Problem Set 2

Se Hyun Kim

Problem 1A

$\pi(X_i) \equiv Pr(D_i = 1|X_i)$ The propensity score is the probability of getting the treatment given (conditional on) X_i (which in this case is a covariate vector).

Problem 1B

$\mathbb{E}[Y_i|D_i = 1, \pi(X_i)] - \mathbb{E}[Y_i|D_i = 0, \pi(X_i)]$ is the difference of expectations of the treated and control potential outcomes given a propensity score $\pi(X_i)$. $\pi(X_i)$ is the probability density function of the propensity score. The whole integral is weighted sum of total ATE, which is the expected ATE.
The identification result holds under the assumptions of the conditional ignorability (treatment assignment D is independent of the potential outcomes $Y_i(1)$ (treated) and $Y_i(0)$ (untreated) given a propensity score $\pi(X_i)$) and common support (every unit can be treated or not treated. not 100%).
The assumption can be expressed in terms of $\pi(X_i)$ as follows
1. Conditional Ignorability: $Y_i(0), Y_i(1) \perp D_i|\pi(X_i)$
2. Common Support: $0 < Pr(D_i = 1|\pi(X_i)) < 1$

Problem 2A

Under the assumption of conditional ignorability and common support, the ATT can be identified in the following way.
$\tau_{ATT} = E[Y_i(1) - Y_i(0)|D_i = 1]$
Based on the law of iterated expectation, this equation can be reformulated as follows. $\tau_{ATT} = E[E[Y_i(1) - Y_i(0)|D_i = 1]D_i = 1]$ This in turn, can be reformulated using the definition of $\mathbb{E}$ as $\tau_{ATT} = \int E[Y_i(1) - Y_i(0)|D_i = 1]f(x|D_i)dx$.
$\tau_{ATT} = \int [E[Y_i|X_i = x, D_i = 1] - E[Y_i|X_i = x, D_i = 0]f(x|D_i = 1)dx$
$\tau_{ATT} = E[\hat{\tau}|D_i = 1]$

Problem 2B

One would pick matching over regression if one wants to avoid assuming certain forms. Under matching, the treatment effect is nonparametrically identified. However, regression assumes linearity. If units have constant treatment effects, regression estimator is unbiased.

Problem 2C

Matching and regression will give the same estimates of the ATT if the control and treatment groups are balanced.

Problem 3A

The Euclidean distance between observations i and j is $d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + ... + (x_{iP} - x_{jP})^2}$ where P is the number of pre-treatment covariates.

Problem 3B

```
# If perseverance had a form, this is what it would look like.
X1 <- rnorm(500, 0, 1)
X2 <- rnorm(500, 0, 1)
X3 <- rnorm(500, 0, 1)
X4 <- rnorm(500, 0, 1)
X5 <- rnorm(500, 0, 1)
X6 <- rnorm(500, 0, 1)
X7 <- rnorm(500, 0, 1)
X8 <- rnorm(500, 0, 1)
X9 <- rnorm(500, 0, 1)
X10 <- rnorm(500, 0, 1)
X11 <- rnorm(500, 0, 1)
X12 <- rnorm(500, 0, 1)
X13 <- rnorm(500, 0, 1)
X14 <- rnorm(500, 0, 1)
X15 <- rnorm(500, 0, 1)
X16 <- rnorm(500, 0, 1)
X17 <- rnorm(500, 0, 1)
X18 <- rnorm(500, 0, 1)
X19 <- rnorm(500, 0, 1)
X20 <- rnorm(500, 0, 1)
X <- replicate(20, 0)
b3 <- cbind(X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X13, X14, X15, X16, X17, X18, X19, X20)
b3 <- rbind(b3, X) %>%
  as_tibble()


dis_closest1 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2)
  if(dis_cov < dis_closest1){
    dis_closest1 <- dis_cov
  } else {
    dis_closest1 <- dis_closest1
  }
}

dis_closest2 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2)
  if(dis_cov < dis_closest2){
    dis_closest2 <- dis_cov
  } else {
    dis_closest2 <- dis_closest2
  }
```

