Homework 2

Jon Campbell

```
library(tidyverse)
library(knitr)
library(PSweight)
library(lme4)
```

Part 1

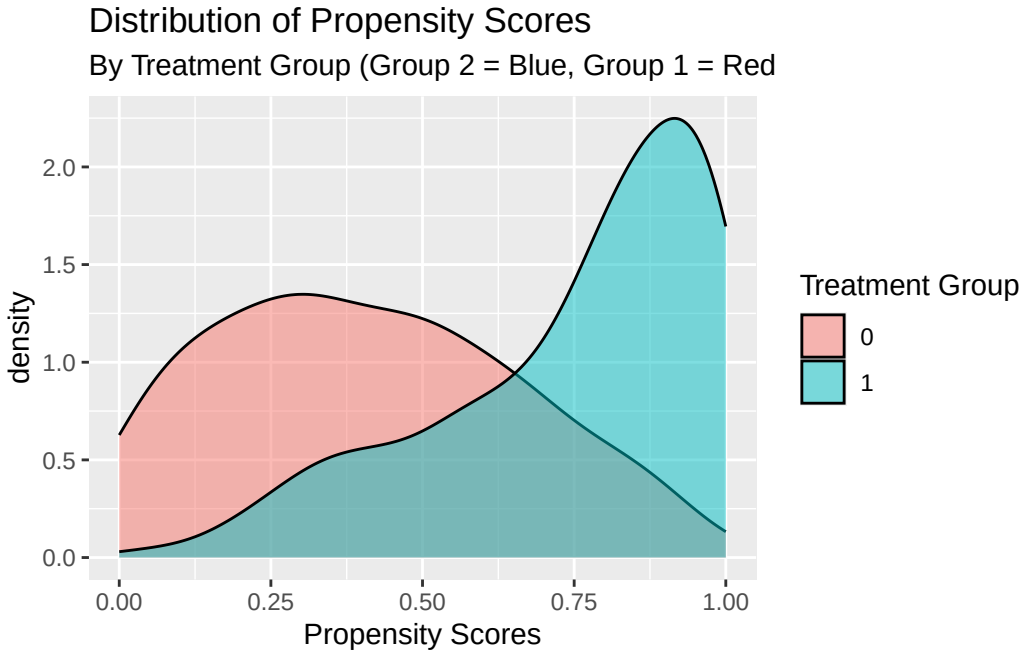
Question 1

```
hsr <- read.table("data/hw2/HSR.txt", header = TRUE)
hsr <- hsr |>
  mutate(pg = if_else(pg == 2,1,0)) |> #Z = 1 (group 2), 0 (Group 1)
  mutate(across(-c(i_age,com_t,mcs_sd,pcs_sd), factor)) |>
  rename(z = pg, y = i_aqoc)
```

