PSET 5

Sarah Van Alsten

2/18/2020

Question 1. The Curse of Dimensionality

a. General Expression to Calculate Euclidean Distance between Two Points:

$$Distance = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 ... + (x_{iP} - x_{jP})^2}$$

```
#write a function that will give us the distance for a given # of covariates
getEuclidean <- function(numCov){

  #take the numCov columns to compare to
  compareDat <- as.data.frame(dat[, 1:numCov])


  #compute euclidean distance for each obs
  compareDat <- compareDat %>%
    #squared distance from 0 for each column
    mutate_all(.funs = ~((0 - .)^2)) %>%
    #create new column as sqrt of sum of these squared distances
    mutate(euclidean = sqrt(rowSums(.)))

  #return the minimum euclidean distance
  return(min(compareDat$euclidean))

}
```

b.

```
#Generate dataset X of 500 obs, 20 covariates (normally distributed 0-1)
set.seed(02139)
dat <- as.data.frame(replicate(n = 20, expr = rnorm(n = 500, mean = 0, sd = 1)))

#make a blank dataframe to hold results
euc.res <- as.data.frame(cbind(1:20, rep(NA, 20)))

#get euclidean distance for 1 covariate
getEuclidean(numCov = 1)
```

```
## [1] 0.001368214
```

```
#add this to the result data
euc.res$V2[1] <- getEuclidean(numCov = 1)

#now do this for the 2:20 covariates
for (i in 2:20){
```

```
  euc.res$V2[i] <- getEuclidean(numCov = i)
}

#plot results
euc.res %>%
  ggplot(aes(x = V1, y = V2)) +
  geom_point() +
  geom_path() +
  theme_bw() +
  labs(x = "Number of Covariates",
       y = "Minimum Euclidean Distance") +
  ggtitle("Euclidean Distance by Number Of Covariates")
```

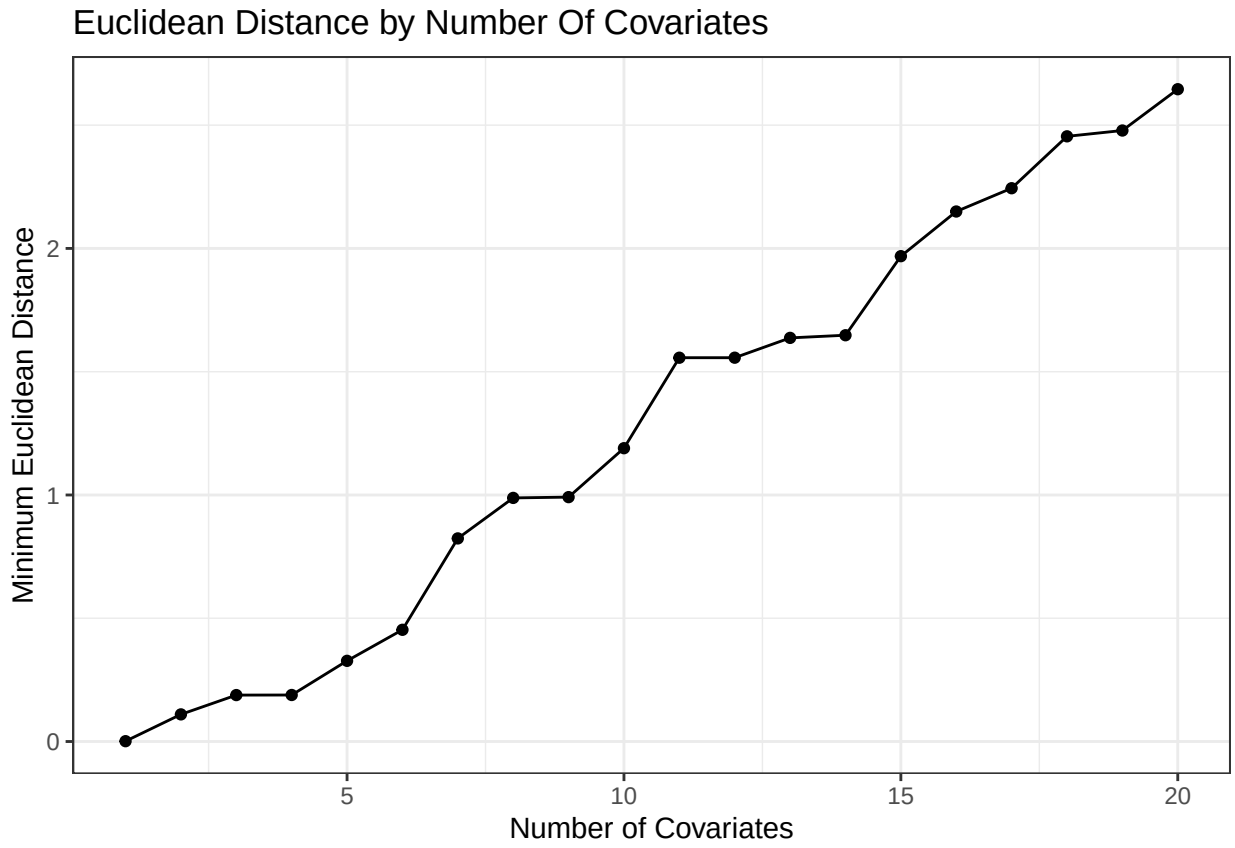


Figure 1: Euclidean Distance by Covariate Number

c. These results demonstrate that as we add more dimensions or covariates to match on, the dissimilarity (distance) between even the best match and the index obervation increases (i.e. it becomes more difficult to find a very close/similar match).

Question 2

a.

```r
#unbiased estiamte of ATE
nsw.t <- t.test(re78 ~ nsw, data = nsw)

#ate and se
(ate.t <- nsw.t$estimate[2] - nsw.t$estimate[1])
```

```
## mean in group 1
##        1794.343
```

```r
(stderr.t <- nsw.t$stderr)
```

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

```r
#re-estimate using linear regression
mod.nsw <- lm(re78 ~ nsw + age + educ + black + hisp + married + re74 + u74,
              data = nsw)

#get results
summary(mod.nsw)
```

```
##
## Call:
## lm(formula = re78 ~ nsw + age + educ + black + hisp + married +
##     re74 + u74, data = nsw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9846  -4399  -1601   3167  54033
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.447e+02  2.603e+03   0.056  0.95569
## nsw          1.721e+03  6.325e+02   2.721  0.00678 **
## age          5.296e+01  4.513e+01   1.174  0.24122
## educ         4.149e+02  1.765e+02   2.351  0.01916 *
## black       -2.166e+03  1.158e+03  -1.870  0.06210 .
## hisp         2.554e+02  1.551e+03   0.165  0.86928
## married     -6.608e+01  8.563e+02  -0.077  0.93852
## re74         1.303e-01  7.685e-02   1.696  0.09065 .
## u74          5.283e+02  9.350e+02   0.565  0.57233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6505 on 436 degrees of freedom
## Multiple R-squared:  0.05519,    Adjusted R-squared:  0.03786
## F-statistic: 3.184 on 8 and 436 DF,  p-value: 0.001602
```

```r
#account for randomziation, use robust SE
sqrt(diag(sandwich::vcovHC(mod.nsw, type="HC2")))
```

```
##  (Intercept)          nsw          age         educ        black         hisp
```

```
## 2869.8421657  677.9793381   40.1987165  164.2376269 1021.4320063 1412.0002640
##      married         re74          u74
##  840.0627653    0.1201542 1094.0567552
```

The unadjusted estimate of the average treatment effect of NSW participation on 1978 earnings was an increase of $1794 dollars (se = $671). After adjusting for age, race, ethnicity, education, marital status, earnings in 1974, and employment in 1974, the average treatment effect of NSW participation was somewhat lower (increase of $1721, se = $678). This suggests that there may have been some confounding by the covariates, but not a substantial amount.

b.

```
#calculate naive ATE using non-experimental data
psid.t <- t.test(re78 ~ nsw, data = psid)

#ate and se
(ate.psid <- psid.t$estimate[2] - psid.t$estimate[1])
```

```
## mean in group 1
##       -15204.78
```

```
(stderr.psid <- psid.t$stderr)
```

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

```
#re-estimate using regression
mod.psid <- lm(re78 ~ nsw + age + educ + black + hisp + married + re74 + u74,
               data = psid)

#get results
summary(mod.psid)
```

```
##
## Call:
## lm(formula = re78 ~ nsw + age + educ + black + hisp + married +
##     re74 + u74, data = psid)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -66010  -5198   -248   4462 109278
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.544e+02  1.474e+03   0.173 0.862950
## nsw         -1.460e+03  1.080e+03  -1.351 0.176687
## age         -8.611e+01  2.345e+01  -3.672 0.000245 ***
## educ         6.619e+02  8.164e+01   8.107 7.81e-16 ***
## black       -8.346e+02  5.342e+02  -1.563 0.118284
## hisp         1.149e+03  1.173e+03   0.979 0.327563
## married      1.453e+03  6.297e+02   2.307 0.021145 *
## re74         7.715e-01  2.086e-02  36.986  < 2e-16 ***
```

```
## u74          2.363e+03  8.407e+02   2.811 0.004968 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10830 on 2666 degrees of freedom
## Multiple R-squared:  0.5217, Adjusted R-squared:  0.5203
## F-statistic: 363.5 on 8 and 2666 DF,  p-value: < 2.2e-16
```

```
#account for randomziation, use robust SE
sqrt(diag(sandwich::vcovHC(mod.psid, type="HC2")))
```

```
##  (Intercept)          nsw          age         educ        black         hisp
## 1.503747e+03 9.327112e+02 2.262536e+01 8.649220e+01 4.717086e+02 1.316119e+03
##      married         re74          u74
## 5.312366e+02 3.238084e-02 1.082312e+03
```

We are estimating the average treatment effect conditional on all the covariates (age, education, race, ethnicity, marital status, earnings in 1974, and unemployment in 1974). The only difference is that now, we are using data from the general population, so we don't expect these variables to be balanced. These methods do not recover the experimental results because, even conditional on all the covariates we adjusted for, the treated (those who were given the work program) and untreated (the general population) are not exchangeable (we don't have conditional ignorability). This indicates lack of balance on other unobserved/not included covariates, or that treated and control units do not tend to have overlap in covariate values, so we lack positivity when conditioning on these.

