Problem Set 3 Solutions

Question 1

Consider the following dataset

```r
q1 <- tibble(group = c("Turnout Rate among those contacted",
    "Turnout Rate among those not contacted", "Overall Turnout Rate"),
    treatment = c(54.43, 36.48, 41.38), t_numbers = c(395, 1050,
        1445), control = c(NA, 37.54, 37.54), c_numbers = c(NA,
        5645, 5645))
```

Estimate the following quantities. Explain what each one means substantively.

$E[d_i(1)] = \frac{395}{1445}$

$E[Y_i(0)|d_i(1) = 0] = 36.48$

$E[Y_i(0)|d_i(1) = 1] = \frac{37.54-(.727*36.48)}{.273} = 40.36$

$E[Y_i(1)|d_i(1) = 1] = 54.43$

Question 2

Explain whether each of the following statements is true or false for the case of one-sided noncompliance. Assume that the experiment satisfies non-interference and excludability assumptions.

a) If the ITT is negative, the CACE must be negative.

TRUE. The ITT can be written as $E[D(1)]CACE$. If this quantity is negative, then since the first term must be non-negative, the CACE must be negative

b) The smaller the ITT_D, the larger the CACE

FALSE, There is no necessary relationship between the ITT_D and the CACE. It is true that mechanically the CACE will be larger than the ITT. However, changing the rate of compliance may also change the ATE among those who now comply.

c) One cannot identify the CACE if no one in the experiment receives treatment.

TRUE. If no one receives treatment it is impossible to estimate the effect of the treatment

Question 3

Consider the following paragraph describing an experiment:

"Researchers were interested the effect of encouraging college students at Berkeley to eat healthier. Canvassers visited dorms and encouraged students to eat more vegetables. Outcomes were measured by whether a student added at least one serving of vegetables to their next meal in the dining halls. When implementing the experiment, researchers encountered one-sided non-compliance. 1015 of the 1849 students assigned treatment were successfully contacted. None of the 1430 students assigned to control were contacted. 591 students in the treatment group added vegetables to their next meal, as opposed to 377 in the control group. 429 of the 1015 students who were successfully contacted add vegetables to their meal as opposed to 539 of the 2264 students that were not canvassed."

a) Estimate the ITT. What does the ITT mean in this context?

$\frac{591}{1849} - \frac{377}{1430} = 0.32 - 0.264 = 0.056$. *Assignment to being contacted caused an estimated 5.6% increase in vegetable taking.*

b) Estimate the ITT_D. What does it mean here?

$\frac{1015}{1849} = 0.549$. The estimated probability that a subject randomly assigned to treatment will be a complier is 54.9%

c) Estimate the CACE and interpret the results.

$\frac{0.056}{0.549} = 0.102$. *The estimated average increase in vegetable taking among compliers when treated versus not is 10.2%*

d) Explain why comparing the vegetable eating rates of the treated and untreated subjects will produce misleading estimates of the ATE.

*Comparisons between the treated and untreated groups conflate the effect of treatment with other differences. This is a problem of selection bias. Who decides to be treated is not-random and post the treatment assignment.

Question 4

Consider the general case of two-sided non-compliance. Assume that excludability and non-interference assumptions hold. State whether the following statements are true or false and why.

a) Among Compliers, the ITT equals the ATE.

TRUE. For Compliers, treatment assignment equals treatment received and so the ITT = ATE

b) Among Defiers, the ITT equals the ATE.

FALSE, for Defiers, treatment assignment is the opposite of treatment received. As a result the ITT = -ATE

c) Among Always-Takers and Never-Takers, the ITT and ATE are zero.

FALSE, For always-takers and never-takers the ITT is zero because they respond the same to both experimental assignments. The ATE may be not zero, but is never able to seen empirically. Remember the ITT is measuring a different estimand than the ATE in general.

Question 5

Suppose a sample contains 30% Always-Takers, 40% Never-Takers, 15% Compliers and 15% Defiers. What is the ITT_D?

$$ITT_D = \pi_c + \pi_A T - (\pi_d + \pi_A T) = \pi_C - \pi_D$$

. *Doing appropriate substition, the* $ITT_D = 0$.