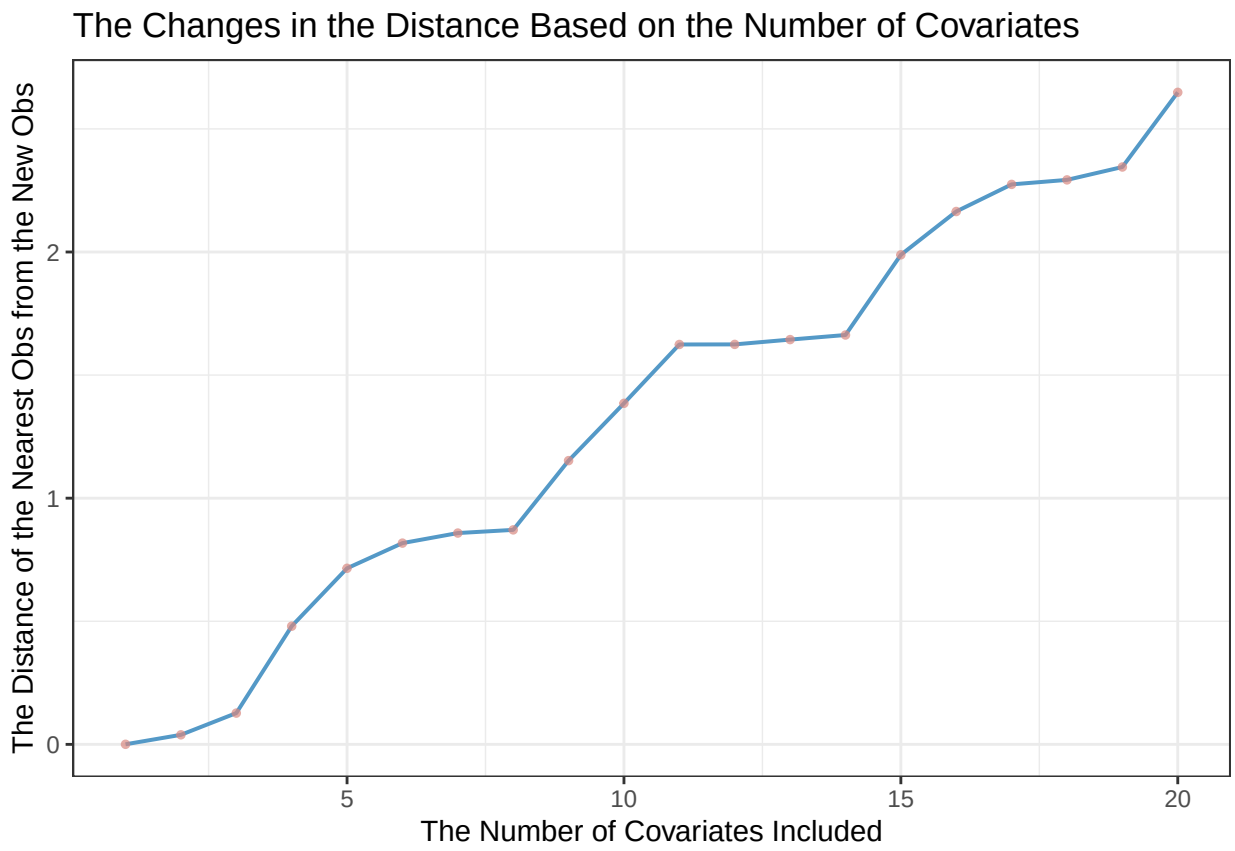
```
}

dis_closest3 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2)
  if(dis_cov < dis_closest3){
    dis_closest3 <- dis_cov
  } else {
    dis_closest3 <- dis_closest3
  }
}

dis_closest4 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2)
  if(dis_cov < dis_closest4){
    dis_closest4 <- dis_cov
  } else {
    dis_closest4 <- dis_closest4
  }
}

dis_closest5 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                  + b3[[i,5]]^2)
  if(dis_cov < dis_closest5){
    dis_closest5 <- dis_cov
  } else {
    dis_closest5 <- dis_closest5
  }
}

dis_closest6 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                  + b3[[i,5]]^2 + b3[[i,6]]^2)
  if(dis_cov < dis_closest6){
    dis_closest6 <- dis_cov
  } else {
    dis_closest6 <- dis_closest6
  }
}

dis_closest7 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                  + b3[[i,5]]^2 + b3[[i,6]]^2 + b3[[i,7]]^2)
  if(dis_cov < dis_closest7){
    dis_closest7 <- dis_cov
  } else {
    dis_closest7 <- dis_closest7
  }
```

```
}

dis_closest8 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                  + b3[[i,5]]^2 + b3[[i,6]]^2 + b3[[i,7]]^2 + b3[[i,8]]^2)
  if(dis_cov < dis_closest8){
    dis_closest8 <- dis_cov
  } else {
    dis_closest8 <- dis_closest8
  }
}

dis_closest9 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                  + b3[[i,5]]^2 + b3[[i,6]]^2 + b3[[i,7]]^2 + b3[[i,8]]^2
                  + b3[[i,9]]^2)
  if(dis_cov < dis_closest9){
    dis_closest9 <- dis_cov
  } else {
    dis_closest9 <- dis_closest9
  }
}

dis_closest10 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                  + b3[[i,5]]^2 + b3[[i,6]]^2 + b3[[i,7]]^2 + b3[[i,8]]^2
                  + b3[[i,9]]^2 + b3[[i,10]]^2)
  if(dis_cov < dis_closest10){
    dis_closest10 <- dis_cov
  } else {
    dis_closest10 <- dis_closest10
  }
}

dis_closest11 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                  + b3[[i,5]]^2 + b3[[i,6]]^2 + b3[[i,7]]^2 + b3[[i,8]]^2
                  + b3[[i,9]]^2 + b3[[i,10]]^2 + b3[[i,11]]^2)
  if(dis_cov < dis_closest11){
    dis_closest11 <- dis_cov
  } else {
    dis_closest11 <- dis_closest11
  }
}

dis_closest12 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                  + b3[[i,5]]^2 + b3[[i,6]]^2 + b3[[i,7]]^2 + b3[[i,8]]^2
```

```
                + b3[[i,9]]^2 + b3[[i,10]]^2 + b3[[i,11]]^2 + b3[[i,12]]^2)
  if(dis_cov < dis_closest12){
    dis_closest12 <- dis_cov
  } else {
    dis_closest12 <- dis_closest12
  }
}

dis_closest13 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                + b3[[i,5]]^2 + b3[[i,6]]^2 + b3[[i,7]]^2 + b3[[i,8]]^2
                + b3[[i,9]]^2 + b3[[i,10]]^2 + b3[[i,11]]^2 + b3[[i,12]]^2
                + b3[[i,13]]^2)
  if(dis_cov < dis_closest13){
    dis_closest13 <- dis_cov
  } else {
    dis_closest13 <- dis_closest13
  }
}

dis_closest14 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                + b3[[i,5]]^2 + b3[[i,6]]^2 + b3[[i,7]]^2 + b3[[i,8]]^2
                + b3[[i,9]]^2 + b3[[i,10]]^2 + b3[[i,11]]^2 + b3[[i,12]]^2
                + b3[[i,13]]^2 + b3[[i,14]]^2)
  if(dis_cov < dis_closest14){
    dis_closest14 <- dis_cov
  } else {
    dis_closest14 <- dis_closest14
  }
}

dis_closest15 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                + b3[[i,5]]^2 + b3[[i,6]]^2 + b3[[i,7]]^2 + b3[[i,8]]^2
                + b3[[i,9]]^2 + b3[[i,10]]^2 + b3[[i,11]]^2 + b3[[i,12]]^2
                + b3[[i,13]]^2 + b3[[i,14]]^2 + b3[[i,15]]^2)
  if(dis_cov < dis_closest15){
    dis_closest15 <- dis_cov
  } else {
    dis_closest15 <- dis_closest15
  }
}

dis_closest16 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                + b3[[i,5]]^2 + b3[[i,6]]^2 + b3[[i,7]]^2 + b3[[i,8]]^2
                + b3[[i,9]]^2 + b3[[i,10]]^2 + b3[[i,11]]^2 + b3[[i,12]]^2
                + b3[[i,13]]^2 + b3[[i,14]]^2 + b3[[i,15]]^2 + b3[[i,16]]^2)
```

```
  if(dis_cov < dis_closest16){
    dis_closest16 <- dis_cov
  } else {
    dis_closest16 <- dis_closest16
  }
}

dis_closest17 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                  + b3[[i,5]]^2 + b3[[i,6]]^2 + b3[[i,7]]^2 + b3[[i,8]]^2
                  + b3[[i,9]]^2 + b3[[i,10]]^2 + b3[[i,11]]^2 + b3[[i,12]]^2
                  + b3[[i,13]]^2 + b3[[i,14]]^2 + b3[[i,15]]^2 + b3[[i,16]]^2
                  + b3[[i,17]]^2)
  if(dis_cov < dis_closest17){
    dis_closest17 <- dis_cov
  } else {
    dis_closest17 <- dis_closest17
  }
}

dis_closest18 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                  + b3[[i,5]]^2 + b3[[i,6]]^2 + b3[[i,7]]^2 + b3[[i,8]]^2
                  + b3[[i,9]]^2 + b3[[i,10]]^2 + b3[[i,11]]^2 + b3[[i,12]]^2
                  + b3[[i,13]]^2 + b3[[i,14]]^2 + b3[[i,15]]^2 + b3[[i,16]]^2
                  + b3[[i,17]]^2 + b3[[i,18]]^2)
  if(dis_cov < dis_closest18){
    dis_closest18 <- dis_cov
  } else {
    dis_closest18 <- dis_closest18
  }
}

dis_closest19 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                  + b3[[i,5]]^2 + b3[[i,6]]^2 + b3[[i,7]]^2 + b3[[i,8]]^2
                  + b3[[i,9]]^2 + b3[[i,10]]^2 + b3[[i,11]]^2 + b3[[i,12]]^2
                  + b3[[i,13]]^2 + b3[[i,14]]^2 + b3[[i,15]]^2 + b3[[i,16]]^2
                  + b3[[i,17]]^2 + b3[[i,18]]^2 + b3[[i,19]]^2)
  if(dis_cov < dis_closest19){
    dis_closest19 <- dis_cov
  } else {
    dis_closest19 <- dis_closest19
  }
}

dis_closest20 <- 100000000000000
for (i in 1:500) {
  dis_cov <- sqrt(b3[[i,1]]^2 + b3[[i,2]]^2 + b3[[i,3]]^2 + b3[[i,4]]^2
                  + b3[[i,5]]^2 + b3[[i,6]]^2 + b3[[i,7]]^2 + b3[[i,8]]^2
```

```
                  + b3[[i,9]]^2 + b3[[i,10]]^2 + b3[[i,11]]^2 + b3[[i,12]]^2
                  + b3[[i,13]]^2 + b3[[i,14]]^2 + b3[[i,15]]^2 + b3[[i,16]]^2
                  + b3[[i,17]]^2 + b3[[i,18]]^2 + b3[[i,19]]^2 + b3[[i,20]]^2)
  if(dis_cov < dis_closest20){
    dis_closest20 <- dis_cov
  } else {
    dis_closest20 <- dis_closest20
  }
}

b3_a <- c(1:20)
b3_b <- c(dis_closest1, dis_closest2, dis_closest3, dis_closest4,
          dis_closest5, dis_closest6, dis_closest7, dis_closest8,
          dis_closest9, dis_closest10, dis_closest11, dis_closest12,
          dis_closest13, dis_closest14, dis_closest15, dis_closest16,
          dis_closest17, dis_closest18, dis_closest19, dis_closest20)
b3_c <- cbind(b3_a, b3_b)%>%
  as_tibble()

b3_c %>%
  ggplot(aes(x = b3_a, y = b3_b)) +
  geom_line(color = "#5499C7",
            size = 0.7) +
  geom_point(alpha = 0.7,
             color = "#D98880",
             size = 1) +
  labs(title = "The Changes in the Distance Based on the Number of Covariates",
       x = "The Number of Covariates Included",
       y = "The Distance of the Nearest Obs from the New Obs") +
  theme_bw()
```