c.

```
#estimate the propensity scores

mb  <- MatchBalance(nsw ~ age + educ + black +
                    hisp + married + re74 + re75  +
                    u74 + u75 + u78,
                    data = psid,
                    nboots = 10)
```

```
##
## ***** (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) 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
##
##
## ***** (V8) 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
##
##
## ***** (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
##
##
## ***** (V10) u78 *****
## before matching:
## mean treatment........ 0.24324
## mean control.......... 0.11486
```

```
## std mean diff......... 29.842
##
## mean raw eQQ diff..... 0.12432
## med  raw eQQ diff..... 0
## max  raw eQQ diff..... 1
##
## mean eCDF diff........ 0.064192
## med  eCDF diff........ 0.064192
## max  eCDF diff........ 0.12838
##
## var ratio (Tr/Co)..... 1.8197
## T-test p-value........ 9.7001e-05
##
##
## Before Matching Minimum p.value: < 2.22e-16
## Variable Name(s): age educ black married re74 re75 u74 u75  Number(s): 1 2 3 5 6 7 8 9
```

```
#make a nice balance table
btab <- tableone::CreateTableOne(vars = c("age", "educ", "black",
                    "hisp" , "married", "re74" , "re75", "u74", "u75", "u78"),
                    data = psid,
                    strata = "nsw",
                    factorVars = c("black",
                    "hisp", "married",
                    "u74", "u75" , "u78"))

#also get the K-S tests for distributions from the match balance
#pvalues for the KS tests for all variables
ks.pval <- c(NA,
             mb$BeforeMatching[[1]][9][[1]][[1]][[1]],
             mb$BeforeMatching[[2]][9][[1]][[1]][[1]],
             "--", #black is categorical
             "--", #hisp is categorical
             "--", #married is categorical
             mb$BeforeMatching[[6]][9][[1]][[1]][[1]],
             mb$BeforeMatching[[7]][9][[1]][[1]][[1]],
             "--", #u74 is categorical
             "--", #u75 is categorical
             "--") #u78 is categorical

#replace 0 with "<0.001"
ks.pval <- str_replace(ks.pval, pattern = "0", replacement ="< 0.001")


#print balance table with KS pvalues
as.data.frame(cbind(print(btab), ks.pval)) %>%
  kableExtra::kable() %>%
  kableExtra::kable_styling("striped")
```

```
##                    Stratified by nsw
##                     0                  1                  p      test
##   n                     2490               185
##   age (mean (SD))     34.85 (10.44)      25.82 (7.16)    <0.001
##   educ (mean (SD))    12.12 (3.08)       10.35 (2.01)    <0.001
```

```
##   black = 1 (%)          624 (25.1)          156 (84.3)    <0.001
##   hisp = 1 (%)            81 ( 3.3)           11 ( 5.9)     0.084
##   married = 1 (%)       2157 (86.6)           35 (18.9)    <0.001
##   re74 (mean (SD)) 19428.75 (13406.88) 2095.57 (4886.62) <0.001
##   re75 (mean (SD)) 19063.34 (13596.95) 1532.06 (3219.25) <0.001
##   u74 = 1 (%)            215 ( 8.6)          131 (70.8)    <0.001
##   u75 = 1 (%)            249 (10.0)          111 (60.0)    <0.001
##   u78 = 1 (%)            286 (11.5)           45 (24.3)    <0.001
```

	0	1	p	test	ks.pval
n	2490	185			NA
age (mean (SD))	34.85 (10.44)	25.82 (7.16)	<0.001		< 0.001
educ (mean (SD))	12.12 (3.08)	10.35 (2.01)	<0.001		< 0.001
black = 1 (%)	624 (25.1)	156 (84.3)	<0.001		–
hisp = 1 (%)	81 (3.3)	11 (5.9)	0.084		–
married = 1 (%)	2157 (86.6)	35 (18.9)	<0.001		–
re74 (mean (SD))	19428.75 (13406.88)	2095.57 (4886.62)	<0.001		< 0.001
re75 (mean (SD))	19063.34 (13596.95)	1532.06 (3219.25)	<0.001		< 0.001
u74 = 1 (%)	215 (8.6)	131 (70.8)	<0.001		–
u75 = 1 (%)	249 (10.0)	111 (60.0)	<0.001		–
u78 = 1 (%)	286 (11.5)	45 (24.3)	<0.001		–

Based on the balance table, all covariates except for Hispanic ethnicity differed between the treated and control groups. In particular, unemployment in 1974 and 1975, being Black, being unmarried, and having low real earnings in 1974 and 1975 were most strongly associated with treatment status.

d.

```r
#estimate propensity scores using logistic regression in experimetn
pscore_model_exp <- glm(nsw ~ age + educ + black +
                    hisp + married + re74 + re75  +
                    u74 + u75 + u78,
                    data = nsw,
                    family = binomial(link = logit))


pscore_exp <- predict(pscore_model_exp, type = "response")

#plot distributions of pscores
nsw$pscore <- pscore_exp

nsw %>%
  ggplot(aes(x = pscore_exp, fill = factor(nsw),alpha = .9)) +
  geom_density() +
  scale_fill_discrete(name = "NSW")+
  scale_alpha_continuous(guide = F) +
  ggtitle("Propensity Scores Experimental") +
  theme_bw()
```

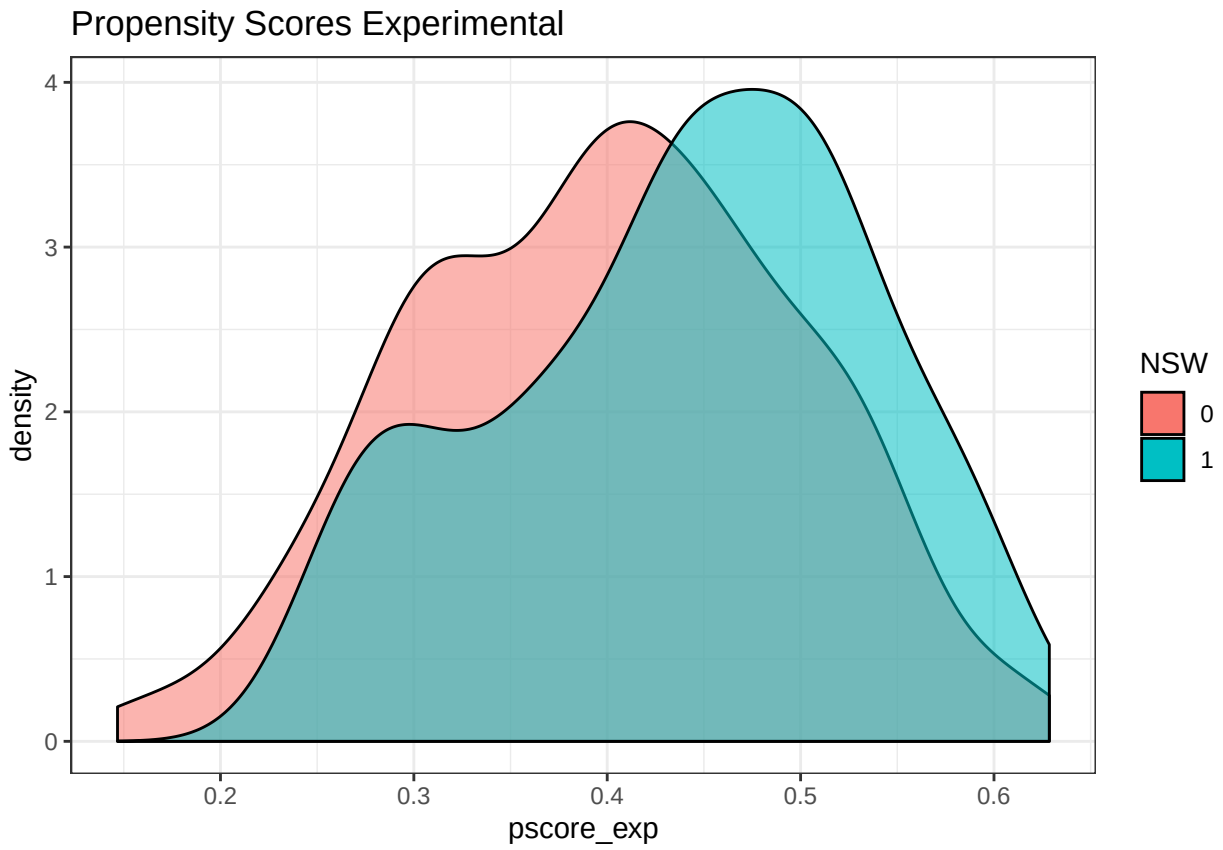


Figure 2: Propensity Score Distributions

```r
#estimate propensity scores using logistic regression in non-experiment
pscore_model <- glm(nsw ~ age + educ + black +
                    hisp + married + re74 + re75  +
                    u74 + u75 + u78,
                    data = psid,
                    family = binomial(link = logit))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
pscore <- predict(pscore_model, type = "response")

#plot distributions of pscores
psid$pscore <- pscore

psid %>%
  ggplot(aes(x = pscore, fill = factor(nsw), alpha = .3)) +
  geom_density() +
  scale_fill_discrete(name = "NSW")+
  scale_alpha_continuous(guide = F) +
  ggtitle("Propensity Scores Non-Experimental") +
  theme_bw()
```

