e-16 ***
## pre_score    1.11403    0.01066  104.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.008 on 997 degrees of freedom
## Multiple R-squared:  0.9459, Adjusted R-squared:  0.9458
## F-statistic:  8719 on 2 and 997 DF,  p-value: < 2.2e-16
```

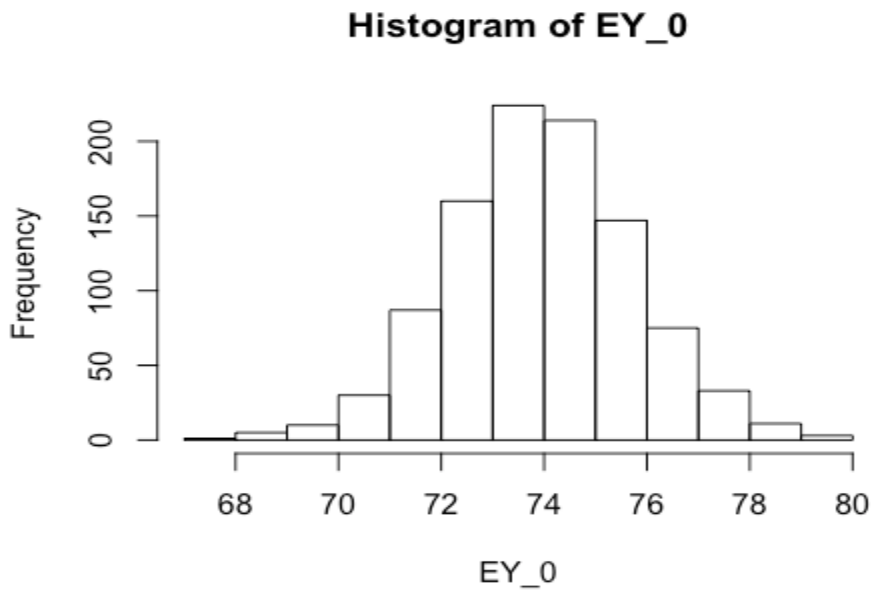
Challege Question

(a) Data simulation

```
# data simulation
EY_0 <- 35 + 0.6 * pre_score
Y0 <- EY_0 + error_0

EY_1 <- 15 + 1 * pre_score
Y1 <- EY_1 + error_1

# plots
hist(EY_0)
```



```
hist(EY_1)
```