Experimental Problem Set

Brock Wilson

4/10/2021

Problem 1

In class, we discussed the Roy (1951) model of selection based on comparative advantage. In this problem, we will simulate a slight extension of the Roy Model to better understand non-compliance and local average treatment effects. Specifically, we will assume people make participation decisions entirely based on their earnings with or without the training less any costs of the training. The extension is accounting for the fact that different choices might have different costs.

For this exercise, assume we are evaluating the impact of earning a certificate from a community college on a worker's earnings. If someone is in the treatment group, they receive the training for free. But someone in the control group can pay to enroll in the program on their own for 1,000. Simulate a sample of 10,000 observations from the following data generating process:

$$Y_0 \sim N(20000, 7000^2)$$
$$Y_1 \sim N(21500, 8000^2)$$

```
set.seed(1)
size = 10000

y_0 = rnorm(n = size, mean = 20000, sd = 7000)
y_1 = rnorm(n = size, mean = 21500, sd = 8000)

df = data.frame(y_0, y_1)
```

1. What is the average treatment effect in your sample? How does it compare to the true average treatment effect?

True Average Treatment Effect is 1500

Average Treatment Effect:

```
#Average Treatment Effect = E[y_1i - y_0i]

ate = mean(df$y_1 - df$y_0)
ate
```

```
## [1] 1512.239
```

Difference between True Average Treatment Effect and Sample Average Treatment Effect

```
1500 - ate
```

```
## [1] -12.23875
```

2. What is the distribution of compliers, always takers, and never takers in your sample (i.e. what is P(A), P(C), and P(N))?

```
#Always Takers
#Always goes to college regardless of treatment status
#Always Takers will go to college if in treatment and Y_1 > Y_0
#Always Takers will go to college if not in treatment and Y_1 - 1000 > Y_0
#If Y_1 - Y_0 > 1000 and Y_1 - Y_0 > 0, then individuals are always takers
#Thus if Y_1 - Y_0 > 1000, individuals are always takers
percent = df %>%
  filter(y_1 - y_0 > 1000) %>%
  summarize(always_takers = n()/size)


#Never Takers
#Never goes to college regardless of treatment status
#Never Takers will not go to college if in treatment and Y_1 < Y_0
#Never Takers will not go to college if not in treatment and Y_1 - 1000 < Y_0
#If Y_1 - Y_0 < 1000 & Y_1 - Y_0 < 0, then individuals are never takers
#Thus if Y_1 - Y_0 < 0, individuals are never takers
percent[1,2] = df %>%
  filter(y_1 - y_0 < 0) %>%
  summarize(never_takers = n()/size)

#Compliers
#Complies with Treatment
#Compliers go to college if in treatment and Y_1 > Y_0
#Compliers do not go to college in not in treatment and Y_1 - 1000 < Y_0
#If 0 < Y_1 - Y_0 < 1000, then individuals are compliers
percent[1,3] = df %>%
  filter(y_1 - y_0 <= 1000) %>%
  filter(y_1 - y_0 >= 0) %>%
  summarize(compliers = n()/size)

percent$sum = sum(percent[1,])

percent
```

```
##   always_takers never_takers compliers sum
## 1        0.5206       0.4409    0.0385   1
```

3. What is the average impact of the training for compliers, always takers, and never takers in your sample?

```
#Average Impact of Training

impact = df %>%
  filter(y_1 - y_0 > 1000) %>%
  summarize(always_takers = mean(y_1) - mean(y_0))

impact[1,2] = df %>%
  filter(y_1 - y_0 < 0) %>%
  summarize(never_takers = mean(y_1) - mean(y_0))

impact[1,3] = df %>%
  filter(y_1 - y_0 <= 1000) %>%
```

```
  filter(y_1 - y_0 >= 0) %>%
  summarize(compliers = mean(y_1) - mean(y_0))

impact
```

```
##   always_takers never_takers compliers
## 1      9626.272     -7977.59  470.6768
```

```
impact[1,1] * percent[1,1] + impact[1,2] * percent[1,2] + impact[1,3] * percent[1,3]
```

```
## [1] 1512.239
```

4. Why is it reasonable to assume there are no defiers given our assumptions about how people are making participation decisions?