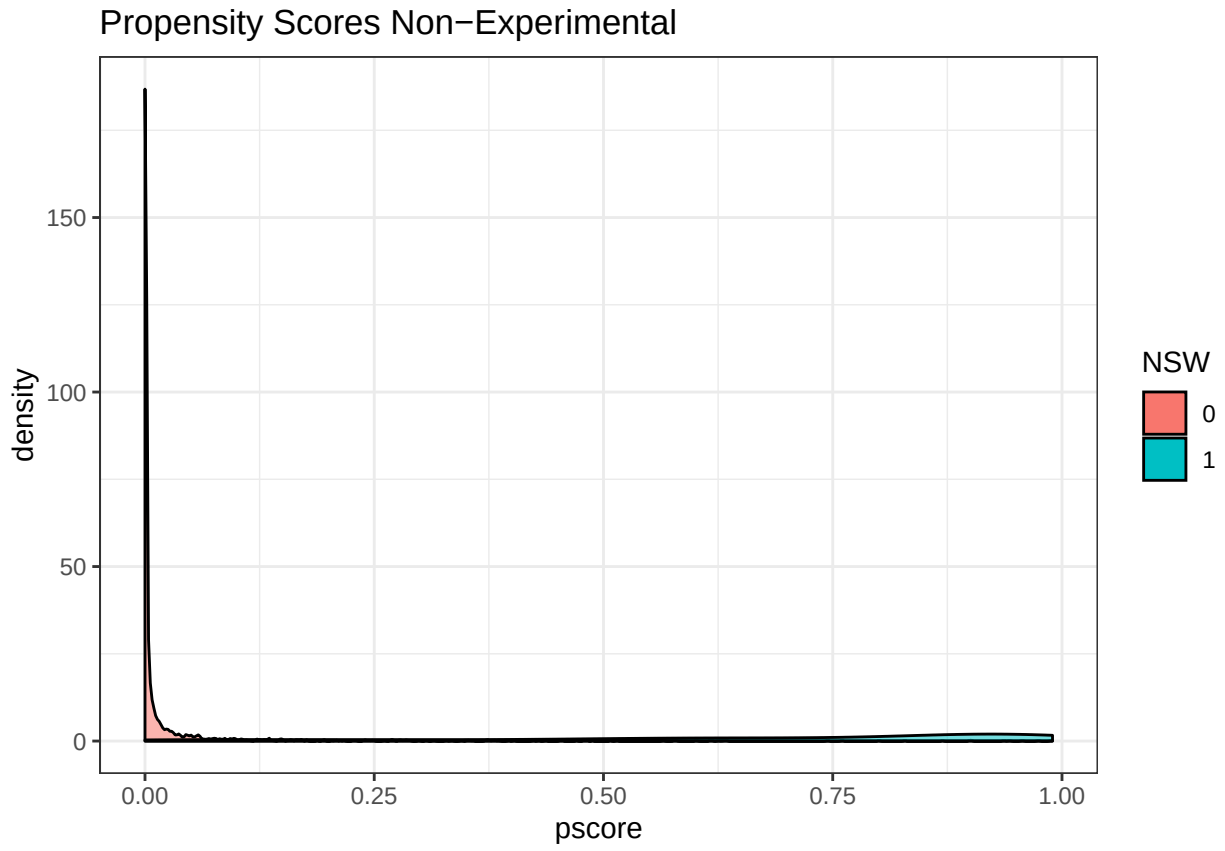


Figure 3: Propensity Score Distributions

```r
#to see better, also plot separately
psid %>%
  ggplot(aes(x = pscore, fill = factor(nsw), alpha = .3)) +
  geom_density() +
  scale_fill_discrete(name = "NSW") +
  facet_wrap(~nsw, scales = "free")  +
  theme_bw() +
  scale_alpha_continuous(guide = F)
```

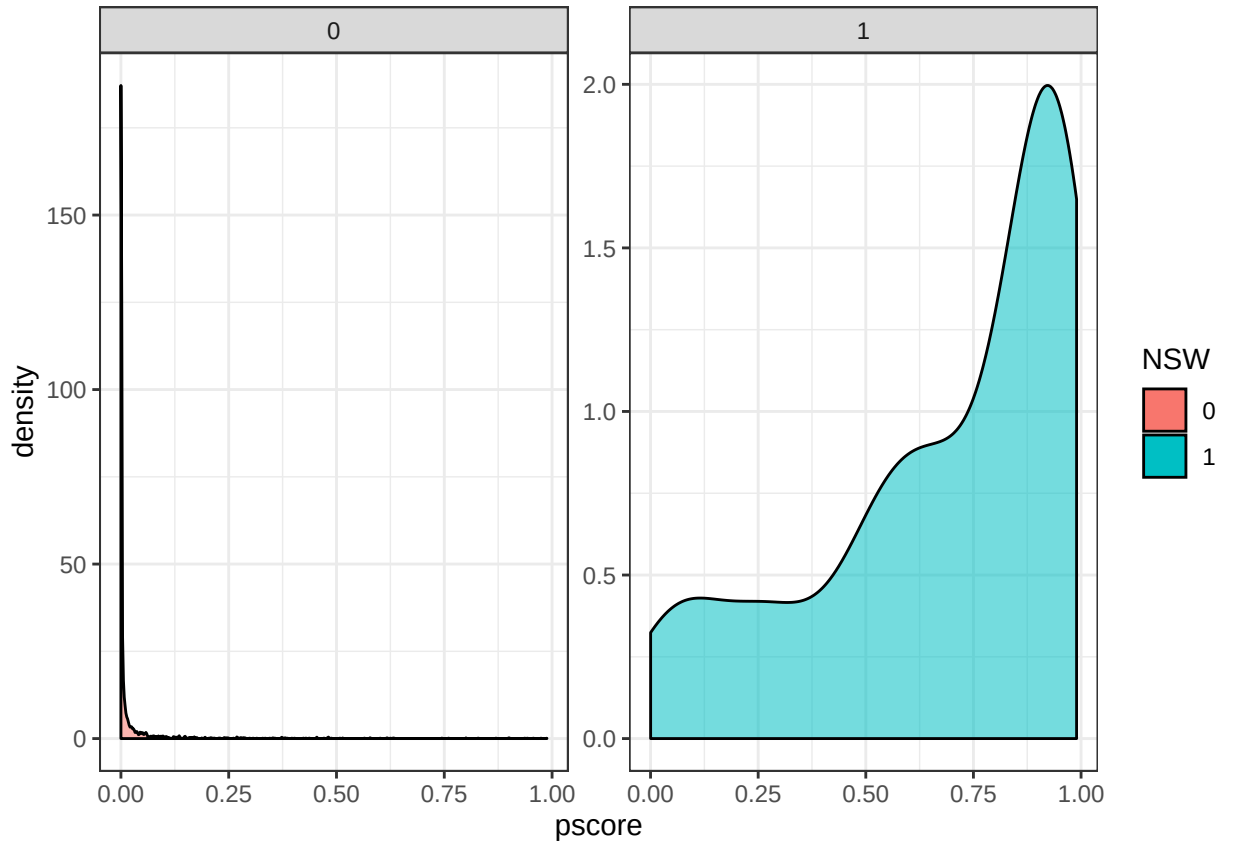


Figure 4: Propensity Score Distributions

In the experimental data, there is substantial overlap in the distributions of propensity scores - those who got treatment seem to have slightly more observations with probability of treatment greater than 0.5 than the control, but otherwise they are very similar.

Conversely, there is very little overlap between the two propensity score distributions in the non-experimental data. Propensity scores in the control units are heavily skewed towards 0, and those in the treatment units are skewed towards 1. This indicates that control units had almost no probability of receiving treatment, whereas a large portion of those in the treatment group had almost complete probability of receiving treatment.

The two differ because the process of randomization in the experimental data helped ensure balance between the two groups AND similar propensity scores (if propensity = probability of treatment and treatment was randomly assigned with equal probability, then propensity distributions should also be comporable), whereas in non-experimental data, pre-treatment covariates were not balanced, especially given the nature of the the treatment (a training program is most likley offered to those with low employment or job skills)