Problem 3C

It demonstrates that as the number of covariates increase, the distance a covariate and the reference point increases. This demonstrates the curse of dimensionality. As the number of covariates increase (dimension increases), the data sparsity exponentially increases.

Problem 4A

```
ne_t <- filter(ne, nsw == 1)
ne_c <- filter(ne, nsw == 0)
nsw_effect4a <- mean(ne_t$re78) - mean(ne_c$re78)
nsw_sd4a <- sqrt(var(ne_t$re78)/length(ne_t$re78) + var(ne_c$re78)/length(ne_c$re78))
nsw_effect4a
```

```
## [1] 1794.343
```

```
nsw_sd4a
```

```
## [1] 670.9967
```

```
a4 <- lm(re78 ~ nsw + age + educ + hisp +black + married + re74 + u74, data = ne)
coeftest(a4, vcov = vcovHC(a4, type = "HC2"))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept)   144.71167  2869.84217  0.0504  0.95981
## nsw          1720.75446   677.97934  2.5381  0.01149 *
## age            52.95678    40.19872  1.3174  0.18840
## educ          414.94025   164.23763  2.5265  0.01187 *
## hisp          255.40632  1412.00026  0.1809  0.85654
## black       -2165.79026  1021.43201 -2.1203  0.03454 *
## married       -66.08068   840.06277 -0.0787  0.93734
## re74            0.13032     0.12015  1.0846  0.27870
## u74           528.30376  1094.05676  0.4829  0.62942
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# using the robust std. error for possible heterockedasticity
```

The difference between the coefficient values from the difference-in-means and the regression are similar (1794.343 vs. 1720.754) because the treatment was randomized. The small difference could be explained by (1) it is unlikely that everything will be 100% balanced and (2) certain confounders.

