Causal Inference TA Section (2): Matching

shengqiaolin@utexas.edu (mailto:shengqiaolin@utexas.edu)

2023-02-08

Today’s Goal

- Calculating Mahalanobis distance
- Propensity Score Matching (*Matching* (https://cran.r-project.org/web/packages/Matching/index.html))
- Coarsened Exact Matching (*MatchIt* (https://cran.r-project.org/web/packages/MatchIt/))
- Some other resources

Calculating Mahalanobis distance

$$\mathrm{Mahalanobis\ Distance}_{ij} = (X_i - X_j)'\Sigma^{-1}(X_i - X_j)$$

where Σ stands var-covar matrix of X.

For example, if $X_i = (0,0), X_j = (1,0), \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

```
mahalanobis(c(0,0), c(1,0), cov=matrix(c(1,0,0,1),2,2))
```

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

Matching

This example is adopted from Jasjeet S. Sekhon.The 2004 Florida Optical Voting Machine Controversy: A Causal Analysis Using Matching (http://sekhon.berkeley.edu/papers/SekhonOpticalMatch.pdf).

```
library(Matching)
load("FLopticalData.RData")
attach(data)
```

The unit of analysis is county in Florida. The treatment is *etouch* i.e. voting machine for electronic voting, the outcome is *bush04*, the vote share for Bush in 2004. We start with looking at the data.

```
dim(data)
```

```
## [1] 67 19
```

```
names(data)
```

```
##  [1] "county"        "bush04"        "regTot04.rep"  "regTot04.dem" 
##  [5] "regTot04.ind"  "etouch"        "income"        "votePer96.dem"
##  [9] "votePer96.rep" "votePer00.dem" "votePer00.rep" "regPer00.dem" 
## [13] "regPer00.rep"  "turnout00"     "hisp00"        "white00"      
## [17] "black00"       "lowEduc00"     "foreignBorn00"
```

These counties are very heterogeneous on many variables: partisanship, races, education, turnout etc.

```
summary(data)
```

```
##       county       bush04        regTot04.rep     regTot04.dem   
##  Alachua : 1   Min.   :0.2980   Min.   :   260   Min.   :  3323  
##  Baker   : 1   1st Qu.:0.5326   1st Qu.:  3116   1st Qu.:  8810  
##  Bay     : 1   Median :0.5861   Median : 25897   Median : 23777  
##  Bradford: 1   Mean   :0.5952   Mean   : 53536   Mean   : 58451  
##  Brevard : 1   3rd Qu.:0.6783   3rd Qu.: 72664   3rd Qu.: 64509  
##  Broward : 1   Max.   :0.7773   Max.   :343772   Max.   :489113  
##  (Other) :61                                                     
##   regTot04.ind        etouch           income      votePer96.dem   
##  Min.   :    97   Min.   :0.0000   Min.   :26032   Min.   :0.2580  
##  1st Qu.:  1036   1st Qu.:0.0000   1st Qu.:30029   1st Qu.:0.3909  
##  Median : 11175   Median :0.0000   Median :33779   Median :0.4304  
##  Mean   : 24340   Mean   :0.2239   Mean   :35385   Mean   :0.4319  
##  3rd Qu.: 23506   3rd Qu.:0.0000   3rd Qu.:40249   3rd Qu.:0.4715  
##  Max.   :202245   Max.   :1.0000   Max.   :52244   Max.   :0.6627  
##                                                                     
##  votePer96.rep    votePer00.dem    votePer00.rep     regPer00.dem   
##  Min.   :0.2687   Min.   :0.2398   Min.   :0.3093   Min.   :0.2374  
##  1st Qu.:0.3920   1st Qu.:0.3715   1st Qu.:0.4974   1st Qu.:0.3997  
##  Median :0.4417   Median :0.4285   Median :0.5465   Median :0.4834  
##  Mean   :0.4446   Mean   :0.4261   Mean   :0.5489   Mean   :0.5558  
##  3rd Qu.:0.4825   3rd Qu.:0.4785   3rd Qu.:0.6068   3rd Qu.:0.7328  
##  Max.   :0.6453   Max.   :0.6741   Max.   :0.7370   Max.   :0.9358  
##                                                                     
##   regPer00.rep       turnout00          hisp00           white00      
##  Min.   :0.04424   Min.   :0.4755   Min.   :0.01514   Min.   :0.4158  
##  1st Qu.:0.19717   1st Qu.:0.6470   1st Qu.:0.02715   1st Qu.:0.8045  
##  Median :0.35667   Median :0.6855   Median :0.04888   Median :0.8500  
##  Mean   :0.31979   Mean   :0.6803   Mean   :0.08533   Mean   :0.8311  
##  3rd Qu.:0.42368   3rd Qu.:0.7219   3rd Qu.:0.09394   3rd Qu.:0.8955  
##  Max.   :0.57081   Max.   :0.8270   Max.   :0.57325   Max.   :0.9558  
##                                                                      
##     black00          lowEduc00       foreignBorn00    
##  Min.   :0.02151   Min.   :0.01762   Min.   :0.01123  
##  1st Qu.:0.08187   1st Qu.:0.03470   1st Qu.:0.02404  
##  Median :0.11600   Median :0.04685   Median :0.05269  
##  Mean   :0.14186   Mean   :0.05282   Mean   :0.07494  
##  3rd Qu.:0.17171   3rd Qu.:0.06493   3rd Qu.:0.09134  
##  Max.   :0.57358   Max.   :0.14686   Max.   :0.50936  
## 
```

If we simply looks at the difference in means, using electronical voting decreased Bush’s share

```
table(etouch)
```

```
## etouch
##  0  1
## 52 15
```

```
tapply(bush04, etouch, mean)
```

