The role of propensity scores in observational study

Submission Instructions

Homework 4 will be manually graded. You may work with 1 partner and turn in a single submission for both group members. Make sure to include the names of both group members below:

First and Last names of both group members: Xinyu Wang

Turn in 1 submission for both students

Objective

This assignment will give you the opportunity to practice several different propensity score approaches to causal inference. In addition you will be asked to interpret the resulting output and discuss the assumptions necessary for causal inference.

R Packages

You will need to use an R package that you may not already have installed, arm.

```
if(isFALSE('arm' %in% installed.packages())){
  install.packages('arm')
}

library(arm)
```

```
## Loading required package: MASS

## Loading required package: Matrix

## Loading required package: lme4

##
## arm (Version 1.13-1, built: 2022-8-25)

## Working directory is /Users/zhenyan/Downloads
```

Problem Statement

In this assignment you will use data from a constructed observational study. The data and an associated data dictionary are available in the assignment information. For this assignment imagine the funders of the IHDP program asked you to conduct an evaluation of whether the IHDP program actually led to improved developmental outcomes at age 3.

The treatment group for the study that the data are drawn from is the group of children who participated in the IHDP intervention discussed in class. The research question of interest focuses on the effect of the IHDP intervention on age 3 IQ scores for the children that participated in it. The data for the comparison sample of children was pulled from the National Longitudinal Study of Youth during a similar period of time that the data were collected for the IHDP study.

In the data the outcome variable is `ppvtr.36` and the treatment variable is `treat`. For the assignment on the computational track you can assume all variables are pre-treatment variables.

Question 1: Load the data and choose confounders (5 points) Load the data from the IHDP.csv file on brightspace and choose the covariates you want to use as confounders. To avoid making unnecessary parametric assumptions you may want to choose binary indicators of unordered categorical variables (rather than a variable labeled e.g. as 1, 2, 3 for different levels of a categorical variable).

Create a new data frame for analysis that includes the outcome in the 1st column, the treatment indicator in the 2nd column, and the covariates in the remaining columns. Be thoughtful about your choices with respect to the nature of the covariates (e.g. is an unordered categorical being represented as such) and timing (don't control for post-treatment variables!). Provide your code and a list of the variable names for the confounder variables chosen.

Now reduce this data frame to include only observations for children whose birthweight is less than 3000 grams.

```
# load data
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df <- read.csv("IHDP.csv")
df <- df[,c(ncol(df),(ncol(df)-1),3:ncol(df)-2,2,1)]
ihdp <- df[,1:(ncol(df)-2)]
head(ihdp)
```

```
##   ppvtr.36 treat momage b.marr momed work.dur prenatal cig sex   bw bwg preterm
## 1      111     1     33      1     4        1        1   0   1 1559   0      10
## 2       81     1     22      0     1        0        1   0   1 2240   1       3
## 3       92     1     13      0     1        0        1   0   1 1900   0       6
## 4      103     1     25      1     4        1        1   0   1 1550   0       8
## 5       81     1     19      0     1        0        1   1   1 2270   1       5
## 6       94     1     19      0     2        1        1   1   0 1550   0       4
##   black hispanic white lths hs ltcoll college dayskidh income
## 1     0        0     1    0  0      0       1       31  42500
## 2     1        0     0    1  0      0       0        4   5000
## 3     1        0     0    1  0      0       0        9  12500
## 4     1        0     0    0  0      0       1       50  42500
## 5     1        0     0    1  0      0       0        4   5000
## 6     1        0     0    0  1      0       0       13  12500
```

```
# code to reduce data to include only observations for children whose birthweight is less than 3000 gra
ihdp <- ihdp[ihdp$bw<3000,]
```

```
# print out the names of all your confounders
covs <- 3:ncol(ihdp)
cov_names <- colnames(ihdp)[3:ncol(ihdp)]
```

Question 2: Estimate the propensity score (5 points) Estimate the propensity score. That is, fit a propensity score model and save the predicted scores. For now use a logistic regression with all confounders as predictors.

```
# code for initial p.score model
propensity_model <- glm(treat ~ momage + b.marr + factor(momed) + work.dur + prenatal + cig + sex + bw

ihdp$initial_pscore <- predict(propensity_model,type = "response")
#ihdp$initial_pscore
```

Question 3: Create a weight variable that will let you perform an analysis on a dataset using matching with replacement. Part a (5 points) Before creating the weight variable you need to determine your estimand. Given the description above about the research question, what is the estimand of interest? (1-word will do)

Ans: ATT

Part b (5 points) Now perform *one-to-one nearest neighbor matching with replacement* using your estimated propensity score from Question 2. Perform this matching using the matching command in the arm package. The "cnts" variable in the output reflects the number of times each control observation was used as a match.

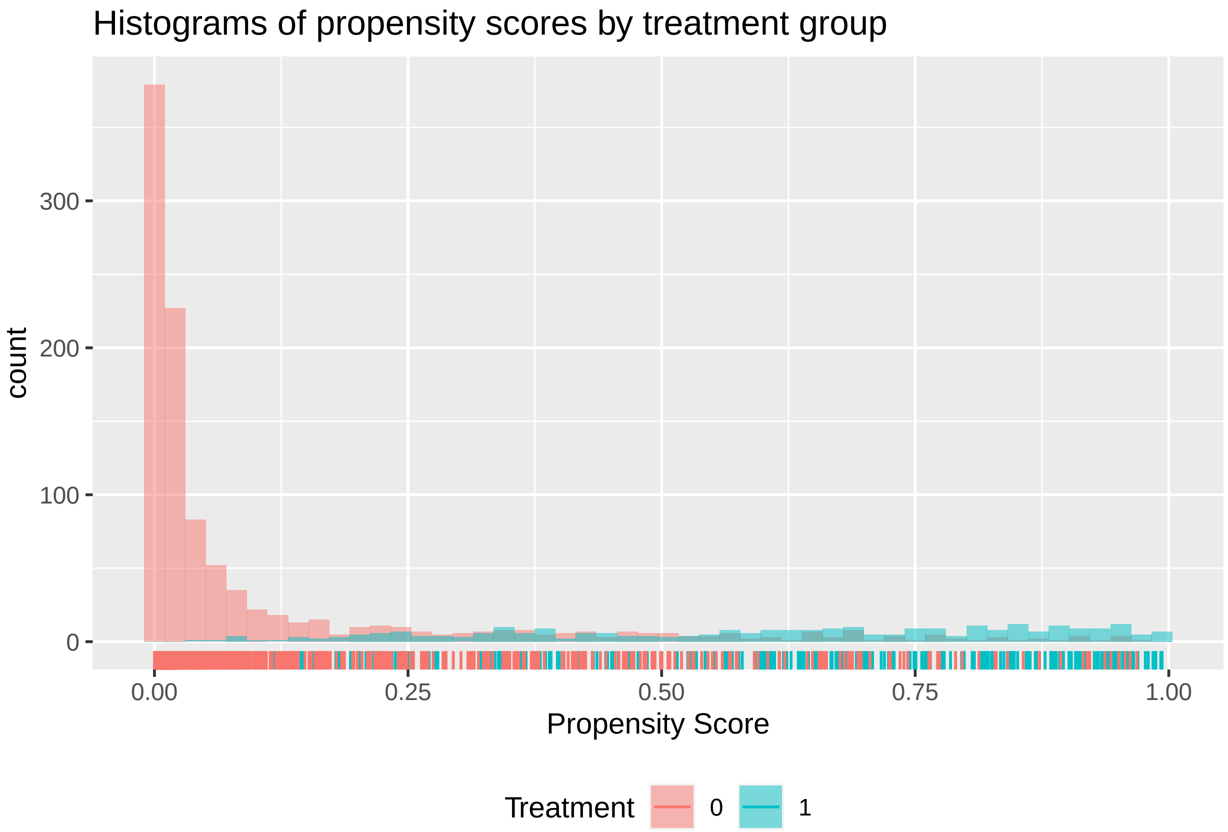
```
# code for matching here
library(arm)

matches <- matching(z=ihdp$treat, score=ihdp$initial_pscore,replace=TRUE)
matched <- matches$cnts
ihdp$matched<- matched
#ihdp$nearest_neighbor_pscore
```