e.

```
#make matches
match.mah <- Match(Y = psid$re78,
                   Tr = psid$nsw,
                   X = psid[, c("age", "black", "u74", "educ", "married")],
                   M = 1, estimand="ATT",
                   Weight = 2)

summary(match.mah)
```

```
##
## Estimate...  1151.7
## AI SE......  1754.8
## T-stat.....  0.65631
## p.val......  0.51162
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  311
```

```
#check balance on all covariates (matched + unmatched)
mb2 <- MatchBalance(nsw ~ age + educ + black + u74 +
             married + hisp + re74 + re75  +
             u75 + u78,
             data = psid,
             match.out = match.mah,
             nboots = 10)
```

```
##
## ***** (V1) age *****
##                        Before Matching        After Matching
## mean treatment........     25.816             25.816
## mean control..........     34.851             24.793
## std mean diff.........    -126.27             14.304
##
## mean raw eQQ diff.....     9.0432             1.3505
## med  raw eQQ diff.....          8                  1
## max  raw eQQ diff.....         17                  6
##
## mean eCDF diff........    0.23165           0.039949
## med  eCDF diff........    0.25299           0.028939
## max  eCDF diff........    0.37714            0.16077
##
## var ratio (Tr/Co).....    0.46963             1.4027
## T-test p-value........ < 2.22e-16          0.0021774
## KS Bootstrap p-value.. < 2.22e-16         < 2.22e-16
## KS Naive p-value...... < 2.22e-16         0.00064553
## KS Statistic..........    0.37714            0.16077
##
##
## ***** (V2) educ *****
```

```
##                         Before Matching        After Matching
## mean treatment........     10.346             10.346
## mean control..........     12.117             10.683
## std mean diff.........    -88.077            -16.758
##
## mean raw eQQ diff.....     1.8595            0.33119
## med  raw eQQ diff.....          2                  0
## max  raw eQQ diff.....          5                  3
##
## mean eCDF diff........     0.1091           0.025476
## med  eCDF diff........    0.01944           0.012862
## max  eCDF diff........    0.40289           0.099678
##
## var ratio (Tr/Co).....    0.42549             1.3757
## T-test p-value........ < 2.22e-16         8.9227e-05
## KS Bootstrap p-value.. < 2.22e-16         < 2.22e-16
## KS Naive p-value...... < 2.22e-16           0.090992
## KS Statistic..........    0.40289           0.099678
##
##
## ***** (V3) black *****
##                         Before Matching        After Matching
## mean treatment........    0.84324            0.84324
## mean control..........     0.2506            0.84324
## std mean diff.........     162.56                  0
##
## mean raw eQQ diff.....    0.58919                  0
## med  raw eQQ diff.....          1                  0
## max  raw eQQ diff.....          1                  0
##
## mean eCDF diff........    0.29632                  0
## med  eCDF diff........    0.29632                  0
## max  eCDF diff........    0.59264                  0
##
## var ratio (Tr/Co).....    0.70739                  1
## T-test p-value........ < 2.22e-16                  1
##
##
## ***** (V4) u74 *****
##                         Before Matching        After Matching
## mean treatment........    0.70811            0.70811
## mean control..........   0.086345            0.70811
## std mean diff.........     136.39                  0
##
## mean raw eQQ diff.....    0.62162                  0
## med  raw eQQ diff.....          1                  0
## max  raw eQQ diff.....          1                  0
##
## mean eCDF diff........    0.31088                  0
## med  eCDF diff........    0.31088                  0
## max  eCDF diff........    0.62176                  0
##
## var ratio (Tr/Co).....     2.6332                  1
## T-test p-value........ < 2.22e-16                  1
```

```
##
##
## ***** (V5) married *****
##                          Before Matching        After Matching
## mean treatment........    0.18919           0.18919
## mean control..........    0.86627           0.18919
## std mean diff.........    -172.41                 0
##
## mean raw eQQ diff.....    0.67568                 0
## med  raw eQQ diff.....          1                 0
## max  raw eQQ diff.....          1                 0
##
## mean eCDF diff........    0.33854                 0
## med  eCDF diff........    0.33854                 0
## max  eCDF diff........    0.67708                 0
##
## var ratio (Tr/Co).....     1.3308                 1
## T-test p-value........ < 2.22e-16                 1
##
##
## ***** (V6) hisp *****
##                          Before Matching        After Matching
## mean treatment........   0.059459          0.059459
## mean control..........    0.03253         0.00045045
## std mean diff.........     11.357            24.885
##
## mean raw eQQ diff.....   0.027027           0.11576
## med  raw eQQ diff.....          0                 0
## max  raw eQQ diff.....          1                 1
##
## mean eCDF diff........   0.013465          0.057878
## med  eCDF diff........   0.013465          0.057878
## max  eCDF diff........   0.026929           0.11576
##
## var ratio (Tr/Co).....     1.7859            124.21
## T-test p-value........    0.13173         0.00080928
##
##
## ***** (V7) re74 *****
##                          Before Matching        After Matching
## mean treatment........     2095.6            2095.6
## mean control..........      19429            3339.2
## std mean diff.........    -354.71           -25.449
##
## mean raw eQQ diff.....      17663              3681
## med  raw eQQ diff.....      18417            1388.7
## max  raw eQQ diff.....     102109             12874
##
## mean eCDF diff........    0.46806           0.14758
## med  eCDF diff........    0.54766            0.1672
## max  eCDF diff........    0.72924           0.26688
##
## var ratio (Tr/Co).....    0.13285           0.50154
## T-test p-value........ < 2.22e-16         0.00072855
```

```
## KS Bootstrap p-value.. < 2.22e-16         < 2.22e-16
## KS Naive p-value...... < 2.22e-16         4.7964e-10
## KS Statistic..........    0.72924            0.26688
##
##
## ***** (V8) re75 *****
##                         Before Matching        After Matching
## mean treatment........     1532.1             1532.1
## mean control..........      19063             4798.3
## std mean diff.........    -544.58            -101.46
##
## mean raw eQQ diff.....      17978             5084.7
## med  raw eQQ diff.....      17903             4277.3
## max  raw eQQ diff.....     131511              14819
##
## mean eCDF diff........    0.46947            0.21111
## med  eCDF diff........    0.53317            0.21865
## max  eCDF diff........    0.77362            0.41801
##
## var ratio (Tr/Co).....   0.056057            0.20048
## T-test p-value........ < 2.22e-16         4.6326e-10
## KS Bootstrap p-value.. < 2.22e-16         < 2.22e-16
## KS Naive p-value...... < 2.22e-16         < 2.22e-16
## KS Statistic..........    0.77362            0.41801
##
##
## ***** (V9) u75 *****
##                         Before Matching        After Matching
## mean treatment........        0.6                0.6
## mean control..........        0.1            0.51599
## std mean diff.........     101.79             17.102
##
## mean raw eQQ diff.....     0.4973           0.038585
## med  raw eQQ diff.....          0                  0
## max  raw eQQ diff.....          1                  1
##
## mean eCDF diff........       0.25           0.019293
## med  eCDF diff........       0.25           0.019293
## max  eCDF diff........        0.5           0.038585
##
## var ratio (Tr/Co).....     2.6801            0.96098
## T-test p-value........ < 2.22e-16           0.019946
##
##
## ***** (V10) u78 *****
##                         Before Matching        After Matching
## mean treatment........    0.24324            0.24324
## mean control..........    0.11486            0.32342
## std mean diff.........     29.842            -18.638
##
## mean raw eQQ diff.....    0.12432          0.0032154
## med  raw eQQ diff.....          0                  0
## max  raw eQQ diff.....          1                  1
##
```

```
## mean eCDF diff........   0.064192          0.0016077
## med  eCDF diff........   0.064192          0.0016077
## max  eCDF diff........    0.12838          0.0032154
##
## var ratio (Tr/Co).....     1.8197            0.84122
## T-test p-value........ 9.7001e-05            0.10376
##
##
## Before Matching Minimum p.value: < 2.22e-16
## Variable Name(s): age educ black u74 married re74 re75 u75  Number(s): 1 2 3 4 5 7 8 9
##
## After Matching Minimum p.value: < 2.22e-16
## Variable Name(s): age educ re74 re75  Number(s): 1 2 7 8
```

```
#vector of covariates after matching
after.tr <- c(mb2$AfterMatching[[1]][3],
           mb2$AfterMatching[[2]][3],
           mb2$AfterMatching[[3]][3],
           mb2$AfterMatching[[4]][3],
           mb2$AfterMatching[[5]][3],
           mb2$AfterMatching[[6]][3],
           mb2$AfterMatching[[7]][3],
           mb2$AfterMatching[[8]][3],
           mb2$AfterMatching[[9]][3],
           mb2$AfterMatching[[10]][3])

#vector of values in control after matching
after.co <- c(mb2$AfterMatching[[1]][4],
           mb2$AfterMatching[[2]][4],
           mb2$AfterMatching[[3]][4],
           mb2$AfterMatching[[4]][4],
           mb2$AfterMatching[[5]][4],
           mb2$AfterMatching[[6]][4],
           mb2$AfterMatching[[7]][4],
           mb2$AfterMatching[[8]][4],
           mb2$AfterMatching[[9]][4],
           mb2$AfterMatching[[10]][4])

#sd in treated and control
var.tr <- c(mb2$AfterMatching[[1]][5],
           mb2$AfterMatching[[2]][5],
           mb2$AfterMatching[[3]][5],
           mb2$AfterMatching[[4]][5],
           mb2$AfterMatching[[5]][5],
           mb2$AfterMatching[[6]][5],
           mb2$AfterMatching[[7]][5],
           mb2$AfterMatching[[8]][5],
           mb2$AfterMatching[[9]][5],
           mb2$AfterMatching[[10]][5])

var.co <- c(mb2$AfterMatching[[1]][6],
           mb2$AfterMatching[[2]][6],
           mb2$AfterMatching[[3]][6],
           mb2$AfterMatching[[4]][6],
```

```
           mb2$AfterMatching[[5]][6],
           mb2$AfterMatching[[6]][6],
           mb2$AfterMatching[[7]][6],
           mb2$AfterMatching[[8]][6],
           mb2$AfterMatching[[9]][6],
           mb2$AfterMatching[[10]][6])

#make numeric and get sqrts
var.tr <- as.numeric(var.tr)
var.co <- as.numeric(var.co)

sd.tr <- sqrt(var.tr)
sd.co <- sqrt(var.co)


#vector of pvalues comparing groups
match.p <- c(mb2$AfterMatching[[1]][7],
           mb2$AfterMatching[[2]][7],
           mb2$AfterMatching[[3]][7],
           mb2$AfterMatching[[4]][7],
           mb2$AfterMatching[[5]][7],
           mb2$AfterMatching[[6]][7],
           mb2$AfterMatching[[7]][7],
           mb2$AfterMatching[[8]][7],
           mb2$AfterMatching[[9]][7],
           mb2$AfterMatching[[10]][7])

#also p values for KS tests
ks.pval2 <- c(mb2$AfterMatching[[1]][9][[1]][[1]][[1]],
             mb2$AfterMatching[[2]][9][[1]][[1]][[1]],
             "--", #black is categorical
             "--", #u74 is categorical
             "--", #married is categorical
             "--", #hisp is categorical
             mb2$AfterMatching[[7]][9][[1]][[1]][[1]], #re74
             mb2$AfterMatching[[8]][9][[1]][[1]][[1]], #re75
             "--", #u75 is categorical
             "--") #u78 is categorical

#make format nicer
ks.pval2 <- sprintf(as.numeric(ks.pval2), fmt = "%.3f")
```

```
## Warning in sprintf(as.numeric(ks.pval2), fmt = "%.3f"): NAs introduced by
## coercion
```

```
#bind cols to make table
df <- as.data.frame(cbind(sprintf(after.tr, fmt = "%.3f"),
                     sprintf(after.co, fmt = "%.3f"),
                     sprintf(sd.tr, fmt = "%.3f"),
                     sprintf(sd.co, fmt = "%.3f"),
                     sprintf(match.p, fmt = "%.3f"),
                     ks.pval2))
```

```r
#add labels
names(df) <- c("mean.tr", "mean.co",
               "sd.tr", "sd.co",
               "pvalue", "KSpvalue")

rownames(df) <- c("age", "educ", "black", "u74",
                  "married", "hisp", "re74", "re75", "u75", "u78")

#print table
df %>% kableExtra::kable() %>% kableExtra::kable_styling("striped")
```

	mean.tr	mean.co	sd.tr	sd.co	pvalue	KSpvalue
age	25.816	24.793	7.136	6.025	0.002	0.001
educ	10.346	10.683	2.005	1.710	0.000	0.091
black	0.843	0.843	0.364	0.364	1.000	NA
u74	0.708	0.708	0.455	0.455	1.000	NA
married	0.189	0.189	0.392	0.392	1.000	NA
hisp	0.059	0.000	0.236	0.021	0.001	NA
re74	2095.574	3339.171	4873.398	6881.430	0.001	0.000
re75	1532.056	4798.256	3210.539	7170.346	0.000	0.000
u75	0.600	0.516	0.490	0.500	0.020	NA
u78	0.243	0.323	0.429	0.468	0.104	NA

I chose to match participants on unemployment in 1974, race, marital status, age, and level of education. While all covariates except ethnicity were significantly (marginally) different between individuals who did and did not receive treatment, some of the variables may be causally post to treatment assignment. For instance, after receiving the NSW program, an individual would likely have different employment status in 1975 and 1978, so I did not match/condition on these. Similarly, wages in 1974 and 1975 are likely to be causally post to treatment assignment (assuming 1974 wages are measured at the year's end and treatment is given near the beginning). Conversely, being unemployed in 1974 is a confounder, because it directly affects your likelihood of getting treatment AND later earnings (if you have gaps in employment, likely your earnings will be lower even 4 years down the road), and similar for race, marital status, age and education. After matching on unemployment in 1974, race, marital status, age, and education, the average treatment effect of participation in the NSW program was a nonsignificant increase of \$1151.70 (se = \$1754.80). After matching, participants were not balanced on 1975 earnings, 1978 earnings, unemployment in 1975, and unemployment in 1978, but were balanced on other variables (however, the distributions of age and education differed by treatment status, even if means were balanced).