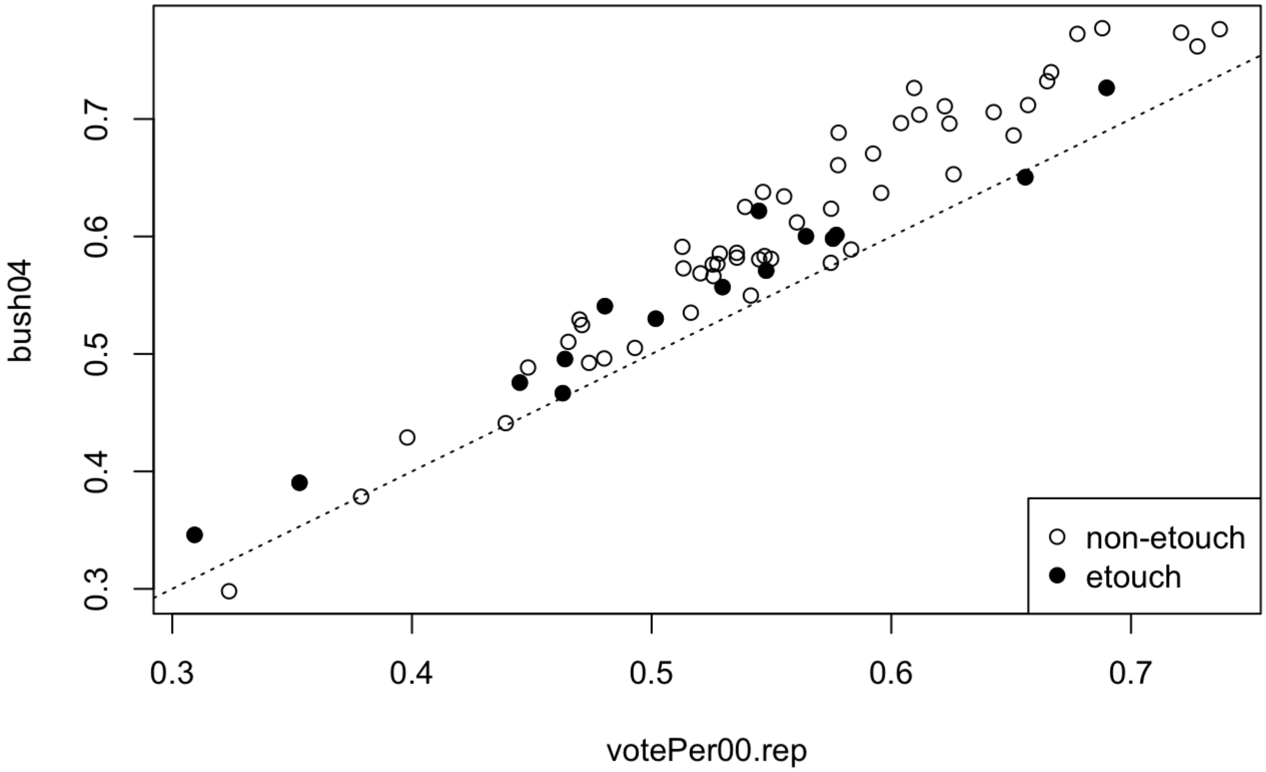
```
##         0         1
## 0.6097499 0.5447324
```

```
summary(lm(bush04 ~ etouch))
```

```
##
## Call:
## lm(formula = bush04 ~ etouch)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.311761 -0.054553 -0.004049  0.077816  0.181817
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.60975    0.01441  42.327   <2e-16 ***
## etouch      -0.06502    0.03045  -2.136   0.0365 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1039 on 65 degrees of freedom
## Multiple R-squared:  0.06556,    Adjusted R-squared:  0.05118
## F-statistic:  4.56 on 1 and 65 DF,  p-value: 0.03649
```

What if we add some covariates given the heterogeneity among counties. Let’s start with vote shares for Bush in 2000.

```
plot(bush04 ~ votePer00.rep)
points(bush04[etouch==1] ~ votePer00.rep[etouch==1], pch=19)
abline(0,1, lty=3)
legend("bottomright", c("non-etouch", "etouch"), pch=c(1,19))
```



```
summary(lm(bush04 ~ etouch + votePer00.rep))
```

```
## 
## Call:
## lm(formula = bush04 ~ etouch + votePer00.rep)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.052780 -0.017312  0.000566  0.021303  0.060632 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   -0.01215    0.02124  -0.572   0.5692    
## etouch        -0.01417    0.00815  -1.739   0.0868 .  
## votePer00.rep  1.11227    0.03739  29.748   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.02719 on 64 degrees of freedom
## Multiple R-squared:  0.937,  Adjusted R-squared:  0.935 
## F-statistic: 475.8 on 2 and 64 DF,  p-value: < 2.2e-16
```