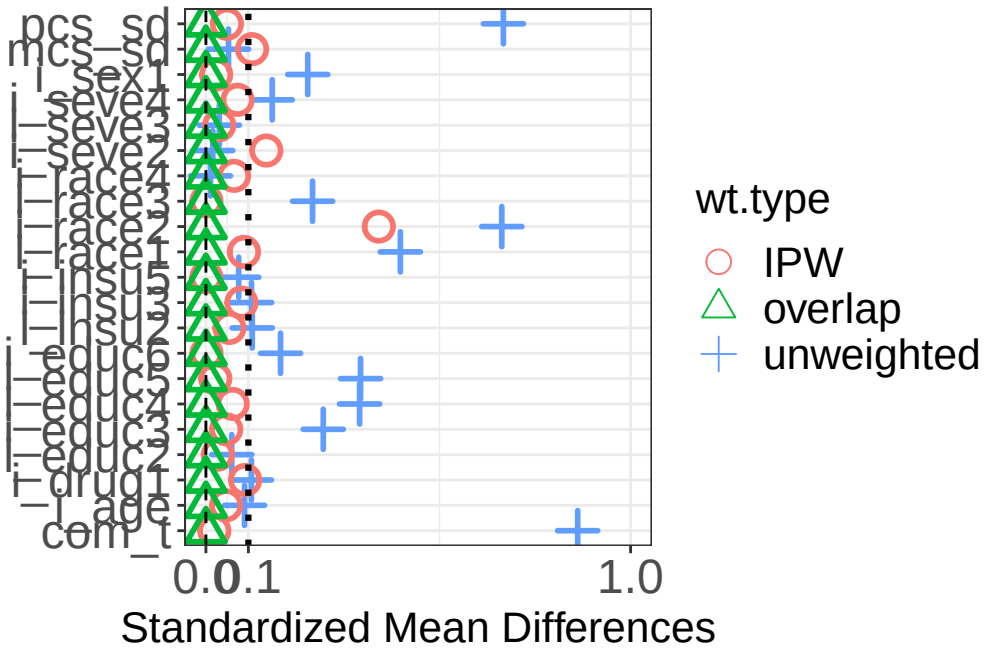
a)

```
ps.form <- z ~ i_age + i_sex + i_race + i_educ + i_insu + i_drug + i_seve + com_t + pcs_sd

pscore_summary <- SumStat(ps.formula = ps.form
                          , weight = c('IPW','overlap')
                          ,data = hsr)
hsr$pscore <- pscore_summary$propensity[,2]

ggplot(hsr, aes(x = pscore, fill = z)) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribution of Propensity Scores",
       subtitle = "By Treatment Group (Group 2 = Blue, Group 1 = Red",
```

```
        x = "Propensity Scores",
        fill = "Treatment Group")
```




```
plot(pscore_summary, metric = "ASD")
```

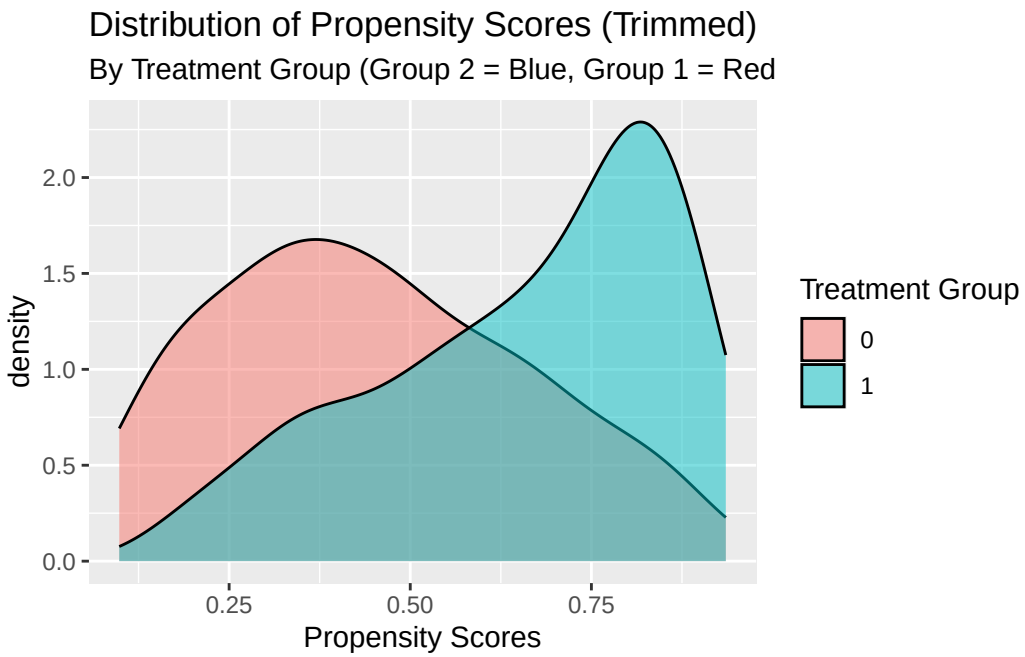


There is pretty good overlap from the distribution of propensity scores by treatment group. From the love plot we see that most of the covariates have an ASD above 0.1 when unweighted but perform much better with IPW. mcs_sd and severity2 are slightly above 0.1 and race2 is around 0.35 under IPW. The covariates are perfected balanced with overlap weighing.

```
#With trimming set at 0.05
hsr.trim <- PStrim(ps.formula = ps.form
                   ,data = hsr, delta = 0.1)[["data"]] |>
  ungroup()

pscore_summary <- SumStat(ps.formula = ps.form
                          , weight = c('IPW','overlap')
                          ,data = hsr.trim)

hsr.trim <- hsr.trim |>
  mutate(pscore = pscore_summary$propensity[,2])
```

