293\_hw1\_v4

library(MatchIt)

## Warning: package 'MatchIt' was built under R version 3.3.2

library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

library(ggplot2)  
library(glmnet)

## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-5

# Remove the list   
rm(list = ls())

# load data  
setwd("/Users/xinli/Desktop/stanford/2017S/293MachineLearning/HW1/")  
dta<-read.csv("socialneighbor.csv")

## No.1 Preliminary Estimation

# Estimate average treatment effects(ATE) before beginning

avg\_treat = mean(dta$outcome\_voted[dta$treat\_neighbors==1])  
avg\_control = mean(dta$outcome\_voted[dta$treat\_neighbors==0])  
ate = abs(avg\_treat - avg\_control)  
ate # preliminary ATE

## [1] 0.08639696

# Estimate 95% confidence interval of the treated group

treat\_se = sd((dta$outcome\_voted[dta$treat\_neighbors==1]))/ sqrt(length((dta$outcome\_voted[dta$treat\_neighbors==1])))  
CI\_T\_plus= 1.96\*treat\_se + avg\_treat  
CI\_T\_minus= avg\_treat - 1.96\*treat\_se   
  
CI\_treatment = c(CI\_T\_minus,avg\_treat,CI\_T\_plus)  
CI\_treatment

## [1] 0.383539 0.390300 0.397061

treat\_se # standard error of the treatment estimation

## [1] 0.003449477

# Estimate 95% confidence interval of the control group

control\_se = sd((dta$outcome\_voted[dta$treat\_neighbors==0]))/ sqrt(length((dta$outcome\_voted[dta$treat\_neighbors==0])))  
CI\_C\_plus= 1.96\*control\_se + avg\_control  
CI\_C\_minus= avg\_control - 1.96\*control\_se   
  
CI\_control = c(CI\_C\_minus,avg\_control,CI\_C\_plus)  
CI\_control

## [1] 0.3010523 0.3039030 0.3067538

control\_se # standard error of the control estimation

## [1] 0.001454477

## No.2 Dropping observations using designed rules/ breaking into sampling data

# Format the data

# The variables "g2004" and "oneperhh" are dropped because they are always 1 in the dataset

working = dta  
covariates.names = names(dta)  
drops <- c("outcome\_voted","treat\_neighbors", "treatment\_dum", "treat\_hawthorne", "treat\_civic",  
 "treat\_self", "g2004", "oneperhh", "outcome\_voted.1", "treatment\_dum.1", "treat\_hawthorne.1",   
 "treat\_civic.1", "treat\_neighbors.1" ,"treat\_self.1")  
select <- !(names(working) %in% drops)  
covariate.names = covariates.names[select]  
  
names(working)[names(working)=="outcome\_voted"] <- "Y"

# Extract the dependent variable

Y <- working[["Y"]]

# The treatment is whether they received the "your neighbors are voting" letter

names(working)[names(working)=="treat\_neighbors"] <- "W"

# Extract treatment variable & covariates

W <- working[["W"]]  
covariates <- working[covariate.names]

# some algorithms require our covariates be scaled

# scale, with default settings, will calculate the mean and standard deviation of the entire vector,

# then "scale" each element by those values by subtracting the mean and dividing by the sd

covariates.scaled <- scale(covariates)  
processed.unscaled <- data.frame(Y, W, covariates)  
processed.scaled <- data.frame(Y, W, covariates.scaled)

# Creating functions

sumx = paste(covariate.names, collapse = " + ") # "X1 + X2 + X3 + ..." for substitution later  
interx = paste(" (",sumx, ")^2", sep="") # "(X1 + X2 + X3 + ...)^2" for substitution later

# W ~ X1 + X2 + X3 + ...

propscr\_formula <- paste("W",sumx, sep=" ~ ")  
propscr\_formula <- as.formula(propscr\_formula)  
propscr\_formula

## W ~ sex + yob + g2000 + g2002 + p2000 + p2002 + cluster + votedav +   
## dem + nov + aug + city + hh\_id + hh\_size + totalpopulation\_estimate +   
## percent\_male + percent\_female + median\_age + percent\_under5years +   
## percent\_5to9years + percent\_10to14years + percent\_15to19years +   
## percent\_20to24years + percent\_25to34years + percent\_35to44years +   
## percent\_45to54years + percent\_55to59years + percent\_60to64years +   
## percent\_65to74years + percent\_75to84years + percent\_85yearsandolder +   
## percent\_18yearsandolder + percent\_21yearsandover + percent\_62yearsandover +   
## percent\_65yearsandover + percent\_white + percent\_black +   
## percent\_amindian\_alaskan + percent\_asian + percent\_nativeandother +   
## percent\_other\_nativeandother + percent\_hispanicorlatino +   
## percent\_race\_other + median\_income + mean\_income + employ\_16 +   
## unemploy\_16 + unemploy\_20to64 + employ\_20to64 + employ\_rename\_20to64 +   
## hsorhigher + bach\_orhigher + less9thgrade + grade9to12 +   
## highschool + somecollege + assoc + bachelors + grad + randn +   
## p2004

# Y ~ X1 + X2 + X3 + ...

linearnotreat <- paste("Y",sumx, sep=" ~ ")  
linearnotreat <- as.formula(linearnotreat)  
linearnotreat

## Y ~ sex + yob + g2000 + g2002 + p2000 + p2002 + cluster + votedav +   
## dem + nov + aug + city + hh\_id + hh\_size + totalpopulation\_estimate +   
## percent\_male + percent\_female + median\_age + percent\_under5years +   
## percent\_5to9years + percent\_10to14years + percent\_15to19years +   
## percent\_20to24years + percent\_25to34years + percent\_35to44years +   
## percent\_45to54years + percent\_55to59years + percent\_60to64years +   
## percent\_65to74years + percent\_75to84years + percent\_85yearsandolder +   
## percent\_18yearsandolder + percent\_21yearsandover + percent\_62yearsandover +   
## percent\_65yearsandover + percent\_white + percent\_black +   
## percent\_amindian\_alaskan + percent\_asian + percent\_nativeandother +   
## percent\_other\_nativeandother + percent\_hispanicorlatino +   
## percent\_race\_other + median\_income + mean\_income + employ\_16 +   
## unemploy\_16 + unemploy\_20to64 + employ\_20to64 + employ\_rename\_20to64 +   
## hsorhigher + bach\_orhigher + less9thgrade + grade9to12 +   
## highschool + somecollege + assoc + bachelors + grad + randn +   
## p2004

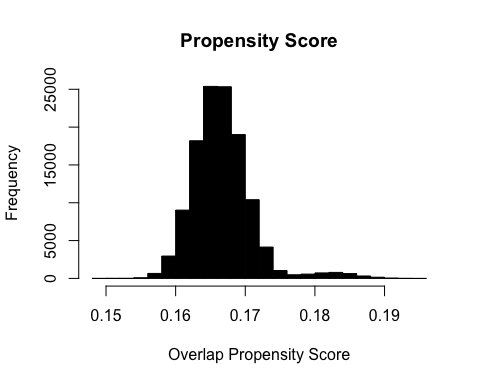
# Y ~ W + X1 + X2 + X3 + ...