f.

```r
#f.
#use exact matching on education, race, ethnicity and married.

#make matches
match.mahF <- Match(Y = psid$re78,
                    Tr = psid$nsw,
                    X = psid[, c("hisp", "black", "educ", "married")],
                    M = 1, estimand="ATT",
                    Weight = 2,
                    exact = TRUE)
#print ATT
summary(match.mahF)
```

```
##
## Estimate...  -5821.6
## AI SE......  840.65
## T-stat.....  -6.9251
## p.val......  4.3572e-12
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  178
## Matched number of observations  (unweighted).  5320
##
## Number of obs dropped by 'exact' or 'caliper'  7
```

```
#check balance on all covariates (matched + unmatched)
mbF <- MatchBalance(nsw ~ age + educ + black + u74 +
             married + hisp + re74 + re75  +
             u75 + u78,
             data = psid,
             match.out = match.mahF,
             nboots = 10)
```

```
##
## ***** (V1) age *****
##                          Before Matching        After Matching
## mean treatment........     25.816             25.882
## mean control.........     34.851             30.467
## std mean diff.........    -126.27            -63.429
##
## mean raw eQQ diff.....     9.0432             3.5624
## med  raw eQQ diff.....          8                  2
## max  raw eQQ diff.....         17                 13
##
## mean eCDF diff........    0.23165           0.091344
## med  eCDF diff........    0.25299           0.092105
## max  eCDF diff........    0.37714            0.23195
##
## var ratio (Tr/Co).....    0.46963            0.49389
## T-test p-value........ < 2.22e-16         9.4773e-07
## KS Bootstrap p-value.. < 2.22e-16         < 2.22e-16
## KS Naive p-value...... < 2.22e-16         < 2.22e-16
## KS Statistic..........    0.37714            0.23195
##
##
## ***** (V2) educ *****
##                          Before Matching        After Matching
## mean treatment........     10.346             10.371
## mean control.........     12.117             10.371
## std mean diff.........    -88.077                  0
##
## mean raw eQQ diff.....     1.8595                   0
## med  raw eQQ diff.....          2                   0
## max  raw eQQ diff.....          5                   0
##
## mean eCDF diff........     0.1091                   0
```

```
## med  eCDF diff........    0.01944                 0
## max  eCDF diff........    0.40289                 0
##
## var ratio (Tr/Co).....    0.42549                 1
## T-test p-value........ < 2.22e-16                 1
## KS Bootstrap p-value.. < 2.22e-16                 1
## KS Naive p-value...... < 2.22e-16                 1
## KS Statistic..........    0.40289                 0
##
##
## ***** (V3) black *****
##                         Before Matching        After Matching
## mean treatment........    0.84324             0.8764
## mean control..........     0.2506             0.8764
## std mean diff.........     162.56                  0
##
## mean raw eQQ diff.....    0.58919                 0
## med  raw eQQ diff.....          1                 0
## max  raw eQQ diff.....          1                 0
##
## mean eCDF diff........    0.29632                 0
## med  eCDF diff........    0.29632                 0
## max  eCDF diff........    0.59264                 0
##
## var ratio (Tr/Co).....    0.70739                 1
## T-test p-value........ < 2.22e-16                 1
##
##
## ***** (V4) u74 *****
##                         Before Matching        After Matching
## mean treatment........    0.70811            0.71348
## mean control..........   0.086345           0.083173
## std mean diff.........     136.39             139.02
##
## mean raw eQQ diff.....    0.62162            0.58534
## med  raw eQQ diff.....          1                  1
## max  raw eQQ diff.....          1                  1
##
## mean eCDF diff........    0.31088            0.29267
## med  eCDF diff........    0.31088            0.29267
## max  eCDF diff........    0.62176            0.58534
##
## var ratio (Tr/Co).....     2.6332             2.6808
## T-test p-value........ < 2.22e-16         < 2.22e-16
##
##
## ***** (V5) married *****
##                         Before Matching        After Matching
## mean treatment........    0.18919            0.19663
## mean control..........    0.86627            0.19663
## std mean diff.........    -172.41                  0
##
## mean raw eQQ diff.....    0.67568                 0
## med  raw eQQ diff.....          1                 0
```

```
## max  raw eQQ diff.....          1                 0
##
## mean eCDF diff........    0.33854                 0
## med  eCDF diff........    0.33854                 0
## max  eCDF diff........    0.67708                 0
##
## var ratio (Tr/Co).....     1.3308                 1
## T-test p-value........ < 2.22e-16                 1
##
##
## ***** (V6) hisp *****
##                         Before Matching        After Matching
## mean treatment........   0.059459           0.022472
## mean control..........    0.03253           0.022472
## std mean diff.........     11.357                  0
##
## mean raw eQQ diff.....   0.027027                 0
## med  raw eQQ diff.....          0                 0
## max  raw eQQ diff.....          1                 0
##
## mean eCDF diff........   0.013465                 0
## med  eCDF diff........   0.013465                 0
## max  eCDF diff........   0.026929                 0
##
## var ratio (Tr/Co).....     1.7859                 1
## T-test p-value........    0.13173                 1
##
##
## ***** (V7) re74 *****
##                         Before Matching        After Matching
## mean treatment........     2095.6             2053.9
## mean control..........      19429              11255
## std mean diff.........    -354.71            -188.51
##
## mean raw eQQ diff.....      17663              10702
## med  raw eQQ diff.....      18417              11958
## max  raw eQQ diff.....     102109             102109
##
## mean eCDF diff........    0.46806            0.33972
## med  eCDF diff........    0.54766            0.37782
## max  eCDF diff........    0.72924            0.61635
##
## var ratio (Tr/Co).....    0.13285            0.33285
## 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            0.61635
##
##
## ***** (V8) re75 *****
##                         Before Matching        After Matching
## mean treatment........     1532.1             1502.9
## mean control..........      19063              10980
## std mean diff.........    -544.58            -295.15
```

```
##
## mean raw eQQ diff.....      17978              11328
## med  raw eQQ diff.....      17903              12532
## max  raw eQQ diff.....     131511             131511
##
## mean eCDF diff........    0.46947            0.36656
## med  eCDF diff........    0.53317            0.36917
## max  eCDF diff........    0.77362            0.68195
##
## var ratio (Tr/Co).....   0.056057            0.14765
## 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.........    0.77362            0.68195
##
##
## ***** (V9) u75 *****
##                         Before Matching        After Matching
## mean treatment........        0.6              0.60674
## mean control..........        0.1             0.095378
## std mean diff.........     101.79               104.39
##
## mean raw eQQ diff.....     0.4973               0.4656
## med  raw eQQ diff.....          0                    0
## max  raw eQQ diff.....          1                    1
##
## mean eCDF diff........       0.25               0.2328
## med  eCDF diff........       0.25               0.2328
## max  eCDF diff........        0.5               0.4656
##
## var ratio (Tr/Co).....     2.6801               2.7655
## T-test p-value........ < 2.22e-16         < 2.22e-16
##
##
## ***** (V10) u78 *****
##                         Before Matching        After Matching
## mean treatment........    0.24324              0.25281
## mean control..........    0.11486              0.15058
## std mean diff.........     29.842               23.456
##
## mean raw eQQ diff.....    0.12432             0.087406
## med  raw eQQ diff.....          0                    0
## max  raw eQQ diff.....          1                    1
##
## mean eCDF diff........   0.064192             0.043703
## med  eCDF diff........   0.064192             0.043703
## max  eCDF diff........    0.12838             0.087406
##
## var ratio (Tr/Co).....     1.8197               1.4769
## T-test p-value........ 9.7001e-05             0.014622
##
##
## Before Matching Minimum p.value: < 2.22e-16
## Variable Name(s): age educ black u74 married re74 re75 u75  Number(s): 1 2 3 4 5 7 8 9
```

```
##
## After Matching Minimum p.value: < 2.22e-16
## Variable Name(s): age u74 re74 re75 u75  Number(s): 1 4 7 8 9
```

```
#make a nicer looking balance table
#vector of covariates after matching
after.tr <- c(mbF$AfterMatching[[1]][3],
           mbF$AfterMatching[[2]][3],
           mbF$AfterMatching[[3]][3],
           mbF$AfterMatching[[4]][3],
           mbF$AfterMatching[[5]][3],
           mbF$AfterMatching[[6]][3],
           mbF$AfterMatching[[7]][3],
           mbF$AfterMatching[[8]][3],
           mbF$AfterMatching[[9]][3],
           mbF$AfterMatching[[10]][3])

#vector of values in control after matching
after.co <- c(mbF$AfterMatching[[1]][4],
           mbF$AfterMatching[[2]][4],
           mbF$AfterMatching[[3]][4],
           mbF$AfterMatching[[4]][4],
           mbF$AfterMatching[[5]][4],
           mbF$AfterMatching[[6]][4],
           mbF$AfterMatching[[7]][4],
           mbF$AfterMatching[[8]][4],
           mbF$AfterMatching[[9]][4],
           mbF$AfterMatching[[10]][4])

#sd in treated and control
var.tr <- c(mbF$AfterMatching[[1]][5],
           mbF$AfterMatching[[2]][5],
           mbF$AfterMatching[[3]][5],
           mbF$AfterMatching[[4]][5],
           mbF$AfterMatching[[5]][5],
           mbF$AfterMatching[[6]][5],
           mbF$AfterMatching[[7]][5],
           mbF$AfterMatching[[8]][5],
           mbF$AfterMatching[[9]][5],
           mbF$AfterMatching[[10]][5])

var.co <- c(mbF$AfterMatching[[1]][6],
           mbF$AfterMatching[[2]][6],
           mbF$AfterMatching[[3]][6],
           mbF$AfterMatching[[4]][6],
           mbF$AfterMatching[[5]][6],
           mbF$AfterMatching[[6]][6],
           mbF$AfterMatching[[7]][6],
           mbF$AfterMatching[[8]][6],
           mbF$AfterMatching[[9]][6],
           mbF$AfterMatching[[10]][6])

#make numeric and get sqrts
var.tr <- as.numeric(var.tr)
```

```r
var.co <- as.numeric(var.co)

sd.tr <- sqrt(var.tr)
sd.co <- sqrt(var.co)


#vector of pvalues comparing groups
match.p <- c(mbF$AfterMatching[[1]][7],
           mbF$AfterMatching[[2]][7],
           mbF$AfterMatching[[3]][7],
           mbF$AfterMatching[[4]][7],
           mbF$AfterMatching[[5]][7],
           mbF$AfterMatching[[6]][7],
           mbF$AfterMatching[[7]][7],
           mbF$AfterMatching[[8]][7],
           mbF$AfterMatching[[9]][7],
           mbF$AfterMatching[[10]][7])

#also p values for KS tests
ks.pval2 <- c(mbF$AfterMatching[[1]][9][[1]][[1]][[1]],
             mbF$AfterMatching[[2]][9][[1]][[1]][[1]],
             "--", #black is categorical
             "--", #u74 is categorical
             "--", #married is categorical
             "--", #hisp is categorical
             mbF$AftereMatching[[7]][9][[1]][[1]][[1]], #re74
             mbF$AfterMatching[[8]][9][[1]][[1]][[1]], #re75
             "--", #u75 is categorical
             "--") #u78 is categorical

#replace 0 with "<0.001"
ks.pval2 <- str_replace(ks.pval2, pattern = "0", replacement ="< 0.001")


#print balance table with KS pvalues
df.F <- as.data.frame(cbind(sprintf(after.tr, fmt = "%.3f"),
                    sprintf(after.co, fmt = "%.3f"),
                    sprintf(sd.tr, fmt = "%.3f"),
                    sprintf(sd.co, fmt = "%.3f"),
                    sprintf(match.p, fmt = "%.3f"),
                    ks.pval2))
```

```
## Warning in cbind(sprintf(after.tr, fmt = "%.3f"), sprintf(after.co, fmt = "%.
## 3f"), : number of rows of result is not a multiple of vector length (arg 6)
```

```r
names(df.F) <- c("mean.tr", "mean.co",
               "sd.tr", "sd.co",
               "pvalue", "KSpvalue")

rownames(df.F) <- c("age", "educ", "black", "u74",
                  "married", "hisp", "re74", "re75", "u75", "u78")

df.F %>% kableExtra::kable() %>% kableExtra::kable_styling("striped")
```