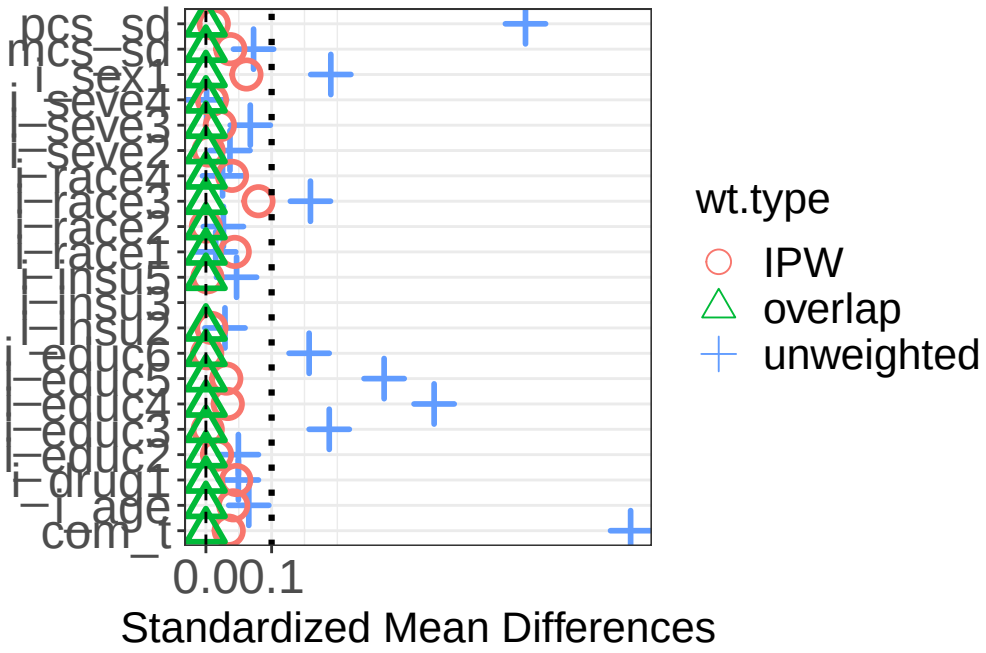
b)

```
ggplot(hsr.trim, aes(x = pscore, fill = z)) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribution of Propensity Scores (Trimmed)",
       subtitle = "By Treatment Group (Group 2 = Blue, Group 1 = Red",
       x = "Propensity Scores",
       fill = "Treatment Group")
```




```
#checking balance via ASD
plot(pscore_summary, metric = "ASD")
```

```
Warning: Removed 3 rows containing missing values (`geom_point()`).
```



From the love plot we see better performance from IPW weighting after trimming 5% of the data. All the ASDs are below 0.1. Not much change in overlap, it is still good.

c)

```
#Inverse Probability Weights
hsr.trim <- hsr.trim |>
  mutate(w = ifelse(z == 1, 1/pscore, 1/(1-pscore)))



table1 <- cbind(
  pscore_summary[["unweighted.sumstat"]][,1:2],
  pscore_summary[["IPW.sumstat"]][,1:2],
  pscore_summary[["overlap.sumstat"]][,1:2]
)
colnames(table1) = c("Original.0", "Original.trt"
                     ,"IPW.0","IPW.trt"
                     , "Overlap.0","Overlap.trt")
table1 |>round(4)
```

```
      Original.0 Original.trt   IPW.0 IPW.trt Overlap.0 Overlap.trt
```