Question 4: Check overlap and balance. Part a (5 points) Examining Overlap. Check overlap on the raw data (that is the data before matching) using some diagnostic plots. Check overlap for the propensity scores as well as two other covariates. Choose two covaraites that you believe are most likely to have lack of overlap. Note that it may be necessary to exclude some observations from the plots if they are being obscured in ways similar to the example discussed in class.

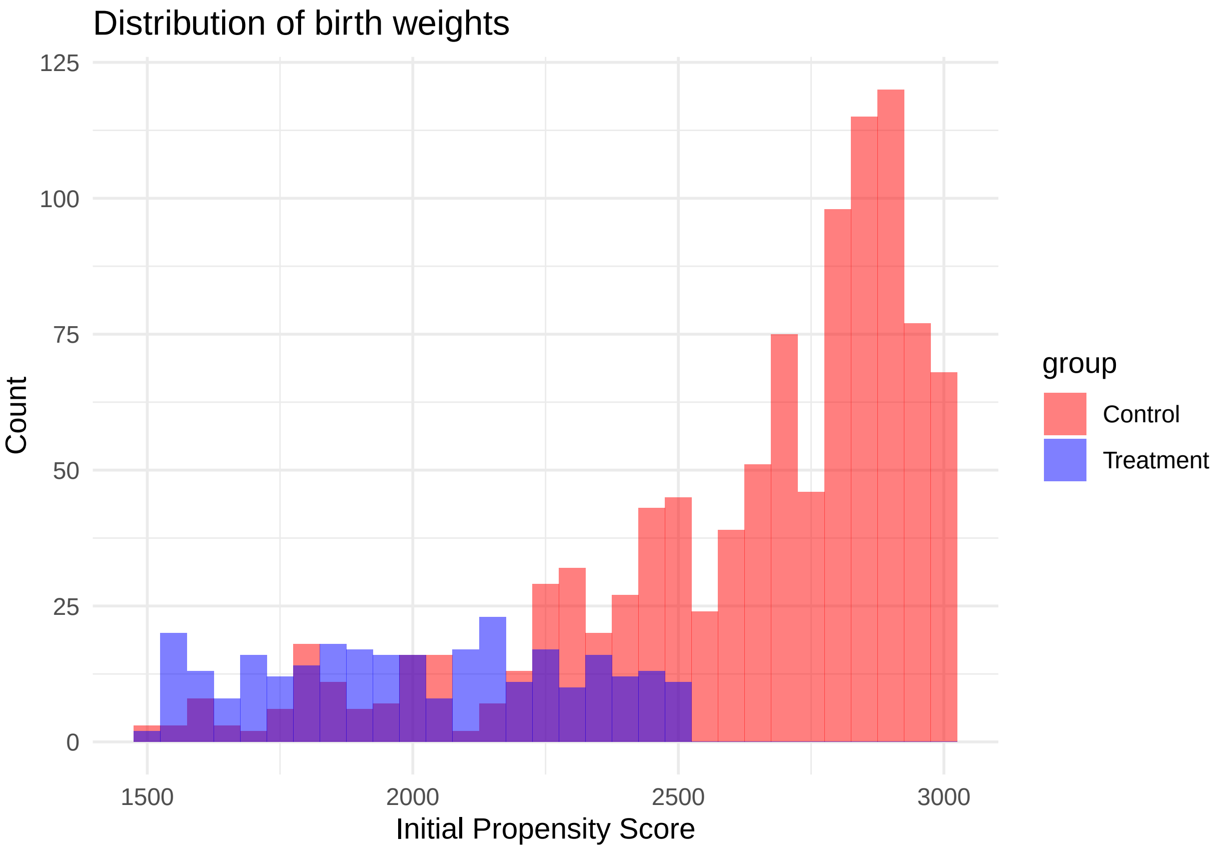
```
# code to check overlap of p.score
library(personalized)
```

```
## Loading required package: glmnet

## Loaded glmnet 4.1-8

## Loading required package: mgcv

## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse

## The following object is masked from 'package:lme4':
##
##     lmList

## This is mgcv 1.8-42. For overview type 'help("mgcv-package")'.

## Loading required package: ggplot2
## Loading required package: plotly

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:MASS':
##
##     select

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

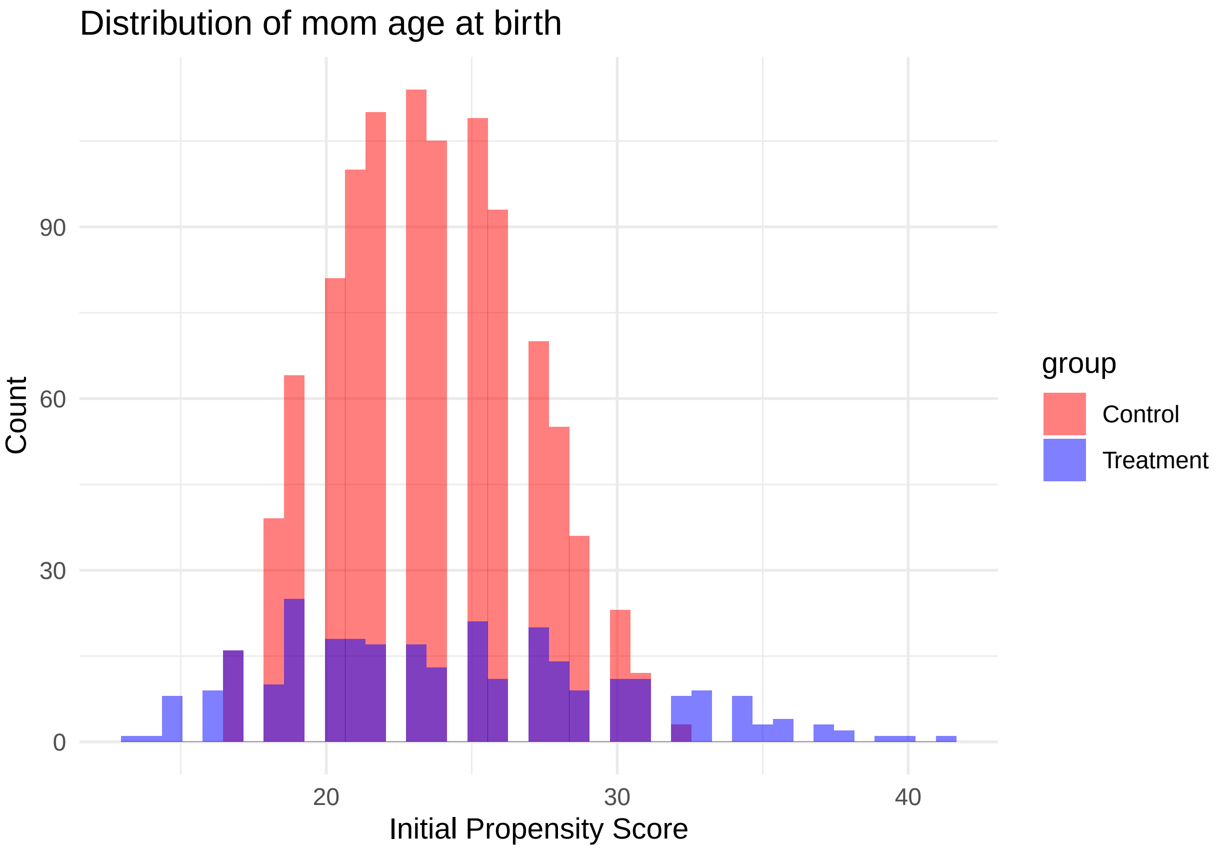
```r
prop.func <- function(x, trt){
    propensity_model <- glm(treat ~ momage + b.marr + factor(momed) + work.dur + prenatal + cig + sex +
                                     bw + bwg + preterm + black + hispanic + white + lths +
                                     hs + ltcoll + college + dayskidh + income,family = binomial,data=i