Problem 4B

```
npw_t <- filter(npw, nsw == 1)
npw_c <- filter(npw, nsw == 0)
nsw_effect4b <- mean(npw_t$re78) - mean(npw_c$re78)
nsw_sd4b <- sqrt(var(npw_t$re78)/length(npw_t$re78) + var(npw_c$re78)/length(npw_c$re78))
nsw_effect4b
```

```
## [1] -15204.78
```

```
nsw_sd4b
```

```
## [1] 657.0765
```

```
b4 <- lm(re78 ~ nsw + age + educ + hisp +black + married + re74 + u74, data = npw)
coeftest(b4, vcov = vcovHC(b4, type = "HC2")) # same reason as above
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  2.5443e+02  1.5037e+03  0.1692 0.8656541
## nsw         -1.4596e+03  9.3271e+02 -1.5649 0.1177216
```

```
## age         -8.6113e+01  2.2625e+01 -3.8061 0.0001444 ***
## educ         6.6188e+02  8.6492e+01  7.6524 2.737e-14 ***
## hisp         1.1488e+03  1.3161e+03  0.8728 0.3828306
## black       -8.3465e+02  4.7171e+02 -1.7694 0.0769397 .
## married      1.4526e+03  5.3124e+02  2.7344 0.0062899 **
## re74         7.7154e-01  3.2381e-02 23.8271 < 2.2e-16 ***
## u74          2.3634e+03  1.0823e+03  2.1837 0.0290716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient values from the difference-in-means and the regression are different (-15204.78 vs. -1459.6). This shows that the treatment has not been randomized since it is from an observational data. Since the treatment is not randomized, we should match the covariates to make the treatment "as-if" random.

Problem 4C

```
mb_4c <- MatchBalance(nsw ~ age + educ + black + hisp + married +
              re74 + u74 + re75 + u75, data = npw)
```

```
##
## ***** (V1) age *****
## before matching:
## mean treatment........ 25.816
## mean control.......... 34.851
## std mean diff......... -126.27
##
## mean raw eQQ diff..... 9.0432
## med  raw eQQ diff..... 8
## max  raw eQQ diff..... 17
##
## mean eCDF diff........ 0.23165
## med  eCDF diff........ 0.25299
## max  eCDF diff........ 0.37714
##
## var ratio (Tr/Co)..... 0.46963
## T-test p-value........ < 2.22e-16
## KS Bootstrap p-value.. < 2.22e-16
## KS Naive p-value...... 0
## KS Statistic.......... 0.37714
##
##
## ***** (V2) educ *****
## before matching:
## mean treatment........ 10.346
## mean control.......... 12.117
## std mean diff......... -88.077
##
## mean raw eQQ diff..... 1.8595
## med  raw eQQ diff..... 2
## max  raw eQQ diff..... 5
##
```

```
## mean eCDF diff........ 0.1091
## med  eCDF diff........ 0.01944
## max  eCDF diff........ 0.40289
##
## var ratio (Tr/Co)..... 0.42549
## T-test p-value........ < 2.22e-16
## KS Bootstrap p-value.. < 2.22e-16
## KS Naive p-value...... 0
## KS Statistic.......... 0.40289
##
##
## ***** (V3) black *****
## before matching:
## mean treatment........ 0.84324
## mean control.......... 0.2506
## std mean diff......... 162.56
##
## mean raw eQQ diff..... 0.58919
## med  raw eQQ diff..... 1
## max  raw eQQ diff..... 1
##
## mean eCDF diff........ 0.29632
## med  eCDF diff........ 0.29632
## max  eCDF diff........ 0.59264
##
## var ratio (Tr/Co)..... 0.70739
## T-test p-value........ < 2.22e-16
##
##
## ***** (V4) hisp *****
## before matching:
## mean treatment........ 0.059459
## mean control.......... 0.03253
## std mean diff......... 11.357
##
## mean raw eQQ diff..... 0.027027
## med  raw eQQ diff..... 0
## max  raw eQQ diff..... 1
##
## mean eCDF diff........ 0.013465
## med  eCDF diff........ 0.013465
## max  eCDF diff........ 0.026929
##
## var ratio (Tr/Co)..... 1.7859
## T-test p-value........ 0.13173
##
##
## ***** (V5) married *****
## before matching:
## mean treatment........ 0.18919
## mean control.......... 0.86627
## std mean diff......... -172.41
##
## mean raw eQQ diff..... 0.67568
```

```
## med  raw eQQ diff..... 1
## max  raw eQQ diff..... 1
##
## mean eCDF diff........ 0.33854
## med  eCDF diff........ 0.33854
## max  eCDF diff........ 0.67708
##
## var ratio (Tr/Co)..... 1.3308
## T-test p-value........ < 2.22e-16
##
##
## ***** (V6) re74 *****
## before matching:
## mean treatment........ 2095.6
## mean control.......... 19429
## std mean diff......... -354.71
##
## mean raw eQQ diff..... 17663
## med  raw eQQ diff..... 18417
## max  raw eQQ diff..... 102109
##
## mean eCDF diff........ 0.46806
## med  eCDF diff........ 0.54766
## max  eCDF diff........ 0.72924
##
## var ratio (Tr/Co)..... 0.13285
## T-test p-value........ < 2.22e-16
## KS Bootstrap p-value.. < 2.22e-16
## KS Naive p-value...... 0
## KS Statistic.......... 0.72924
##
##
## ***** (V7) u74 *****
## before matching:
## mean treatment........ 0.70811
## mean control.......... 0.086345
## std mean diff......... 136.39
##
## mean raw eQQ diff..... 0.62162
## med  raw eQQ diff..... 1
## max  raw eQQ diff..... 1
##
## mean eCDF diff........ 0.31088
## med  eCDF diff........ 0.31088
## max  eCDF diff........ 0.62176
##
## var ratio (Tr/Co)..... 2.6332
## T-test p-value........ < 2.22e-16
##
##
## ***** (V8) re75 *****
## before matching:
## mean treatment........ 1532.1
## mean control.......... 19063
```

```
## std mean diff......... -544.58
##
## mean raw eQQ diff..... 17978
## med  raw eQQ diff..... 17903
## max  raw eQQ diff..... 131511
##
## mean eCDF diff........ 0.46947
## med  eCDF diff........ 0.53317
## max  eCDF diff........ 0.77362
##
## var ratio (Tr/Co)..... 0.056057
## T-test p-value........ < 2.22e-16
## KS Bootstrap p-value.. < 2.22e-16
## KS Naive p-value...... 0
## KS Statistic.......... 0.77362
##
##
## ***** (V9) u75 *****
## before matching:
## mean treatment........ 0.6
## mean control.......... 0.1
## std mean diff......... 101.79
##
## mean raw eQQ diff..... 0.4973
## med  raw eQQ diff..... 0
## max  raw eQQ diff..... 1
##
## mean eCDF diff........ 0.25
## med  eCDF diff........ 0.25
## max  eCDF diff........ 0.5
##
## var ratio (Tr/Co)..... 2.6801
## T-test p-value........ < 2.22e-16
##
##
## Before Matching Minimum p.value: < 2.22e-16
## Variable Name(s): age educ black married re74 u74 re75 u75  Number(s): 1 2 3 5 6 7 8 9
```