linear <- paste("Y",paste("W",sumx, sep=" + "), sep=" ~ ")  
linear <- as.formula(linear)  
linear

## Y ~ W + sex + yob + g2000 + g2002 + p2000 + p2002 + cluster +   
## votedav + dem + nov + aug + city + hh\_id + hh\_size + totalpopulation\_estimate +   
## percent\_male + percent\_female + median\_age + percent\_under5years +   
## percent\_5to9years + percent\_10to14years + percent\_15to19years +   
## percent\_20to24years + percent\_25to34years + percent\_35to44years +   
## percent\_45to54years + percent\_55to59years + percent\_60to64years +   
## percent\_65to74years + percent\_75to84years + percent\_85yearsandolder +   
## percent\_18yearsandolder + percent\_21yearsandover + percent\_62yearsandover +   
## percent\_65yearsandover + percent\_white + percent\_black +   
## percent\_amindian\_alaskan + percent\_asian + percent\_nativeandother +   
## percent\_other\_nativeandother + percent\_hispanicorlatino +   
## percent\_race\_other + median\_income + mean\_income + employ\_16 +   
## unemploy\_16 + unemploy\_20to64 + employ\_20to64 + employ\_rename\_20to64 +   
## hsorhigher + bach\_orhigher + less9thgrade + grade9to12 +   
## highschool + somecollege + assoc + bachelors + grad + randn +   
## p2004

# inital propensity scores

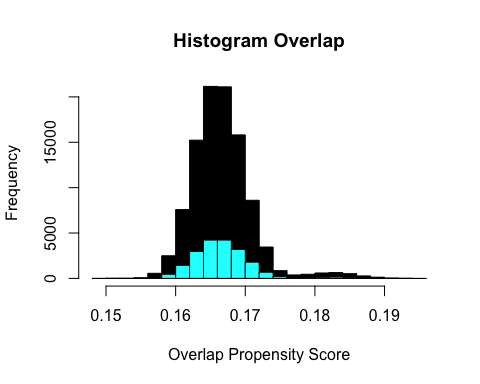
prop\_score = glm(propscr\_formula ,data = processed.scaled, family = "binomial")  
prs\_df <- data.frame(pr\_score = predict(prop\_score, type = "response"),  
 W = prop\_score$model$W)  
overlap = prs\_df[prs\_df$pr\_score <= 0.975 & prs\_df$pr\_score >= 0.025, ]  
hist(overlap$pr\_score, col = 1, xlab = "Overlap Propensity Score", main = "Propensity Score")



print(mean(overlap$pr\_score))

## [1] 0.1666681

data\_overlap = processed.scaled[prs\_df$pr\_score <= 0.975 & prs\_df$pr\_score >= 0.025, ]  
prop\_treated = overlap[overlap$W == 1, ]  
prop\_untreated = overlap[overlap$W == 0, ]  
hist(prop\_untreated$pr\_score, col = 1, xlab = "Overlap Propensity Score", main = "Histogram Overlap")  
hist(prop\_treated$pr\_score, col = 5,add = T)

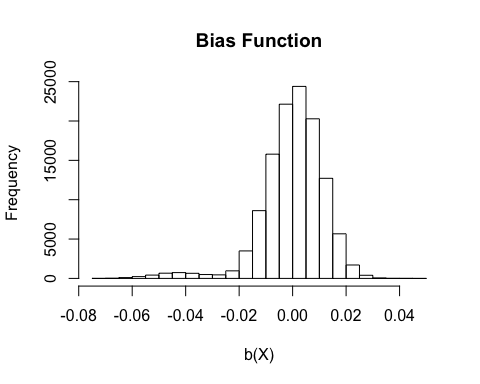


# conditional mean of potential outcomes

cond\_mean = glm(linearnotreat, data = data\_overlap, family = "binomial")  
condmean\_df <- data.frame(cond\_mean = predict(cond\_mean, type = "response"),  
 Y = cond\_mean$model$Y)

# bias function

p = mean(data\_overlap$W)  
mu\_0 = mean(data\_overlap$Y[data\_overlap$W == 0])  
mu\_1 = mean(data\_overlap$Y[data\_overlap$W == 1])  
mu\_0x = mu\_0\*condmean\_df$cond\_mean/sum(condmean\_df$cond\_mean)  
mu\_1x = mu\_1\*condmean\_df$cond\_mean/sum(condmean\_df$cond\_mean)  
  
bias\_fxn = (overlap$pr\_score - p)/(p\*(1-p)) \* (p\*(mu\_0x - mu\_0) + (1 - p)\*(mu\_1x - mu\_1))  
hist(bias\_fxn, xlab = "b(X)", main = "Bias Function")



### Dropping data

# prelim\_model = glm(linear,data = processed.scaled, family = "binomial")

# Covariates chosen based on the ones with high coefficients in the prelim\_model

drop\_confounders = (working$percent\_female >= 52 | working$percent\_male >= 50 |   
 working$cluster >= 5000 | working$bachelors >= 30 | working$grad >= 10) & (W == 1)  
data\_new = processed.scaled[!drop\_confounders, ]

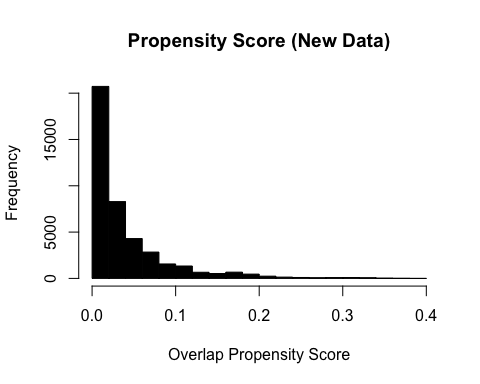
# New point estimate of ATE

avg\_treat\_n = mean(data\_new$Y[data\_new$W==1])  
avg\_control\_n = mean(data\_new$Y[data\_new$W==0])  
ate\_n = abs(avg\_treat\_n - avg\_control\_n)  
ate\_n

## [1] 0.1311655

# New propensity scores

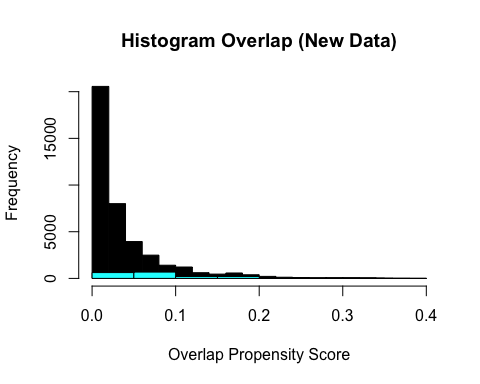
prop\_score\_n = glm(propscr\_formula, data = data\_new, family = "binomial")  
prs\_df\_n <- data.frame(pr\_score = predict(prop\_score\_n, type = "response"),  
 W = prop\_score\_n$model$W)  
overlap\_n = prs\_df\_n[prs\_df\_n$pr\_score <= 0.975 & prs\_df\_n$pr\_score >= 0.005, ]  
hist(overlap\_n$pr\_score, col = 1, xlab = "Overlap Propensity Score", main = "Propensity Score (New Data)")



print(mean(overlap\_n$pr\_score))