It is reasonable to assume there are no defiers given our assumption because individuals are either better off with treatment (always-takers), without treatment (never-takers), or with treatment if provided (compliers). Specifically to be a defier, it must be the case that:

Defiers go to college if not in treatment which implies $Y_1 - 1000 > Y_0$

Defiers choose to not go to college if in treatment which implies $Y_1 < Y_0$

Thus to be a defier, it must be that $0 > Y_1 - Y_0 > 1000$ which is impossible.

So far, we have been using the full sample because we observe both potential outcomes. Now, let's pretend we are in the real world and only observe the outcome that results from someone's participation decision. To this end, randomly assign half of your sample to a treatment group and half to a control group. Generate an indicator P that equals 1 if someone receives the training and 0 otherwise. Remember: we have assumed people make participation decisions entirely based on their earnings with or without the training less any costs of the training. This should depend on the observations treatment status.

Generate a variable Y equal to observed earnings using the following formula:

$$Y = PY_1 + (1 - P)Y_0$$

```
df$treatment = rbinom(n = size, size = 1, prob = 0.5)

#Option 1

# df$decision = if (df$treatment == 1) {
#   ifelse(df$y_1-1000>df$y_0, 1, 0)
# } else {
#   ifelse(df$y_1>df$y_0, 1, 0)
# }

#Option 2

df$decision = df$treatment #Compliers
df$decision = ifelse(df$y_1 < df$y_0, 0, 1) #Never Takers
df$decision = ifelse(df$y_1 - 1000 > df$y_0, 1, 0) #Always Takers

# a = sum(ifelse(df$y_1 < df$y_0, 0, 1))
# b = sum(ifelse(df$y_1 - 1000 > df$y_0, 1, 0))
# a+b

df = df %>%
  mutate(y = decision*y_1 + (1-decision)*y_0)
```

5. Use a regression to estimate the intent-to-treat effect in your sample. What is the point estimate and the 95% confidence interval around the estimate?

```
sum1 = summary(lm(data = df, y ~ treatment))

b = sum1$coefficients[2,1]
c = sum1$coefficients[2,1] + 2*sum1$coefficients[2,2]
a = sum1$coefficients[2,1] - 2*sum1$coefficients[2,2]

table = cbind(a,b,c)
colnames(table) = c("Lower Bound", "Estimate", "Upper Bound")
table
```

```
##      Lower Bound  Estimate Upper Bound
## [1,]   -411.2639 -157.6294    96.00498
```

6. Use two-stage least squares to estimate the local average treatment effect in your sample. Comment on the point estimate and the 95% confidence interval around the estimate. How does this compare to the effects we estimated earlier in this problem?

Solution:

The confidence interval and point estimate are significantly different from problem 5. This is due to a weak instrument. The problem is that we are instrumenting decision with treatment, however treatment is a weak instrument. We can see this one of two ways. In the summary of the IV regression, we see that treatment is a weak instrument. Additionally, we can take a look at the first stage regression. Both show the same result, that treatment is weak instrument. This is a problem because a weak instrument will be biased, but be consistent as n approaches infinity. Additionally, the estimated standard error will be too small. I add in a grab to show the distribution of the estimates which has a wider distribution compared to the OLS estimates.

```
#2SLS Regression
sum1 = summary(ivreg(data = df, y ~ decision | treatment))
sum1
```

```
##
## Call:
## ivreg(formula = y ~ decision | treatment, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29441.0  -9851.8    270.6   9282.4  34604.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    33921      13055   2.598  0.00938 **
## decision      -17202      25076  -0.686  0.49273
##
## Diagnostic tests:
##                   df1  df2 statistic p-value
## Weak instruments    1 9998     0.841   0.359
## Wu-Hausman          1 9997     1.995   0.158
## Sargan              0   NA        NA      NA
##
## Residual standard error: 11490 on 9998 degrees of freedom
## Multiple R-Squared: -2.283,  Adjusted R-squared: -2.283
## Wald test: 0.4706 on 1 and 9998 DF,  p-value: 0.4927
```