```
bt_4c <- baltest.collect(matchbal.out = mb_4c,
                                        var.names = colnames(npw)[-c(1,9,12)],
                                        after = FALSE)

bt_4c[,-8:-10] %>%
  round(3) %>%
  kable(col.names = c("Mean Treatment","Mean Control","Std Dev","Std Dev Pooled","Var Ratio","T P-v","KS
        caption = "Chekcing Covariate Balance",
        align = "c")
```

Every covariate, except **hisp**, seems to be an important factor in treatment. The differences between the Mean Treatment and Mean Control are significant for all other covariates (except **hisp**). The difference in means is not by chance because the low p-value shows that the null hypothesis can be rejected. The covariates are not balanced.

Table 1: Chekcing Covariate Balance

	Mean Treatment	Mean Control	Std Dev	Std Dev Pooled	Var Ratio	T P-v	KS P-v
age	25.816	34.851	-126.266	-100.943	0.470	0.000	0
educ	10.346	12.117	-88.077	-68.052	0.425	0.000	0
black	0.843	0.251	162.564	147.980	0.707	0.000	NA
hisp	0.059	0.033	11.357	12.859	1.786	0.132	NA
married	0.189	0.866	-172.406	-184.233	1.331	0.000	NA
re74	2095.574	19428.746	-354.707	-171.783	0.133	0.000	0
re75	0.708	0.086	136.391	164.210	2.633	0.000	NA
u74	1532.056	19063.338	-544.576	-177.437	0.056	0.000	0
u75	0.600	0.100	101.786	122.842	2.680	0.000	NA

Problem 4D

```
# for experimental data
# The following two lines won't knit...
covar1 <- ne[,c(2,3,4,5,6,7,10)]
pscore.fmla1 <- as.formula(paste("nsw ~", paste(names(covar1), collapse = "+")))
pscore_model1 <- glm(pscore.fmla1, data = ne, family = binomial(link = logit))
pscore1 <- predict(pscore_model1, type = "response")
match.pscore1 <- Match(Tr = ne$nsw, X = pscore1, M = 1, estimand = "ATT")
MatchBalance(pscore.fmla1, data = npw, match.out = match.pscore1)
```

```
##
## ***** (V1) age *****
##                          Before Matching        After Matching
## mean treatment........     25.816              35.93
## mean control.........     34.851             27.441
## std mean diff.........    -126.27             77.077
##
## mean raw eQQ diff.....     9.0432             8.5468
## med  raw eQQ diff.....          8                  7
## max  raw eQQ diff.....         17                 20
##
## mean eCDF diff........    0.23165            0.22492
## med  eCDF diff........    0.25299            0.23263
## max  eCDF diff........    0.37714            0.41088
##
## var ratio (Tr/Co).....    0.46963             1.9901
## T-test p-value........ < 2.22e-16         3.1086e-15
## KS Bootstrap p-value.. < 2.22e-16         < 2.22e-16
## KS Naive p-value...... < 2.22e-16         < 2.22e-16
## KS Statistic.........    0.37714            0.41088
##
##
## ***** (V2) educ *****
##                          Before Matching        After Matching
## mean treatment........     10.346             13.746
## mean control.........     12.117             13.062
## std mean diff.........    -88.077             35.765
```

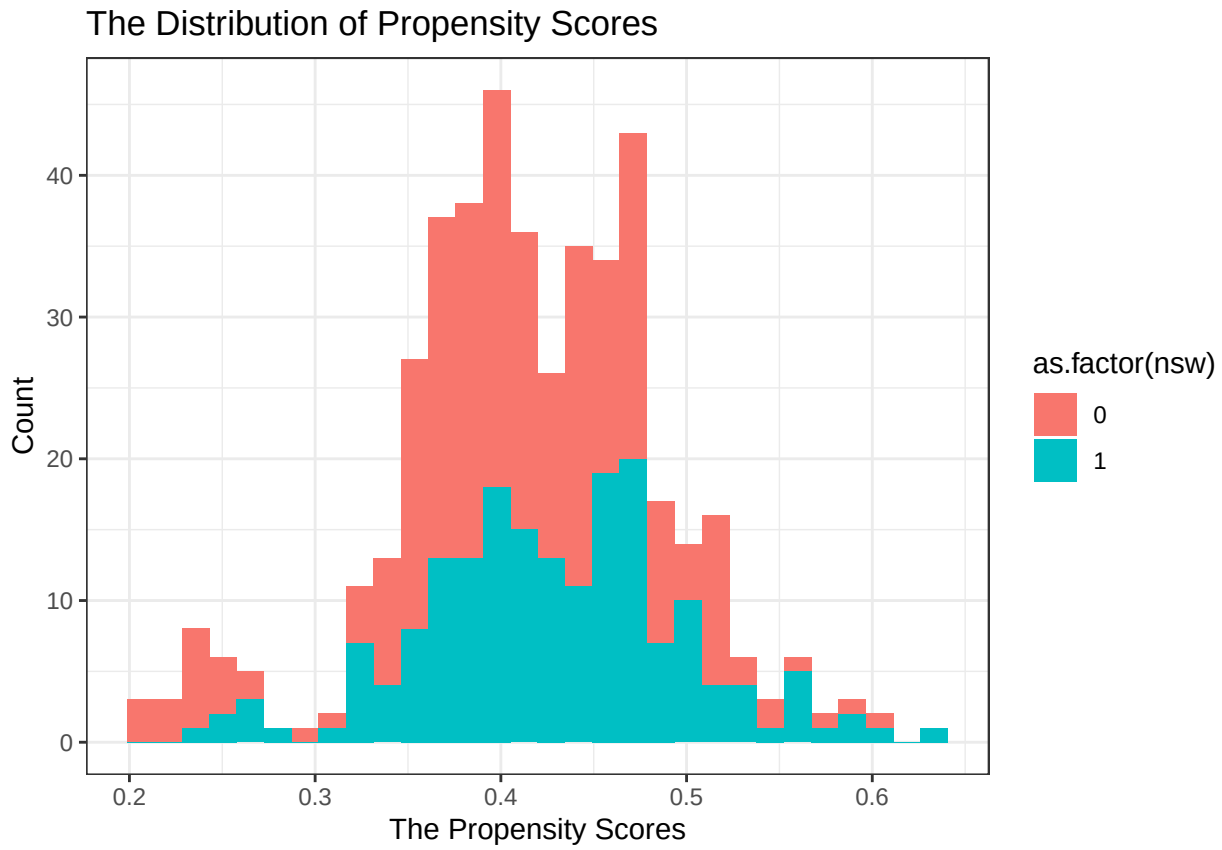
```
##
## mean raw eQQ diff.....     1.8595            0.92145
## med  raw eQQ diff.....          2                  1
## max  raw eQQ diff.....          5                  3
##
## mean eCDF diff........     0.1091            0.15358
## med  eCDF diff........    0.01944            0.16314
## max  eCDF diff........    0.40289            0.27492
##
## var ratio (Tr/Co).....    0.42549             1.3499
## T-test p-value........ < 2.22e-16         0.00043932
## KS Bootstrap p-value.. < 2.22e-16         < 2.22e-16
## KS Naive p-value...... < 2.22e-16         2.7277e-11
## KS Statistic.........    0.40289            0.27492
##
##
## ***** (V3) black *****
##                         Before Matching        After Matching
## mean treatment........    0.84324            0.11351
## mean control..........     0.2506            0.25578
## std mean diff.........     162.56            -44.726
##
## mean raw eQQ diff.....    0.58919            0.20544
## med  raw eQQ diff.....          1                  0
## max  raw eQQ diff.....          1                  1
##
## mean eCDF diff........    0.29632            0.10272
## med  eCDF diff........    0.29632            0.10272
## max  eCDF diff........    0.59264            0.20544
##
## var ratio (Tr/Co).....    0.70739            0.52863
## T-test p-value........ < 2.22e-16         0.00020455
##
##
## ***** (V4) hisp *****
##                         Before Matching        After Matching
## mean treatment........   0.059459           0.021622
## mean control..........    0.03253           0.043243
## std mean diff.........     11.357            -14.826
##
## mean raw eQQ diff.....   0.027027           0.021148
## med  raw eQQ diff.....          0                  0
## max  raw eQQ diff.....          1                  1
##
## mean eCDF diff........   0.013465           0.010574
## med  eCDF diff........   0.013465           0.010574
## max  eCDF diff........   0.026929           0.021148
##
## var ratio (Tr/Co).....     1.7859             0.5113
## T-test p-value........    0.13173            0.24801
##
##
## ***** (V5) married *****
##                         Before Matching        After Matching
```

