# Denis Ostroushko - HW4

## Introduction

### Imputation and Variable Processing

For the purpose of this assignment we retain the same imputation schemes we used in the previous two assignments. We will use imputation with the median of observed values and replace missing values with modes for categorical predictors.

Since we will need to create two models that have all possible confounders in the data set, we need to be careful with variable inclusion. We drop variable `Hisp` because it is highly correlated with other variables that contain race and ethnic information. We also `Drug.Add` due to the issues with its imputation. When imputed with the most common level "No", which indicated no drug use, this variable has one unique level. Such zero variance predictors can cause problems with fitting models, so we will avoid using it in our analyses.

Variables `BMI`, `BL.Cig.Day`, `BL.Drks.Day`, `N.living.kids` are imputed with medians like `N.prev.preg`, `Birthweight` in the previous assignments.

`Use.Alc` is imputed with a mode like `Race_ethnicity`, `Use.Tob`

**Note: I left small steps and explanations for myself for future use**

## Problem 1

### 1 - A: regression adjustment

Code chunk below produces average treatment effect among treated for the pre-term pregnancy reduction reduction. The basic idea is:

1. Develop and 'outcome' regression model using all available data
2. Subset the data to those who received treatment

3. Take the difference between average of model estimated counterfactuals under the condition that everyone received treatment and average of model estimated counterfactuals under the condition of no treatment for everyone.

```
pregnancy_model <-
  glm(
    `Preg.ended...37.wk` ~
      Group + Race_ethnicity + Public.Asstce + Use.Tob + N.prev.preg +
      Live.PTB + BL.GE + BL..BOP + BL..PD.4 + BL..CAL.3,

    data = data,
    family = "binomial"
  )

all_no_treat <- data %>%
  filter(Group == "T") %>%
  select(-Group) %>%
  mutate(Group = "C")

all_no_treat$porential_no_trt <- predict(pregnancy_model, all_no_treat , type = "response"

all_treat <- data %>%
  filter(Group == "T") %>%
  select(-Group) %>%
  mutate(Group = "T")

all_treat$porential_trt <- predict(pregnancy_model, all_treat , type = "response")

preg_att <- mean(all_treat$porential_trt, na.rm = T) - mean(all_no_treat$porential_no_trt,
```

Code below shows how to estimate birthweight increase average treatment effect among treated using regression adjustment approach.

```
borthweight_model <-
  lm(
    Birthweight ~
    Group *(Race_ethnicity + Public.Asstce + Use.Tob +
              N.prev.preg + Live.PTB + BL.GE + BL..BOP +
              BL..PD.4 + BL..CAL.3),

    data = data
  )
```

```
all_no_treat <- data %>%
  filter(Group == "T") %>%
  select(-Group) %>%
  mutate(Group = "C")

all_no_treat$porential_no_trt <- predict(borthweight_model, all_no_treat , type = "respons

all_treat <- data %>%
  filter(Group == "T") %>%
  select(-Group) %>%
  mutate(Group = "T")

all_treat$porential_trt <- predict(borthweight_model, all_treat , type = "response")

bwt_att <- mean(all_treat$porential_trt, na.rm = T) - mean(all_no_treat$porential_no_trt,
```

## 1 - B: propensity score regression adjustment

First, we develop a propensity score model using the same approach and a set of covariates as all previous assignments

```
propensity_score_model <- glm(

  I(data$Group == "T") %>% as.numeric() ~
    Race_ethnicity + Public.Asstce +
    Use.Tob + N.prev.preg + Live.PTB +
    BL.GE + BL..BOP + BL..PD.4 + BL..CAL.3,

  data = data,
  family = "binomial"
)

data$propensity_scores <- propensity_score_model$fitted.values
```

Idea of propensity score regression adjustment:

1. Use propensity score as a statistic that summarizes all measured confounding variables as a predictor in the 'outcome' regression model

2. Develop a flexible ‘outcome’ regression model with propensity score, treatment variable, and their interaction if possible. Use all available data, both controls and treated.
3. Compare average of counterfactuals under treatment and no treatment for the entire population/sample of treated people.

*Model Choice:* we covered how we can apply splines to regression scores to obtain non-linear regression effects. I tried several models and concluded that a simple model with only an interaction term has the smallest AIC.