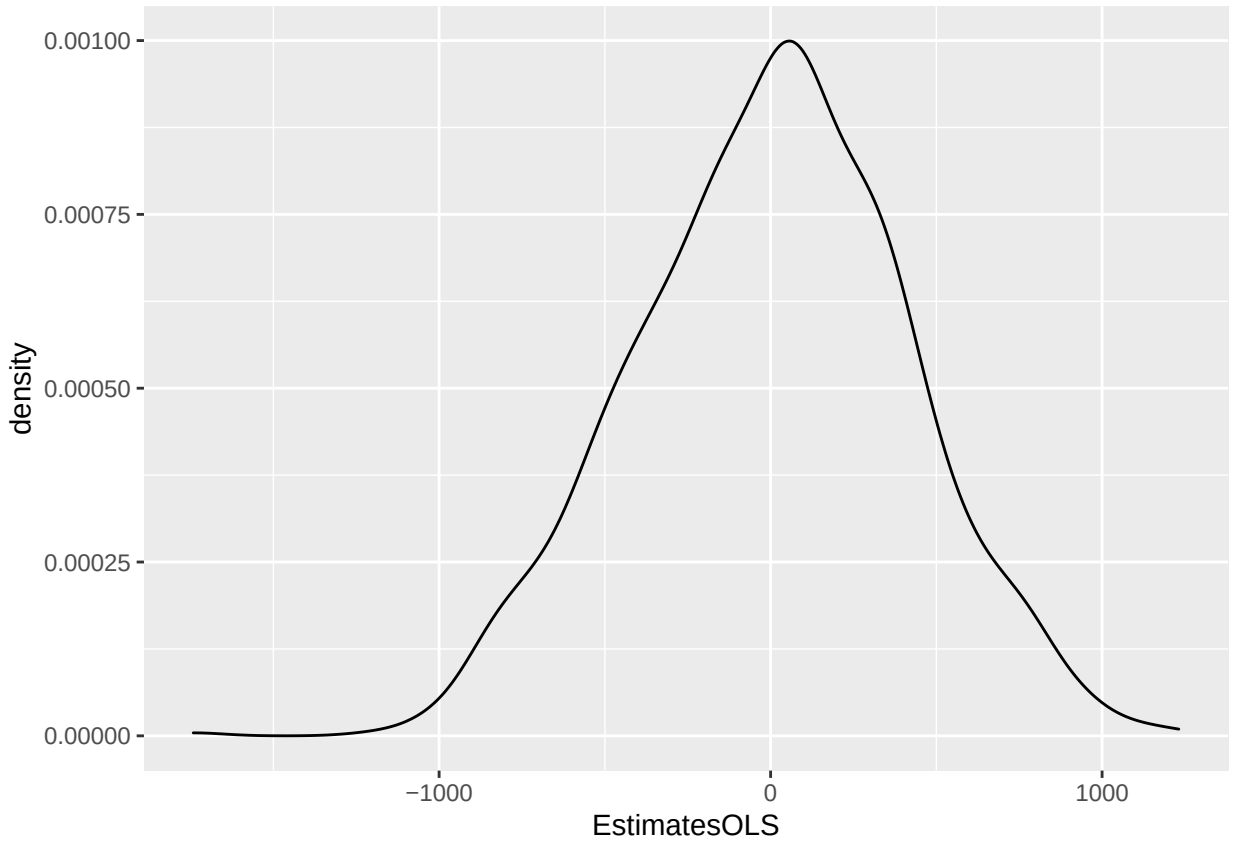
```
summary(lm(data = df, decision ~ treatment))
```

```
##
## Call:
## lm(formula = decision ~ treatment, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5251 -0.5159  0.4749  0.4841  0.4841
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.515925   0.007139  72.272   <2e-16 ***
## treatment   0.009163   0.009994   0.917    0.359
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4996 on 9998 degrees of freedom
## Multiple R-squared:  8.407e-05,  Adjusted R-squared:  -1.594e-05
## F-statistic: 0.8406 on 1 and 9998 DF,  p-value: 0.3592
```