i_age	40.6444	40.0763	41.2248	40.8633	40.9453	40.9453
i_sex1	0.7556	0.6695	0.6731	0.7019	0.7054	0.7054
i_race1	0.8444	0.8390	0.8477	0.8315	0.8440	0.8440
i_race2	0.0111	0.0085	0.0086	0.0086	0.0109	0.0109
i_race3	0.0444	0.0169	0.0351	0.0514	0.0404	0.0404
i_race4	0.0778	0.0847	0.0827	0.0720	0.0770	0.0770
i_educ2	0.0111	0.0169	0.0126	0.0146	0.0154	0.0154
i_educ3	0.0889	0.0424	0.0607	0.0600	0.0690	0.0690
i_educ4	0.3778	0.2203	0.2884	0.2733	0.3008	0.3008
i_educ5	0.2556	0.3814	0.3204	0.3351	0.3109	0.3109
i_educ6	0.2667	0.3390	0.3178	0.3169	0.3039	0.3039
i_insu2	0.2667	0.2797	0.2761	0.2799	0.2545	0.2545
i_insu3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
i_insu5	0.0333	0.0254	0.0228	0.0224	0.0251	0.0251
i_drug1	0.9889	0.9831	0.9930	0.9885	0.9923	0.9923
i_seve2	0.2444	0.2288	0.2269	0.2252	0.2442	0.2442
i_seve3	0.5000	0.4661	0.4464	0.4570	0.4636	0.4636
i_seve4	0.1444	0.1441	0.1629	0.1663	0.1582	0.1582
com_t	2.6778	1.8983	2.2422	2.1997	2.2906	2.2906
pcs_sd	43.3774	47.7123	45.5611	45.6737	45.1796	45.1796
mcs_sd	48.5509	49.2238	48.7339	49.0888	48.7561	48.7561

d)

```
hsr.trimmed <- data.frame(hsr.trim) |>
  mutate(across(c(1,3:8,12),function(x){as.numeric(levels(x))[x]})) |>
  mutate(across(c(2,9:11),scale))

ate.ipw<- summary(lm(y ~ z,data=hsr.trimmed
                     ,weights=hsr.trimmed$w))$coefficients[2,1:2]

ate.ipw.norm <- summary(PSweight(ps.formula = ps.form, yname = "y"
                                 ,data = hsr.trimmed, weight = "IPW"))$estimates[,1:2]
ato <- summary(PSweight(ps.formula = ps.form, yname = "y", data = hsr.trimmed
                    , weight = "overlap"))$estimates[,1:2]

res <- rbind(
  ate.ipw,
  ate.ipw.norm,
  ato
)
```

```r
rownames(res) <- c("ATE NonNormalized","ATE Normalized", "ATO")
res
```

```
                  Estimate Std. Error
ATE NonNormalized -0.1596368 0.06079223
ATE Normalized    -0.1409276 0.06330441
ATO               -0.1225200 0.06513518
```

Question 2

a)

```r
ps.form <- z ~ i_age + i_sex + i_race + i_educ + i_insu + i_drug + i_seve + com_t + pcs_sd

ps <- glm(ps.form, family = "binomial"
                   , data = hsr.trim)$fitted.values
hsr.trim <- hsr.trim |> mutate(pscore = ps)
hsr_subsetted <- subset(hsr.trim, i_sex == 1)

hsr_subsetted$w <- ifelse(hsr_subsetted$z == 1, 1/hsr_subsetted$pscore
                          , 1/(1-hsr_subsetted$pscore))
z <- as.numeric(levels(hsr_subsetted$z))[hsr_subsetted$z]
y <- as.numeric(levels(hsr_subsetted$y))[hsr_subsetted$y]
w <- hsr_subsetted$w

att.subgroup <- mean(y*z*w - y*(1-z)*w)
att.subgroup
```

```
[1] -0.09841427
```

b)

```r
ps.form <- z ~ i_age + i_race + i_educ + i_insu + i_drug + i_seve + com_t + pcs_sd + mcs_s

ps <- glm(ps.form, family = "binomial"
                   , data = hsr.trim)$fitted.values

hsr.trim <- hsr.trim |> mutate(pscore = ps)
hsr_subsetted <- subset(hsr.trim, i_sex == 1)
```

```
hsr_subsetted$w <- ifelse(hsr_subsetted$z == 1, 1/hsr_subsetted$pscore
                          , 1/(1-hsr_subsetted$pscore))
z <- as.numeric(levels(hsr_subsetted$z))[hsr_subsetted$z]
y <- as.numeric(levels(hsr_subsetted$y))[hsr_subsetted$y]
w <- hsr_subsetted$w

att.subgroup <- mean(y*z*w - y*(1-z)*w)
att.subgroup
```

```
[1] -0.1126802
```