Propensity score regression adjustment for the pregnancy ATT is given below:

```
# simple model has the smallest AIC
preg_prop_regression <-
  glm(`Preg.ended...37.wk` ~ Group*propensity_scores, data = data, family = "binomial")


all_no_treat <- data %>%
  filter(Group == "T") %>%
  select(-Group) %>%
  mutate(Group = "C")

all_no_treat$porential_no_trt <- predict(preg_prop_regression, all_no_treat , type = "resp

all_treat <- data %>%
  filter(Group == "T") %>%
  select(-Group) %>%
  mutate(Group = "T")

all_treat$porential_trt <- predict(preg_prop_regression, all_treat , type = "response")

preg_att_psr <- mean(all_treat$porential_trt, na.rm = T) - mean(all_no_treat$porential_no_
```

Propensity score regression adjustment for the birthweight ATT is given below. I also selected a more simple model based on AIC.

```
# simple model has the smallest AIC
bw_prop_regression <-
  lm(Birthweight ~ Group*propensity_scores, data = data)

all_no_treat <- data %>%
  filter(Group == "T") %>%
  select(-Group) %>%
  mutate(Group = "C")
```

```
all_no_treat$porential_no_trt <- predict(bw_prop_regression, all_no_treat , type = "respon

all_treat <- data %>%
  filter(Group == "T") %>%
  select(-Group) %>%
  mutate(Group = "T")

all_treat$porential_trt <- predict(bw_prop_regression, all_treat , type = "response")

bw_att_psr <- mean(all_treat$porential_trt, na.rm = T) - mean(all_no_treat$porential_no_tr
```

## 1 - C: propensity score stratification

Estimation idea: use 'bucket assignment' and weights for quintiles/'buckets' obtained from the treated sample only. But, we apply this quintile cutoffs and quintile weights to both treated and untreated controls. Since we do not have a way of estimating counteractions under no treatment for treated subjects in this approach, we hope that by bucketing treated and controls with similar propensity score we compare similar people. Untreated controls with similar characteristics, on average, provide $E[Y^0]$ for treated, i.e. what their average would be had they not received treatment.

Estimation procedure: 1. Develop a propensity score model using all available data 2. Find cut-off values for propensity score quintiles using only treated subjects 3. Find weights of each quintile using the distribution of treated subjects into buckets 4. Compare average observed outcome between treated and controls in each quintile/group 5. Take the weighted average of differences.

Estimation of PSS ATT for pregnancy outcome:

```
# so for ATT using PRS we need to use quintiles cutoffs of controls, and apply them to the

ps_quintile <-
  cut(data$propensity_scores,
      breaks = c(0,
                 quantile(data[data$Group == "T", ]$propensity_score, p = c(0.2, 0.4, 0.6,
                 1),
      labels = 1:5)

nA <- nrow(data[data$Group == "T", ])
nAj <- table(ps_quintile[data$Group == "T"])
```

```r
te_quintile <-

  tapply(ifelse(data$`Preg.ended...37.wk`[data$Group == "T"] == "Yes",1,0),  ## outcomes i
         ps_quintile[data$Group == "T"], ## quintiles in the treated group
         mean) -  # apply mean within each quintile: get proportion of yeses in each strat

    tapply(ifelse(data$`Preg.ended...37.wk`[data$Group == "C"] == "Yes",1,0),  # outcomes
           ps_quintile[data$Group == "C"],   # average outcomes for controls within quinti
           mean)

preg_pss_att <- sum(te_quintile *nAj/nA)
```

Estimation of PSS birht weight ATT:

```r
# so for ATT using PRS we need to use quintiles cutoffs of controls, and apply them to the

te_quintile <-

  tapply(data$Birthweight[data$Group == "T"],  ## outcomes in the treated gorup
         ps_quintile[data$Group == "T"], ## quintiles in the treated group
         mean) -  # apply mean within each quintile: get proportion of yeses in each strat

    tapply(data$Birthweight[data$Group == "C"],  # outcomes for controls only
           ps_quintile[data$Group == "C"],   # average outcomes for controls within quinti
           mean)

bw_pss_att <- sum(te_quintile *nAj/nA)
```

## 1 - D

Estimation idea: estimating ATE using IPW involved weights $\frac{A_i}{\pi_i}$ for treated and $\frac{1-A_i}{1-\pi_i}$ for controls. Now, we intend to estimate ATT, so we focus on the population of treated. Therefore, similarly to PSS, we treat outcomes for controls as if these are outcomes that treated people would have had they not received treatment. When estimating ATT using such observational data, we need to additionally multiply weights by $\pi_i$.