    initial_pscore <- predict(propensity_model,type = "response")
    initial_pscore
}
check.overlap(x = ihdp,
              trt = ihdp$treat,
              propensity.func = prop.func)
```



```
# overlap of one covariate
# Combine the treatment and control datasets and create a variable to indicate the group
ihdp$group <- ifelse(ihdp$treat == 1, "Treatment", "Control")

# Plotting with ggplot2
ggplot(ihdp, aes(x = bw, fill = group)) +
  geom_histogram(position="identity",alpha = 0.5, binwidth = 50) +
  labs(title = "Distribution of birth weights",
       x = "Initial Propensity Score",
       y = "Count") +
  scale_fill_manual(values = c("Treatment" = "blue", "Control" = "red")) + # Colors for each group
  theme_minimal()
```



```r
# overlap of another covariate
ggplot(ihdp, aes(x = momage, fill = group)) +
  geom_histogram(position="identity",alpha = 0.5, binwidth = 0.7) +
  labs(title = "Distribution of mom age at birth",
       x = "Initial Propensity Score",
       y = "Count") +
  scale_fill_manual(values = c("Treatment" = "blue", "Control" = "red")) + # Colors for each group
  theme_minimal()
```



Part b (5 points)
Interpreting Overlap. What do these plots reveal about the overlap required to estimate our estimand of interest?

The plots above show the imbalance of the raw data before matching and there are overlaps for the p score and the two corvariate(child's birth weights and mom age of birth). We need to do the pscore matching later to decrease the influence of the covariate in casual inference and decrease the bias of estimation of estimand of interest, in this case, ATT.

Part c (5 points) Examining Balance. You will build your own function to check balance! This function should take as inputs (at least) the data frame created in Question 1, the vector with the covariate names chosen in Question 1, and the weights created in Question 3. It should output the following:

1) Mean in the pre-match treatment group
2) Mean in the pre-match control group
3) Mean in the matched treatment group*
4) Mean in the matched control group
5) Pre-match mean difference (standardized for continuous variables, not standardized for binary variables)
6) Matched mean difference (standardized for continuous variables, not standardized for binary variables)
7) Ratio of standard deviations across pre-match groups (control/treated)
8) Ratio of standard deviations across matched groups (control/treated)

I provide a "unit test" of this function below to help ensure that you are doing the right thing.

This will only differ from column (1) if you restrict your dataset to observations with common support.

```
is_binary <- function(covariates) {
  length(unique(covariates)) == 2
}
```

```r
check_balance <- function(data, covariates, weights) {
  # Split the original data into treatment and control groups
  treated <- data[data$treat == 1, ]
  control <- data[data$treat == 0, ]

  n_treated <- nrow(treated)
  n_control <- nrow(control)
  # Split the matched data similarly
  treated_matched <- treated
  # you must subset 'weights' just like you did with 'control'.
  control_weights <- weights[data$treat == 0]

# Let's also make sure that 'control_weights' is a whole number since you can't replicate rows fraction
# If weights are floating-point numbers, they should be very close to integer values, and you can round
  control_weights <- round(control_weights)

# Now, replicate the indices of 'control' based on 'control_weights'.
  indices_to_repeat <- rep(seq_along(control_weights), times = control_weights)

# Subset 'control' based on these indices to create your matched control set.
  control_matched <- control[indices_to_repeat, ]

  # Functions to calculate means and standard deviations
  calc_means <- function(df) sapply(df[covariates], mean, na.rm = TRUE)
  calc_sds <- function(df) sapply(df[covariates], sd, na.rm = TRUE)

  calc_pooled_sd <- function(sd_treat, sd_control, n_treat, n_control) {
  sqrt(((n_treat - 1) * sd_treat^2 + (n_control - 1) * sd_control^2) / (n_treat + n_control - 2))
  }

  # Calculate the pre-matching and post-matching statistics
  pre_means_treated <- calc_means(treated)
  pre_means_control <- calc_means(control)
  post_means_treated <- calc_means(treated_matched)
  post_means_control <- calc_means(control_matched)

  pre_sds_treated <- calc_sds(treated)
  pre_sds_control <- calc_sds(control)
  post_sds_treated <- calc_sds(treated_matched)
  post_sds_control <- calc_sds(control_matched)

  pre_pooled_sd <- calc_pooled_sd(pre_sds_treated, pre_sds_control, n_treated, n_control)
  post_pooled_sd <- calc_pooled_sd(post_sds_treated, post_sds_control, n_treated, n_control)

  # Calculate mean differences and standard deviation ratios
  calculate_differences <- function(pre_treated, pre_control, post_treated, post_control) {
    pre_diff <- pre_treated - pre_control
    post_diff <- post_treated - post_control
    list(pre = pre_diff, post = post_diff)
  }

  # Correcting the calculation of mean differences and standard deviation ratios
  pre_mean_diff <- pre_means_treated - pre_means_control
  post_mean_diff <- post_means_treated -post_means_control
  # Calculate the ratios of standard deviations
  pre_ratio_std <- pre_sds_control / pre_sds_treated
  post_ratio_std <- post_sds_control / post_sds_treated

  binary_flags <- sapply(data[covariates], is_binary)
  # Combine everything into a data frame, ensuring that we're using list elements correctly
  balance_table <- data.frame(
    variable = covariates,
    mn1 = round(pre_means_treated,3),
    mn0 = round(pre_means_control,3),
    mn1.m = round(post_means_treated,3),
    mn0.m = round(post_means_control,3),
    diff = round(ifelse(binary_flags,pre_mean_diff,pre_mean_diff/pre_pooled_sd),3),
    diff.m = round(ifelse(binary_flags,post_mean_diff,post_mean_diff/post_pooled_sd),3),
    ratio = round(pre_ratio_std,3),
    ratio.m = round(post_ratio_std,3)
  )

  return(balance_table)
}
```

Unit Test. **Show the results of your balance function on a simple example where the propensity score is fit using logistic regression on bw and b.marr and the matching is performed using 1-1 nearest neighbor matching with replacement.** If your results match these you can be reasonably sure you built the function correctly.