	mean.tr	mean.co	sd.tr	sd.co	pvalue	KSpvalue
age	25.882	30.467	7.207	10.256	0.000	< 0.001
educ	10.371	10.371	2.030	2.030	1.000	1
black	0.876	0.876	0.329	0.329	1.000	–
u74	0.713	0.083	0.452	0.276	0.000	–
married	0.197	0.197	0.397	0.397	1.000	–
hisp	0.022	0.022	0.148	0.148	1.000	–
re74	2053.873	11254.765	4867.153	8436.313	0.000	< 0.001
re75	1502.864	10980.124	3201.949	8332.952	0.000	–
u75	0.607	0.095	0.488	0.294	0.000	–
u78	0.253	0.151	0.435	0.358	0.015	< 0.001

The ATT using the exact matching procedure is a significant *decrease* of \$5821.60 (se = \$840.65). In this round of matching, there is balance on education, race, marital status, and ethnicity (what we matched on), but not on the other variables. The results are also different in that, among the matched variables, p values are =1 rather than simply > 0.05, reflecting the exact matching process. We have fewer observations that were able to be matched, and the estimate of the effect is in the opposite direction. The results differ both because of the different covariates that we matched on and because of the exact matching process: we threw away additional observations that didn't have exact matches, and limited the variation of the covariates rather than just reducing it.

g.

```
#g. Re-estimate att with bias corrections

#match using M =1, bias corrected
m.biascorr1 <- Match(Y = psid$re78,
                     Tr = psid$nsw,
                     X = psid[, c("hisp", "black", "educ", "married")],
                     M = 1,
                     estimand="ATT",
                     Weight = 2,
                     BiasAdjust = TRUE)
summary(m.biascorr1)
```

```
##
## Estimate...  -5452.7
## AI SE......  891.28
## T-stat.....  -6.1179
## p.val......  9.4823e-10
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  5328
```

```
#match using M =4, bias corrected
m.biascorr4 <- Match(Y = psid$re78,
                     Tr = psid$nsw,
                     X = psid[, c("hisp", "black", "educ", "married")],
                     M = 4,
                     estimand="ATT",
```

```
                    Weight = 2,
                    BiasAdjust = TRUE)
summary(m.biascorr4)
```

```
##
## Estimate...  -5477.2
## AI SE......  874.68
## T-stat.....  -6.2619
## p.val......  3.8025e-10
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  5417
```

```
#match using M =10, bias corrected
m.biascorr10 <- Match(Y = psid$re78,
                    Tr = psid$nsw,
                    X = psid[, c("hisp", "black", "educ", "married")],
                    M = 10,
                    estimand="ATT",
                    Weight = 2,
                    BiasAdjust = TRUE)
summary(m.biascorr10)
```

```
##
## Estimate...  -6055.9
## AI SE......  869.11
## T-stat.....  -6.9679
## p.val......  3.2165e-12
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  6438
```

```
##################################################
#without bias correction
#match using M =1, bias corrected
m.nocorr1 <- Match(Y = psid$re78,
                    Tr = psid$nsw,
                    X = psid[, c("hisp", "black", "educ", "married")],
                    M = 1,
                    estimand="ATT",
                    Weight = 2,
                    BiasAdjust = F)
summary(m.nocorr1)
```

```
##
## Estimate...  -5437.3
## AI SE......  894.38
## T-stat.....  -6.0794
```

```
## p.val......  1.2064e-09
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  5328
```

```
#match using M =4, bias corrected
m.nocorr4 <- Match(Y = psid$re78,
                   Tr = psid$nsw,
                   X = psid[, c("hisp", "black", "educ", "married")],
                   M = 4,
                   estimand="ATT",
                   Weight = 2,
                   BiasAdjust = F)
summary(m.nocorr4)
```

```
##
## Estimate...  -5522.9
## AI SE......  873.45
## T-stat.....  -6.3231
## p.val......  2.5633e-10
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  5417
```

```
#match using M =10, bias corrected
m.nocorr10 <- Match(Y = psid$re78,
                    Tr = psid$nsw,
                    X = psid[, c("hisp", "black", "educ", "married")],
                    M = 10,
                    estimand="ATT",
                    Weight = 2,
                    BiasAdjust = F)
summary(m.nocorr10)
```

```
##
## Estimate...  -6216.9
## AI SE......  866.8
## T-stat.....  -7.1722
## p.val......  7.3785e-13
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  6438
```

The estimates in the bias-adjusted estimates are less extreme and standard errors larger than those that aren't bias adjusted. At the same time, as the number of matches increases, the number of observations and the magnitude of the effect both increase (i.e. the ATT gets more extreme, in this case more negative).

The differences may be accounted by sample size and covariance between matching variables. I trust the 1:1 matching and the bias-adjusted matches more because matching 1:1 should, on average, make the matches more similar since we only match to the closest observation rather than including several that are less similar to the treated unit. The bias-adjustment should perform better than the uncorrected match because it accounts for confounders and specifically the conditonal expectation of those confounders/matching variables given other included matching variables.

h.

```
#h.
#match on the propensity score from part d
pscore.match <- Match(Y = psid$re78,
                      Tr = psid$nsw,
                      X = psid$pscore,
                      Weight = 2,
                      estimand = "ATT",
                      M = 1)

summary(pscore.match)
```

```
##
## Estimate...  923.98
## AI SE......  1693.8
## T-stat.....  0.5455
## p.val......  0.58541
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  2012
```

When matching on propensity scores, the ATT is a nonsignificant increase of \$923.98 (se = \$1693.80)

i.

```
#i.
#set seed
set.seed(01239)

#make function that can be bootstapped to get se of propensity score
calc_pscore <- function(){

  #take a random sample of rows from the data frame
  psid.sub <- psid[sample(1:nrow(psid), size = 1000, replace = F),]

  #estimate the propensity scores
  ps.mod <- glm(nsw ~ hisp + black + educ + married,
                data = psid.sub)

  #get the scores
  psid.sub$ps.scores <- predict(ps.mod, type = "response")
```

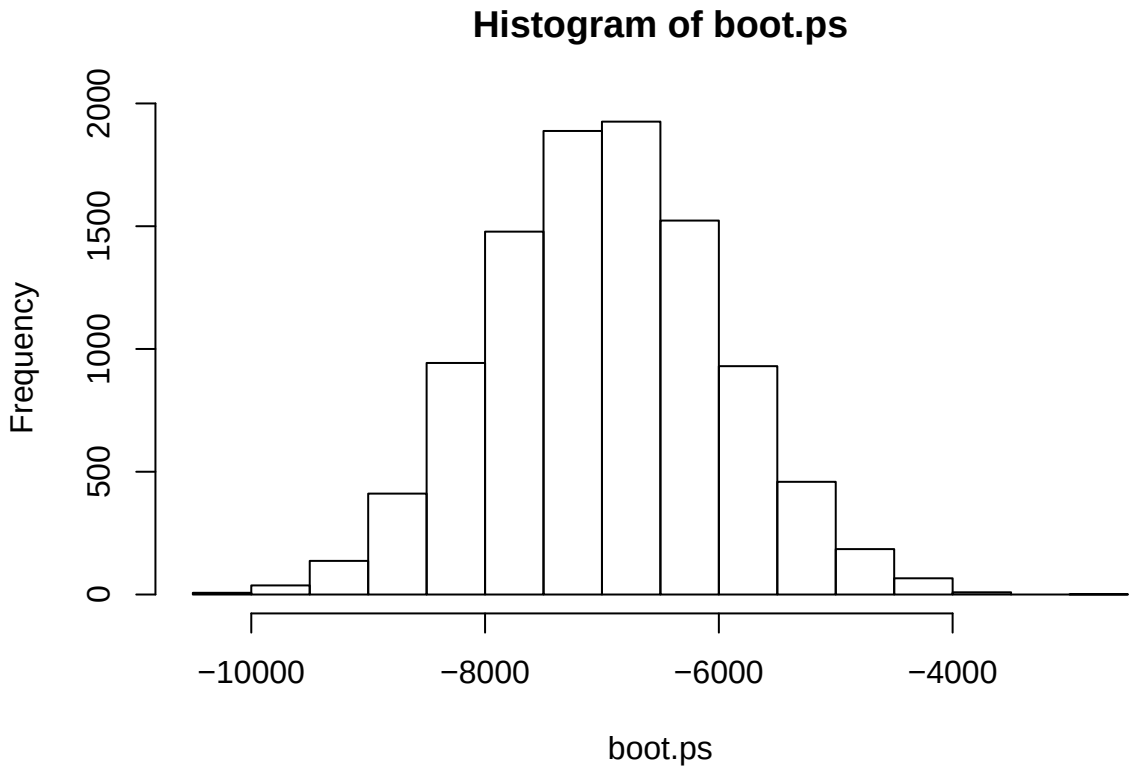
```r
  #define the PS weights for control group to get the ATT
  psid.sub$ps.weight <- ifelse(psid.sub$nsw == 0,
                               (psid.sub$ps.scores/(1 - psid.sub$ps.scores)), 0)

  #get the ATT
  ATT <- (1/sum(psid.sub$nsw == 1))*
    sum((psid.sub$nsw * psid.sub$re78) - (psid.sub$re78*(1 - psid.sub$nsw)* psid.sub$ps.weight), na.rm =

  #return the ATT
  return(ATT)
}

#bootstap the estimate
boot.ps <- replicate(n = 10000, expr = calc_pscore(), simplify = "vector")

#take a look at it
hist(boot.ps)
```




```r
#get estimate and standard error of the estimate
mean(boot.ps)
```

```
## [1] -6970.25
```

```r
sqrt(var(boot.ps))
```

```
## [1] 996.5845
```