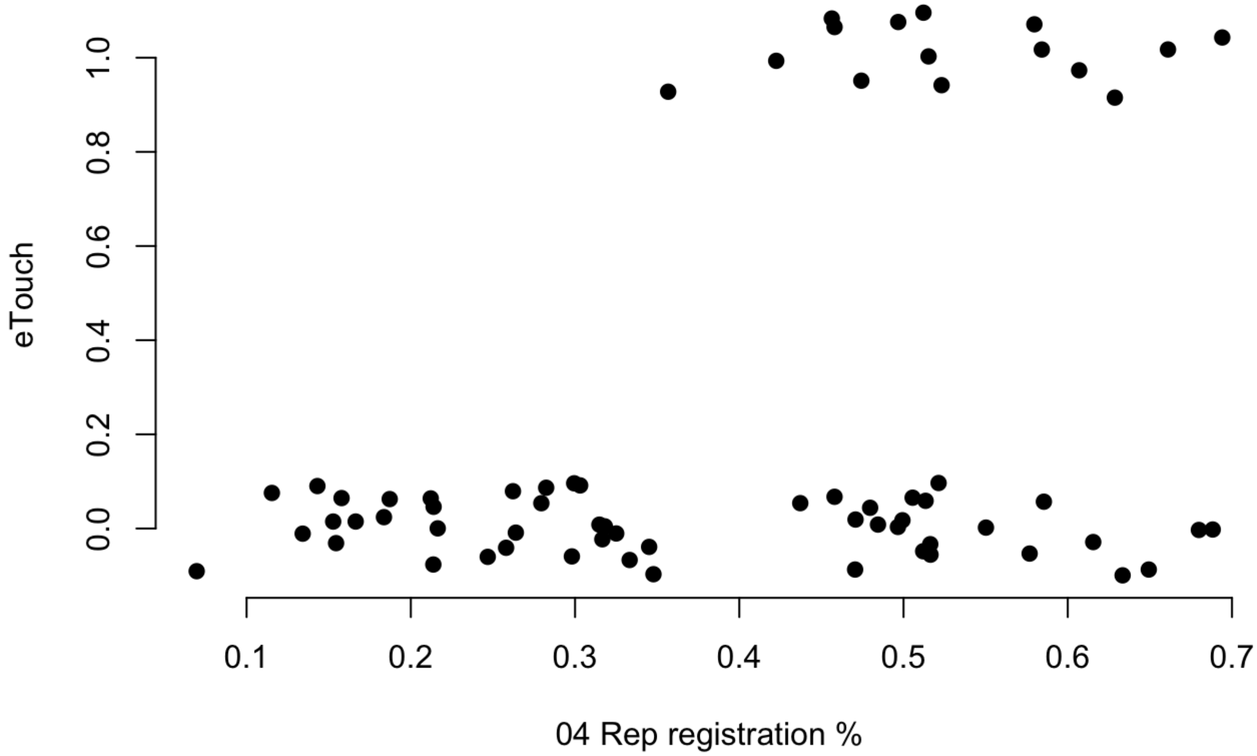
We then create variables for the share of both parties in 2004 registration and add them into the model:

```
reg2pty04.rep <- regTot04.rep / (regTot04.rep + regTot04.dem)
reg2pty04.dem <- regTot04.dem / (regTot04.rep + regTot04.dem)
summary(lm(bush04 ~ etouch + votePer00.rep + reg2pty04.rep))
```

```
## 
## Call:
## lm(formula = bush04 ~ etouch + votePer00.rep + reg2pty04.rep)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.060715 -0.012574  0.000669  0.018052  0.044165 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   -0.0034137  0.0196353  -0.174 0.862536    
## etouch        -0.0002508  0.0084093  -0.030 0.976298    
## votePer00.rep  1.1453036  0.0354938  32.268  < 2e-16 ***
## reg2pty04.rep -0.0745826  0.0206343  -3.614 0.000599 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.02494 on 63 degrees of freedom
## Multiple R-squared:  0.9478, Adjusted R-squared:  0.9453 
## F-statistic: 381.3 on 3 and 63 DF,  p-value: < 2.2e-16
```

Note that p value is 0.98, which indicates a super strong **nonfinding**. The changes in the significance of *etouch* suggest the results are very model-based. Matching might help, for

```
par(mfrow=c(1,1))
plot(reg2pty04.rep, jitter(etouch, factor=.5), pch=19,
     ylab="eTouch", bty="n", xlab="04 Rep registration %")
```

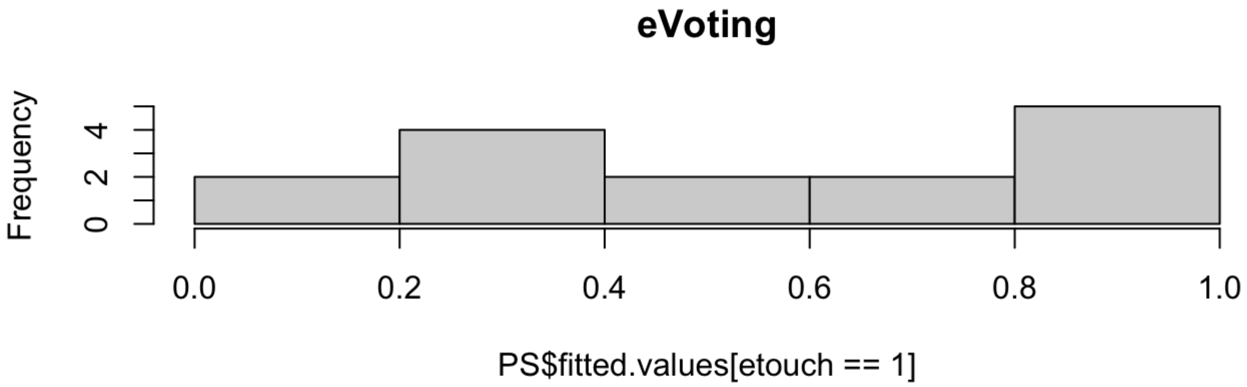


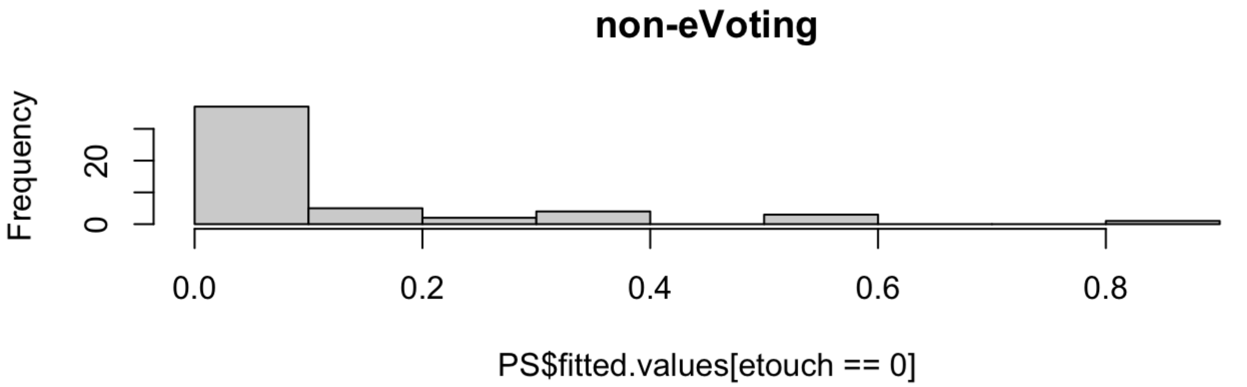
Propensity Score Matching

- Run a logit/probit model to estimate propensity scores for each observation
- Matching treated and control groups based on the propensity scores
- Show balance of your matching
- Estimate the treatment effects (ATT)