```
              mn1       mn0     mn1.m     mn0.m    diff  diff.m  ratio  ratio.m
bw       2008.648  2629.482  2008.648  2001.838  -2.191   0.024  1.175    1.044
b.marr      0.431     0.595     0.431     0.486  -0.164  -0.055  0.000    0.000
```

```
# show balance function matches unit test here
propensity1 <- glm(treat ~ b.marr + bw,family = binomial,data=ihdp)
pscore1 <- predict(propensity1,type = "response")

matches1 <- matching(z=ihdp$treat, score=pscore1,replace=TRUE)
matched1 <- matches1$cnts

cov_names1 <- c("bw","b.marr")
check_balance(data = ihdp, covariates = cov_names1, weights = matched1)
```

```
##        variable      mn1      mn0    mn1.m    mn0.m   diff diff.m ratio ratio.m
## bw           bw 2008.648 2629.482 2008.648 2001.838 -1.924  0.023 1.175   1.044
## b.marr   b.marr    0.431    0.595    0.431    0.486 -0.164 -0.055 0.990   1.009
```

Part d (5 points) Using your new balance function, check of the balance for your confounders. Make sure to print your balance statistics.

```
#print balance of all confounders
check_balance(ihdp,cov_names,matched)
```

```
##          variable      mn1      mn0    mn1.m    mn0.m   diff diff.m ratio
## momage     momage   24.445   23.541   24.445   23.934  0.228  0.126 0.552
## b.marr     b.marr    0.431    0.595    0.431    0.552 -0.164 -0.121 0.990
## momed       momed    1.966    1.946    1.966    1.972  0.022 -0.007 0.822
## work.dur work.dur    0.590    0.578    0.590    0.572  0.012  0.017 1.003
## prenatal prenatal     0.955     0.976     0.955     0.986 -0.021 -0.031 0.743
## cig           cig     0.352     0.428     0.352     0.366 -0.076 -0.014 1.035
## sex           sex     0.507     0.544     0.507     0.541 -0.037 -0.034 0.995
## bw             bw  2008.648  2629.482  2008.648  2000.698 -1.924  0.027 1.175
## bwg           bwg     0.490     0.928     0.490     0.510 -0.439 -0.021 0.516
## preterm   preterm     6.072     2.406     6.072     6.413  1.543 -0.125 1.295
## black       black     0.503     0.377     0.503     0.431  0.127  0.072 0.968
## hispanic hispanic     0.093     0.185     0.093     0.128 -0.092 -0.034 1.336
## white       white     0.403     0.438     0.403     0.441 -0.034 -0.038 1.010
## lths         lths     0.434     0.341     0.434     0.383  0.094  0.052 0.955
## hs             hs     0.283     0.422     0.283     0.362 -0.140 -0.079 1.095
## ltcoll     ltcoll     0.166     0.187     0.166     0.155 -0.022  0.010 1.049
## college   college     0.117     0.050     0.117     0.100  0.068  0.017 0.674
## dayskidh dayskidh    14.686     6.021    14.686    13.143  0.910  0.105 0.794
## income     income 21347.394 27330.257 21347.394 24315.154 -0.084 -0.038 3.822
##          ratio.m
## momage     0.576
## b.marr     1.004
## momed      0.937
## work.dur   1.006
## prenatal   0.564
## cig        1.009
## sex        0.997
## bw         1.064
## bwg        1.000
## preterm    1.521
## black      0.990
## hispanic   1.148
## white      1.012
## lths       0.981
## hs         1.067
## ltcoll     0.974
## college    0.933
## dayskidh   1.377
## income     4.234
```

Part e (5 points) How do you interpret the resulting balance? In particular what are your concerns with regard to covariates that are not well balanced (3-4 sentences at most).

Some covariates has made the ratio of the standard deviation after matching much more close to 1 than ratio of standard deviation and the mean difference after matching is much more close to 0 than the mean difference before matching, which means some covariates has become a little bit more balanced. However, three of the covariates are not well balanced, namely, momage,b.marr and preterm. These three have more deviations from 1 of ratio after matching than before matching and this means these three covariates are imbalanced.

Question 5: Creating a better matching model It is rare that your first specification of the propensity score model or choice of matching method is the best. Your goal in this assignment is to achieve an absolute value standardized difference in means of lower than .11 for all confounders. Note in practice you would want to get the best balance possible but for this assignment only you can use .11 as the goal. You will lose 2 points for each confounder that is equal or above .11. *note there are 125 possible points in this assignment.*

```
propensity_model <- glm(treat ~ momage + b.marr + factor(momed) + work.dur + prenatal + cig + sex + bw

ihdp$pscore <- predict(propensity_model,type = "response")
```

```
library(MatchIt)

# Perform radius matching with a specified caliper
#caliper_val <- 0.05
match_out <- matchit(treat ~ pscore, data = ihdp, method = "full",distance="logit", caliper=0.01)

check_balance(ihdp,cov_names,match_out$weights)
```

```
##          variable       mn1       mn0     mn1.m     mn0.m   diff diff.m ratio
## momage     momage    24.445    23.541    24.445    23.757  0.228  0.173 0.552
## b.marr     b.marr     0.431     0.595     0.431     0.495 -0.164 -0.064 0.990
## momed       momed     1.966     1.946     1.966     1.901  0.022  0.066 0.822
## work.dur work.dur     0.590     0.578     0.590     0.535  0.012  0.055 1.003
## prenatal prenatal     0.955     0.976     0.955     0.979 -0.021 -0.023 0.743
## cig           cig     0.352     0.428     0.352     0.380 -0.076 -0.028 1.035
## sex           sex     0.507     0.544     0.507     0.535 -0.037 -0.028 0.995
## bw             bw  2008.648  2629.482  2008.648  2004.072 -1.924  0.015 1.175
## bwg           bwg     0.490     0.928     0.490     0.535 -0.439 -0.045 0.516
## preterm   preterm     6.072     2.406     6.072     6.089  1.543 -0.006 1.295
## black       black     0.503     0.377     0.503     0.457  0.127  0.046 0.968
## hispanic hispanic     0.093     0.185     0.093     0.144 -0.092 -0.051 1.336
## white       white     0.403     0.438     0.403     0.398 -0.034  0.005 1.010
## lths         lths     0.434     0.341     0.434     0.414  0.094  0.020 0.955
## hs             hs     0.283     0.422     0.283     0.364 -0.140 -0.081 1.095
## ltcoll     ltcoll     0.166     0.187     0.166     0.128 -0.022  0.037 1.049
## college   college     0.117     0.050     0.117     0.094  0.068  0.024 0.674
## dayskidh dayskidh    14.686     6.021    14.686    13.595  0.910  0.070 0.794
## income     income 21347.394 27330.257 21347.394 27982.643 -0.084 -0.069 3.822
##          ratio.m
## momage     0.556
## b.marr     1.009
## momed      0.921
## work.dur   1.014
## prenatal   0.699
## cig        1.016
## sex        0.997
## bw         1.056
## bwg        0.997
## preterm    1.532
## black      0.996
## hispanic   1.209
## white      0.998
## lths       0.993
## hs         1.068
## ltcoll     0.900
## college    0.905
## dayskidh   1.480
## income     5.219
```

```
# for imbalanced covariates
match_out_mahalanobis <- matchit(treat ~ momage, data = ihdp, distance="logit",method = "nearest")

# check these covariates again
check_balance(ihdp, c("momage"), match_out_mahalanobis$weights)
```

```
##        variable    mn1    mn0  mn1.m  mn0.m  diff diff.m ratio ratio.m
## momage   momage 24.445 23.541 24.445 23.938 0.228  0.105 0.552   0.769
```

Part a (5 points) In part a you will explore fitting different propensity score models and/or using different matching techniques to improve the balance. This will likely take many attempts. Report the code you used to fit your final propensity score model (i.e. the one that creates the best balance in your estimation) and create matches using this estimated score.