```
## mean treatment........    0.18919            0.82162
## mean control..........    0.86627            0.77815
## std mean diff.........    -172.41             11.324
##
## mean raw eQQ diff.....    0.67568          0.0090634
## med  raw eQQ diff.....          1                  0
## max  raw eQQ diff.....          1                  1
##
## mean eCDF diff........    0.33854          0.0045317
## med  eCDF diff........    0.33854          0.0045317
## max  eCDF diff........    0.67708          0.0090634
##
## var ratio (Tr/Co).....     1.3308            0.84898
## T-test p-value........ < 2.22e-16            0.30365
##
##
## ***** (V6) re74 *****
##                         Before Matching        After Matching
## mean treatment........     2095.6              510.26
## mean control..........      19429                9782
## std mean diff.........    -354.71             -929.75
##
## mean raw eQQ diff.....      17663              8967.6
## med  raw eQQ diff.....      18417              9629.8
## max  raw eQQ diff.....     102109               11560
##
## mean eCDF diff........    0.46806            0.66197
## med  eCDF diff........    0.54766            0.79154
## max  eCDF diff........    0.72924                  1
##
## var ratio (Tr/Co).....    0.13285            0.20697
## T-test p-value........ < 2.22e-16          < 2.22e-16
## KS Bootstrap p-value.. < 2.22e-16          < 2.22e-16
## KS Naive p-value...... < 2.22e-16          < 2.22e-16
## KS Statistic..........    0.72924                  1
##
##
## ***** (V7) u74 *****
##                         Before Matching        After Matching
## mean treatment........    0.70811            0.75135
## mean control..........   0.086345                  0
## std mean diff.........     136.39             173.36
##
## mean raw eQQ diff.....    0.62162            0.78248
## med  raw eQQ diff.....          1                  1
## max  raw eQQ diff.....          1                  1
##
## mean eCDF diff........    0.31088            0.39124
## med  eCDF diff........    0.31088            0.39124
## max  eCDF diff........    0.62176            0.78248
##
## var ratio (Tr/Co).....     2.6332                Inf
## T-test p-value........ < 2.22e-16          < 2.22e-16
##
```

```
##
## Before Matching Minimum p.value: < 2.22e-16
## Variable Name(s): age educ black married re74 u74  Number(s): 1 2 3 5 6 7
##
## After Matching Minimum p.value: < 2.22e-16
## Variable Name(s): age educ re74 u74  Number(s): 1 2 6 7
```

```
ggplot(data = ne,
       aes(x = pscore1,
           group = as.factor(nsw),
           fill = as.factor(nsw))) +
  geom_histogram(alpha = .6) +
  labs(title = "The Distribution of Propensity Scores",
       x = "The Propensity Scores",
       y = "Count") +
  stat_bin(bins = 30) +
  theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# for non-experimental data
covar <- npw[,c(2,3,4,5,6,7,10)]
pscore.fmla <- as.formula(paste("nsw ~", paste(names(covar), collapse = "+")))
pscore_model <- glm(pscore.fmla, data = npw, family = binomial(link = logit))
pscore <- predict(pscore_model, type = "response")
```