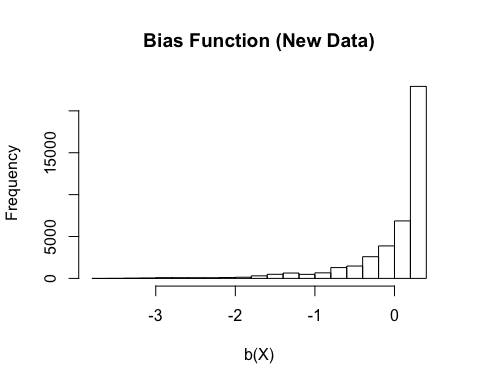
## [1] 0.0411822

data\_new = data\_new[prs\_df\_n$pr\_score <= 0.975 & prs\_df\_n$pr\_score >= 0.005, ]  
prop\_treated\_n = overlap\_n[overlap\_n$W == 1, ]  
prop\_untreated\_n = overlap\_n[overlap\_n$W == 0, ]  
hist(prop\_untreated\_n$pr\_score, col = 1, xlab = "Overlap Propensity Score", main = "Histogram Overlap (New Data)")  
hist(prop\_treated\_n$pr\_score, col = 5,add = T)



# New bias function

cond\_mean\_n = glm(linearnotreat ,data = data\_new, family = "binomial")  
condmean\_df\_n <- data.frame(cond\_mean = predict(cond\_mean\_n, type = "response"),  
 Y = cond\_mean\_n$model$Y)  
  
p = mean(data\_new$W)  
mu\_0 = mean(data\_new$Y[data\_new$W == 0])  
mu\_1 = mean(data\_new$Y[data\_new$W == 1])  
mu\_0x = mu\_0\*condmean\_df\_n$cond\_mean/sum(condmean\_df\_n$cond\_mean)  
mu\_1x = mu\_1\*condmean\_df\_n$cond\_mean/sum(condmean\_df\_n$cond\_mean)  
  
bias\_fxn\_n = (overlap\_n$pr\_score - p)/(p\*(1-p)) \* (p\*(mu\_0x - mu\_0) + (1 - p)\*(mu\_1x - mu\_1))  
hist(bias\_fxn\_n, xlab = "b(X)", main = "Bias Function (New Data)")



#summary(bias\_fxn\_n)

## No.3 In the modified dataset, testing traditional methods for estimating the ATE:

# Redefining the dataset

W = data\_new$W  
Y = data\_new$Y  
save(data\_new, overlap\_n, overlap, W, Y, linear, linearnotreat, propscr\_formula,   
 bias\_fxn, bias\_fxn\_n, ate\_n, ate, file="new\_dataset")  
rm(list = ls())  
load("new\_dataset")

# 3.1 a) Propensity Score weighting

psw = overlap\_n$pr\_score  
psw[W == 1] = (1/overlap\_n$pr\_score[W==1])  
psw[W == 0] = (1/(1-overlap\_n$pr\_score[W==0]))  
psw\_model = lm(linear, data=data\_new, weights=psw)  
coef(psw\_model)["W"] # ate\_psw #ATE is the coefficient of W

## W   
## 0.1104609

summary(psw\_model)

##   
## Call:  
## lm(formula = linear, data = data\_new, weights = psw)  
##   
## Weighted Residuals:  
## Min 1Q Median 3Q Max   
## -11.6948 -0.3593 -0.2007 0.5116 8.6373   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3439208 0.0085591 40.182 < 2e-16 \*\*\*  
## W 0.1104609 0.0047688 23.163 < 2e-16 \*\*\*  
## sex 0.0035778 0.0022048 1.623 0.104651   
## yob 0.0078288 0.0024462 3.200 0.001374 \*\*   
## g2000 -0.0032864 0.0024885 -1.321 0.186622   
## g2002 -0.0137420 0.0034078 -4.032 5.53e-05 \*\*\*  
## p2000 0.0272529 0.0024565 11.094 < 2e-16 \*\*\*  
## p2002 0.0321101 0.0027704 11.591 < 2e-16 \*\*\*  
## cluster 9.9765724 21.9825873 0.454 0.649946   
## votedav 0.0989013 0.0021760 45.451 < 2e-16 \*\*\*  
## dem -0.0067959 0.0022705 -2.993 0.002763 \*\*   
## nov 0.0637412 0.0037770 16.876 < 2e-16 \*\*\*  
## aug 0.0553982 0.0039725 13.946 < 2e-16 \*\*\*  
## city -0.0004668 0.0028109 -0.166 0.868118   
## hh\_id -9.9818026 21.9825185 -0.454 0.649774   
## hh\_size -0.0064301 0.0024493 -2.625 0.008661 \*\*   
## totalpopulation\_estimate -0.0024833 0.0052363 -0.474 0.635328   
## percent\_male -0.0078079 0.0046403 -1.683 0.092456 .   
## percent\_female NA NA NA NA   
## median\_age -0.0150458 0.0205355 -0.733 0.463763   
## percent\_under5years 0.1735736 0.0385397 4.504 6.69e-06 \*\*\*  
## percent\_5to9years 0.1667305 0.0380587 4.381 1.18e-05 \*\*\*  
## percent\_10to14years 0.1815740 0.0380857 4.768 1.87e-06 \*\*\*  
## percent\_15to19years 0.2244384 0.0453982 4.944 7.69e-07 \*\*\*  
## percent\_20to24years 0.2901008 0.0767494 3.780 0.000157 \*\*\*  
## percent\_25to34years 0.3056577 0.0735689 4.155 3.26e-05 \*\*\*  
## percent\_35to44years 0.2020628 0.0556173 3.633 0.000280 \*\*\*  
## percent\_45to54years 0.2323161 0.0584686 3.973 7.10e-05 \*\*\*  
## percent\_55to59years 0.1437101 0.0356631 4.030 5.60e-05 \*\*\*  
## percent\_60to64years 0.1648947 0.0390645 4.221 2.44e-05 \*\*\*  
## percent\_65to74years 0.2055611 0.0997843 2.060 0.039399 \*   
## percent\_75to84years 0.1346264 0.0718140 1.875 0.060847 .   
## percent\_85yearsandolder 0.0879909 0.0409606 2.148 0.031705 \*   
## percent\_18yearsandolder 0.2018413 0.0168009 12.014 < 2e-16 \*\*\*  
## percent\_21yearsandover -0.0897308 0.0233399 -3.845 0.000121 \*\*\*  
## percent\_62yearsandover -0.0928004 0.0307309 -3.020 0.002531 \*\*   
## percent\_65yearsandover 0.0836431 0.2029097 0.412 0.680182   
## percent\_white 0.0524518 0.0294563 1.781 0.074974 .   
## percent\_black 0.0437486 0.0218788 2.000 0.045551 \*   
## percent\_amindian\_alaskan 0.0270297 0.0059812 4.519 6.23e-06 \*\*\*  
## percent\_asian -0.0259966 0.0144064 -1.805 0.071157 .   
## percent\_nativeandother 0.0027566 0.0027245 1.012 0.311644   
## percent\_other\_nativeandother 0.0438467 0.0038912 11.268 < 2e-16 \*\*\*  
## percent\_hispanicorlatino 0.0270501 0.0073538 3.678 0.000235 \*\*\*  
## percent\_race\_other -0.0722351 0.0102390 -7.055 1.75e-12 \*\*\*  
## median\_income 0.0213751 0.0174778 1.223 0.221342   
## mean\_income -0.0203875 0.0254090 -0.802 0.422342   
## employ\_16 -0.1233365 0.0221910 -5.558 2.75e-08 \*\*\*  
## unemploy\_16 -0.0723103 0.0147184 -4.913 9.01e-07 \*\*\*  
## unemploy\_20to64 0.0379297 0.0140922 2.692 0.007115 \*\*   
## employ\_20to64 0.0511395 0.0175334 2.917 0.003540 \*\*   
## employ\_rename\_20to64 -0.0486188 0.0054071 -8.992 < 2e-16 \*\*\*  
## hsorhigher -0.5521551 0.2566383 -2.151 0.031443 \*   
## bach\_orhigher 1.7510584 0.8865240 1.975 0.048252 \*   
## less9thgrade 0.3283036 0.0900636 3.645 0.000267 \*\*\*  
## grade9to12 0.4817695 0.1478123 3.259 0.001118 \*\*   
## highschool 2.7354941 0.4271044 6.405 1.52e-10 \*\*\*  
## somecollege 1.1062172 0.1717248 6.442 1.19e-10 \*\*\*  
## assoc 0.5749625 0.0894367 6.429 1.30e-10 \*\*\*  
## bachelors 1.4520884 0.4024767 3.608 0.000309 \*\*\*  
## grad 1.3357719 0.3862822 3.458 0.000545 \*\*\*  
## randn 0.0144257 0.0022834 6.318 2.68e-10 \*\*\*  
## p2004 0.0519879 0.0028527 18.224 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6314 on 42206 degrees of freedom  
## Multiple R-squared: 0.174, Adjusted R-squared: 0.1728   
## F-statistic: 145.8 on 61 and 42206 DF, p-value: < 2.2e-16