```
# final pscore model and matching code
propensity_match <- function(data, treat_formula, covariate_names, distance_method = "logit", caliper_va

  # Fit the propensity score model
  propensity_model <- glm(treat_formula, family = binomial, data = data)

  # Compute propensity scores
  data$pscore <- predict(propensity_model, type = "response")

  # Perform matching using the full method
  match_out <- matchit(treat ~ pscore, data = data, method = "full", distance = distance_method, calipe

  # Perform nearest neighbor matching for imbalanced covariates
  match_out_mahalanobis <- matchit(treat ~ momage, data = data, distance = distance_method, method = "ne

  # Check balance for "momage" covariate
  balance_check_full <- check_balance(data,cov_names,match_out$weights)
  balance_check_momage <- check_balance(data, c("momage"), match_out_mahalanobis$weights)

  return(list(
    match_out = match_out,
    match_out_momage = match_out_mahalanobis,
    balance_full = balance_check_full,
    balance_momage = balance_check_momage
  ))
}

result <- propensity_match(ihdp, treat ~ momage + b.marr + factor(momed) + work.dur + prenatal + cig + s
```

Part b (20 points) Using your balance function, print the balance of all your confounders using your final propensity score model from part a to create the propensity score and subsequent matches.

```
# print balance
print(result$balance_full)
```

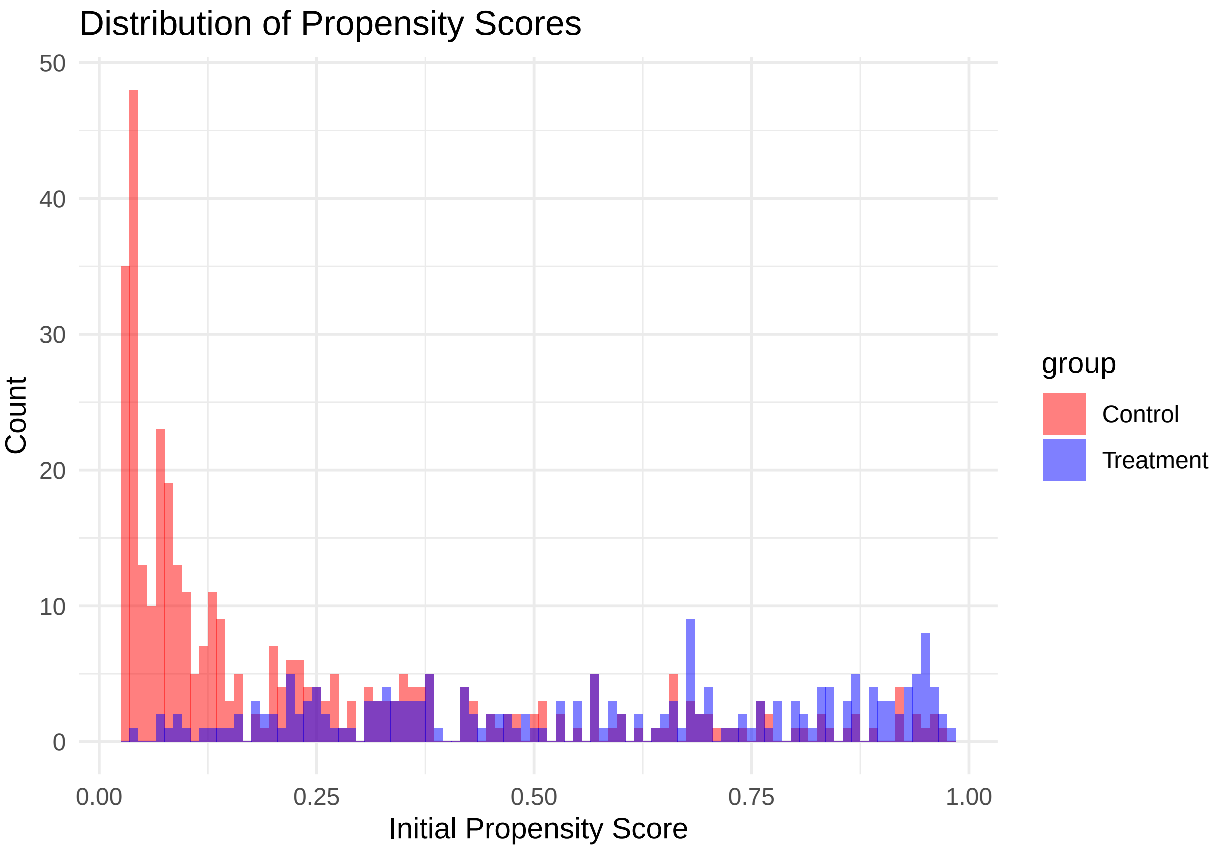
```
##          variable      mn1      mn0    mn1.m    mn0.m   diff diff.m ratio
## momage     momage   24.445   23.541   24.445   23.757  0.228  0.173 0.552
## b.marr     b.marr    0.431    0.595    0.431    0.495 -0.164 -0.064 0.990
## momed       momed    1.966    1.946    1.966    1.901  0.022  0.066 0.822
## work.dur work.dur    0.590    0.578    0.590    0.535  0.012  0.055 1.003
## prenatal prenatal    0.955    0.976    0.955    0.979 -0.021 -0.023 0.743
## cig           cig    0.352    0.428    0.352    0.380 -0.076 -0.028 1.035
## sex           sex    0.507    0.544    0.507    0.535 -0.037 -0.028 0.995
## bw             bw 2008.648 2629.482 2008.648 2004.072 -1.924  0.015 1.175
## bwg           bwg    0.490    0.928    0.490    0.535 -0.439 -0.045 0.516
## preterm   preterm    6.072    2.406    6.072    6.089  1.543 -0.006 1.295
## black       black    0.503    0.377    0.503    0.457  0.127  0.046 0.968
## hispanic hispanic     0.093     0.185     0.093     0.144 -0.092 -0.051 1.336
## white       white     0.403     0.438     0.403     0.398 -0.034  0.005 1.010
## lths         lths     0.434     0.341     0.434     0.414  0.094  0.020 0.955
## hs             hs     0.283     0.422     0.283     0.364 -0.140 -0.081 1.095
## ltcoll     ltcoll     0.166     0.187     0.166     0.128 -0.022  0.037 1.049
## college   college     0.117     0.050     0.117     0.094  0.068  0.024 0.674
## dayskidh dayskidh    14.686     6.021    14.686    13.595  0.910  0.070 0.794
## income     income 21347.394 27330.257 21347.394 27982.643 -0.084 -0.069 3.822
##          ratio.m
## momage     0.556
## b.marr     1.009
## momed      0.921
## work.dur   1.014
## prenatal   0.699
## cig        1.016
## sex        0.997
## bw         1.056
## bwg        0.997
## preterm    1.532
## black      0.996
## hispanic   1.209
## white      0.998
## lths       0.993
## hs         1.068
## ltcoll     0.900
## college    0.905
## dayskidh   1.480
## income     5.219
```