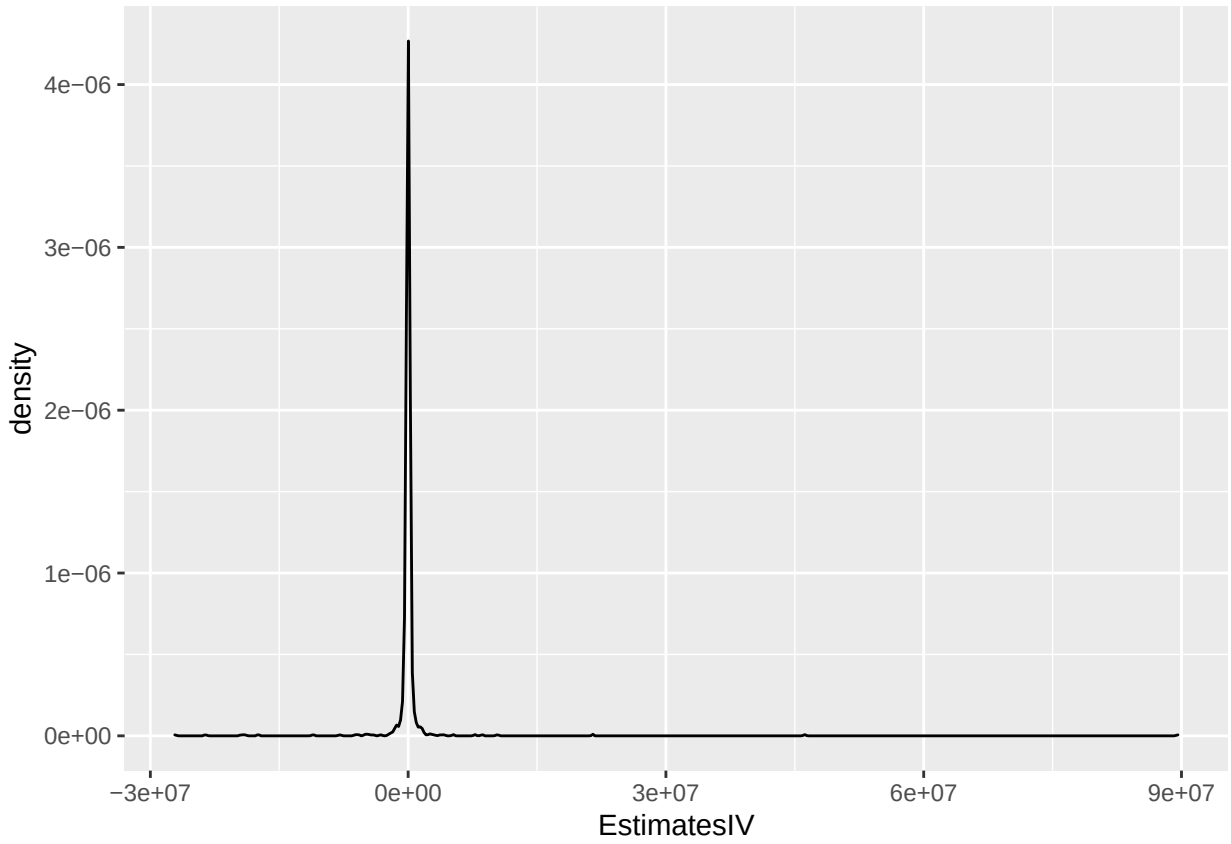
```r
b = sum1$coefficients[2,1]
c = sum1$coefficients[2,1] + 2*sum1$coefficients[2,2]
a = sum1$coefficients[2,1] - 2*sum1$coefficients[2,2]

table = cbind(a,b,c)
colnames(table) = c("Lower Bound", "Estimate", "Upper Bound")
table
```

```
##      Lower Bound  Estimate Upper Bound
## [1,]   -67354.79 -17202.19    32950.41
```

```r
size = 1000
dist1 = data.frame(matrix(nrow = size, ncol = 1, data = 0))
colnames(dist1) = "EstimatesOLS"
dist2 = data.frame(matrix(nrow = size, ncol = 1, data = 0))
colnames(dist2) = "EstimatesIV"

for (i in 1:1000){
y_0 = rnorm(n = size, mean = 20000, sd = 7000)
y_1 = rnorm(n = size, mean = 21500, sd = 8000)
df = data.frame(y_0, y_1)
df$treatment = rbinom(n = size, size = 1, prob = 0.5)

df$decision = df$treatment #Compliers
df$decision = ifelse(df$y_1 < df$y_0, 0, 1) #Never Takers
df$decision = ifelse(df$y_1 - 1000 > df$y_0, 1, 0) #Always Takers

df = df %>%
  mutate(y = decision*y_1 + (1-decision)*y_0)
a = summary(lm(data = df, y ~ treatment))
b = ivreg(data = df, y ~ treatment | decision)

dist1[i,1] = a$coefficients[2,1]
dist2[i,1] = b$coefficients[2]
}

ggplot(dist1, aes(x=EstimatesOLS)) +
  geom_density()
```




```
ggplot(dist2, aes(x=EstimatesIV)) +
  geom_density()
```