Using the bootstrapped sample, the estimated effect of treatment on the treated is a decrease of $6970 (se = $996.58).

j.

```r
#j.
#use genmatch


#The covariates we want to obtain balance on
BalanceMat <- cbind(psid$age, psid$educ, psid$black, psid$hisp,
                    psid$married, psid$u74, psid$u75, psid$re75,
                    psid$re74)

#get weight matrix
genout <- GenMatch(Tr=psid$nsw,
                   X=psid[, c("hisp", "black", "educ", "married",
                              "u74", "u74", "re75", "re74", "age")],
                   BalanceMatrix = BalanceMat,
                   estimand="ATT",
                   M=1,
                   pop.size=1000,
                   max.generations=10,
                   wait.generations=1)
```

```
## Loading required namespace: rgenoud
```

```
##
##
## Sat Feb 22 13:39:51 2020
## Domains:
##  0.000000e+00   <=  X1   <=    1.000000e+03
##  0.000000e+00   <=  X2   <=    1.000000e+03
##  0.000000e+00   <=  X3   <=    1.000000e+03
##  0.000000e+00   <=  X4   <=    1.000000e+03
##  0.000000e+00   <=  X5   <=    1.000000e+03
##  0.000000e+00   <=  X6   <=    1.000000e+03
##  0.000000e+00   <=  X7   <=    1.000000e+03
##  0.000000e+00   <=  X8   <=    1.000000e+03
##  0.000000e+00   <=  X9   <=    1.000000e+03
##
## Data Type: Floating Point
## Operators (code number, name, population)
##  (1) Cloning.....................  122
##  (2) Uniform Mutation............  125
##  (3) Boundary Mutation...........  125
##  (4) Non-Uniform Mutation........  125
##  (5) Polytope Crossover..........  125
```

```
##  (6) Simple Crossover.................  126
##  (7) Whole Non-Uniform Mutation........  125
##  (8) Heuristic Crossover..............  126
##  (9) Local-Minimum Crossover...........  0
##
## SOFT Maximum Number of Generations: 10
## Maximum Nonchanging Generations: 1
## Population size       : 1000
## Convergence Tolerance: 1.000000e-03
##
## Not Using the BFGS Derivative Based Optimizer on the Best Individual Each Generation.
## Not Checking Gradients before Stopping.
## Using Out of Bounds Individuals.
##
## Maximization Problem.
## GENERATION: 0 (initializing the population)
## Lexical Fit..... 2.272099e-02  5.319785e-02  2.226246e-01  3.173158e-01  3.173158e-
01  3.563232e-01  5.139095e-01  5.865731e-01  5.865731e-01  6.637987e-01  6.723649e-
01  9.647361e-01  1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00
## #unique......... 1000, #Total UniqueCount: 1000
## var 1:
## best............ 1.817564e+02
## mean............ 4.880627e+02
## variance........ 8.390839e+04
## var 2:
## best............ 3.060854e+02
## mean............ 5.029369e+02
## variance........ 8.608517e+04
## var 3:
## best............ 3.774918e+01
## mean............ 5.052404e+02
## variance........ 8.331925e+04
## var 4:
## best............ 5.180914e+02
## mean............ 5.038556e+02
## variance........ 8.197554e+04
## var 5:
## best............ 6.095814e+02
## mean............ 4.891010e+02
## variance........ 8.457126e+04
## var 6:
## best............ 9.089981e+02
## mean............ 5.140775e+02
## variance........ 7.969952e+04
## var 7:
## best............ 3.999536e+02
## mean............ 5.019544e+02
## variance........ 8.103764e+04
## var 8:
## best............ 1.427068e+02
## mean............ 5.153542e+02
## variance........ 8.116842e+04
## var 9:
## best............ 9.681594e+02
```

```
## mean..... 4.983121e+02
## variance..... 8.264051e+04
##
## GENERATION: 1
## Lexical Fit..... 3.047234e-02  6.924680e-02  2.727149e-01  3.209609e-01  4.558922e-
01  7.829853e-01  7.857581e-01  7.857581e-01  9.447220e-01  9.877578e-01  1.000000e+00  1.000000e+00  1
## #unique..... 766, #Total UniqueCount: 1766
## var 1:
## best..... 1.435342e+02
## mean..... 3.003175e+02
## variance..... 5.723892e+04
## var 2:
## best..... 4.403986e+02
## mean..... 4.165797e+02
## variance..... 3.472968e+04
## var 3:
## best..... 5.252147e+01
## mean..... 9.888003e+01
## variance..... 1.799153e+04
## var 4:
## best..... 5.334053e+02
## mean..... 5.411624e+02
## variance..... 2.795320e+04
## var 5:
## best..... 3.092803e+02
## mean..... 5.255273e+02
## variance..... 6.014128e+04
## var 6:
## best..... 7.628114e+02
## mean..... 7.503077e+02
## variance..... 4.787083e+04
## var 7:
## best..... 2.685156e+02
## mean..... 3.655923e+02
## variance..... 3.062583e+04
## var 8:
## best..... 1.688072e+02
## mean..... 2.508209e+02
## variance..... 4.348569e+04
## var 9:
## best..... 7.444920e+02
## mean..... 7.742057e+02
## variance..... 4.358248e+04
##
## GENERATION: 2
## Lexical Fit..... 3.047234e-02  8.924281e-02  3.304106e-01  3.502179e-01  4.753205e-
01  6.806208e-01  6.806208e-01  8.052096e-01  8.860512e-01  9.877578e-01  1.000000e+00  1.000000e+00  1
## #unique..... 752, #Total UniqueCount: 2518
## var 1:
## best..... 1.527144e+02
## mean..... 1.950450e+02
## variance..... 9.973208e+03
## var 2:
## best..... 5.221980e+02
```

```
## mean..... 4.162362e+02
## variance..... 6.620128e+03
## var 3:
## best..... 5.467111e+01
## mean..... 6.756474e+01
## variance..... 6.683324e+03
## var 4:
## best..... 5.937397e+02
## mean..... 5.502806e+02
## variance..... 1.189836e+04
## var 5:
## best..... 3.195374e+02
## mean..... 4.331309e+02
## variance..... 3.796711e+04
## var 6:
## best..... 7.599888e+02
## mean..... 7.836407e+02
## variance..... 9.845598e+03
## var 7:
## best..... 2.671457e+02
## mean..... 2.932288e+02
## variance..... 6.837642e+03
## var 8:
## best..... 1.750932e+02
## mean..... 1.802388e+02
## variance..... 6.688833e+03
## var 9:
## best..... 7.451105e+02
## mean..... 7.796190e+02
## variance..... 1.730113e+04
##
## GENERATION: 3
## Lexical Fit..... 3.047234e-02  8.924281e-02  3.304106e-01  3.502179e-01  4.753205e-
01  6.806208e-01  6.806208e-01  8.052096e-01  8.860512e-01  9.877578e-01  1.000000e+00  1.000000e+00  1
## #unique..... 736, #Total UniqueCount: 3254
## var 1:
## best..... 1.527144e+02
## mean..... 1.814274e+02
## variance..... 1.244950e+04
## var 2:
## best..... 5.221980e+02
## mean..... 4.946320e+02
## variance..... 6.686050e+03
## var 3:
## best..... 5.467111e+01
## mean..... 6.776135e+01
## variance..... 5.209990e+03
## var 4:
## best..... 5.937397e+02
## mean..... 5.770441e+02
## variance..... 9.038555e+03
## var 5:
## best..... 3.195374e+02
## mean..... 3.580654e+02
```

```
## variance........ 2.271481e+04
## var 6:
## best............ 7.599888e+02
## mean............ 7.629931e+02
## variance........ 5.605512e+03
## var 7:
## best............ 2.671457e+02
## mean............ 2.678824e+02
## variance........ 2.391310e+03
## var 8:
## best............ 1.750932e+02
## mean............ 1.849858e+02
## variance........ 4.544963e+03
## var 9:
## best............ 7.451105e+02
## mean............ 7.416791e+02
## variance........ 5.191376e+03
##
## GENERATION: 4
## Lexical Fit..... 3.047234e-02  8.924281e-02  3.304106e-01  3.502179e-01  4.753205e-
01  6.806208e-01  6.806208e-01  8.052096e-01  8.860512e-01  9.877578e-01  1.000000e+00  1.000000e+00  1
## #unique......... 647, #Total UniqueCount: 3901
## var 1:
## best............ 1.527144e+02
## mean............ 1.703798e+02
## variance........ 7.614570e+03
## var 2:
## best............ 5.221980e+02
## mean............ 5.201065e+02
## variance........ 3.936668e+03
## var 3:
## best............ 5.467111e+01
## mean............ 7.186412e+01
## variance........ 8.515208e+03
## var 4:
## best............ 5.937397e+02
## mean............ 5.804390e+02
## variance........ 4.163829e+03
## var 5:
## best............ 3.195374e+02
## mean............ 3.301523e+02
## variance........ 6.113196e+03
## var 6:
## best............ 7.599888e+02
## mean............ 7.501328e+02
## variance........ 5.070217e+03
## var 7:
## best............ 2.671457e+02
## mean............ 2.745080e+02
## variance........ 3.720346e+03
## var 8:
## best............ 1.750932e+02
## mean............ 1.832603e+02
## variance........ 3.110296e+03
```

```
## var 9:
## best..... 7.451105e+02
## mean..... 7.364873e+02
## variance........ 4.364942e+03
##
## 'wait.generations' limit reached.
## No significant improvement in 1 generations.
##
## Solution Lexical Fitness Value:
## 3.047234e-02  8.924281e-02  3.304106e-01  3.502179e-01  4.753205e-01  6.806208e-
01  6.806208e-01  8.052096e-01  8.860512e-01  9.877578e-01  1.000000e+00  1.000000e+00  1.000000e+00  1
##
## Parameters at the Solution:
##
##  X[ 1] : 1.527144e+02
##  X[ 2] : 5.221980e+02
##  X[ 3] : 5.467111e+01
##  X[ 4] : 5.937397e+02
##  X[ 5] : 3.195374e+02
##  X[ 6] : 7.599888e+02
##  X[ 7] : 2.671457e+02
##  X[ 8] : 1.750932e+02
##  X[ 9] : 7.451105e+02
##
## Solution Found Generation 2
## Number of Generations Run 4
##
## Sat Feb 22 13:42:21 2020
## Total run time : 0 hours 2 minutes and 30 seconds
```

```
#match
match.gen <- Match(Y=psid$re78,
                   Tr=psid$nsw,
                   X=psid[, c("hisp", "black", "educ", "married",
                              "u74", "u74", "re75", "re74", "age")],
                   estimand="ATT",
                   Weight.matrix=genout)
summary(match.gen)
```

```
##
## Estimate...  2189.3
## AI SE......  1811.1
## T-stat.....  1.2088
## p.val......  0.22674
##
## Original number of observations..............  2675
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  201
```

```
#check balance
MatchBalance(nsw ~ age + educ + black + u74 +
             married + hisp + re74 + re75  +
```

```
            u75 + u78,
            data = psid,
            match.out = match.gen,
            nboots = 10)
```

```