```r
print(result$balance_momage)
```

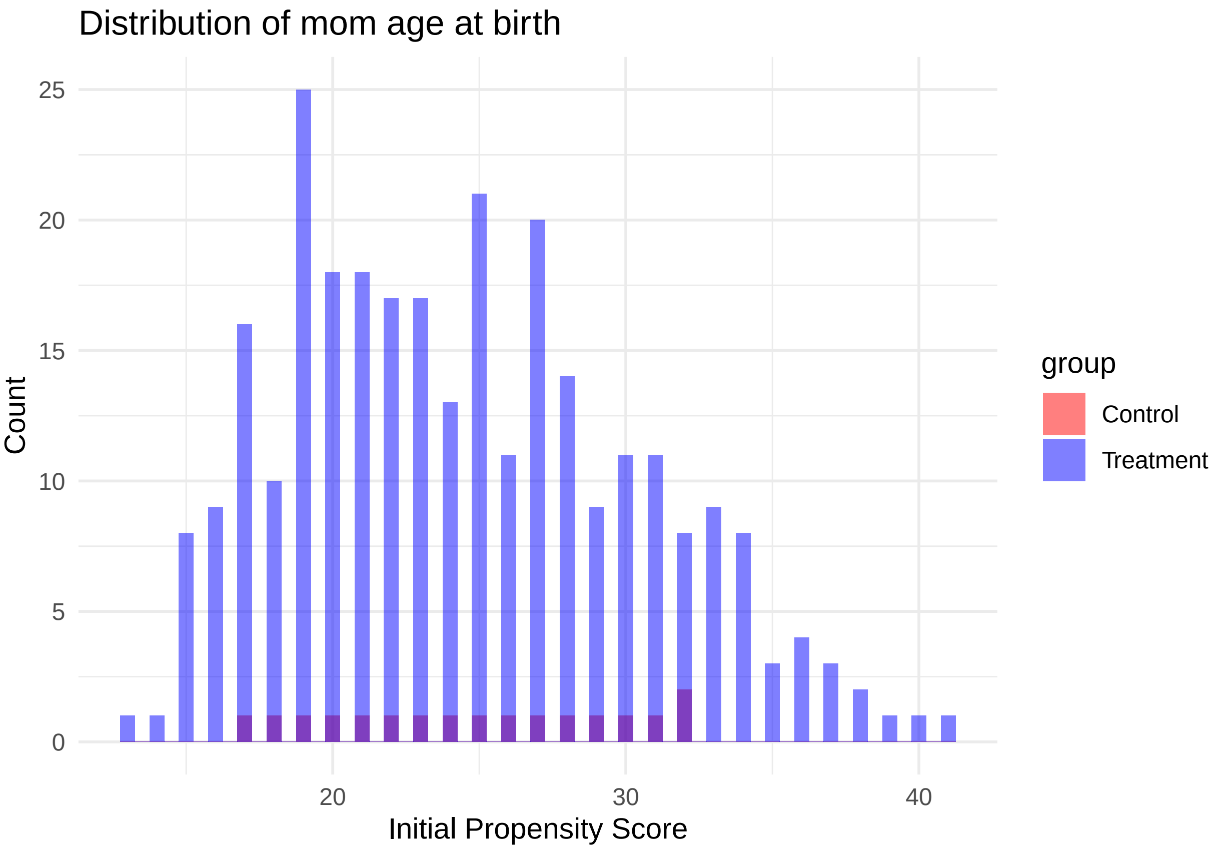
```
##        variable    mn1    mn0  mn1.m mn0.m  diff diff.m ratio ratio.m
## momage   momage 24.445 23.541 24.445  24.5 0.228  -0.01 0.552   0.878
```

Part c (5 points) Examining Overlap of matched data. Check overlap on the matched data (that is the data after matching) using some diagnostic plots. Check overlap for the propensity scores as well as the same two covariates from earlier . Note that it may be necessary to exclude some observations from the plots if they are being obscured in ways similar to the example discussed in class.

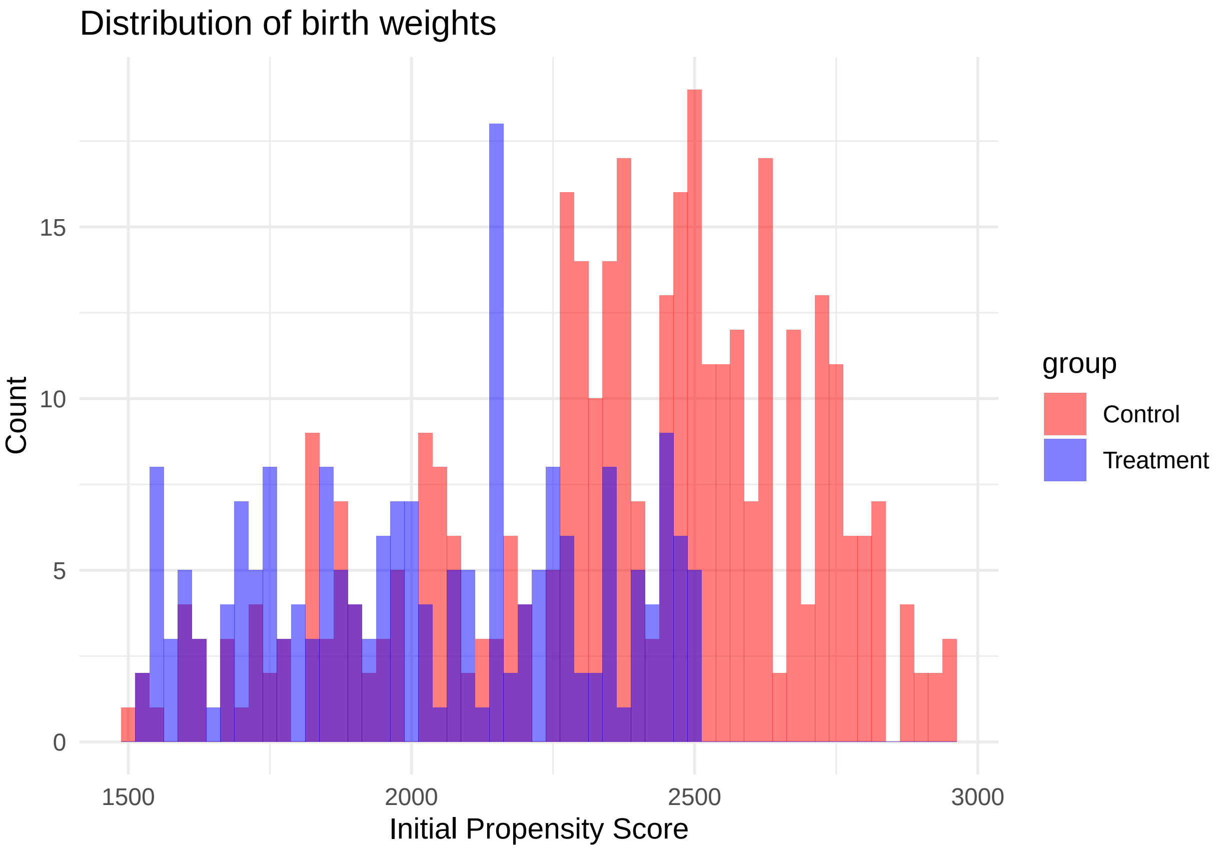
```r
# overlap of p.score
library(ggplot2)

result <- propensity_match(ihdp, treat ~ momage + b.marr + factor(momed) + work.dur + prenatal + cig +
matched_data <- match.data(result$match_out)  # Extract matched data from match_out object
matched_data_momage <- match.data(result$match_out_momage)


ggplot(matched_data, aes(x = initial_pscore, fill = group)) +
  geom_histogram(position="identity",alpha = 0.5, binwidth = 0.01) +
  labs(title = "Distribution of Propensity Scores",
       x = "Initial Propensity Score",
       y = "Count") +
  scale_fill_manual(values = c("Treatment" = "blue", "Control" = "red")) + # Colors for each group
  theme_minimal()
```