*Motivation for doing this procedure other than mathematical way of obtaining $E[Y^1|A = 1]$ from $E[Y^1]$ are not yet clear to me*

```
w1 <- ifelse(data$Group == "T", 1, 0)
w0 <- (1 - ifelse(data$Group == "T", 1, 0))/(1 - data$propensity_scores) * data$propensity

preg_ipw_att <-
  weighted.mean(ifelse(data$Preg.ended...37.wk == "Yes", 1, 0), w = w1) -
  weighted.mean(ifelse(data$Preg.ended...37.wk == "Yes", 1, 0), w = w0)
```

```
bw_ipw_att <-
  weighted.mean(data$Birthweight, w = w1) -
  weighted.mean(data$Birthweight, w = w0)
```

## 1 - E

### 1:1 propensity score matching

*1:1 and 2:1 matching is self explanatory*

1. David suggested we could 'exploit' correlation, stemming from the fact that matched controls and treated subjects *should* have similar covariates, through similar propensity scores, therefore we may argue that their outcomes are correlated. We could use a matched t-test to get a C.I. in a 1:1 matched sample, however, I am not sure how to proceed in a 2:1 matched sample, so I will not employ this estimation method.

2. We will see that there are still some covariates that are not balanced in the 1:1 and 2:1 matched samples. In order to achieve higher degree of balance, I attmented to match exactly on imbalanced covariates, but found that:

   - software was not able to find matched for everyone, especially in the 2:1 matched sample, and I did not want to discard the data.
   - even after exact match on those imbalanced covariates, I still was not able to achive SMD of less than or around 0.1

```
data <- data %>% arrange(PID)
rownames(data) <- 1:nrow(data)

mod_match <-
  matchit(
    # use the same variables to match as a propensity score model
    Group ~ Race_ethnicity + Public.Asstce + Use.Tob + N.prev.preg + Live.PTB +
                          BL.GE + BL..BOP + BL..PD.4 + BL..CAL.3,
    distance = "logit",
      method = "nearest",
    data = data,
    ratio = 1
    )

# mod_match$match.matrix -- this list contains a list of all rows matched to the controls
```

```
p1 <- mean(ifelse(data$Preg.ended...37.wk[data$Group == "T"] == "Yes", 1, 0)) # observed a
p0 <- mean(ifelse(data$Preg.ended...37.wk == "Yes", 1, 0)[as.numeric(mod_match$match.matri

n1 <- data %>% filter(Group == "T") %>% nrow() # number of treated
n0 <- length(as.numeric(mod_match$match.matrix)) # number of controls matched to the treat
#n0 <- table(imai$PHN.C1)[1]

preg_match.1.1_att = p1 -  p0
SE_preg.1.1 <- sqrt(p1*(1-p1)/n1 + p0*(1-p0)/n0) # Var(TRT Estimate) + Var(Matched Est) =
                                        # proportions are Bernoulli random variable
```

```
m1 <- mean(data$Birthweight[data$Group == "T"]) # observed average in the group of treated
m0 <- mean(data$Birthweight[as.numeric(mod_match$match.matrix)]) # average in the macthed

bw_match.1.1_att = m1 -  m0
SE_bw.1.1 <- sqrt((sd(data$Birthweight[data$Group == "T"])^2)/n1 +
             (sd(data$Birthweight[as.numeric(mod_match$match.matrix)])^2)/n0
           ) # Var(TRT Estimate) + Var(Matched Est) = Var(ATT)
              # These are continuous so, we just need to calculate their variances
```