Question 6

Hyde (2010) considers the effect of international election observers on monitoring election fraud. Due to difficulty in reaching villages and time constraints, observers monitored 68 of the 409 polling places assigned to treatment. Observers monitored 21 of the 1562 stations assigned to the control group. The primary outcome measure is the number of ballots declared invalid by polling station officials.

The dataset for this problem is called "Hyde_POP_2012.csv"

a) Is monotonicity a plausible assumption in this application?

Monotonicity implies that there are no defiers. Here defiers would be polling stations that are monitored if and only if they are assigned to the control group. In this case it is reasonable to assume because monitors are likely to monitor stations closer to them regardless of treatment status.

b) Under the assumption of monotonicity, what proportion of polling locations are Compliers, Never-Takers, and Always-Takers?

From the problem and the assumption of monotonicity $\pi_D = 0, \pi_A T = 0.013, \pi_C = \frac{68}{409} - \frac{21}{1562} = 0.153$. *Never takers make up the remainder which is .834 or 83.4%*

c) Using the dataset, estimate the ITT and CACE. Interpret the results.

```
ITT <- lm_robust(invalidballots ~ Sample, data = hyde) %>%
    tidy() %>%
    filter(term == "Sample") %>%
    select(estimate) %>%
    pull()
ITT
```

```
## [1] 4.824097
```

The ITT shows that compared to control stations, the effect of the assignment of monitoring leads to a change of about 5 ballots per polling station.

```
ITT_D <- lm_robust(observed ~ Sample, data = hyde) %>%
    tidy() %>%
    filter(term == "Sample") %>%
    select(estimate) %>%
    pull()
ITT_D
```

```
## [1] 0.1528149
```

The proportion of compliers is about 15.3%

```
CACE <- ITT/ITT_D
CACE
```

```
## [1] 31.56824
```

Assuming non-interference, excludability, and monotonicity, the CACE implies that an actual visit by monitors leads to a change in about 32 ballots among Compliers

d) Use randomization to test the sharp null hypothesis that there is no intent-to-treat effect for any polling location. Interpret the results. Explain why testing this null hypothesis that the ITT is zero for all units is the same as testing the null hypothesis that the ATE is zero for all compliers.

```
set.seed(1234567)  # Not required here, but will give exact match

### Step 1, figure out the treatment allocation scheme
N_treat <- sum(hyde$observed == 1)
N_control <- nrow(hyde) - N_treat
N_total <- N_treat + N_control
N_treat
```

```
## [1] 89
```

```
N_control
```

```
## [1] 1882
```

```
N_total
```

```
## [1] 1971
```

```r
## There are 89 treated units and 1882 control units Adapt
## our treatment assignment function
get_treatment_assignment <- function() {
    random_treat <- sample(x = c(rep(1, 409), rep(0, 1562)),
        size = 1971, replace = F)
}

## Nothing needs to change with these functions because
## they are general enough. We just need to be
get_ate <- function(df, y, d) {

    ## Get groups
    y1 <- df[[y]][d == 1]
    y0 <- df[[y]][d == 0]

    ## Conditional Expected Values
    E_Y1 <- mean(y1, na.rm = T)
    E_Y0 <- mean(y0, na.rm = T)

    # Return the difference in means
    return(E_Y1 - E_Y0)
}

sim_dm <- function(df, y) {
    d <- get_treatment_assignment()
    get_ate(df, y, d = d)
}

dm <- NULL
num_perms <- 10000
for (i in 1:num_perms) {
    dm[i] <- sim_dm(hyde, "invalidballots")
}

sum(abs(dm) >= abs(ITT))/num_perms
```

```
## [1] 0.4957
```

the p-value should be highly non-significant. Testing the null hypothesis that the ITT is zero is the same as testing the null that the CACE is zero for all compliers because ITT is in the numerator of the CACE.

Alternatively, if you've done some googling there's a randomization inference package called ri2

```r
library(ri2)
declaration <- declare_ra(N = nrow(hyde), m = 409)

ri2_out <- conduct_ri(formula = invalidballots ~ Sample, declaration = declaration,
    assignment = "Sample", sharp_hypothesis = 0, data = hyde)

summary(ri2_out)
```

```
##     term estimate two_tailed_p_value
## 1 Sample 4.824097              0.476
```

This will give you functionally the same answer.