Replication 2

Zayd Abdalla

Question 1

```
# Load in data
nsw_cps <- read_csv("nsw_cps.csv")
```

Part A. You will fit a linear probability model (OLS) for one of the following and you will fit a logit for the second.

```
# LPM (OLS) Quadratic Model
lpm <- lm(treat ~ age + agesq + educ + educsq + marr + nodegree +
            black + hisp + re75 + u75,
          data = nsw_cps)

# Logit Cubic Model
logit <- glm(treat ~ age + agesq + agecb + educ + educsq + marr + nodegree +
               black + hisp + re75 + u75,
             family = binomial(link = "logit"), data = nsw_cps)
```

Part B. Fit one propensity score using up to a quadratic for each variable for one set of analysis, and a cubic for a separate set of analysis.

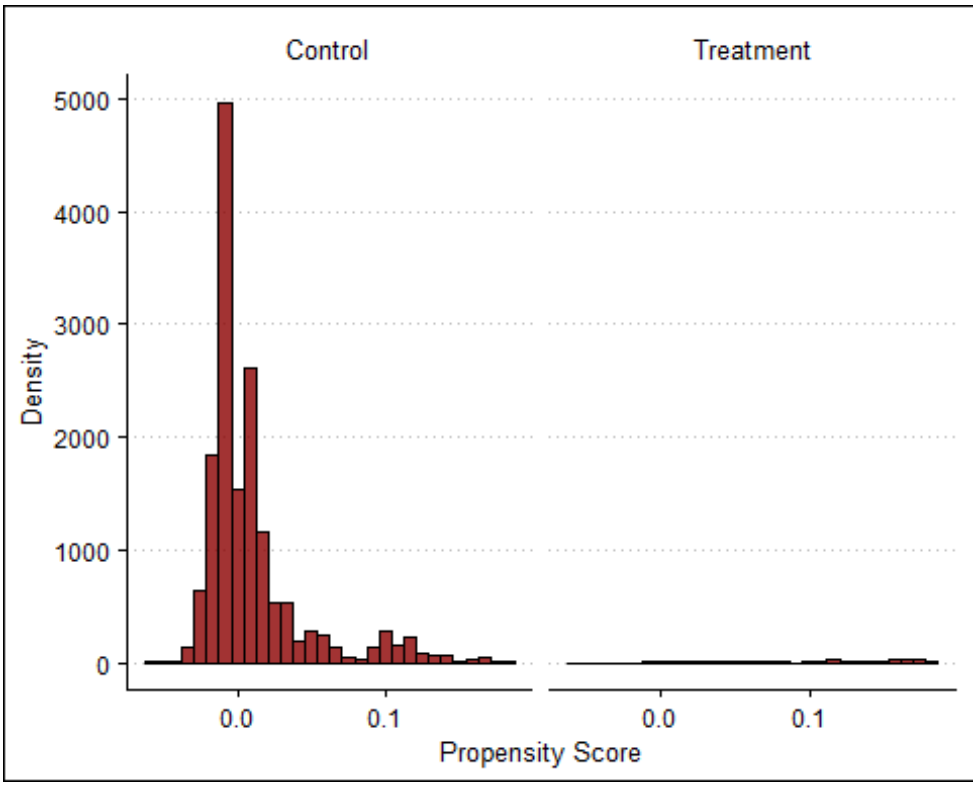
```
# LPM  Propensity Score
prs_lpm_df <-
  tibble(
    pr_score = predict(lpm, type = "response"),
    treat = lpm$model$treat
  )

# Logit Propensity Score
prs_logit_df <-
  tibble(
    pr_score = predict(logit, type = "response"),
    treat = logit$model$treat
  )
```

Part C. Create a histogram showing the distribution of the propensity score for the treatment and control group. What is the max and min values of the propensity score for the treatment group? What is it for the control group?

```
ctrl_trt_labs <- c("Control", "Treatment")
names(ctrl_trt_labs) <- c("0", "1")
```