# 3.1 b) Direct regression analysis

reg\_model = lm(linear, data=data\_new)  
coef(reg\_model)["W"] # ate\_ols

## W   
## 0.09681301

summary(reg\_model)

##   
## Call:  
## lm(formula = linear, data = data\_new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.8591 -0.3319 -0.1947 0.4994 1.0945   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.442e-01 6.817e-03 50.494 < 2e-16 \*\*\*  
## W 9.681e-02 1.068e-02 9.062 < 2e-16 \*\*\*  
## sex -5.363e-03 2.130e-03 -2.518 0.011814 \*   
## yob 2.843e-03 2.421e-03 1.174 0.240357   
## g2000 -1.263e-02 2.327e-03 -5.426 5.78e-08 \*\*\*  
## g2002 -4.273e-03 3.331e-03 -1.283 0.199547   
## p2000 1.665e-02 2.365e-03 7.043 1.91e-12 \*\*\*  
## p2002 2.855e-02 2.699e-03 10.578 < 2e-16 \*\*\*  
## cluster -1.045e+01 2.122e+01 -0.493 0.622208   
## votedav 1.101e-01 2.054e-03 53.612 < 2e-16 \*\*\*  
## dem -4.793e-03 2.210e-03 -2.168 0.030151 \*   
## nov 2.762e-02 3.680e-03 7.507 6.16e-14 \*\*\*  
## aug 7.623e-02 3.814e-03 19.988 < 2e-16 \*\*\*  
## city 1.070e-02 2.711e-03 3.948 7.91e-05 \*\*\*  
## hh\_id 1.043e+01 2.122e+01 0.492 0.622989   
## hh\_size 3.285e-03 2.394e-03 1.372 0.169975   
## totalpopulation\_estimate -1.189e-02 4.559e-03 -2.609 0.009091 \*\*   
## percent\_male -4.656e-03 3.646e-03 -1.277 0.201543   
## percent\_female NA NA NA NA   
## median\_age -3.100e-02 1.859e-02 -1.668 0.095345 .   
## percent\_under5years 5.451e-03 3.526e-02 0.155 0.877130   
## percent\_5to9years 3.087e-02 3.509e-02 0.880 0.379048   
## percent\_10to14years 2.459e-02 3.493e-02 0.704 0.481400   
## percent\_15to19years 2.223e-02 4.178e-02 0.532 0.594676   
## percent\_20to24years -1.395e-03 6.947e-02 -0.020 0.983976   
## percent\_25to34years -1.957e-02 6.743e-02 -0.290 0.771588   
## percent\_35to44years -1.886e-02 5.118e-02 -0.368 0.712533   
## percent\_45to54years -1.195e-02 5.351e-02 -0.223 0.823285   
## percent\_55to59years -6.636e-03 3.253e-02 -0.204 0.838373   
## percent\_60to64years -1.205e-02 3.542e-02 -0.340 0.733812   
## percent\_65to74years 6.097e-01 9.424e-02 6.470 9.91e-11 \*\*\*  
## percent\_75to84years 4.579e-01 6.802e-02 6.732 1.69e-11 \*\*\*  
## percent\_85yearsandolder 2.489e-01 3.850e-02 6.467 1.01e-10 \*\*\*  
## percent\_18yearsandolder 1.279e-01 1.601e-02 7.986 1.43e-15 \*\*\*  
## percent\_21yearsandover -3.255e-02 2.185e-02 -1.490 0.136322   
## percent\_62yearsandover -5.354e-02 2.635e-02 -2.031 0.042224 \*   
## percent\_65yearsandover -1.115e+00 1.934e-01 -5.767 8.14e-09 \*\*\*  
## percent\_white 2.121e-02 2.522e-02 0.841 0.400421   
## percent\_black 1.333e-02 1.831e-02 0.728 0.466636   
## percent\_amindian\_alaskan 6.717e-04 4.933e-03 0.136 0.891700   
## percent\_asian 4.812e-03 1.279e-02 0.376 0.706713   
## percent\_nativeandother -1.641e-03 2.555e-03 -0.642 0.520573   
## percent\_other\_nativeandother 2.541e-02 3.761e-03 6.757 1.43e-11 \*\*\*  
## percent\_hispanicorlatino 2.431e-02 6.494e-03 3.743 0.000182 \*\*\*  
## percent\_race\_other -6.040e-03 9.256e-03 -0.653 0.514051   
## median\_income -1.267e-02 1.533e-02 -0.826 0.408544   
## mean\_income -2.649e-04 2.167e-02 -0.012 0.990246   
## employ\_16 -1.279e-01 1.916e-02 -6.676 2.48e-11 \*\*\*  
## unemploy\_16 -1.481e-01 1.269e-02 -11.666 < 2e-16 \*\*\*  
## unemploy\_20to64 1.125e-01 1.218e-02 9.239 < 2e-16 \*\*\*  
## employ\_20to64 5.626e-02 1.600e-02 3.516 0.000439 \*\*\*  
## employ\_rename\_20to64 -3.252e-02 5.023e-03 -6.475 9.60e-11 \*\*\*  
## hsorhigher 3.536e-02 2.323e-01 0.152 0.878995   
## bach\_orhigher -3.200e+00 8.011e-01 -3.994 6.51e-05 \*\*\*  
## less9thgrade -1.881e-01 8.238e-02 -2.284 0.022385 \*   
## grade9to12 -3.085e-01 1.346e-01 -2.293 0.021880 \*   
## highschool -9.623e-01 3.668e-01 -2.623 0.008708 \*\*   
## somecollege -3.988e-01 1.476e-01 -2.701 0.006917 \*\*   
## assoc -2.093e-01 7.691e-02 -2.721 0.006514 \*\*   
## bachelors 9.045e-01 3.744e-01 2.416 0.015714 \*   
## grad 8.020e-01 3.603e-01 2.226 0.026040 \*   
## randn 3.531e-04 2.232e-03 0.158 0.874287   
## p2004 3.747e-02 2.774e-03 13.507 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4353 on 42206 degrees of freedom  
## Multiple R-squared: 0.157, Adjusted R-squared: 0.1557   
## F-statistic: 128.8 on 61 and 42206 DF, p-value: < 2.2e-16

