Problem Set Assignment No. 4

Aria Muchhal

02 April, 2023

```
## Keep this line always
knitr::opts_chunk$set(echo = TRUE,
                      collapse = TRUE,
                      warning = FALSE, message = FALSE,
                      fig.align = 'center')
```

0.

Task A.

```
library(knitr)
library(tinytex)
library(purrr)
library(lmtest)
library(sandwich)
library(fixest)
library(modelsummary)
library(lattice)
```

1.

Task A. Creating the Data

```
set.seed(2024)
N=1000
tau=4
hID<-1:N
type<-as.integer(rbernoulli(N,.25))
df<-data.frame(hID, type)
consWithout=rpois(N,20+8*df$type)
df$consWithout<-consWithout
consWith=rpois(N, 20-tau+8*df$type)
df$consWith<-consWith
```

Question A. Creating the Data

a. How many columns are in the data?

4.

b. What is the average of consWith for type==1? What about type==0?

23.56863 for type==1, 15.9906 for type==0.

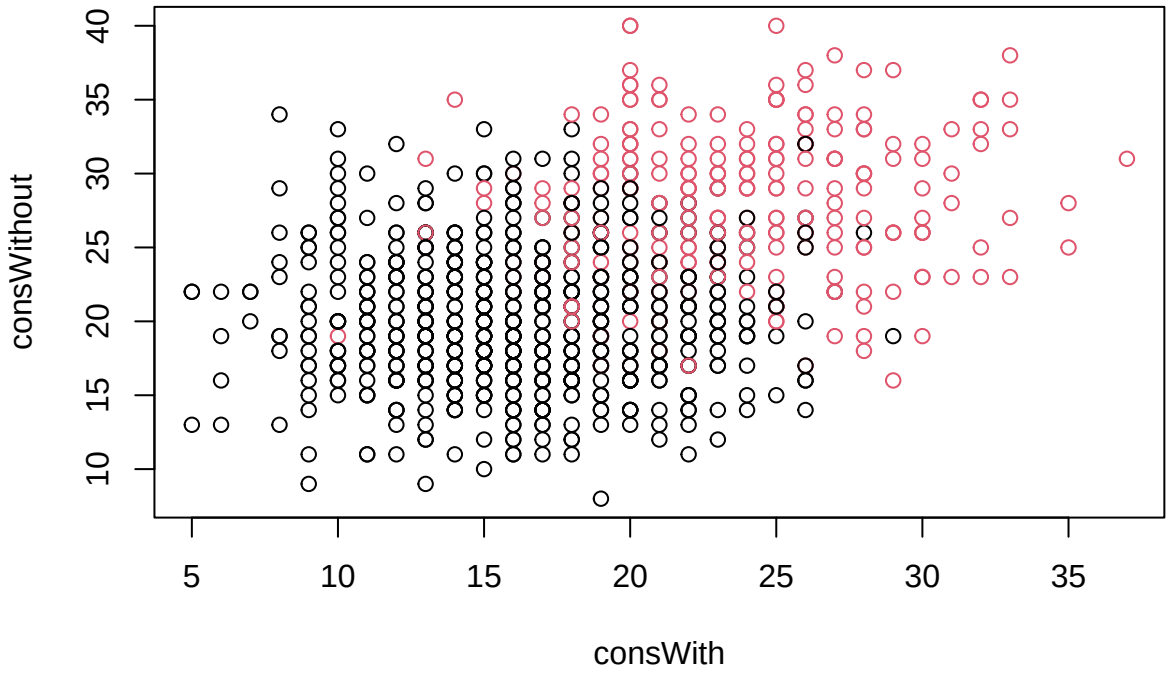
c. What is the average difference between consWith and consWithout? Does this match what you expect from the data creation?

4.022, which makes sense, as the intended difference between the two was tau=4.