LPM Histogram, min/max values

```
# LPM Histogram
prs_lpm_df %>%
  ggplot() +
  geom_histogram(aes(x = pr_score), fill = 'dark red', color = 'black', alpha
= 0.8) +
  labs(x = "Propensity Score", y = "Density") +
  theme_clean() +
  facet_grid(. ~ treat, labeller = labeller(treat = ctrl_trt_labs))
```



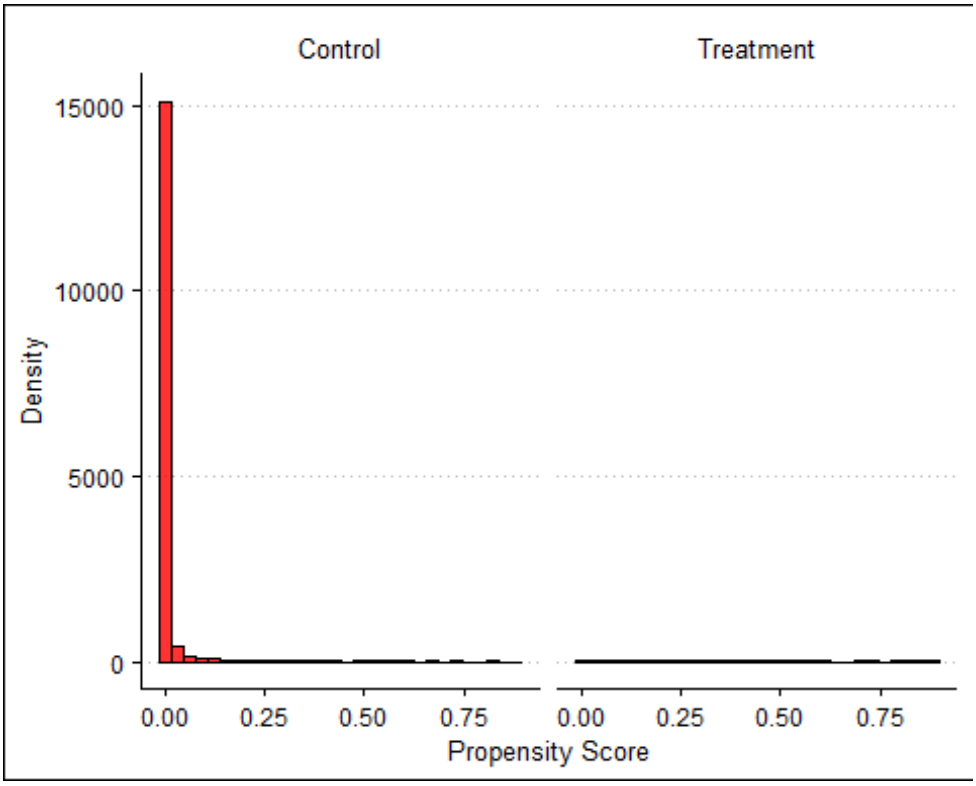

```
# LPM min and max
prs_lpm_df %>%
  group_by(treat) %>%
  summarise(Max = round(max(pr_score), 4), Min = round(min(pr_score), 4)) %>%
  mutate(treat = if_else(treat == 0, "Control", "Treatment")) %>%
  rename(Treat = treat) %>%
  kbl("pipe")
```

Treat	Max	Min
Control	0.1823	-0.0573
Treatment	0.1824	-0.0087

Logit Histogram, min/max values

```
# Logit -- Histogram
prs_logit_df %>%
```

```
  ggplot() +
  geom_histogram(aes(x = pr_score), fill = 'red', color = 'black', alpha =
0.8) +
  labs(x = "Propensity Score", y = "Density") +
  theme_clean() +
  facet_grid(. ~ treat, labeller = labeller(treat = ctrl_trt_labs))
```