```
# overlap of a covariate
ggplot(matched_data_momage, aes(x = momage, fill = group)) +
  geom_histogram(position="identity",alpha = 0.5, binwidth = 0.5) +
  labs(title = "Distribution of mom age at birth",
       x = "Initial Propensity Score",
       y = "Count") +
  scale_fill_manual(values = c("Treatment" = "blue", "Control" = "red")) + # Colors for each group
  theme_minimal()
```



```r
# overlap of another covariate
ggplot(matched_data, aes(x = bw, fill = group)) +
  geom_histogram(position="identity",alpha = 0.5, binwidth = 25) +
  labs(title = "Distribution of birth weights",
       x = "Initial Propensity Score",
       y = "Count") +
  scale_fill_manual(values = c("Treatment" = "blue", "Control" = "red")) + # Colors for each group
  theme_minimal()
```



Question 6: Using IPTW. Part a Model (5 points)
Estimate propensity scores and use this pscore model to create IPTW weights. Show all your code used to create your weights.

Make sure that you create weights specific to the correct estimand.

```
# code for IPTW model
library(MatchIt)
ihdp$weights <- ifelse(ihdp$treat == 1,
                        1,
                        ihdp$pscore / (1 - ihdp$initial_pscore))
```

Part b Balance (5 points) Using your balance function, check the balance from your IPTW model

```
# IPTW balance
check_balance(ihdp,cov_names,ihdp$weights)
```

```
##            variable      mn1      mn0    mn1.m    mn0.m   diff diff.m ratio
## momage       momage   24.445   23.541   24.445   24.167  0.228  0.067 0.552
## b.marr       b.marr    0.431    0.595    0.431    0.423 -0.164  0.008 0.990
## momed         momed    1.966    1.946    1.966    1.981  0.022 -0.016 0.822
## work.dur   work.dur    0.590    0.578    0.590    0.450  0.012  0.140 1.003
## prenatal   prenatal    0.955    0.976    0.955    0.989 -0.021 -0.034 0.743
## cig             cig    0.352    0.428    0.352    0.251 -0.076  0.100 1.035
## sex             sex    0.507    0.544    0.507    0.418 -0.037  0.089 0.995
## bw               bw 2008.648 2629.482 2008.648 1845.891 -1.924  0.621 1.175
## bwg             bwg    0.490    0.928    0.490    0.312 -0.439  0.177 0.516
## preterm     preterm    6.072    2.406    6.072    7.691  1.543 -0.581 1.295
## black       black     0.503     0.377     0.503     0.550  0.127 -0.047 0.968
## hispanic hispanic     0.093     0.185     0.093     0.101 -0.092 -0.007 1.336
## white       white     0.403     0.438     0.403     0.349 -0.034  0.054 1.010
## lths         lths     0.434     0.341     0.434     0.389  0.094  0.046 0.955
## hs             hs     0.283     0.422     0.283     0.333 -0.140 -0.051 1.095
## ltcoll     ltcoll     0.166     0.187     0.166     0.185 -0.022 -0.020 1.049
## college   college     0.117     0.050     0.117     0.093  0.068  0.025 0.674
## dayskidh dayskidh    14.686     6.021    14.686    15.119  0.910 -0.032 0.794
## income     income 21347.394 27330.257 21347.394 17703.928 -0.084  0.190 3.822
##          ratio.m
## momage     0.597
## b.marr     0.997
## momed      0.940
## work.dur   1.011
## prenatal   0.494
## cig        0.908
## sex        0.986
## bw         0.904
## bwg        0.927
## preterm    1.552
## black      0.995
## hispanic   1.034
## white      0.971
## lths       0.983
## hs         1.046
## ltcoll     1.045
## college    0.901
## dayskidh   1.257
## income     0.894
```

Question 7: Matching vs IPTW (5 points) Which approach would you choose, your matching model from Question 5 or your IPTW model from question 6, justify your choice? (1 paragraph at most) I prefer to choose matching model from Question 5 because this model has made the absolute values of standardized difference in means less than 0.11 but for IPTW model from Question 6, work.dr,bw,bwg,preterm and income have the standardized difference in means more than 0.11 which means that these five covariates are not well matched. #### Question 8: Estimate the treatment effect with IPTW (5 points)

Estimate the treatment effect for the correct causal estimand using IPTW. Report your point estimate and a corrected standard error.

```r
# outcome model using IPTW
library(MatchIt)
library(survey)
```

```
## Loading required package: grid

## Loading required package: survival

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
##
##     dotchart
```

```r
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
lm_model<-lm(ppvtr.36 ~ treat + momage + b.marr + factor(momed) + work.dur + prenatal + cig + sex + bw +

#summary of the model
summary(lm_model)
```

```
##
## Call:
## lm(formula = ppvtr.36 ~ treat + momage + b.marr + factor(momed) +
##     work.dur + prenatal + cig + sex + bw + bwg + preterm + black +
##     hispanic + white + lths + hs + ltcoll + college + dayskidh +
##     income, data = ihdp, weights = ihdp$weights)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -71.794  -1.651   0.505   2.870 137.484
##
## Coefficients: (5 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.060e+02  7.326e+00  14.475  < 2e-16 ***
## treat            1.028e+01  9.546e-01  10.773  < 2e-16 ***
## momage          -2.897e-02  1.157e-01  -0.250 0.802341
## b.marr           2.035e-01  1.147e+00   0.177 0.859207
## factor(momed)2   3.565e+00  1.158e+00   3.079 0.002123 **
## factor(momed)3   5.679e+00  1.476e+00   3.847 0.000125 ***
## factor(momed)4   1.811e+01  1.909e+00   9.491  < 2e-16 ***
## work.dur         5.054e+00  1.050e+00   4.812 1.66e-06 ***
## prenatal        -3.596e+00  2.917e+00  -1.233 0.217869
## cig              1.694e+00  1.056e+00   1.605 0.108844
## sex             -2.739e+00  9.450e-01  -2.898 0.003817 **
## bw              -9.120e-03  2.987e-03  -3.054 0.002305 **
## bwg              3.598e+00  1.809e+00   1.989 0.046963 *
## preterm          4.451e-01  1.997e-01   2.228 0.026030 *
## black           -1.589e+01  1.181e+00 -13.450  < 2e-16 ***
## hispanic        -1.105e+01  1.662e+00  -6.652 4.24e-11 ***
## white                   NA         NA      NA       NA
## lths                    NA         NA      NA       NA
## hs                      NA         NA      NA       NA
## ltcoll                  NA         NA      NA       NA
## college                 NA         NA      NA       NA
## dayskidh        -3.021e-01  4.298e-02  -7.029 3.35e-12 ***
## income           4.201e-05  1.343e-05   3.128 0.001801 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.84 on 1302 degrees of freedom
## Multiple R-squared:  0.4032, Adjusted R-squared:  0.3954
## F-statistic: 51.75 on 17 and 1302 DF,  p-value: < 2.2e-16
```

```
cat("The point estimate is 10.28 and the corrected standard error is 0.95.")
```

```
## The point estimate is 10.28 and the corrected standard error is 0.95.
```

Question 9: Estimate the treatment effect with Matching (5 points) Estimate the treatment effect for the correct causal estimand using your matching model from Question 5. Report your point estimate and and a corrected standard error.