We start with calculating the propensity of each county to employ voting machines using logistic regression. Here we use the share of each party in voter registration in 2004 and the population structure (share of each race). The distributions look quite different. Treated groups are those that very likely to be treated and control groups are those unlikely to be treated.

```
PS <- glm(etouch ~ regTot04.dem + regTot04.rep + regTot04.ind + hisp00 + white00 + black
00, family=binomial(link="logit"))
par(mfrow=c(2,1))
hist(PS$fitted.values[etouch==1], main="eVoting")
hist(PS$fitted.values[etouch==0], main="non-eVoting")
```



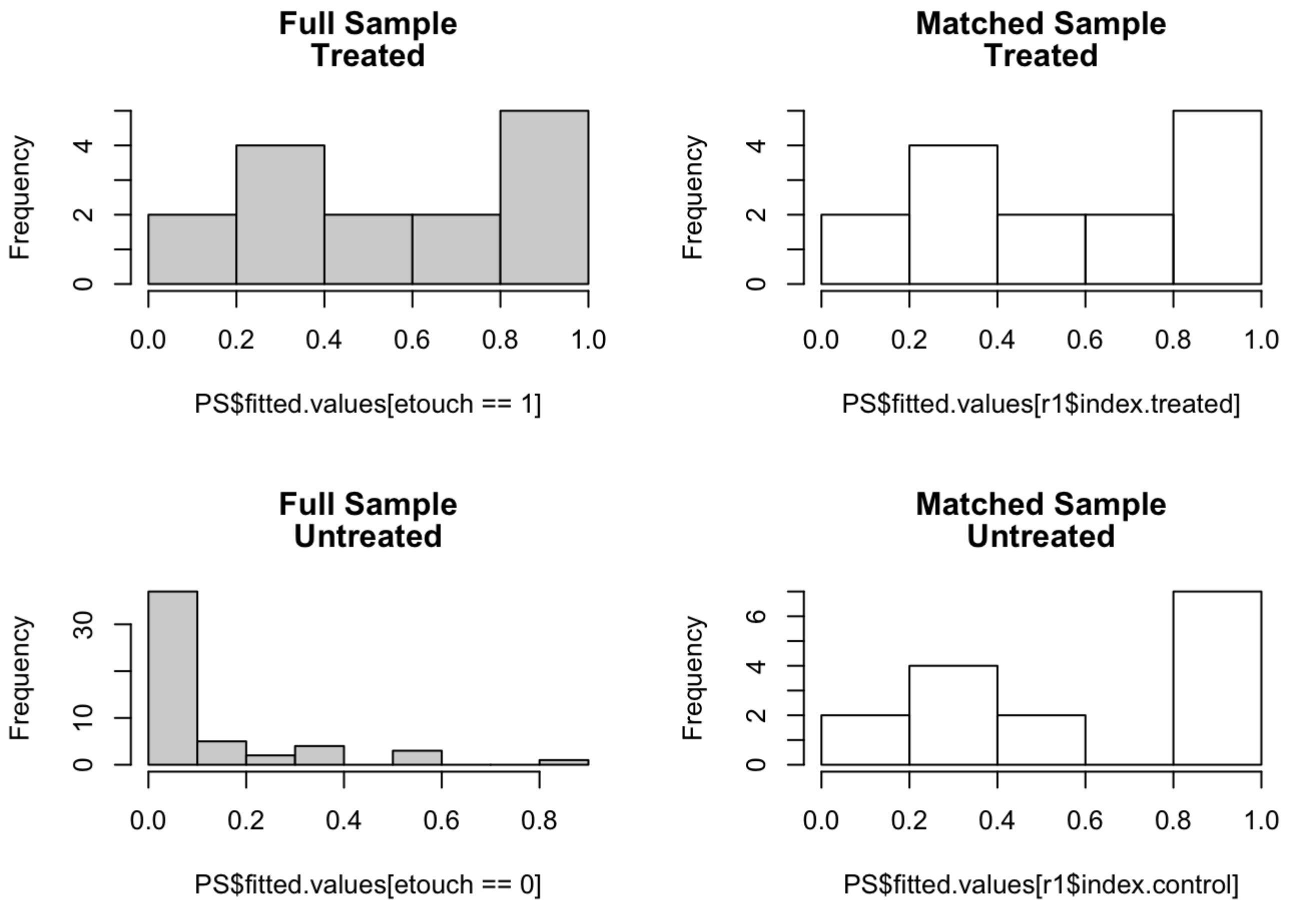


We then match the sample based on propensity scores and estimate the treatment effect (By default, it uses 1:1 matching. All weights=1). The result remains insignificant.