### 2:1 propensity score matching

```
data <- data %>% arrange(PID)
rownames(data) <- 1:nrow(data)

mod_match2 <-
  matchit(
    # use the same variables to match as a propensity score model
    ifelse(Group == "T", 1, 0) ~ Race_ethnicity + Public.Asstce + Use.Tob + N.prev.preg +
                          BL.GE + BL..BOP + BL..PD.4 + BL..CAL.3,
    distance = "logit",
      method = "nearest",
    data = data,
    ratio = 2
    )
# mod_match$match.matrix -- this list contains a list of all rows matched to the controls
```

```
p1 <- mean(ifelse(data$Preg.ended...37.wk[data$Group == "T"] == "Yes", 1, 0)) # observed a
p0 <- mean(ifelse(data$Preg.ended...37.wk == "Yes", 1, 0)[as.numeric(mod_match2$match.matr

n1 <- data %>% filter(Group == "T") %>% nrow() # number of treated
n0 <- length(as.numeric(mod_match2$match.matrix)) # number of controls matched to the trea
#n0 <- table(imai$PHN.C1)[1]

preg_match.1.2_att = p1 -  p0
SE_preg.1.2 <- sqrt(p1*(1-p1)/n1 + p0*(1-p0)/n0) # Var(TRT Estimate) + Var(Matched Est) =
                                        # proportions are Bernoulli random variable
```

```
m1 <- mean(data$Birthweight[data$Group == "T"]) # observed average in the group of treated
m0 <- mean(data$Birthweight[as.numeric(mod_match2$match.matrix)]) # average in the macthed

bw_match.1.2_att = m1 -  m0
SE_bw.1.2 <- sqrt((sd(data$Birthweight[data$Group == "T"])^2)/n1 +
             (sd(data$Birthweight[as.numeric(mod_match2$match.matrix)])^2)/n0
           ) # Var(TRT Estimate) + Var(Matched Est) = Var(ATT)
              # These are continuous so, we just need to calculate their variances
```

Using bootstrap I obtained standard errors for all estimators, except matched estimators. Variance and 95% confidence intervals were calculated directly from the matched sample of data. Figure 1 compares variance all considered ATT estimators. As we can see, they have relatively similar variance and point estimates. The only exception is 1:1 matched sample estimate, which has a notably higher variance, probably due to the fact that we discard some data and lose statistical power.

An interesting observation to me was the fact that 2:1 matching point estimate is close to other methods. In previous assignments we saw that the un-adjusted treatment effect was -0.0332. However, 2:1 also uses almost the entire data set, the ratio of controls to treated is abound 2.2 in the data set. In fact, the rate of unfavorable pregnancy outcomes among non-matched controls is 0.289, which is extremely high compared to the unadjusted effect, or any ATT we estimated.

Figure 2 shows ATT estimates and variance associated with each method of estimation. Same comments apply here.

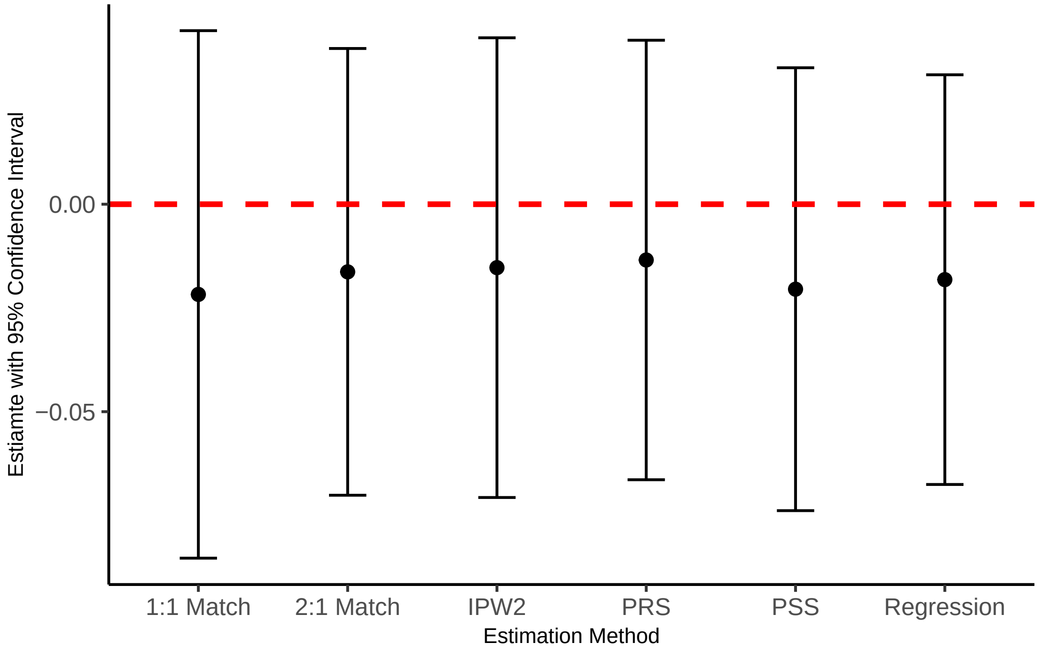

Figure 1: Comparison of Estiamtion Method Varinances for ATT of Pregnancy Outcome