# 3.1 c) Double robust analysis weighting with inverse propensity scores

#Take probabilities  
prop\_matrix <- model.matrix(propscr\_formula, data\_new)[,-1]  
prop\_score\_dbl\_rbst = glm(W~prop\_matrix, family = 'binomial')  
  
#apply inverse  
psw\_double = 1/prop\_score\_dbl\_rbst$fitted.values  
  
#apply weights  
psw\_double[W == 1] = (1/prop\_score\_dbl\_rbst$fitted.values[W==1])  
psw\_double[W == 0] = (1/(1-prop\_score\_dbl\_rbst$fitted.values[W==0]))  
  
  
# estimate robust estimator  
double\_model = lm(linear, data=data\_new, weights =psw )  
coef(double\_model)["W"] # ate\_double robust

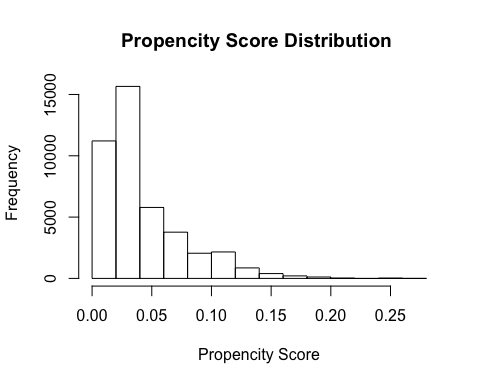
## W   
## 0.1104609

summary(double\_model)

##   
## Call:  
## lm(formula = linear, data = data\_new, weights = psw)  
##   
## Weighted Residuals:  
## Min 1Q Median 3Q Max   
## -11.6948 -0.3593 -0.2007 0.5116 8.6373   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3439208 0.0085591 40.182 < 2e-16 \*\*\*  
## W 0.1104609 0.0047688 23.163 < 2e-16 \*\*\*  
## sex 0.0035778 0.0022048 1.623 0.104651   
## yob 0.0078288 0.0024462 3.200 0.001374 \*\*   
## g2000 -0.0032864 0.0024885 -1.321 0.186622   
## g2002 -0.0137420 0.0034078 -4.032 5.53e-05 \*\*\*  
## p2000 0.0272529 0.0024565 11.094 < 2e-16 \*\*\*  
## p2002 0.0321101 0.0027704 11.591 < 2e-16 \*\*\*  
## cluster 9.9765724 21.9825873 0.454 0.649946   
## votedav 0.0989013 0.0021760 45.451 < 2e-16 \*\*\*  
## dem -0.0067959 0.0022705 -2.993 0.002763 \*\*   
## nov 0.0637412 0.0037770 16.876 < 2e-16 \*\*\*  
## aug 0.0553982 0.0039725 13.946 < 2e-16 \*\*\*  
## city -0.0004668 0.0028109 -0.166 0.868118   
## hh\_id -9.9818026 21.9825185 -0.454 0.649774   
## hh\_size -0.0064301 0.0024493 -2.625 0.008661 \*\*   
## totalpopulation\_estimate -0.0024833 0.0052363 -0.474 0.635328   
## percent\_male -0.0078079 0.0046403 -1.683 0.092456 .   
## percent\_female NA NA NA NA   
## median\_age -0.0150458 0.0205355 -0.733 0.463763   
## percent\_under5years 0.1735736 0.0385397 4.504 6.69e-06 \*\*\*  
## percent\_5to9years 0.1667305 0.0380587 4.381 1.18e-05 \*\*\*  
## percent\_10to14years 0.1815740 0.0380857 4.768 1.87e-06 \*\*\*  
## percent\_15to19years 0.2244384 0.0453982 4.944 7.69e-07 \*\*\*  
## percent\_20to24years 0.2901008 0.0767494 3.780 0.000157 \*\*\*  
## percent\_25to34years 0.3056577 0.0735689 4.155 3.26e-05 \*\*\*  
## percent\_35to44years 0.2020628 0.0556173 3.633 0.000280 \*\*\*  
## percent\_45to54years 0.2323161 0.0584686 3.973 7.10e-05 \*\*\*  
## percent\_55to59years 0.1437101 0.0356631 4.030 5.60e-05 \*\*\*  
## percent\_60to64years 0.1648947 0.0390645 4.221 2.44e-05 \*\*\*  
## percent\_65to74years 0.2055611 0.0997843 2.060 0.039399 \*   
## percent\_75to84years 0.1346264 0.0718140 1.875 0.060847 .   
## percent\_85yearsandolder 0.0879909 0.0409606 2.148 0.031705 \*   
## percent\_18yearsandolder 0.2018413 0.0168009 12.014 < 2e-16 \*\*\*  
## percent\_21yearsandover -0.0897308 0.0233399 -3.845 0.000121 \*\*\*  
## percent\_62yearsandover -0.0928004 0.0307309 -3.020 0.002531 \*\*   
## percent\_65yearsandover 0.0836431 0.2029097 0.412 0.680182   
## percent\_white 0.0524518 0.0294563 1.781 0.074974 .   
## percent\_black 0.0437486 0.0218788 2.000 0.045551 \*   
## percent\_amindian\_alaskan 0.0270297 0.0059812 4.519 6.23e-06 \*\*\*  
## percent\_asian -0.0259966 0.0144064 -1.805 0.071157 .   
## percent\_nativeandother 0.0027566 0.0027245 1.012 0.311644   
## percent\_other\_nativeandother 0.0438467 0.0038912 11.268 < 2e-16 \*\*\*  
## percent\_hispanicorlatino 0.0270501 0.0073538 3.678 0.000235 \*\*\*  
## percent\_race\_other -0.0722351 0.0102390 -7.055 1.75e-12 \*\*\*  
## median\_income 0.0213751 0.0174778 1.223 0.221342   
## mean\_income -0.0203875 0.0254090 -0.802 0.422342   
## employ\_16 -0.1233365 0.0221910 -5.558 2.75e-08 \*\*\*  
## unemploy\_16 -0.0723103 0.0147184 -4.913 9.01e-07 \*\*\*  
## unemploy\_20to64 0.0379297 0.0140922 2.692 0.007115 \*\*   
## employ\_20to64 0.0511395 0.0175334 2.917 0.003540 \*\*   
## employ\_rename\_20to64 -0.0486188 0.0054071 -8.992 < 2e-16 \*\*\*  
## hsorhigher -0.5521551 0.2566383 -2.151 0.031443 \*   
## bach\_orhigher 1.7510584 0.8865240 1.975 0.048252 \*   
## less9thgrade 0.3283036 0.0900636 3.645 0.000267 \*\*\*  
## grade9to12 0.4817695 0.1478123 3.259 0.001118 \*\*   
## highschool 2.7354941 0.4271044 6.405 1.52e-10 \*\*\*  
## somecollege 1.1062172 0.1717248 6.442 1.19e-10 \*\*\*  
## assoc 0.5749625 0.0894367 6.429 1.30e-10 \*\*\*  
## bachelors 1.4520884 0.4024767 3.608 0.000309 \*\*\*  
## grad 1.3357719 0.3862822 3.458 0.000545 \*\*\*  
## randn 0.0144257 0.0022834 6.318 2.68e-10 \*\*\*  
## p2004 0.0519879 0.0028527 18.224 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6314 on 42206 degrees of freedom  
## Multiple R-squared: 0.174, Adjusted R-squared: 0.1728   
## F-statistic: 145.8 on 61 and 42206 DF, p-value: < 2.2e-16