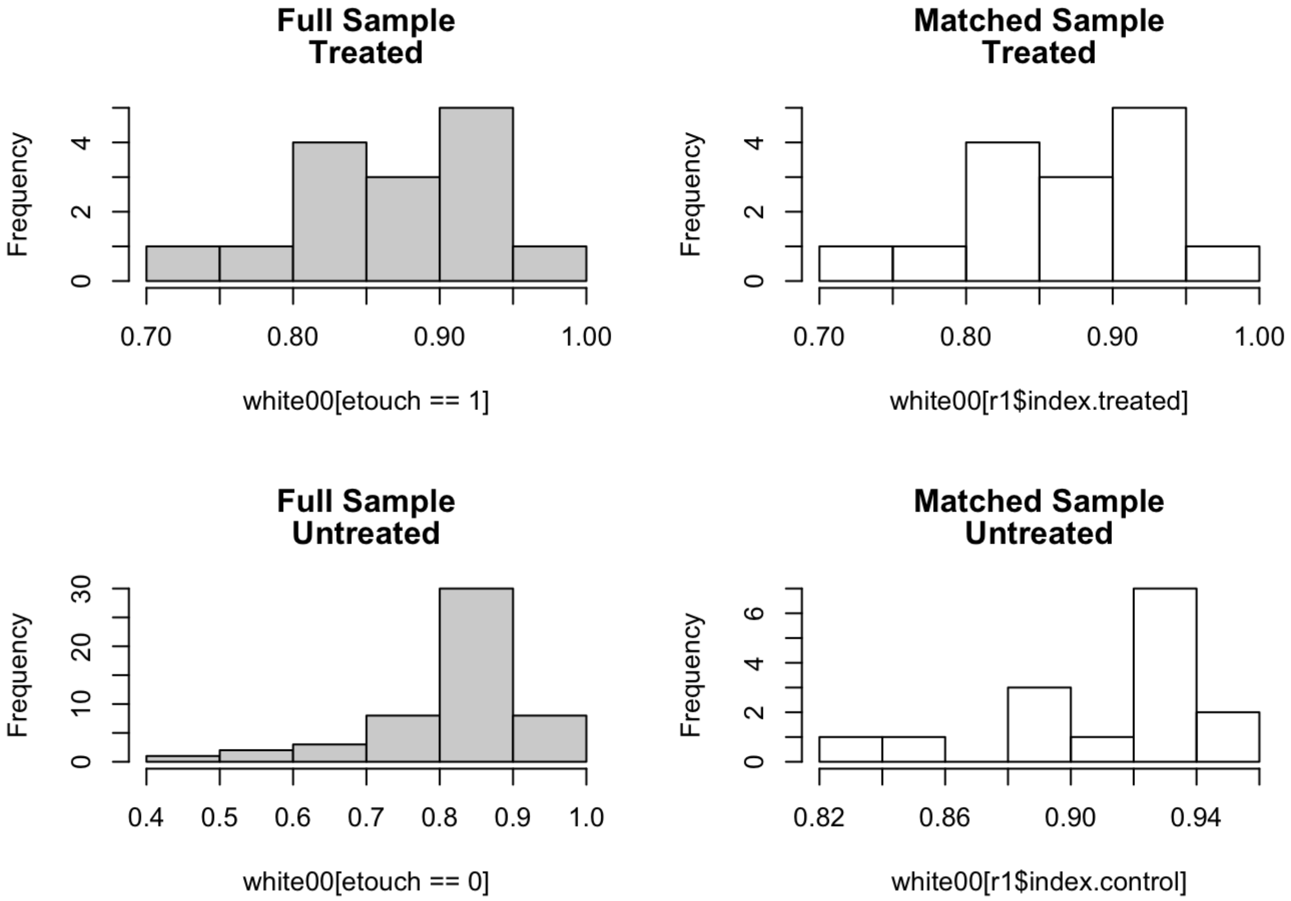
```
r1 <- Match(Y=bush04, Tr=etouch, X=PS$fitted.values)
```

Let check the balance of matching. We take two examples: the propensity and the population structure.

```
library(weights)
par(mfcol=c(2,2))
hist(PS$fitted.values[etouch==1],
     main=c("Full Sample", "Treated"))
hist(PS$fitted.values[etouch==0],
     main=c("Full Sample", "Untreated"))
wtd.hist(PS$fitted.values[r1$index.treated],
         weight=r1$weights,
         main=c("Matched Sample", "Treated"))
wtd.hist(PS$fitted.values[r1$index.control],
         weight=r1$weights,
         main=c("Matched Sample", "Untreated"))
```



```
par(mfcol=c(2,2))
hist(white00[etouch==1],
     main=c("Full Sample", "Treated"))
hist(white00[etouch==0],
     main=c("Full Sample", "Untreated"))
wtd.hist(white00[r1$index.treated],
         weight=r1$weights,
         main=c("Matched Sample", "Treated"))
wtd.hist(white00[r1$index.control],
         weight=r1$weights,
         main=c("Matched Sample", "Untreated"))
```



Full Sample Treated

Matched Sample Treated

Full Sample Untreated

Matched Sample Untreated

We can also see some tables and stats if needed.

```
mb <- MatchBalance(etouch ~ regTot04.dem + regTot04.rep + regTot04.ind + hisp00 + white0
0 + black00,  match.out=r1)
```