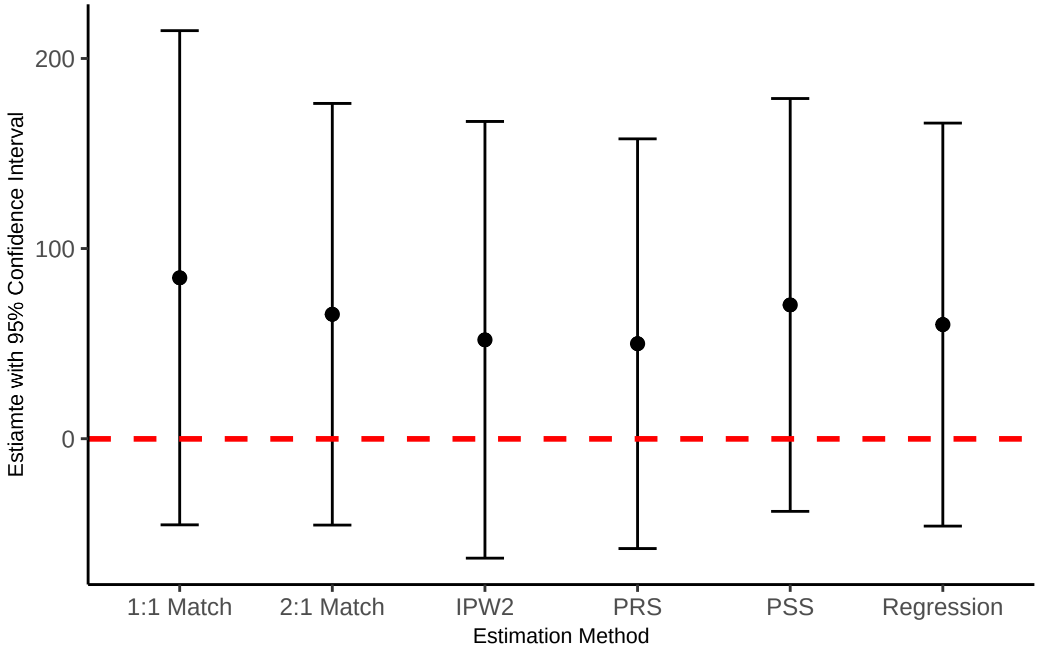

Figure 2: Comparison of Estiamtion Method Varinances for ATT of Birthweight Outcome

## Problem 2

Figure 3 shows SMD between treated and untreated using raw sample data and weighted/matched samples. IPW 'upsampled' groups shows the best balance between covariates incluedd in the propensity score model. Matched samples still show a large degree of imbalance. I tried to balance covariates foe high SMD through exact matching, which was not successful.

Figure 4 shows the distribution of propensity scores before and after matching. It looks like 1:1 achieves a reasonably more balanced distribution, although its shape is different between the groups.

Figure 5 shows the distribution of propensity scores before and after matching. It looks like 2:1 is very similar to full sample differences, because we only about 40 subjects in the 2:1 matching process.

Figure 6 shows distribution of differences in propensity scores for a treated subjects and their matched control. The difference in scores from a treated subject to the first matched control is centered at 0, which implies that every matched person has a very similar propensity score.

As we can see, second matched controls are usually further away from their matched treated subject in terms of propensity score. Sometimes the difference gets as high as 0.2, which is quite high, but to be expected.