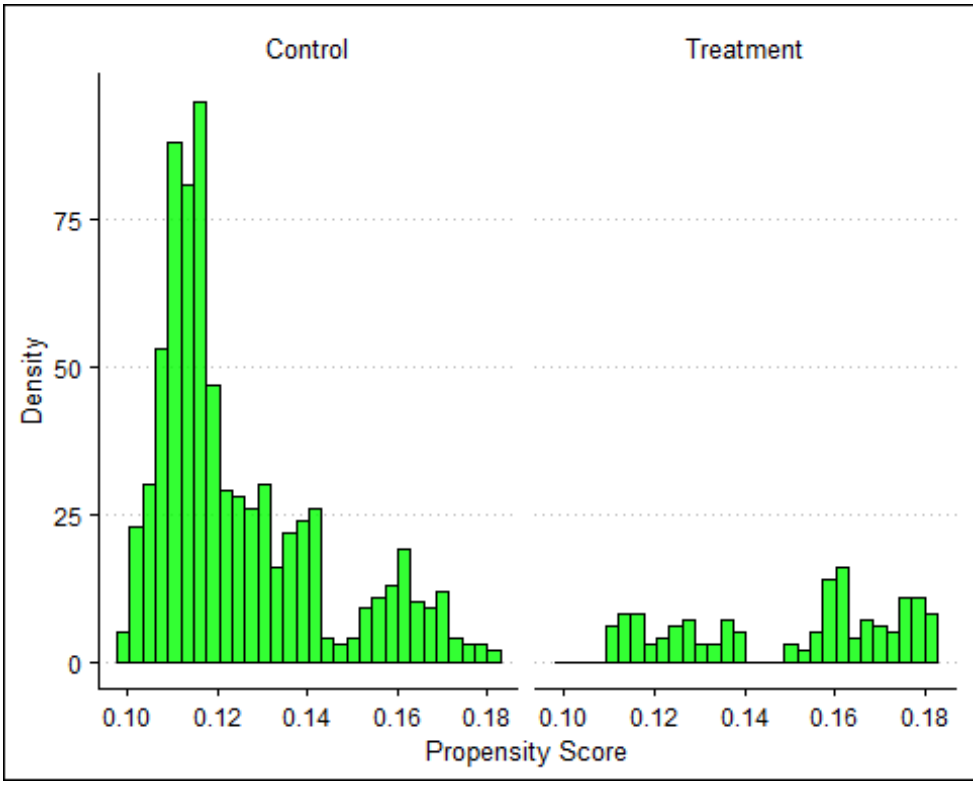

```
# Logit -- max/min
prs_logit_df %>%
  group_by(treat) %>%
  summarise(Max = round(max(pr_score), 4), Min = round(min(pr_score), 4)) %>%
  mutate(treat = if_else(treat == 0, "Control", "Treatment")) %>%
  rename(Treat = treat) %>%
  kbl("pipe")
```

Treat	Max	Min
Control	0.8714	0.0000
Treatment	0.8812	0.0012

Part D. Drop all units whose propensity scores are less than 0.1 and more than 0.9 then repeat 1C

```
# LPM Histogram after filtering out scores
prs_lpm_df %>%
  filter(between(pr_score, 0.1, 0.9)) %>%
  ggplot() +
```

```
  geom_histogram(aes(x = pr_score), fill = 'green', color = 'black', alpha =
0.8) +
  labs(x = "Propensity Score", y = "Density") +
  theme_clean() +
  facet_grid(. ~ treat, labeller = labeller(treat = ctrl_trt_labs))
```