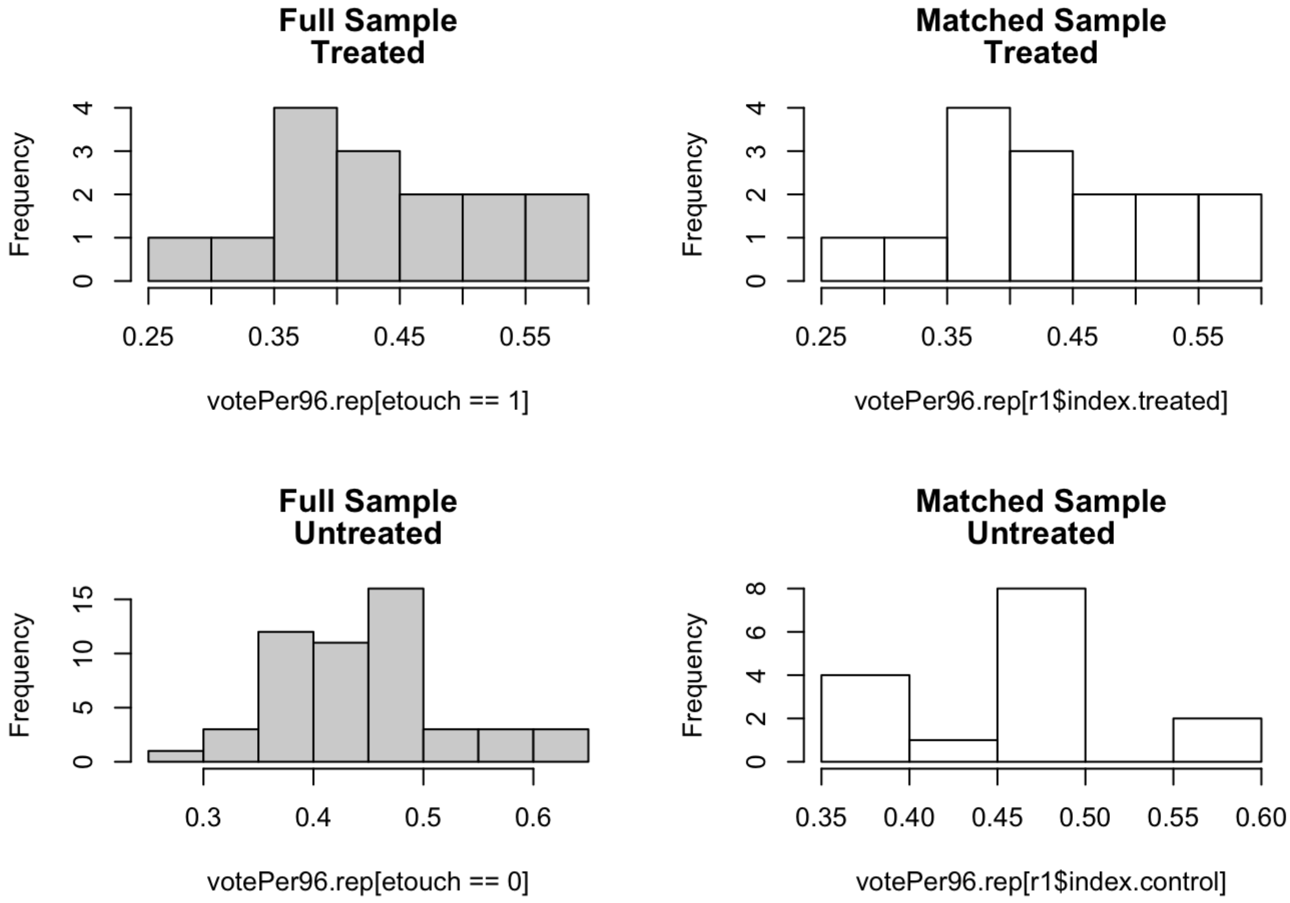
```
## 
## ***** (V1) regTot04.dem *****
##                         Before Matching        After Matching
## mean treatment........     139111            139111 
## mean control..........      35184             63915 
## std mean diff.........     67.161            48.594 
## 
## mean raw eQQ diff.....      99234             87584 
## med  raw eQQ diff.....      39829             20301 
## max  raw eQQ diff.....     294024            373341 
## 
## mean eCDF diff........    0.29615           0.14242 
## med  eCDF diff........     0.3141           0.13333 
## max  eCDF diff........    0.48077           0.33333 
## 
## var ratio (Tr/Co).....     12.041            20.212 
## T-test p-value........   0.021589          0.062429 
## KS Bootstrap p-value..      0.006             0.274 
## KS Naive p-value......  0.0057459           0.37521 
## KS Statistic..........    0.48077           0.33333 
## 
## 
## ***** (V2) regTot04.rep *****
##                         Before Matching        After Matching
## mean treatment........     124571            124571 
## mean control..........      33046             84398 
## std mean diff.........     88.941            39.038 
## 
## mean raw eQQ diff.....      90564             57546 
## med  raw eQQ diff.....      79045             27428 
## max  raw eQQ diff.....     174632            202169 
## 
## mean eCDF diff........    0.33718           0.13636 
## med  eCDF diff........       0.35           0.13333 
## max  eCDF diff........    0.57821           0.33333 
## 
## var ratio (Tr/Co).....     5.3691            5.5937 
## T-test p-value........  0.0041597          0.093614 
## KS Bootstrap p-value.. < 2.22e-16             0.282 
## KS Naive p-value......  0.0003795           0.37521 
## KS Statistic..........    0.57821           0.33333 
## 
## 
## ***** (V3) regTot04.ind *****
##                         Before Matching        After Matching
## mean treatment........      61518             61518 
## mean control..........      13615             32768 
## std mean diff.........     72.752            43.663 
## 
## mean raw eQQ diff.....      45249             40340 
## med  raw eQQ diff.....      20246             18329 
## max  raw eQQ diff.....     146108            150973 
```

```
##
## mean eCDF diff........    0.32505            0.16364
## med  eCDF diff........    0.34231            0.13333
## max  eCDF diff........    0.54872                0.4
##
## var ratio (Tr/Co).....     11.279             17.074
## T-test p-value........   0.014141           0.082163
## KS Bootstrap p-value.. < 2.22e-16              0.124
## KS Naive p-value......  0.0018047             0.1813
## KS Statistic..........    0.54872                0.4
##
##
## ***** (V4) hisp00 *****
##                        Before Matching        After Matching
## mean treatment........    0.12191            0.12191
## mean control..........   0.074775           0.068197
## std mean diff.........     34.627             39.459
##
## mean raw eQQ diff.....    0.04384           0.054206
## med  raw eQQ diff.....   0.032182           0.031623
## max  raw eQQ diff.....    0.17734            0.17734
##
## mean eCDF diff........     0.2061            0.26667
## med  eCDF diff........    0.20256                0.3
## max  eCDF diff........    0.44359                0.6
##
## var ratio (Tr/Co).....     2.5494             2.2167
## T-test p-value........    0.22053            0.22473
## KS Bootstrap p-value..      0.004              0.004
## KS Naive p-value......   0.013913          0.0090332
## KS Statistic..........    0.44359                0.6
##
##
## ***** (V5) white00 *****
##                        Before Matching        After Matching
## mean treatment........    0.86834            0.86834
## mean control..........    0.82038            0.91283
## std mean diff.........     72.623             -67.36
##
## mean raw eQQ diff.....    0.06333           0.046525
## med  raw eQQ diff.....   0.034532           0.043789
## max  raw eQQ diff.....     0.3273           0.085413
##
## mean eCDF diff........    0.14545            0.25758
## med  eCDF diff........    0.12821            0.26667
## max  eCDF diff........    0.36026            0.53333
##
## var ratio (Tr/Co).....    0.40098             2.8571
## T-test p-value........   0.038725           0.031501
## KS Bootstrap p-value..      0.068              0.022
## KS Naive p-value......   0.070883           0.028057
## KS Statistic..........    0.36026            0.53333
```

```
##
##
## ***** (V6) black00 *****
##                        Before Matching        After Matching
## mean treatment........    0.10768            0.10768
## mean control..........    0.15173           0.067459
## std mean diff.........    -72.421             66.116
##
## mean raw eQQ diff.....   0.050061           0.042927
## med  raw eQQ diff.....   0.024652           0.041538
## max  raw eQQ diff.....    0.35486           0.075748
##
## mean eCDF diff........    0.12979            0.23333
## med  eCDF diff........    0.13205                0.2
## max  eCDF diff........    0.28333            0.53333
##
## var ratio (Tr/Co).....    0.33124             2.6379
## T-test p-value........   0.046802           0.032147
## KS Bootstrap p-value..      0.214              0.022
## KS Naive p-value......    0.24759           0.028057
## KS Statistic..........    0.28333            0.53333
##
##
## Before Matching Minimum p.value: < 2.22e-16
## Variable Name(s): regTot04.rep regTot04.ind  Number(s): 2 3
##
## After Matching Minimum p.value: 0.004
## Variable Name(s): hisp00  Number(s): 4
```