# 3.2 Estimate propencity score using LASSO

prop\_matrix <- model.matrix(propscr\_formula, data\_new)[,-1]  
propscr\_lasso = glmnet(prop\_matrix, W, alpha = 1, family = 'binomial')  
prs\_lasso = predict(propscr\_lasso, newx=as.matrix(data\_new[, -(1:2)]),   
 s = 0.001, type = "response")  
hist(prs\_lasso, xlab = "Propencity Score", main = "Propencity Score Distribution ")



# Propensity Score weighting

prs\_lasso[W == 1] = (1/prs\_lasso[W==1])  
prs\_lasso[W == 0] = (1/(1-prs\_lasso[W==0]))  
psw\_lasso\_model = lm(linear, data=data\_new, weights=prs\_lasso)  
coef(psw\_lasso\_model)["W"] # ate\_psw

## W   
## 0.1021995

summary(psw\_lasso\_model)

##   
## Call:  
## lm(formula = linear, data = data\_new, weights = prs\_lasso)  
##   
## Weighted Residuals:  
## Min 1Q Median 3Q Max   
## -5.1205 -0.3521 -0.2022 0.5194 5.3569   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.340801 0.008136 41.886 < 2e-16 \*\*\*  
## W 0.102200 0.004734 21.588 < 2e-16 \*\*\*  
## sex -0.002362 0.002190 -1.079 0.280631   
## yob 0.006380 0.002460 2.593 0.009511 \*\*   
## g2000 -0.009911 0.002439 -4.063 4.86e-05 \*\*\*  
## g2002 -0.014578 0.003364 -4.334 1.47e-05 \*\*\*  
## p2000 0.024126 0.002437 9.899 < 2e-16 \*\*\*  
## p2002 0.037726 0.002756 13.690 < 2e-16 \*\*\*  
## cluster 3.693153 21.833808 0.169 0.865681   
## votedav 0.104206 0.002091 49.843 < 2e-16 \*\*\*  
## dem -0.008560 0.002251 -3.802 0.000144 \*\*\*  
## nov 0.052658 0.003763 13.994 < 2e-16 \*\*\*  
## aug 0.057853 0.003926 14.737 < 2e-16 \*\*\*  
## city 0.008072 0.002756 2.929 0.003399 \*\*   
## hh\_id -3.704227 21.833769 -0.170 0.865282   
## hh\_size -0.001974 0.002432 -0.812 0.416921   
## totalpopulation\_estimate -0.003669 0.005021 -0.731 0.464883   
## percent\_male -0.003093 0.004478 -0.691 0.489707   
## percent\_female NA NA NA NA   
## median\_age -0.005038 0.020069 -0.251 0.801808   
## percent\_under5years 0.094084 0.037813 2.488 0.012845 \*   
## percent\_5to9years 0.093858 0.037464 2.505 0.012240 \*   
## percent\_10to14years 0.103198 0.037551 2.748 0.005994 \*\*   
## percent\_15to19years 0.130358 0.044689 2.917 0.003536 \*\*   
## percent\_20to24years 0.169940 0.075264 2.258 0.023956 \*   
## percent\_25to34years 0.186538 0.072354 2.578 0.009937 \*\*   
## percent\_35to44years 0.116576 0.054631 2.134 0.032858 \*   
## percent\_45to54years 0.132412 0.057544 2.301 0.021394 \*   
## percent\_55to59years 0.084957 0.035174 2.415 0.015725 \*   
## percent\_60to64years 0.094954 0.038285 2.480 0.013136 \*   
## percent\_65to74years 0.501996 0.096765 5.188 2.14e-07 \*\*\*  
## percent\_75to84years 0.367974 0.069576 5.289 1.24e-07 \*\*\*  
## percent\_85yearsandolder 0.209501 0.039660 5.282 1.28e-07 \*\*\*  
## percent\_18yearsandolder 0.144341 0.016371 8.817 < 2e-16 \*\*\*  
## percent\_21yearsandover -0.090478 0.023147 -3.909 9.29e-05 \*\*\*  
## percent\_62yearsandover -0.083448 0.030185 -2.765 0.005703 \*\*   
## percent\_65yearsandover -0.636238 0.201070 -3.164 0.001556 \*\*   
## percent\_white 0.051009 0.028513 1.789 0.073625 .   
## percent\_black 0.043692 0.021141 2.067 0.038768 \*   
## percent\_amindian\_alaskan 0.025766 0.005658 4.554 5.28e-06 \*\*\*  
## percent\_asian -0.014082 0.014284 -0.986 0.324223   
## percent\_nativeandother 0.003119 0.002655 1.174 0.240224   
## percent\_other\_nativeandother 0.038209 0.003926 9.733 < 2e-16 \*\*\*  
## percent\_hispanicorlatino 0.021067 0.007250 2.906 0.003663 \*\*   
## percent\_race\_other -0.033126 0.010202 -3.247 0.001168 \*\*   
## median\_income 0.022491 0.016675 1.349 0.177409   
## mean\_income -0.012442 0.024215 -0.514 0.607370   
## employ\_16 -0.139900 0.021662 -6.458 1.07e-10 \*\*\*  
## unemploy\_16 -0.103911 0.014266 -7.284 3.30e-13 \*\*\*  
## unemploy\_20to64 0.068132 0.013724 4.965 6.91e-07 \*\*\*  
## employ\_20to64 0.064618 0.017441 3.705 0.000212 \*\*\*  
## employ\_rename\_20to64 -0.038166 0.005288 -7.218 5.37e-13 \*\*\*  
## hsorhigher 0.230146 0.255426 0.901 0.367578   
## bach\_orhigher 1.553935 0.865381 1.796 0.072555 .   
## less9thgrade 0.380165 0.090904 4.182 2.89e-05 \*\*\*  
## grade9to12 0.599931 0.148779 4.032 5.53e-05 \*\*\*  
## highschool 1.415208 0.412318 3.432 0.000599 \*\*\*  
## somecollege 0.571565 0.165799 3.447 0.000567 \*\*\*  
## assoc 0.292881 0.086378 3.391 0.000698 \*\*\*  
## bachelors 0.423915 0.390783 1.085 0.278024   
## grad 0.340895 0.374729 0.910 0.362981   
## randn 0.006599 0.002288 2.884 0.003932 \*\*   
## p2004 0.050878 0.002858 17.799 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6002 on 42206 degrees of freedom  
## Multiple R-squared: 0.1602, Adjusted R-squared: 0.159   
## F-statistic: 132 on 61 and 42206 DF, p-value: < 2.2e-16