```r
match.pscore <- Match(Tr = npw$nsw, X = pscore, M = 1, estimand = "ATT")
MatchBalance(pscore.fmla, data = npw, match.out = match.pscore)
```

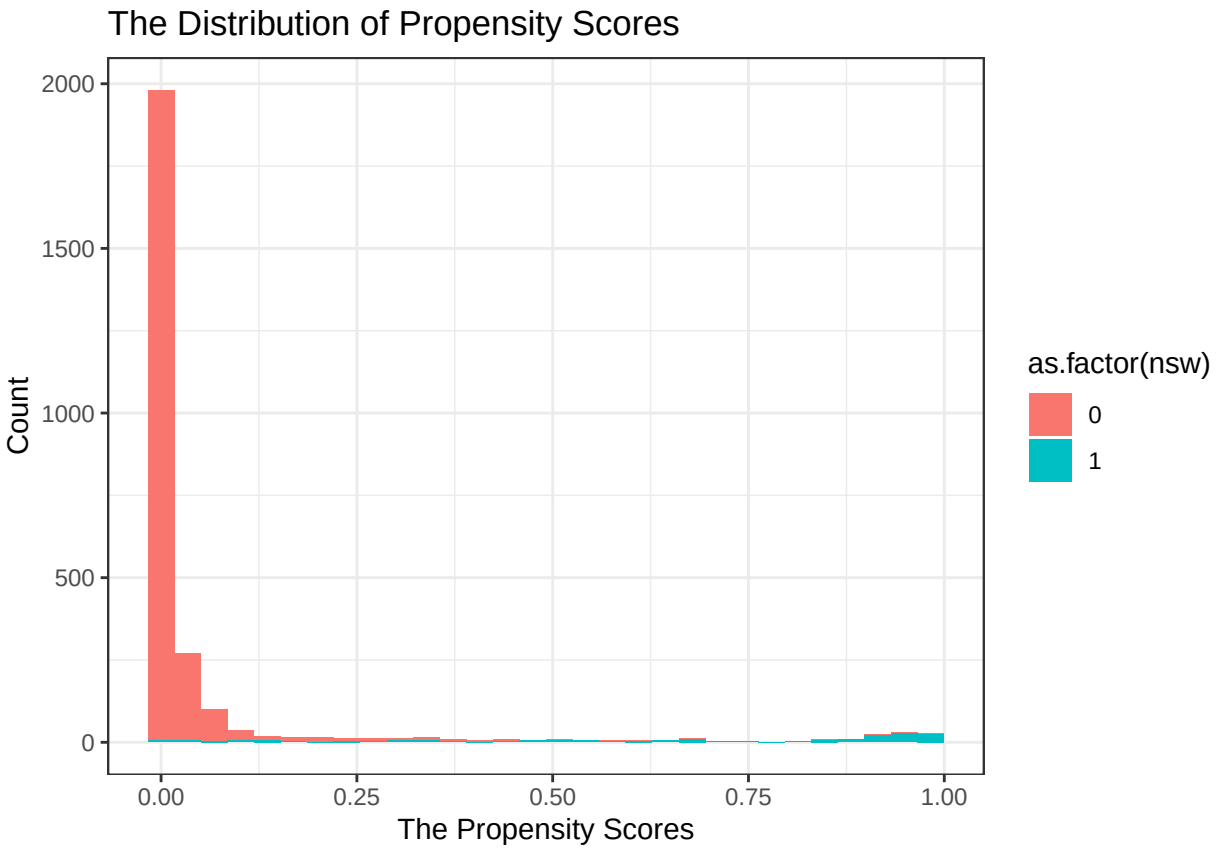
```
##
## ***** (V1) age *****
##                          Before Matching        After Matching
## mean treatment........     25.816             25.816
## mean control.........     34.851             23.713
## std mean diff.........    -126.27             29.402
##
## mean raw eQQ diff.....     9.0432             3.7696
## med  raw eQQ diff.....          8                  2
## max  raw eQQ diff.....         17                 18
##
## mean eCDF diff........    0.23165           0.096656
## med  eCDF diff........    0.25299           0.093058
## max  eCDF diff........    0.37714            0.25849
##
## var ratio (Tr/Co).....    0.46963             1.3318
## T-test p-value........ < 2.22e-16         0.00068309
## KS Bootstrap p-value.. < 2.22e-16         < 2.22e-16
## KS Naive p-value...... < 2.22e-16         < 2.22e-16
## KS Statistic..........    0.37714            0.25849
##
##
## ***** (V2) educ *****
##                          Before Matching        After Matching
## mean treatment........     10.346             10.346
## mean control.........     12.117             10.372
## std mean diff.........    -88.077            -1.2739
##
## mean raw eQQ diff.....     1.8595             1.1477
## med  raw eQQ diff.....          2                  1
## max  raw eQQ diff.....          5                  4
##
## mean eCDF diff........     0.1091           0.067425
## med  eCDF diff........    0.01944           0.059084
## max  eCDF diff........    0.40289            0.29247
##
## var ratio (Tr/Co).....    0.42549            0.65618
## T-test p-value........ < 2.22e-16            0.90171
## KS Bootstrap p-value.. < 2.22e-16         < 2.22e-16
## KS Naive p-value...... < 2.22e-16         < 2.22e-16
## KS Statistic..........    0.40289            0.29247
##
##
## ***** (V3) black *****
##                          Before Matching        After Matching
## mean treatment........    0.84324            0.84324
## mean control.........     0.2506             0.89072
## std mean diff.........     162.56            -13.022
##
## mean raw eQQ diff.....    0.58919            0.13737
```

```
## med  raw eQQ diff.....          1                  0
## max  raw eQQ diff.....          1                  1
##
## mean eCDF diff........    0.29632           0.068685
## med  eCDF diff........    0.29632           0.068685
## max  eCDF diff........    0.59264            0.13737
##
## var ratio (Tr/Co).....    0.70739              1.358
## T-test p-value........ < 2.22e-16            0.14827
##
##
## ***** (V4) hisp *****
##                          Before Matching        After Matching
## mean treatment........   0.059459           0.059459
## mean control..........    0.03253           0.014904
## std mean diff.........     11.357              18.79
##
## mean raw eQQ diff.....   0.027027            0.01034
## med  raw eQQ diff.....          0                  0
## max  raw eQQ diff.....          1                  1
##
## mean eCDF diff........   0.013465          0.0051699
## med  eCDF diff........   0.013465          0.0051699
## max  eCDF diff........   0.026929            0.01034
##
## var ratio (Tr/Co).....     1.7859             3.8091
## T-test p-value........    0.13173           0.025466
##
##
## ***** (V5) married *****
##                          Before Matching        After Matching
## mean treatment........    0.18919            0.18919
## mean control..........    0.86627              0.114
## std mean diff.........    -172.41             19.146
##
## mean raw eQQ diff.....    0.67568           0.081241
## med  raw eQQ diff.....          1                  0
## max  raw eQQ diff.....          1                  1
##
## mean eCDF diff........    0.33854            0.04062
## med  eCDF diff........    0.33854            0.04062
## max  eCDF diff........    0.67708           0.081241
##
## var ratio (Tr/Co).....     1.3308             1.5187
## T-test p-value........ < 2.22e-16          0.0058003
##
##
## ***** (V6) re74 *****
##                          Before Matching        After Matching
## mean treatment........     2095.6             2095.6
## mean control..........      19429             2813.5
## std mean diff.........    -354.71            -14.692
##
## mean raw eQQ diff.....      17663             3475.8
```