Do matching affect other variables? Seems not

```
par(mfcol=c(2,2))
hist(votePer96.rep[etouch==1],
     main=c("Full Sample", "Treated"))
hist(votePer96.rep[etouch==0],
     main=c("Full Sample", "Untreated"))
wtd.hist(votePer96.rep[r1$index.treated],
         weight=r1$weights,
         main=c("Matched Sample", "Treated"))
wtd.hist(votePer96.rep[r1$index.control],
         weight=r1$weights,
         main=c("Matched Sample", "Untreated"))
```


Full Sample Treated

Matched Sample Treated

Full Sample Untreated

Matched Sample Untreated

Finnally, let’s see the estimated ATT.

```
summary(r1)
```

```
##
## Estimate...  -0.01594
## AI SE......  0.043608
## T-stat.....  -0.36552
## p.val......  0.71472
##
## Original number of observations.............  67
## Original number of treated obs..............  15
## Matched number of observations..............  15
## Matched number of observations  (unweighted).  15
```

Matching on variables

Using the covariates to calculate the Mahalanobis distance for each observations

```
r2 <- Match(Y=bush04, Tr=etouch, X=data[,c(3:5, 15:17)])
summary(r2)
```

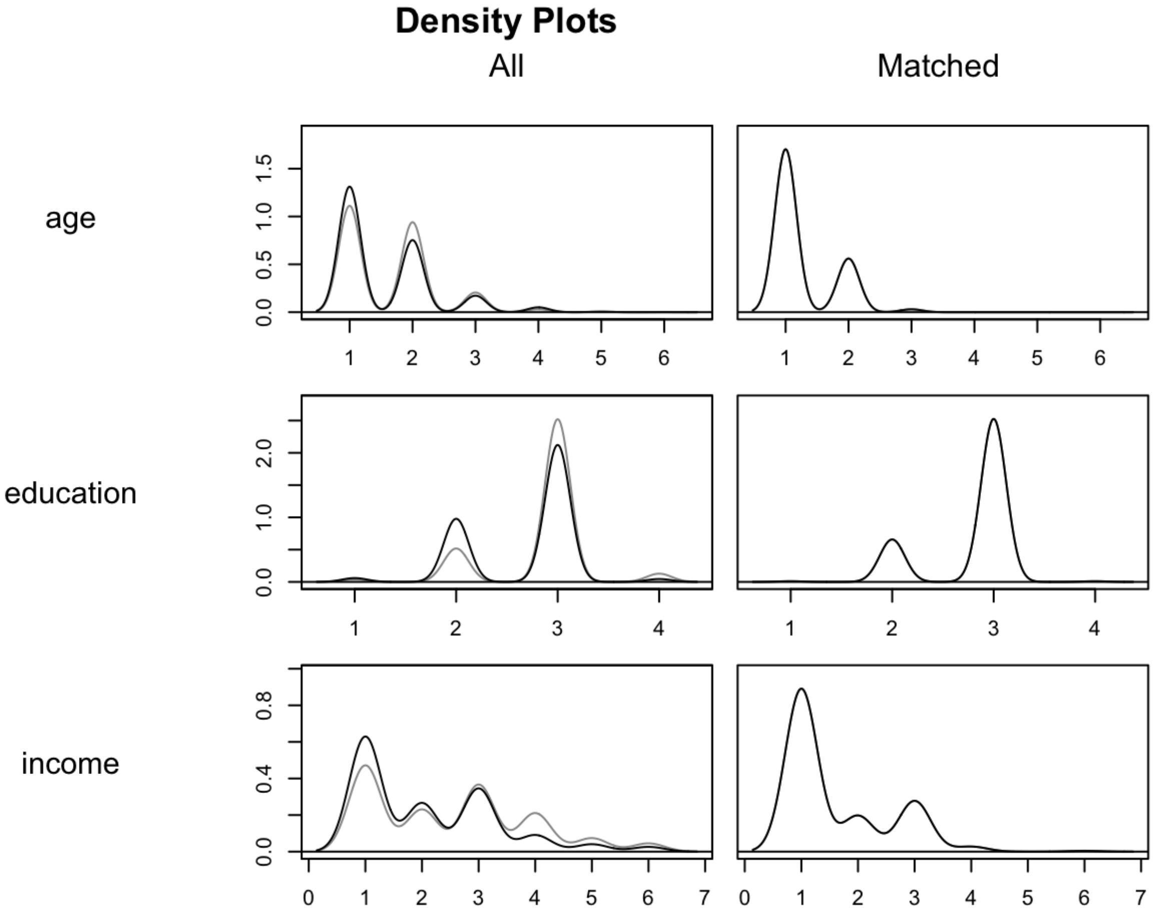
```
##
## Estimate...  -0.035793
## AI SE......  0.032596
## T-stat.....  -1.0981
## p.val......  0.27217
##
## Original number of observations.............  67
## Original number of treated obs..............  15
## Matched number of observations..............  15
## Matched number of observations  (unweighted).  15
```