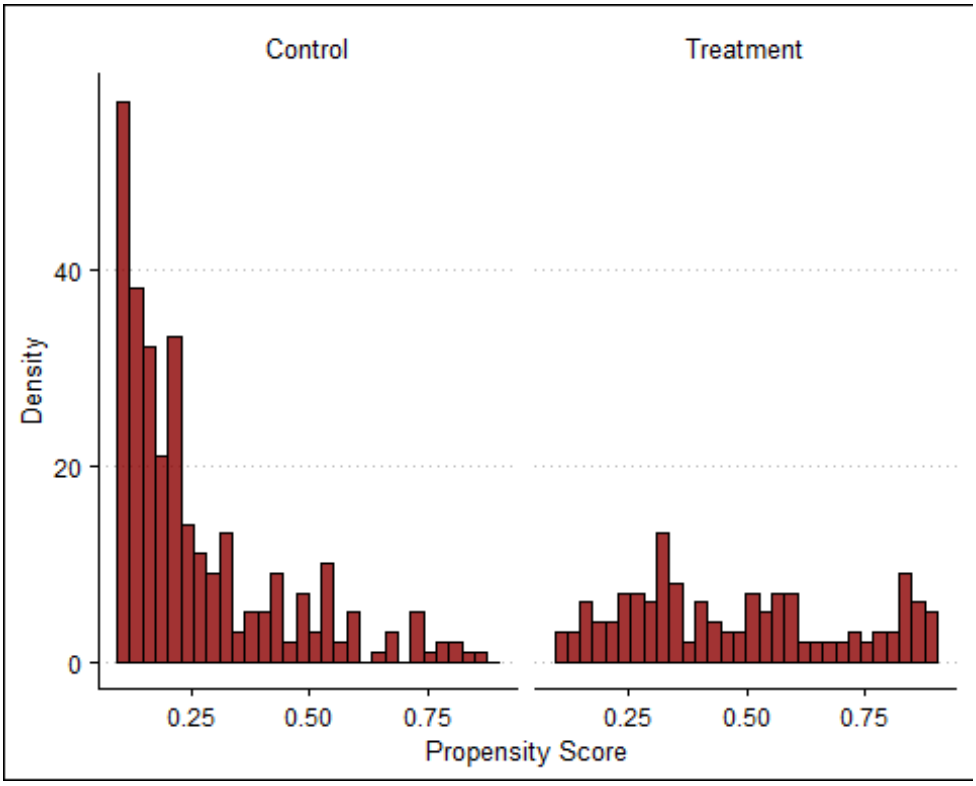



```
# LPM min/max values after filtering out scores
prs_lpm_df %>%
  filter(between(pr_score, 0.1, 0.9)) %>%
  group_by(treat) %>%
  summarise(Max = round(max(pr_score), 4), Min = round(min(pr_score), 4)) %>%
  mutate(treat = if_else(treat == 0, "Control", "Treatment")) %>%
  rename(Treat = treat) %>%
  kbl("pipe")
```

Treat	Max	Min
Control	0.1823	0.1001
Treatment	0.1824	0.1098

```
# Logit Histogram after filtering out scores
prs_logit_df %>%
  filter(between(pr_score, 0.1, 0.9)) %>%
  ggplot() +
  geom_histogram(aes(x = pr_score), fill = 'dark red', color = 'black', alpha
= 0.8) +
  labs(x = "Propensity Score", y = "Density") +
```

```
  theme_clean() +
  facet_grid(. ~ treat, labeller = labeller(treat = ctrl_trt_labs))
```




```
# Logit min/max values after filtering out scores
prs_logit_df %>%
  filter(between(pr_score, 0.1, 0.9)) %>%
  group_by(treat) %>%
  summarise(Max = round(max(pr_score), 4), Min = round(min(pr_score), 4)) %>%
  mutate(treat = if_else(treat == 0, "Control", "Treatment")) %>%
  rename(Treat = treat) %>%
  kbl("pipe")
```

Treat	Max	Min
Control	0.8714	0.1004
Treatment	0.8812	0.1067

Question 2. Calculate a before and after first difference for each unit.

```
# LPM First Difference
lpm_ey1 <- prs_lpm_df %>%
  filter(treat == 1) %>%
  pull(pr_score) %>%
  mean()

lpm_ey0 <- prs_lpm_df %>%
  filter(treat == 0) %>%
  pull(pr_score) %>%
```

```
  mean()

lpm_sdo <- round(lpm_ey1 - lpm_ey0, 4)
glue('The simple difference for the LPM model is {lpm_sdo}.')

## The simple difference for the LPM model is 0.1216.

# Logit First Difference
logit_ey1 <- prs_logit_df %>%
  filter(treat == 1) %>%
  pull(pr_score) %>%
  mean()

logit_ey0 <- prs_logit_df %>%
  filter(treat == 0) %>%
  pull(pr_score) %>%
  mean()

logit_sdo <- round(logit_ey1 - logit_ey0, 4)
glue('The simple difference for the logit model is {logit_sdo}.')

## The simple difference for the logit model is 0.3707.
```