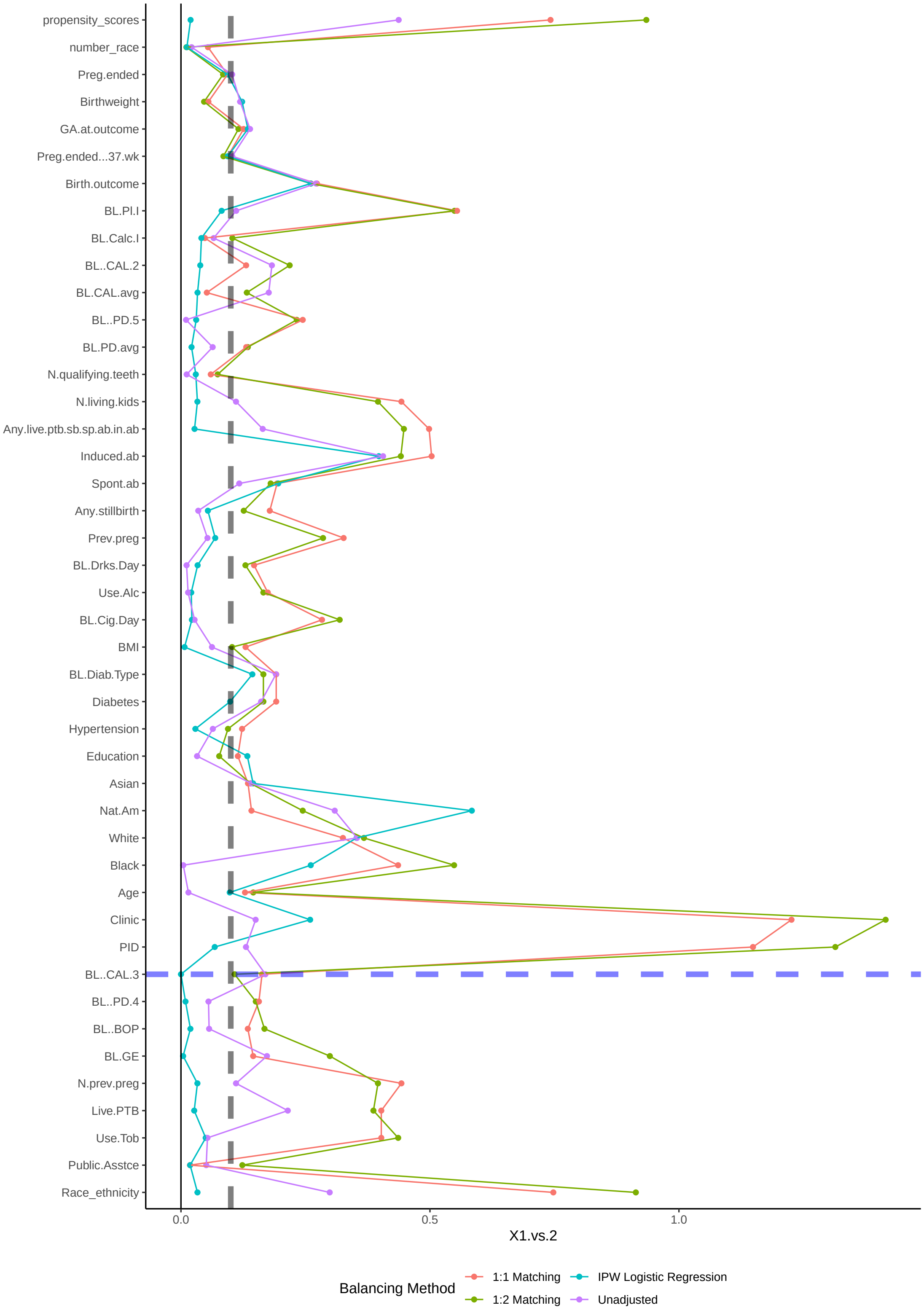

Figure 3: Comparion of confounding adjustment methods and their impact on SMDs

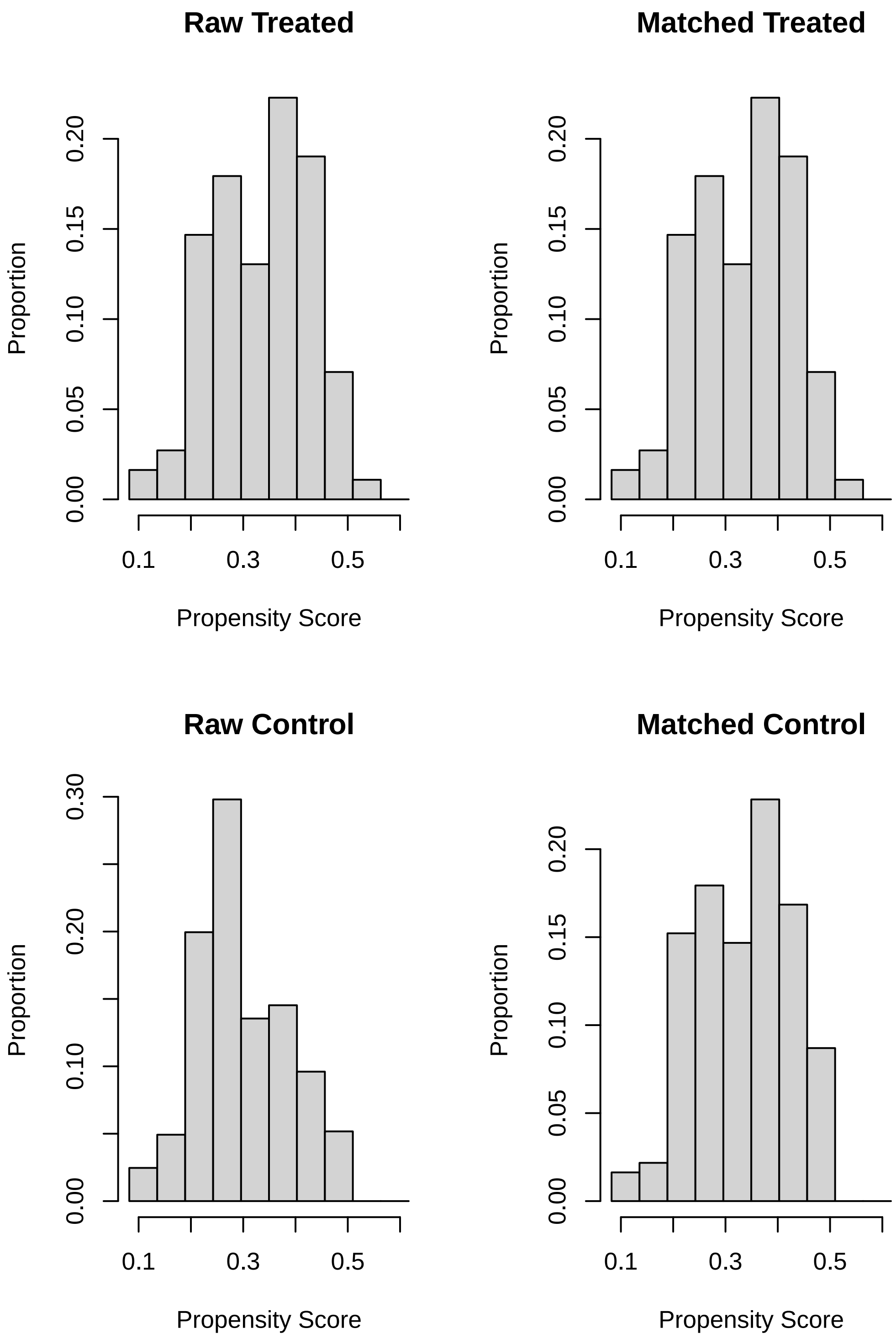