Coarsened Exact Matching

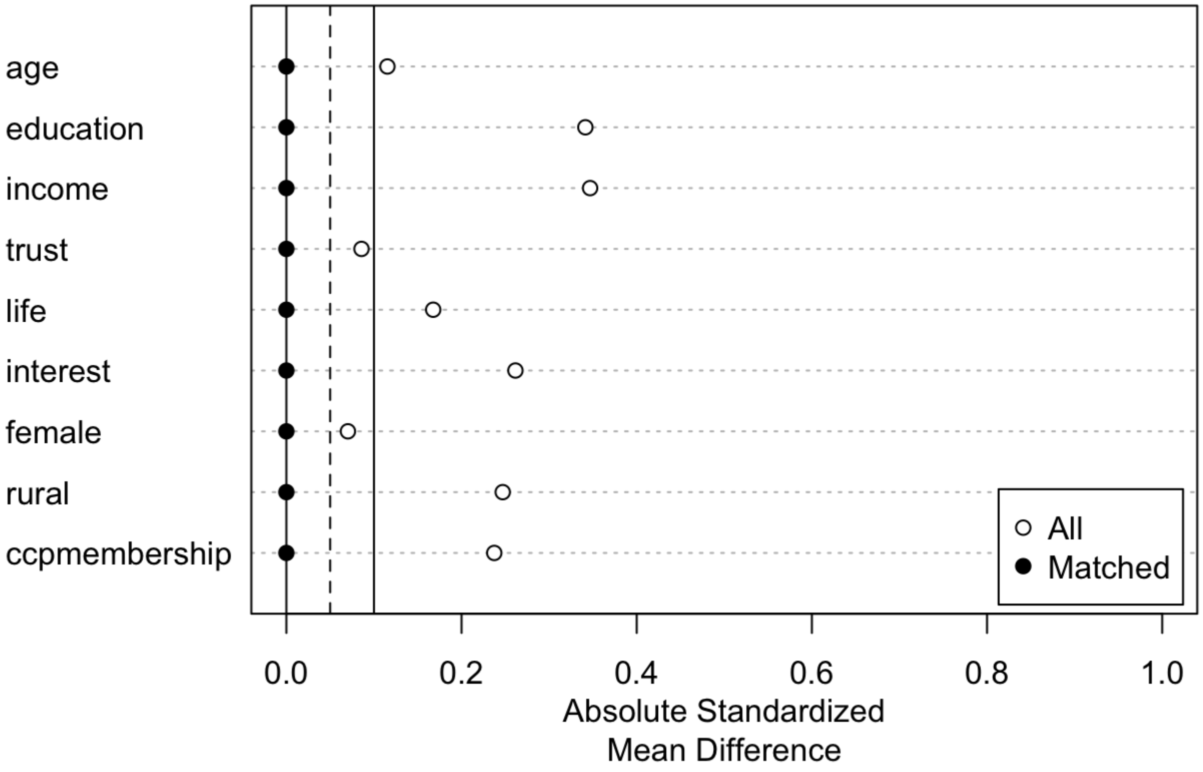
- Select covariates for matching (and the cutoff/threshold, if needed)
- Matching the treated and control groups based on the covarites
- Show balance
- Estimate the effects

In addition to Propensity Score Matching, another widely used method of matching is coarsened exact matching. We use package “MatchIt (https://kosukeimai.github.io/MatchIt/index.html)” to replicate results from Wang and Huang 2020, CPS, When “Fake News” Becomes Real: The Consequences of False Government Denials in an Authoritarian Country (https://doi.org/10.1177/0010414020957672).

```
library(haven)
library(MatchIt)
Data <- read_dta("cps-wave2_clean.dta")
cem1<- matchit(experience~age+education+income+trust+life+interest+female+rural+ccpmembership,method = 'cem',data=Data)
plot(cem1, type = "density", interactive = FALSE,which.xs = ~age+education+income)
```



```
plot(summary(cem1),xlim=c(0,1))
```



```
mdata<- match.data(cem1)
summary(lm(rumor~experience,data=mdata))
```

```
## 
## Call:
## lm(formula = rumor ~ experience, data = mdata)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -47.261 -27.006   2.994  27.994  52.994 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   47.006      1.727  27.215   <2e-16 ***
## experience     1.255      2.504   0.501    0.616    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 31.75 on 643 degrees of freedom
## Multiple R-squared:  0.0003904,  Adjusted R-squared:  -0.001164 
## F-statistic: 0.2512 on 1 and 643 DF,  p-value: 0.6164
```

Other resources

- Many studies now use matching as a robustness check and leave most details of matching process in the appendix. But some does use mataching as a main specification and published in top journal. E.g., Agerberg and Sohlberg, CPS 2021, Personal Proximity and Reactions to Terrorism (https://doi.org/10.1177/0010414021997162). (codes are based on Stata).
- The developers' tutorial (https://kosukeimai.github.io/MatchIt/index.html) for *MatchIt* (https://cran.r-project.org/web/packages/MatchIt/). Although PSM and CEM are widely used and accepted by many top journals, you can still play with other matching methods (sometimes for robustness). The key is to achieve the balance of observed covariates (so that we can be more confident that unobserved confounders are conceled out).
- Most studies use matching in cross-sectional data. For time-series cross-sectional data, see Imai, Kim, and Wang (2021) *PanelMatch* (https://cran.r-project.org/web/packages/PanelMatch/index.html).