##
## ***** (V1) age *****
##                         Before Matching        After Matching
## mean treatment........     25.816             25.816
## mean control..........     34.851             26.077
## std mean diff.........    -126.27            -3.6514
##
## mean raw eQQ diff.....     9.0432             1.5572
## med  raw eQQ diff.....          8                  1
## max  raw eQQ diff.....         17                  7
##
## mean eCDF diff........    0.23165            0.04606
## med  eCDF diff........    0.25299           0.029851
## max  eCDF diff........    0.37714            0.14428
##
## var ratio (Tr/Co).....    0.46963            0.86582
## T-test p-value........ < 2.22e-16            0.35022
## KS Bootstrap p-value.. < 2.22e-16          < 2.22e-16
## KS Naive p-value...... < 2.22e-16           0.030472
## KS Statistic..........    0.37714            0.14428
##
##
## ***** (V2) educ *****
##                         Before Matching        After Matching
## mean treatment........     10.346             10.346
## mean control..........     12.117             10.463
## std mean diff.........    -88.077            -5.8248
##
## mean raw eQQ diff.....     1.8595             0.40796
## med  raw eQQ diff.....          2                  0
## max  raw eQQ diff.....          5                  2
##
## mean eCDF diff........     0.1091            0.02914
## med  eCDF diff........    0.01944           0.012438
## max  eCDF diff........    0.40289            0.12438
##
## var ratio (Tr/Co).....    0.42549            0.84626
## T-test p-value........ < 2.22e-16            0.47532
## KS Bootstrap p-value.. < 2.22e-16          < 2.22e-16
## KS Naive p-value...... < 2.22e-16           0.089243
## KS Statistic..........    0.40289            0.12438
##
##
## ***** (V3) black *****
##                         Before Matching        After Matching
## mean treatment........    0.84324            0.84324
## mean control..........     0.2506            0.84324
## std mean diff.........     162.56                  0
```

```
##
## mean raw eQQ diff.....    0.58919                  0
## med  raw eQQ diff.....          1                  0
## max  raw eQQ diff.....          1                  0
##
## mean eCDF diff........    0.29632                  0
## med  eCDF diff........    0.29632                  0
## max  eCDF diff........    0.59264                  0
##
## var ratio (Tr/Co).....    0.70739                  1
## T-test p-value........ < 2.22e-16                  1
##
##
## ***** (V4) u74 *****
##                          Before Matching        After Matching
## mean treatment........    0.70811            0.70811
## mean control..........   0.086345            0.70811
## std mean diff.........     136.39                  0
##
## mean raw eQQ diff.....    0.62162                  0
## med  raw eQQ diff.....          1                  0
## max  raw eQQ diff.....          1                  0
##
## mean eCDF diff........    0.31088                  0
## med  eCDF diff........    0.31088                  0
## max  eCDF diff........    0.62176                  0
##
## var ratio (Tr/Co).....     2.6332                  1
## T-test p-value........ < 2.22e-16                  1
##
##
## ***** (V5) married *****
##                          Before Matching        After Matching
## mean treatment........    0.18919            0.18919
## mean control..........    0.86627            0.18919
## std mean diff.........    -172.41                  0
##
## mean raw eQQ diff.....    0.67568                  0
## med  raw eQQ diff.....          1                  0
## max  raw eQQ diff.....          1                  0
##
## mean eCDF diff........    0.33854                  0
## med  eCDF diff........    0.33854                  0
## max  eCDF diff........    0.67708                  0
##
## var ratio (Tr/Co).....     1.3308                  1
## T-test p-value........ < 2.22e-16                  1
##
##
## ***** (V6) hisp *****
##                          Before Matching        After Matching
## mean treatment........   0.059459           0.059459
## mean control..........    0.03253           0.059459
## std mean diff.........     11.357                  0
```

```
##
## mean raw eQQ diff.....   0.027027                 0
## med  raw eQQ diff.....          0                 0
## max  raw eQQ diff.....          1                 0
##
## mean eCDF diff........   0.013465                 0
## med  eCDF diff........   0.013465                 0
## max  eCDF diff........   0.026929                 0
##
## var ratio (Tr/Co).....     1.7859                 1
## T-test p-value........    0.13173                 1
##
##
## ***** (V7) re74 *****
##                          Before Matching        After Matching
## mean treatment........     2095.6                 2095.6
## mean control..........      19429                 2126.9
## std mean diff.........    -354.71               -0.64183
##
## mean raw eQQ diff.....      17663                 236.23
## med  raw eQQ diff.....      18417                      0
## max  raw eQQ diff.....     102109                 3274.1
##
## mean eCDF diff........    0.46806               0.013248
## med  eCDF diff........    0.54766              0.0099502
## max  eCDF diff........    0.72924               0.044776
##
## var ratio (Tr/Co).....    0.13285                 1.0678
## T-test p-value........ < 2.22e-16                0.80521
## KS Bootstrap p-value.. < 2.22e-16                   0.8
## KS Naive p-value...... < 2.22e-16                0.98776
## KS Statistic..........    0.72924               0.044776
##
##
## ***** (V8) re75 *****
##                          Before Matching        After Matching
## mean treatment........     1532.1                 1532.1
## mean control..........      19063                 1550.9
## std mean diff.........    -544.58               -0.58517
##
## mean raw eQQ diff.....      17978                 222.05
## med  raw eQQ diff.....      17903                      0
## max  raw eQQ diff.....     131511                 4218.2
##
## mean eCDF diff........    0.46947               0.018274
## med  eCDF diff........    0.53317              0.0099502
## max  eCDF diff........    0.77362               0.094527
##
## var ratio (Tr/Co).....   0.056057                0.89028
## T-test p-value........ < 2.22e-16                0.88605
## KS Bootstrap p-value.. < 2.22e-16                   0.1
## KS Naive p-value...... < 2.22e-16                0.33041
## KS Statistic..........    0.77362               0.094527
##
```

```
##
## ***** (V9) u75 *****
##                         Before Matching        After Matching
## mean treatment........        0.6               0.6
## mean control..........        0.1             0.58378
## std mean diff.........     101.79              3.3012
##
## mean raw eQQ diff.....     0.4973           0.014925
## med  raw eQQ diff.....          0                  0
## max  raw eQQ diff.....          1                  1
##
## mean eCDF diff........       0.25          0.0074627
## med  eCDF diff........       0.25          0.0074627
## max  eCDF diff........        0.5           0.014925
##
## var ratio (Tr/Co).....     2.6801            0.98773
## T-test p-value........ < 2.22e-16            0.68062
##
##
## ***** (V10) u78 *****
##                         Before Matching        After Matching
## mean treatment........    0.24324           0.24324
## mean control..........    0.11486           0.36577
## std mean diff.........     29.842            -28.48
##
## mean raw eQQ diff.....    0.12432           0.12935
## med  raw eQQ diff.....          0                 0
## max  raw eQQ diff.....          1                 1
##
## mean eCDF diff........   0.064192          0.064677
## med  eCDF diff........   0.064192          0.064677
## max  eCDF diff........    0.12838           0.12935
##
## var ratio (Tr/Co).....     1.8197            0.7935
## T-test p-value........ 9.7001e-05          0.015395
##
##
## Before Matching Minimum p.value: < 2.22e-16
## Variable Name(s): age educ black u74 married re74 re75 u75  Number(s): 1 2 3 4 5 7 8 9
##
## After Matching Minimum p.value: < 2.22e-16
## Variable Name(s): age educ  Number(s): 1 2
```

Using genetic matching, the ATT is estimated to be a nonsignificant increase of \$2248.1 (se = \$1784.30). All covariates are balanced except age, education, and unemployment in 1978. While it takes longer, I prefer the genetic matching because of the iterative and automated matching process, and the ability to acheive good balance (at least using this data).

k.

The ATT is identified if there are no open backdoor paths confounding the association between program participation and 1978 earnings (including those by unobserved confounders), if we have not conditioned on any colliders, if there is no measurement error, and if we haven't conditioned on post-treatment covariates.

Matching makes assumptions about confounding/blocking backdoor paths somewhat more plausible, but, as in regression, the success of matching in controlling for confounding depends on appropriate selection of matched covariates. We also cannot match on unobserved covariates/confounders, which may still differ between treatment and control groups after matching; thus, omitted variable bias is still possible. However, matching can be helpful in ensuring balance on what you have measured and positivity, whereas regression can violate positivity if there are no/few observations with given combinations of covariates.