```
# outcome model using matching
results <- propensity_match(ihdp, treat ~ momage + b.marr + factor(momed) + work.dur + prenatal + cig +

my_weights <- result$match_out$weights

weighted_analysis <- lm(ppvtr.36 ~ treat + momage + b.marr + factor(momed) + work.dur + prenatal + cig

summary(weighted_analysis)
```

```
## 
## Call:
## lm(formula = ppvtr.36 ~ treat + momage + b.marr + factor(momed) + 
##     work.dur + prenatal + cig + sex + bw + bwg + preterm + black + 
##     hispanic + white + lths + hs + ltcoll + college + dayskidh + 
##     income, data = ihdp, weights = my_weights)
## 
## Weighted Residuals:
##    Min     1Q Median     3Q    Max 
## -97.56   0.00   0.00   0.00  87.28 
## 
## Coefficients: (5 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     1.125e+02  1.142e+01   9.857  < 2e-16 ***
## treat           1.163e+01  1.541e+00   7.545 1.92e-13 ***
## momage         -2.140e-01  2.039e-01  -1.049 0.294506    
## b.marr          1.271e+00  1.777e+00   0.716 0.474597    
## factor(momed)2  5.262e+00  1.818e+00   2.894 0.003951 ** 
## factor(momed)3  1.134e+01  2.447e+00   4.635 4.48e-06 ***
## factor(momed)4  2.119e+01  3.149e+00   6.727 4.41e-11 ***
## work.dur        5.625e+00  1.652e+00   3.405 0.000710 ***
## prenatal        1.323e+00  4.324e+00   0.306 0.759669    
## cig             3.262e+00  1.687e+00   1.933 0.053724 .  
## sex             4.086e+00  1.481e+00   2.759 0.006001 ** 
## bw             -1.745e-02  4.750e-03  -3.673 0.000263 ***
## bwg             9.006e+00  2.882e+00   3.125 0.001873 ** 
## preterm         3.885e-01  3.428e-01   1.133 0.257680    
## black          -1.533e+01  1.838e+00  -8.341 6.16e-16 ***
## hispanic       -7.517e+00  2.463e+00  -3.052 0.002381 ** 
## white                  NA         NA      NA       NA    
## lths                   NA         NA      NA       NA    
## hs                     NA         NA      NA       NA    
## ltcoll                 NA         NA      NA       NA    
## college                NA         NA      NA       NA    
## dayskidh       -2.845e-01  6.212e-02  -4.580 5.77e-06 ***
## income         -1.068e-05  9.205e-06  -1.160 0.246547    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.07 on 541 degrees of freedom
## Multiple R-squared:  0.3982, Adjusted R-squared:  0.3793
## F-statistic: 21.06 on 17 and 541 DF,  p-value: < 2.2e-16
```

```
cat("The point estimate is 11.63 and the corrected standard error is 1.54.")
```

```
## The point estimate is 11.63 and the corrected standard error is 1.54.
```

Question 10: Causal Interpretation (10 points) Provide a causal interpretation of your estimate of your preferred model (the model fit in Question 8 or 9). Include all relevant causal assumptions.

I prefer the model with IPTW because it has got less corrected standard error and there is some difference between two estimates of ATT. Causal assumptions include: SUVTA(i.e. no interaction effect between covariates), ignorability(i.e. can ignore the effects of underlying variables that has not been included), there are no post-treatment covariates(assume income one year after birth as a pre-treatment variable) in this dataset.

Question 11: Comparison to linear regression **Part a** (5 points) Fit a regression of your outcomes to the treatment indicator and covariates.

```
# fit linear model
lin_model <- lm(ppvtr.36~treat+momage + b.marr + factor(momed) + work.dur + prenatal + cig + sex + bw +
summary(lin_model)
```

```
##
## Call:
## lm(formula = ppvtr.36 ~ treat + momage + b.marr + factor(momed) +
##     work.dur + prenatal + cig + sex + bw + bwg + preterm + black +
##     hispanic + white + lths + hs + ltcoll + college + dayskidh +
##     income, data = ihdp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.331  -9.637   0.757  11.132  54.525
##
## Coefficients: (5 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.976e+01  6.805e+00  11.721  < 2e-16 ***
## treat            1.154e+01  1.488e+00   7.753 1.81e-14 ***
## momage          -1.815e-01  1.289e-01  -1.408 0.159263
## b.marr           2.611e+00  1.094e+00   2.386 0.017154 *
## factor(momed)2   6.329e+00  1.143e+00   5.536 3.75e-08 ***
## factor(momed)3   1.007e+01  1.463e+00   6.881 9.18e-12 ***
## factor(momed)4   1.607e+01  2.244e+00   7.164 1.31e-12 ***
## work.dur         4.327e+00  9.955e-01   4.347 1.49e-05 ***
## prenatal         2.381e+00  2.777e+00   0.857 0.391395
## cig              1.441e+00  1.010e+00   1.427 0.153880
## sex              9.980e-01  9.226e-01   1.082 0.279602
## bw               1.710e-03  2.182e-03   0.784 0.433280
## bwg             -1.208e+00  2.031e+00  -0.595 0.552258
## preterm          4.828e-01  2.393e-01   2.017 0.043872 *
## black           -1.643e+01  1.165e+00 -14.106  < 2e-16 ***
## hispanic        -1.315e+01  1.410e+00  -9.322  < 2e-16 ***
## white                   NA         NA      NA       NA
## lths                   NA         NA      NA       NA
## hs                     NA         NA      NA       NA
## ltcoll                 NA         NA      NA       NA
## college                NA         NA      NA       NA
## dayskidh       -2.338e-01  6.009e-02  -3.891 0.000105 ***
## income          4.904e-06  6.713e-06   0.730 0.465229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.62 on 1302 degrees of freedom
## Multiple R-squared:  0.3426, Adjusted R-squared:  0.334
## F-statistic: 39.91 on 17 and 1302 DF,  p-value: < 2.2e-16
```

Part b (5 points) Interpret the results of the program (coefficient on treat) non-causally. Since the treatment effect with estimand as ATT is positive, the IHDP intervention can increase the value of ppvtr.36, which is the IQ of the children at 3 years old.

Part c (5 points) Why might we prefer the results from the propensity score approach to the linear regression results in terms of identifying a causal effect?

Because the linear regression approach needs the balance of the covariates between treatment group and control group and as can be seen in EDA part, the distribution between the treament group and control group for some covariates are quite different and imbalance. And the propensity score mathcing method can solve the problem directly and reduce the influence of imbalance distribution of covariates to find the casual relationship. That's why we prefer propensity socre matching on this dataset.