Question 3. Construct a weighted difference-in-differences using the first equation at this https://causalinf.substack.com/p/callaway-and-santanna-dd-estimator

```
nsw_dw_dps_ctrl <- nsw_cps %>% cbind(pscore = prs_logit_df$pr_score)

N <- nrow(nsw_dw_dps_ctrl)

# non-normalized weights
nsw_dw_dps_ctrl <- nsw_dw_dps_ctrl %>%
  mutate(d1 = treat/pscore,
         d0 = (1 - treat)/(1 - pscore))

s1 <- sum(nsw_dw_dps_ctrl$d1)
s0 <- sum(nsw_dw_dps_ctrl$d0)


nsw_dw_dps_ctrl <- nsw_dw_dps_ctrl %>%
  mutate(y1 = treat * re78/pscore,
         y0 = (1 - treat) * re78/(1 - pscore),
         ht = y1 - y0)

# normalized weights
nsw_dw_dps_ctrl <- nsw_dw_dps_ctrl %>%
  mutate(y1 = (treat*re78/pscore)/(s1/N),
         y0 = ((1 - treat)*re78/(1 - pscore))/(s0/N),
         norm = y1 - y0)
```

```r
nsw_dw_dps_ctrl %>%
  pull(ht) %>%
  mean()
```

```
## [1] -11564.42
```

```r
nsw_dw_dps_ctrl %>%
  pull(norm) %>%
  mean()
```

```
## [1] -6182.63
```

```r
# Filter propensity scores
nsw_dw_dps_ctrl <- nsw_dw_dps_ctrl %>%
  select(-d1, -d0, -y1, -y0, -ht, -norm) %>%
  filter(!(pscore >= 0.9)) %>%
  filter(!(pscore <= 0.1))

N <- nrow(nsw_dw_dps_ctrl)

# non-normalized weights using filtered data
nsw_dw_dps_ctrl <- nsw_dw_dps_ctrl %>%
  mutate(d1 = treat/pscore,
         d0 = (1 - treat)/(1 - pscore))

s1 <- sum(nsw_dw_dps_ctrl$d1)
s0 <- sum(nsw_dw_dps_ctrl$d0)

nsw_dw_dps_ctrl <- nsw_dw_dps_ctrl %>%
  mutate(y1 = treat * re78/pscore,
         y0 = (1 - treat) * re78/(1 - pscore),
         ht = y1 - y0)

# normalized weights using filtered data
nsw_dw_dps_ctrl <- nsw_dw_dps_ctrl %>%
  mutate(y1 = (treat*re78/pscore)/(s1/N),
         y0 = ((1 - treat)*re78/(1 - pscore))/(s0/N),
         norm = y1 - y0)

nsw_dw_dps_ctrl %>%
  pull(ht) %>%
  mean()
```

```
## [1] 218.7165
```

```r
nsw_dw_dps_ctrl %>%
  pull(norm) %>%
  mean()
```

```
## [1] 755.1361
```

In estimating the treatment effect using an inverse probability weighting and using the non-normalized weighting, I estimated an ATT of roughly 11,564. When using the normalized weights, I estimated an ATT of roughly 6,182.

Once again, I repeated the process of filtering the propensity score, retaining scores of values between 0.1 and 0.9. This results in an estimate of roughly 218 using the non-normalized weights. When using normalized weights, my estimate now became 755.

```
# Callaway and Sant'Anna method
p_dt1 <- mean(nsw_dw_dps_ctrl$pscore)

nsw_dw_dps_ctrl %>%
  mutate(estimator = (re78 - re75)/p_dt1 * (treat - pscore) / (1 - pscore))
%>%
  summarise(mean(estimator),
            sd(estimator)/sqrt(n()))

##   mean(estimator) sd(estimator)/sqrt(n())
## 1        1282.446                1052.451
```

Following the Callaway & Sant'Anna method, I obtained an ATT of roughly 1,282.