7. Re-run your code but drawing a sample of 1,000,000 observations instead of 10,000. How does the estimated LATE compare to the earlier treatment effects now?

The estimated LATE is no better off as we increase the sample size. This is because of the problem before, we have a weak instrument.

```
size = 1000000
y_0 = rnorm(n = size, mean = 20000, sd = 7000)
y_1 = rnorm(n = size, mean = 21500, sd = 8000)
df = data.frame(y_0, y_1)

df$treatment = rbinom(n = size, size = 1, prob = 0.5)

#Option 1

# df$decision = if (df$treatment == 1) {
#   ifelse(df$y_1-1000>df$y_0, 1, 0)
# } else {
#   ifelse(df$y_1>df$y_0, 1, 0)
# }

#Option 2

df$decision = df$treatment #Compliers
df$decision = ifelse(df$y_1 < df$y_0, 0, 1) #Never Takers
df$decision = ifelse(df$y_1 - 1000 > df$y_0, 1, 0) #Always Takers

# a = sum(ifelse(df$y_1 < df$y_0, 0, 1))
```

```r
# b = sum(ifelse(df$y_1 - 1000 > df$y_0, 1, 0))
# a+b
#Option 3
#df$decision = df$treatment



df = df %>%
  mutate(y = decision*y_1 + (1-decision)*y_0)

summary(lm(data = df, y ~ treatment))
```

```
##
## Call:
## lm(formula = y ~ treatment, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -27200  -4311   -196   4104  36031
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 25017.020      8.919 2804.859   <2e-16 ***
## treatment      -1.924     12.616   -0.152    0.879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6308 on 999998 degrees of freedom
## Multiple R-squared:  2.325e-08,  Adjusted R-squared:  -9.768e-07
## F-statistic: 0.02325 on 1 and 999998 DF,  p-value: 0.8788
```

```r
summary(ivreg(data = df, y ~ decision | treatment))
```

```
##
## Call:
## ivreg(formula = y ~ decision | treatment, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28746.7  -4636.4   -218.9   4403.1  37466.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    26562      10836   2.451   0.0142 *
## decision       -2983      20907  -0.143   0.8866
##
## Diagnostic tests:
##                   df1   df2 statistic p-value
## Weak instruments 1e+00 1e+06     0.417   0.519
## Wu-Hausman       1e+00 1e+06     0.076   0.783
## Sargan           0e+00    NA        NA      NA
##
## Residual standard error: 6742 on 999998 degrees of freedom
## Multiple R-Squared: -0.1425, Adjusted R-squared: -0.1425
```

```
## Wald test: 0.02035 on 1 and 999998 DF,  p-value: 0.8866
```

Problem 2

Suppose a researcher is evaluating an experiment using a sample that consists of 50% compliers, 25% always takers, and 25% never takers. The researcher decides to estimate the intent-to-treat effect but dropping people who did not comply with the treatment protocol from the sample. After dropping people who did not comply with their treatment assignment from the sample, what is the distribution of compliers, always takers, and never takers in the treatment group? In the control group? Why is this a problem?