Task B. Plotting

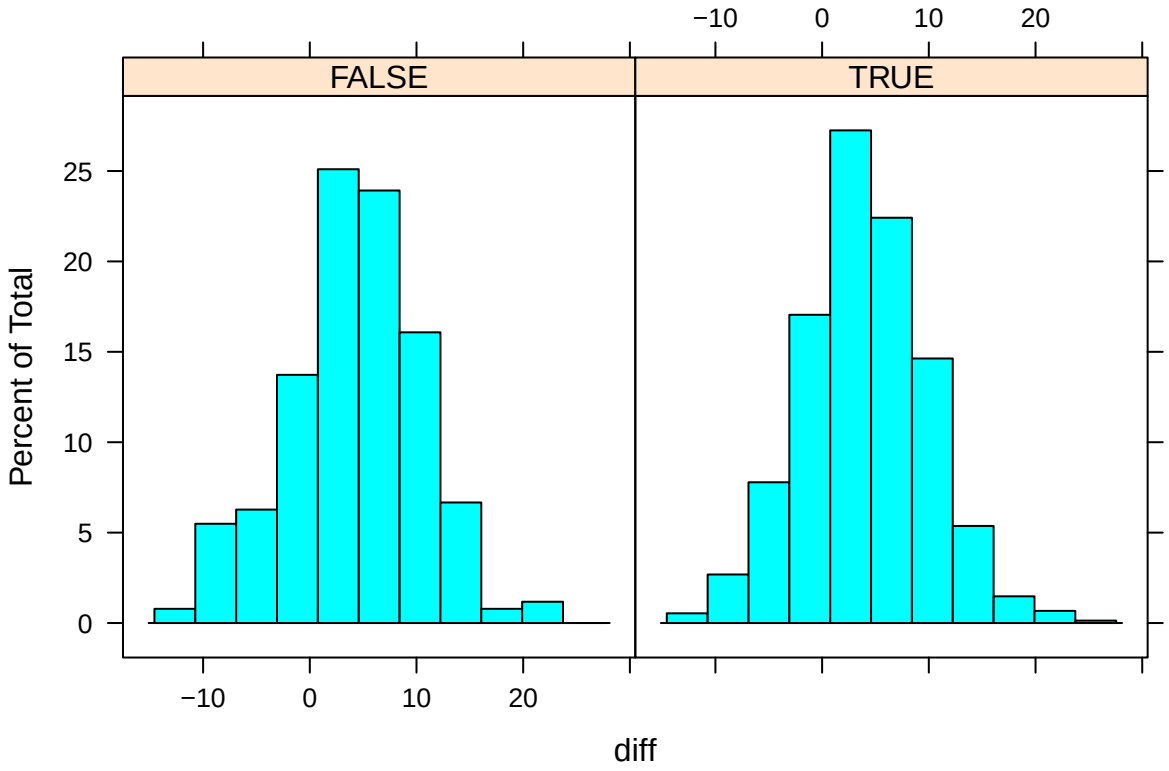
(a)

```
plot(consWith, consWithout, col=as.factor(df$type))
```



(b)

```
diff=consWithout-consWith
df$diff<-diff
histogram(~diff|type==0, data=df)
```



Question B.

a. What does the first plot tell us about consumption by type?

Consumption with the high-consumption type or type==TRUE is higher than with type==FALSE.

b. Is it feasible that some households may have an increase in consumption after doing the conservation program?

It is possible but unlikely.

Task C

(a)

```
set.seed(2025)
df$tmtProb <- with(df, ifelse(type==0, .1, .4))
```

(b)

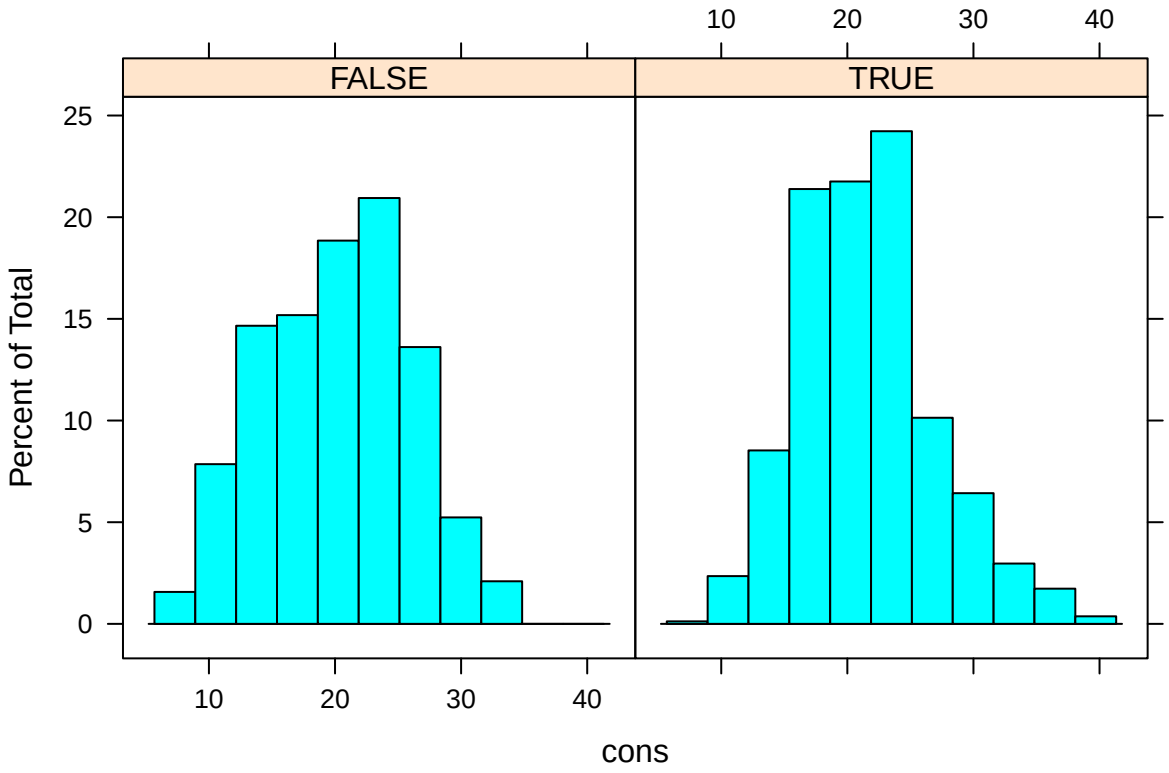
```
df$tmt <- as.integer(rbernoulli(1000, p=df$tmtProb))
```