# Double Robust

#apply inverse  
psw\_double = 1/prs\_lasso  
  
#apply weights  
psw\_double[W == 1] = (1/prs\_lasso[W==1])  
psw\_double[W == 0] = (1/(1-prs\_lasso[W==0]))  
  
  
# estimate robust estimator  
double\_model = lm(linear, data=data\_new, weights =psw )  
coef(double\_model)["W"] # ate\_double robust

## W   
## 0.1104609

summary(double\_model)

##   
## Call:  
## lm(formula = linear, data = data\_new, weights = psw)  
##   
## Weighted Residuals:  
## Min 1Q Median 3Q Max   
## -11.6948 -0.3593 -0.2007 0.5116 8.6373   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3439208 0.0085591 40.182 < 2e-16 \*\*\*  
## W 0.1104609 0.0047688 23.163 < 2e-16 \*\*\*  
## sex 0.0035778 0.0022048 1.623 0.104651   
## yob 0.0078288 0.0024462 3.200 0.001374 \*\*   
## g2000 -0.0032864 0.0024885 -1.321 0.186622   
## g2002 -0.0137420 0.0034078 -4.032 5.53e-05 \*\*\*  
## p2000 0.0272529 0.0024565 11.094 < 2e-16 \*\*\*  
## p2002 0.0321101 0.0027704 11.591 < 2e-16 \*\*\*  
## cluster 9.9765724 21.9825873 0.454 0.649946   
## votedav 0.0989013 0.0021760 45.451 < 2e-16 \*\*\*  
## dem -0.0067959 0.0022705 -2.993 0.002763 \*\*   
## nov 0.0637412 0.0037770 16.876 < 2e-16 \*\*\*  
## aug 0.0553982 0.0039725 13.946 < 2e-16 \*\*\*  
## city -0.0004668 0.0028109 -0.166 0.868118   
## hh\_id -9.9818026 21.9825185 -0.454 0.649774   
## hh\_size -0.0064301 0.0024493 -2.625 0.008661 \*\*   
## totalpopulation\_estimate -0.0024833 0.0052363 -0.474 0.635328   
## percent\_male -0.0078079 0.0046403 -1.683 0.092456 .   
## percent\_female NA NA NA NA   
## median\_age -0.0150458 0.0205355 -0.733 0.463763   
## percent\_under5years 0.1735736 0.0385397 4.504 6.69e-06 \*\*\*  
## percent\_5to9years 0.1667305 0.0380587 4.381 1.18e-05 \*\*\*  
## percent\_10to14years 0.1815740 0.0380857 4.768 1.87e-06 \*\*\*  
## percent\_15to19years 0.2244384 0.0453982 4.944 7.69e-07 \*\*\*  
## percent\_20to24years 0.2901008 0.0767494 3.780 0.000157 \*\*\*  
## percent\_25to34years 0.3056577 0.0735689 4.155 3.26e-05 \*\*\*  
## percent\_35to44years 0.2020628 0.0556173 3.633 0.000280 \*\*\*  
## percent\_45to54years 0.2323161 0.0584686 3.973 7.10e-05 \*\*\*  
## percent\_55to59years 0.1437101 0.0356631 4.030 5.60e-05 \*\*\*  
## percent\_60to64years 0.1648947 0.0390645 4.221 2.44e-05 \*\*\*  
## percent\_65to74years 0.2055611 0.0997843 2.060 0.039399 \*   
## percent\_75to84years 0.1346264 0.0718140 1.875 0.060847 .   
## percent\_85yearsandolder 0.0879909 0.0409606 2.148 0.031705 \*   
## percent\_18yearsandolder 0.2018413 0.0168009 12.014 < 2e-16 \*\*\*  
## percent\_21yearsandover -0.0897308 0.0233399 -3.845 0.000121 \*\*\*  
## percent\_62yearsandover -0.0928004 0.0307309 -3.020 0.002531 \*\*   
## percent\_65yearsandover 0.0836431 0.2029097 0.412 0.680182   
## percent\_white 0.0524518 0.0294563 1.781 0.074974 .   
## percent\_black 0.0437486 0.0218788 2.000 0.045551 \*   
## percent\_amindian\_alaskan 0.0270297 0.0059812 4.519 6.23e-06 \*\*\*  
## percent\_asian -0.0259966 0.0144064 -1.805 0.071157 .   
## percent\_nativeandother 0.0027566 0.0027245 1.012 0.311644   
## percent\_other\_nativeandother 0.0438467 0.0038912 11.268 < 2e-16 \*\*\*  
## percent\_hispanicorlatino 0.0270501 0.0073538 3.678 0.000235 \*\*\*  
## percent\_race\_other -0.0722351 0.0102390 -7.055 1.75e-12 \*\*\*  
## median\_income 0.0213751 0.0174778 1.223 0.221342   
## mean\_income -0.0203875 0.0254090 -0.802 0.422342   
## employ\_16 -0.1233365 0.0221910 -5.558 2.75e-08 \*\*\*  
## unemploy\_16 -0.0723103 0.0147184 -4.913 9.01e-07 \*\*\*  
## unemploy\_20to64 0.0379297 0.0140922 2.692 0.007115 \*\*   
## employ\_20to64 0.0511395 0.0175334 2.917 0.003540 \*\*   
## employ\_rename\_20to64 -0.0486188 0.0054071 -8.992 < 2e-16 \*\*\*  
## hsorhigher -0.5521551 0.2566383 -2.151 0.031443 \*   
## bach\_orhigher 1.7510584 0.8865240 1.975 0.048252 \*   
## less9thgrade 0.3283036 0.0900636 3.645 0.000267 \*\*\*  
## grade9to12 0.4817695 0.1478123 3.259 0.001118 \*\*   
## highschool 2.7354941 0.4271044 6.405 1.52e-10 \*\*\*  
## somecollege 1.1062172 0.1717248 6.442 1.19e-10 \*\*\*  
## assoc 0.5749625 0.0894367 6.429 1.30e-10 \*\*\*  
## bachelors 1.4520884 0.4024767 3.608 0.000309 \*\*\*  
## grad 1.3357719 0.3862822 3.458 0.000545 \*\*\*  
## randn 0.0144257 0.0022834 6.318 2.68e-10 \*\*\*  
## p2004 0.0519879 0.0028527 18.224 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6314 on 42206 degrees of freedom  
## Multiple R-squared: 0.174, Adjusted R-squared: 0.1728   
## F-statistic: 145.8 on 61 and 42206 DF, p-value: < 2.2e-16