c)

```
hsr_subsetted <- subset(hsr.trim, i_sex == 1)

ps.form <- z ~ i_age + i_race + i_educ + i_insu + i_drug + i_seve + com_t + pcs_sd + mcs_s

ps <- glm(ps.form, family = "binomial"
                   , data = hsr_subsetted)$fitted.values

hsr_subsetted <- hsr_subsetted |> mutate(pscore = ps)

hsr_subsetted$w <- ifelse(hsr_subsetted$z == 1, 1/hsr_subsetted$pscore
                          , 1/(1-hsr_subsetted$pscore))
z <- as.numeric(levels(hsr_subsetted$z))[hsr_subsetted$z]
y <- as.numeric(levels(hsr_subsetted$y))[hsr_subsetted$y]
w <- hsr_subsetted$w

att.subgroup <- mean(y*z*w - y*(1-z)*w)
att.subgroup
```

```
[1] -0.08976207
```

Question 3

Option (c) is most likely the best way to estimate $E[Y(1)|V = 1)] - E[Y(2)|V = 1]$ because it adjusts the propensity score estimation to the subgroup. This helps with incorporating information specific to V=1.

Part 2

```
brscn <- read.table("data/hw2/brscn.txt", header = TRUE)


brscn.fact <- brscn |>
  mutate(across(2:12,factor))
```

Question 1

```
form <- black ~ agecat + mcaid + poor + cendivis + model + taxstat + affil + blprop + blpr

#test <- PSweight_cl(form, yname = "brscn", data = brscn)
#$fitted.values
#glmer ; weighted average for each cluster
#test$coefficients
fit <- glmer(form, family = "binomial", data = brscn.fact)
pscores <- predict(fit, type = "response")
brscn.fact <- brscn.fact |>
  mutate(pscore = pscores) |>
  mutate(w = if_else(black == 1, 1/pscore, 1/(1-pscore)))
```

```
brscn <- brscn.fact |>
  mutate(across(c(group,black,brscn),function(x){as.numeric(levels(x))[x]}))

get_trt_effect <- function(num,dat) {
  filt <- dat |> filter(group == num)
  z <- filt$black
  y <- filt$brscn
  w <- filt$w

  res <- list(
    trt.effect = sum(y*z*w)/sum(z*w) - sum(y*(1-z)*w)/sum((1-z)*w),
    grp.weight = sum(w)
  )
  S <- 50
  boot <- numeric(S)
  for (i in 1:S) {
    boot.sample <- filt[sample(nrow(filt)
```

```
                                    , nrow(filt), replace = TRUE), ]
    z <- boot.sample$black
    y <- boot.sample$brscn
    w <- boot.sample$w
    boot[i] <- sum(y*z*w)/sum(z*w) - sum(y*(1-z)*w)/sum((1-z)*w)
  }

  res["se"] <- sum((boot - res$trt.effect)^2)/(length(boot)-1)

  res
}


grps <- unique(brscn$group)

trt.effects <- numeric(length(grps))
grp.weights <- numeric(length(grps))
boot.se <- numeric(length(grps))
for (i in seq_along(grps)) {
  res <- get_trt_effect(grps[i],brscn)
  trt.effects[i] <- res$trt.effect
  grp.weights[i] <- res$grp.weight
  boot.se[i] <- res$se
}

grp.weights.norm <- grp.weights/sum(grp.weights)
ate <- sum(trt.effects*grp.weights.norm)/length(grps)
boot.st.error <- sum(boot.se*grp.weights.norm)/length(grps)

cbind(ate,boot.st.error)
```

```
              ate boot.st.error
[1,] -0.0002584003  4.838788e-05
```