Figure 4: Distribution of Propensity Scores Before and After 1:1 Matching

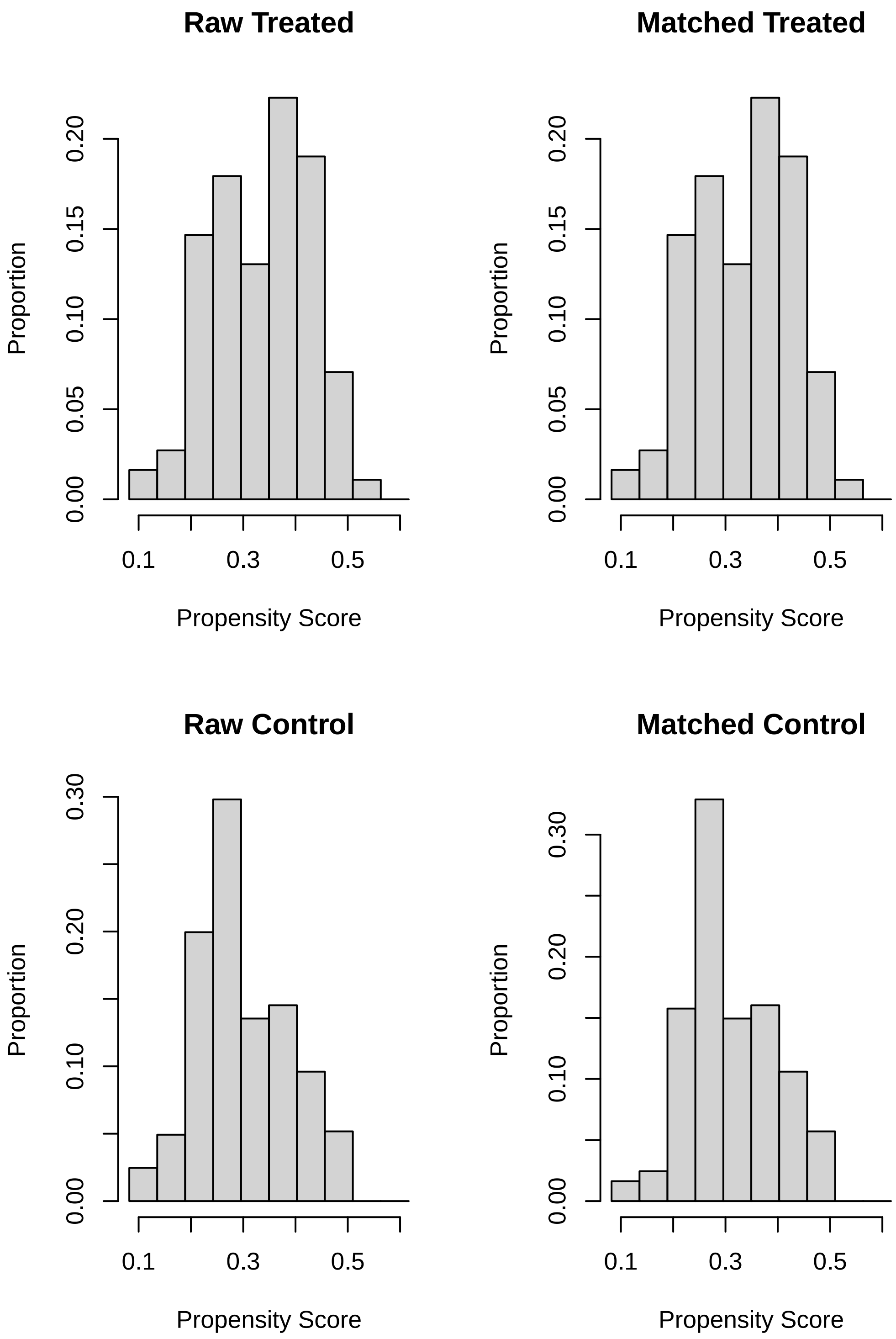

Figure 5: Distribution