```
## med  raw eQQ diff.....      18417             2132.3
## max  raw eQQ diff.....     102109              16427
##
## mean eCDF diff........    0.46806           0.080936
## med  eCDF diff........    0.54766           0.079764
## max  eCDF diff........    0.72924            0.18316
##
## var ratio (Tr/Co).....    0.13285            0.91742
## T-test p-value........ < 2.22e-16           0.053347
## KS Bootstrap p-value.. < 2.22e-16         < 2.22e-16
## KS Naive p-value...... < 2.22e-16         2.7375e-10
## KS Statistic..........    0.72924            0.18316
##
##
## ***** (V7) u74 *****
##                         Before Matching        After Matching
## mean treatment........    0.70811            0.70811
## mean control..........   0.086345             0.6042
## std mean diff.........     136.39             22.793
##
## mean raw eQQ diff.....    0.62162           0.023634
## med  raw eQQ diff.....          1                  0
## max  raw eQQ diff.....          1                  1
##
## mean eCDF diff........    0.31088           0.011817
## med  eCDF diff........    0.31088           0.011817
## max  eCDF diff........    0.62176           0.023634
##
## var ratio (Tr/Co).....     2.6332             0.8643
## T-test p-value........ < 2.22e-16           0.002436
##
##
## Before Matching Minimum p.value: < 2.22e-16
## Variable Name(s): age educ black married re74 u74  Number(s): 1 2 3 5 6 7
##
## After Matching Minimum p.value: < 2.22e-16
## Variable Name(s): age educ re74  Number(s): 1 2 6
```

```
ggplot(data = npw,
        aes(x = pscore,
            group = as.factor(nsw),
            fill = as.factor(nsw))) +
    geom_histogram(alpha = .6) +
  labs(title = "The Distribution of Propensity Scores",
       x = "The Propensity Scores",
       y = "Count") +
  stat_bin(bins = 30) +
  theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The distribution of the propensity scores for the experimental data demonstrates that the treatment has been randomized. On the other hand, the distribution of the propensity scores for the non-experimental data demonstrates the contrary. They the propensity scores for the treated and the untreated are skewed to the opposite direction.

Problem 4E

```
m1 <- Match(Y = npw$re78, Tr = npw$nsw, X = covar, M = 1, Weight = 2, BiasAdjust = TRUE)
summary(m1)
```

```
##
## Estimate...  2773.2
## AI SE......  1600
## T-stat.....  1.7333
## p.val......  0.083048
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  197
```

```
m2 <- Match(Y = npw$re78, Tr = npw$nsw, X = covar, M = 4, Weight = 2, BiasAdjust = TRUE)
summary(m2)
```

```
##
## Estimate...  1733.8
## AI SE......  1688.8
## T-stat.....  1.0266
## p.val......  0.30461
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  804
```

```
m3 <- Match(Y = npw$re78, Tr = npw$nsw, X = covar, M = 10, Weight = 2, BiasAdjust = TRUE)
summary(m3)
```

```
##
## Estimate...  1538.8
## AI SE......  1322.6
## T-stat.....  1.1635
## p.val......  0.24463
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  1858
```

```
m4 <- Match(Y = npw$re78, Tr = npw$nsw, X = covar, M = 1, Weight = 2)
summary(m4)
```

```
##
## Estimate...  2459.2
## AI SE......  1600.3
## T-stat.....  1.5367
## p.val......  0.12437
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  197
```

```
m5 <- Match(Y = npw$re78, Tr = npw$nsw, X = covar, M = 4, Weight = 2)
summary(m5)
```

```
##
## Estimate...  1217.1
## AI SE......  1687.1
## T-stat.....  0.7214
## p.val......  0.47066
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  804
```

```
m6 <- Match(Y = npw$re78, Tr = npw$nsw, X = covar, M = 10, Weight = 2)
summary(m6)
```

```
##
## Estimate...  432.71
## AI SE......  1353.4
## T-stat.....  0.31971
## p.val......  0.74919
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  1858
```

Problem 4F

```
m4f <- Match(Y = npw$re78, Tr = npw$nsw, X = pscore, M = 1, Weight = 2, BiasAdjust = TRUE)
summary(m4f)
```

```
##
## Estimate...  1771.8
## AI SE......  1915.1
## T-stat.....  0.9252
## p.val......  0.35486
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  677
```

Problem 4G

```
g4 <- sum(npw$re78 * (npw$nsw - pscore) / (1 - pscore)) / length(npw_t$nsw) # from 4b filtered for 4b
g4
```

```
## [1] 746.8158
```

The ATT by weighting on the propensity score does not accord with the previous results. The problem is that PS weighting can be biased if the samples are small or there are multiple continuous variables. Bias must be adjusted. Since there are four continuous covariates in our model, PS weighting without bias adjustment will give a biased estimate.

Problem 4H

The NSW program is effective. First, the results from the experimental data demonstrate its effectiveness. The histogram from 4D demonstrates that the treatment in the experimental data was successfully randomized. Second, our results from 4F, matching with propensity score, also demonstrates that the program

is effective (if conditional ignorability and common support are assumed). The ATT with bias corrected matching estimators and `M = 4` gave a similar results but without the treatment being randomized, we can't be certain of its unbiasedness. Therefore, the use of propensity scores on 4F is better. By comparing results from 4F and 4G, we can see that the results for 4F have adjusted for the bias caused by matching.