Solution:

There would be no never takers in the treatment group (since they wouldn't comply with treatment), all of the always takers who were selected for treatment and all of the compliers who were selected for treatment. In the control group, there would be no always takers, some of the never takers who were selected for control and all of the compliers who were selected for treatment.

This is a problem because we are still including the effect of never takers and always takers in our treatment effect. What we truly want to know is the effect of treatment for specifically those who complied. Never takers and always takers may skew our results and/or make our results not as generalizable.

This is illustrated well in Problem 1. If it is the case that by randomization, only never takers are in the control group and compliers in the treatment group, then we will see a negative effect. This is because the never takers have a high income without community college and compliers will often have a lower return from community college.

$$\begin{aligned} E(Y_{1,i} - Y_{0,i}|T_i) &= E(Y_{1,i}|T_i) - E(Y_{0,i}|T_i) \\ &= [P(C) * E(Y_{1,C}|T_i) + P(A) * E(Y_{1,A}|T_i)] - [P(C) * E(Y_{1,C}|T_i) + P(N) * E(Y_{1,N}|T_i)] \end{aligned}$$

CHECK

Problem 3

Table 1 below includes details about the total sample size, probability of treatment, and treatment effect within each of 5 blocks. What would the pooled treatment effect be if estimated using an OLS regression of the outcome on treatment and block fixed effects? How much weight does each block get in the pooled estimate?

CHECK

```
block = seq(1:5)
table = data.frame(block)
table$N = c(100, 100, 200, 200, 300)
table$prob_T = c(0.5, 0.25, 0.5, 0.75, 0.1)
table$treatment_effect = c(-1, 0, 1, 2, 3)

table
```

```
##   block   N prob_T treatment_effect
## 1     1 100   0.50               -1
## 2     2 100   0.25                0
## 3     3 200   0.50                1
## 4     4 200   0.75                2
## 5     5 300   0.10                3
```

```
#Each blocks variance formula
table$block_variance = table$prob_T*(1-table$prob_T)*table$N
#Total variance
total_var = sum(table$block_variance)
#Weights for each block
table$weights = table$block_variance/total_var

#Pooled Treatment Effect
sum(table$weights*table$treatment_effect)
```

```
## [1] 1.14376
```

```
table
```

```
##   block   N prob_T treatment_effect block_variance   weights
## 1     1 100   0.50               -1          25.00 0.1579779
## 2     2 100   0.25                0          18.75 0.1184834
## 3     3 200   0.50                1          50.00 0.3159558
## 4     4 200   0.75                2          37.50 0.2369668
## 5     5 300   0.10                3          27.00 0.1706161
```

Problem 4

In class, we saw how Moulton's design effect can be used to approximate the impact of clustering on our standard error. This problem will give you practice using this adjustment.

Imagine you are considering running a cluster randomized experiment with 10,000 observations split evenly across 118 clusters. You will randomly assign half of clusters to treatment and control. Your main outcome of interest is a binary variable that historically has equaled one for 72.9 percent of observations.

1. What is your minimum detectable effect if the intracluster correlation is 0?
2. What is your minimum detectable effect if the intracluster correlation is 0.2?
3. How do these answers change if you can control for baseline characteristics that explain 20 percent of residual variation?