of Propensity Scores Before and After 2:1 Matching

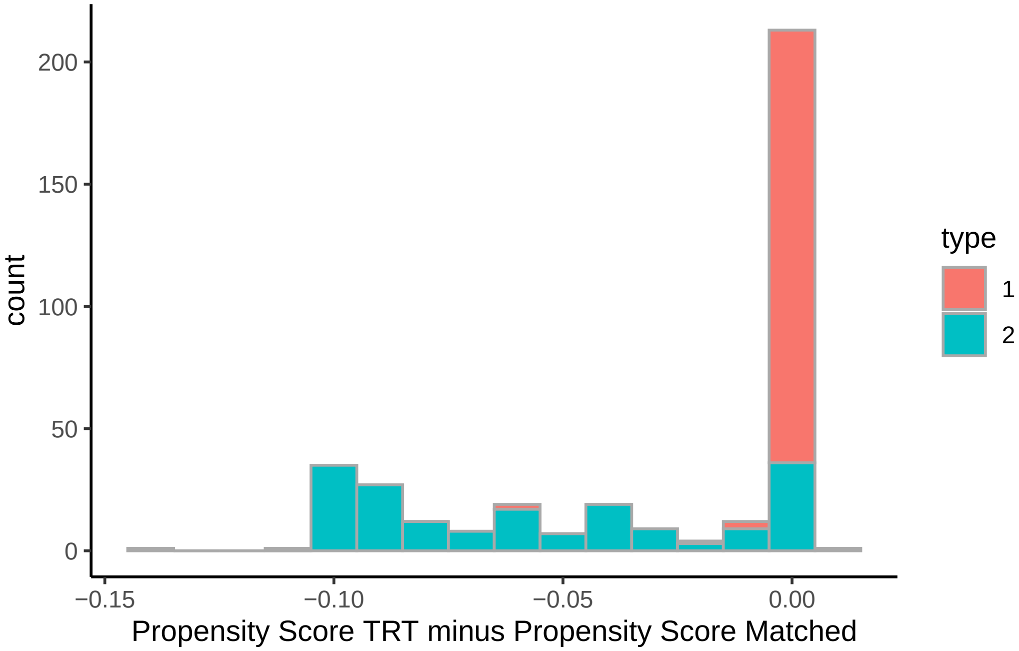

Figure 6: Difference between the Treated ith Matched Control