(c)

```
df$cons <- with(df, ifelse(tmt==1, consWith, consWithout))
```

d)

```
histogram(~cons|tmt==0, data=df)
```



e)

```
df0bs<- data.frame(df$cons, df$tmt)
```

Question C

a. Looking at your histogram, does it look like treatment reduces consumption? Why might this be misleading?

No, it looks like treatment increases it, but this is misleading because high consumption people are more likely to participate in the program, and therefore are more likely to get treatment. Thus, their consumption would be higher than low consumption people who don't sign up for the program.

b. Is treatment correlated with the outcome variable? How do we know?

Yes, as being in the program decreases your consumption, so the treatment of completing the conservation program is correlated with the outcome.

c. We learned that selection bias is $E[Y_0|D=1] - E[Y_0|D=0]$. In our context, what is Y_0? Do we always observe it? Is it in our (simulated) data? Is it in our observed data? What is D in our context?

Y_0 is the consumption of people when D=0, where D is our treatment–completing the conservation program. We only observe Y_0 when D=0, and it is in our simulated and observed data.

d. What is your estimate from the data for $\hat{E}[Y_0|tmt = 0]$?

21.48949

e. What is your estimate from the data for $\hat{E}[Y_0|tmt = 1]$?

20.26178

f. If we did not construct the data ourselves using consWith and consWithout, would we be able to calculate a sample estimate of $E[Y_0|tmt = 1]$?

No.

##2. Regressions

Task A

```
coeftest(lm(df.cons ~ df.tmt, df0bs), vcov = vcovHC, "HC1")
##
## z test of coefficients:
##
##             Estimate Std. Error  z value Pr(>|z|)
## (Intercept) 21.48949    0.19131 112.3289  < 2e-16 ***
## df.tmt      -1.22771    0.46489  -2.6409  0.00827 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question A

a. What is the coefficient on tmt in your regression? Is it statistically significant?

-1.22771, and it is statistically significant

b. What was the true value we originally set and are they close? Does the true value fall in the 95% confidence interval?

No, as the original value was -4, and it would not fall within the 95% confidence interval.

c. Why might the estimate not be very close to the true value?

The true value didn't take into account the increase in probability of completing treatment if you are a high consumption individual and was based on a different seed.

Task B Randomization

(a)

```
set.seed(2026)
df$rtmt<-as.integer(rbernoulli(1000, p=.25))
df$rcons<-with(df, ifelse(rtmt==1, consWith, consWithout))
```

Question B

a. What is $E[Y0|rtmt == 1]$?

18.05833

b. What is $E[Y0|rtmt == 0]$?

21.76184

C. What is the difference between your answers to a and b?

-3.70351

Task C

```
coeftest(lm(rcons ~ rtmt, df), vcov = vcovHC, "HC1")
##
## z test of coefficients:
##
##             Estimate Std. Error  z value  Pr(>|z|)
## (Intercept) 21.76184    0.20368 106.8458 < 2.2e-16 ***
## rtmt        -3.70351    0.42085  -8.8002 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question C

a. What is the estimate of the treatment effect?

-3.70351

b. How does it differ from the estimate we got in Task 1.C, without randomization?

It is significantly closer to the true value.

c. How does it differ from your answer to Question 2.B.c?

It's the same.

d. Did randomization work?

Yes. We found more accurate answers to the true values.

Section 3

I spent 4 continuous hours on this problem set, but it took around half a day with breaks.