Solution:

Part 1 Need to find δ

Given:

$$t_{\alpha/2} = \frac{\hat{\Delta} - 0}{S.E.(\hat{\Delta})}$$

$$-t_{\beta} = \frac{\hat{\Delta} - \delta}{S.E.(\hat{\Delta})}$$

Solving for δ:

$$\delta = \hat{\Delta} + S.E.(\hat{\Delta})$$

$$\delta = (t_{\alpha/2} + t_{\beta}) * \sqrt{\frac{\sigma^2}{N * Var(T_i)}}$$

where $t_{\alpha/2} = 1.96$, $t_{\beta} = 0.84$, $N = 10000$, $Var(T_i) = p * (1-p) = 0.5 * 0.5$, $\sigma^2 = ?$

Lastly, we are interested in a binary variable that historically has equaled one for 72.9 percent of observations. Thus:

$$Var(y_o) = (1-p) * p = 0.729 * (1 - 0.729) = Var(\epsilon) = \sigma^2$$

This means that

$$\delta = (1.96 + 0.84) * \sqrt{\frac{0.729 * (1 - 0.729)}{10000 * 0.25}}$$

$$\delta = 0.02489066$$

Part 2

$$\delta = \hat{\Delta} + S.E.(\hat{\Delta})$$

$$\delta = (t_{\alpha/2} + t_{\beta}) * S.E.(\hat{\Delta})$$

$$\delta = (t_{\alpha/2} + t_{\beta}) * \sqrt{\frac{\sigma^2 * (1 + (n-1) * \rho)}{N * Var(T_i)}}$$

where $t_{\alpha/2} = 1.96$, $t_\beta = 0.84$, $N = 10000$, $Var(T_i) = p * (1-p) = 0.5 * 0.5$, $n = \frac{10000}{118}$, $\rho = 0.2$ and $\sigma^2 = 0.729 * (1 - 0.729)$.

Thus:

$$\delta = (1.96 + 0.84) * \sqrt{\frac{0.729 * (1 - 0.729)(1 + (\frac{10000}{118} - 1) * 0.2)}{10000 * 0.25}}$$

$$\delta = 0.1048637$$

Part 3 Part 1

$$\delta = (t_{\alpha/2} + t_\beta) * \sqrt{\frac{(1 - R^2) * \sigma^2}{N * Var(T_i)}}$$

$$\delta = 0.02489066 * \sqrt{(1 - 0.2)} = 0.02226288$$

Part 2

$$\delta = (t_{\alpha/2} + t_\beta) * \sqrt{\frac{(1 - R^2) * \sigma^2 * (1 + (n-1) * \rho)}{N * Var(T_i)}}$$

$$\delta = 0.1048637 * \sqrt{(1 - 0.2)} = 0.09379294$$

For both of our answers, as our R^2 goes up, then the square rooted term becomes smaller. This means that we can detect a smaller δ. This is good because before we may have detected an effect of 0.10, but now we could detect an effect of 0.09.

CHECK!

Problem 5

We often focus on identifying and estimating average treatment effects like:

$$E[Y|T=1] - E[Y|T=0]E[Y|T=1] - E[Y|T=0]$$

Discuss why this rule implies that the variance of treatment effects is not identified by randomization into treatment or control alone.

Propose an additional assumption that would allow you to identify the variance of treatment effects. Is it plausible?

$$Var(Y_{i,1} - Y_{i,0}|T_i) = Var(Y_{i,1}|T_i) + Var(Y_{i,0}|T_i) - 2Cov(Y_{i,1}, Y_{i,0}|T_i)$$

$$Cov(Y_{i,1}, Y_{i,0}|T_i) = E(Y_{i,1}, Y_{i,0}|T_i) - E(Y_{i,0}|T_i) * E(Y_{i,1}|T_i)$$

Unfortunately to take the expected value of $E(Y_{i,1}, Y_{i,0}|T_i)$, we need to know the joint probability distribution.

To solve this, I propose: $Y_{i,1}, Y_{i,0}$ are independent given T_i. This implies that:

$$E(Y_{i,1}, Y_{i,0}|T_i) = E(Y_{i,0}|T_i) * E(Y_{i,1}|T_i)$$

which implies that $Cov(Y_{i,1}, Y_{i,0}|T_i) = 0$.

Independence of outcomes is a strong assumption to make. Consider the example of treating individuals with medication for obesity. It could be likely that those who lost the most weight from treatment are also the same individuals who would lose the most weight without treatment (maybe due to motivation). This would induce positive correlation and null the assumption. This is one example, but it would be difficult to imagine a scenario where independence is satisfied.