# 3.3 Using lasso to estimate

atelasso\_matrix <- model.matrix(linear, data\_new)[,-1]  
ate\_lasso\_cv = cv.glmnet(atelasso\_matrix, Y, alpha = 1, family = 'binomial', nfolds=5)  
ate\_lasso = glmnet(atelasso\_matrix, Y, alpha = 1, family = 'binomial', lambda = ate\_lasso\_cv$lambda.1se)  
coef(ate\_lasso)[2]

## [1] 0.4527351

# 3.4a Belloni‐Chernozhukov‐Hansen method: Coefficient selection

prop\_matrix <- model.matrix(linear, data\_new)[,-(1:2)]  
BCH\_treatment\_lambda = cv.glmnet(prop\_matrix, W, alpha = 1, family = 'binomial', nfolds=5)  
BCH\_outcome\_lambda = cv.glmnet(prop\_matrix, Y, alpha = 1, family = 'binomial', nfolds=5)  
  
# lasso models  
BCH\_lasso\_treat = glmnet(prop\_matrix, W, alpha = 1, family = 'binomial', lambda = BCH\_treatment\_lambda$lambda.min)  
BCH\_lasso\_out = glmnet(prop\_matrix, Y, alpha = 1, family = 'binomial', lambda = BCH\_outcome\_lambda$lambda.min)  
  
#coefficients  
coef(BCH\_lasso\_treat)

## 62 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) -5.312449007  
## sex -0.015057741  
## yob -0.013319232  
## g2000 0.053857513  
## g2002 -0.025632554  
## p2000 -0.019651136  
## p2002 -0.069485620  
## cluster -0.997709097  
## votedav 0.014459750  
## dem 0.042341038  
## nov -0.004675636  
## aug 0.100500660  
## city 0.025237628  
## hh\_id -0.016954642  
## hh\_size 0.026160756  
## totalpopulation\_estimate 0.370178091  
## percent\_male -0.420183557  
## percent\_female .   
## median\_age -0.336953504  
## percent\_under5years -0.048932610  
## percent\_5to9years -0.083211978  
## percent\_10to14years -0.131182285  
## percent\_15to19years -0.162302503  
## percent\_20to24years 0.001901026  
## percent\_25to34years -0.426508503  
## percent\_35to44years 0.174323314  
## percent\_45to54years 0.048539195  
## percent\_55to59years 0.075958969  
## percent\_60to64years -0.361643207  
## percent\_65to74years 8.504051652  
## percent\_75to84years 5.509589459  
## percent\_85yearsandolder 3.692526287  
## percent\_18yearsandolder -0.325381857  
## percent\_21yearsandover 0.165702487  
## percent\_62yearsandover 2.102623764  
## percent\_65yearsandover -16.103681725  
## percent\_white 2.438068074  
## percent\_black 1.404836298  
## percent\_amindian\_alaskan 0.097645625  
## percent\_asian 1.052324920  
## percent\_nativeandother 0.090353181  
## percent\_other\_nativeandother -0.189294894  
## percent\_hispanicorlatino 0.117457481  
## percent\_race\_other 0.652167397  
## median\_income -0.155504059  
## mean\_income -0.354740152  
## employ\_16 1.144366289  
## unemploy\_16 0.540382889  
## unemploy\_20to64 -0.469206487  
## employ\_20to64 -0.754051478  
## employ\_rename\_20to64 -0.046827936  
## hsorhigher 2.752827817  
## bach\_orhigher .   
## less9thgrade 0.428578447  
## grade9to12 1.129229677  
## highschool -1.959647110  
## somecollege -1.023337495  
## assoc -0.369775713  
## bachelors -1.887838258  
## grad -3.735916971  
## randn .   
## p2004 -0.064301549

coef(BCH\_lasso\_out)

## 62 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) -0.5359041832  
## sex -0.0241755199  
## yob 0.0040176292  
## g2000 -0.0536558550  
## g2002 -0.0178816292  
## p2000 0.0918571334  
## p2002 0.1600019022  
## cluster -0.0948502604  
## votedav 1.4113712354  
## dem -0.0334720805  
## nov 0.2042427730  
## aug 0.3723555531  
## city 0.0444535936  
## hh\_id -0.0172651651  
## hh\_size -0.0046551421  
## totalpopulation\_estimate -0.0443695142  
## percent\_male -0.0001151843  
## percent\_female .   
## median\_age -0.1860752260  
## percent\_under5years 0.0095195548  
## percent\_5to9years 0.1325195879  
## percent\_10to14years 0.0982548457  
## percent\_15to19years 0.1178854337  
## percent\_20to24years 0.0965790887  
## percent\_25to34years -0.0255698818  
## percent\_35to44years -0.0425309199  
## percent\_45to54years 0.0075238874  
## percent\_55to59years 0.0200620481  
## percent\_60to64years -0.0103658250  
## percent\_65to74years 0.0671636042  
## percent\_75to84years 0.1455981159  
## percent\_85yearsandolder 0.0529385334  
## percent\_18yearsandolder 0.5506084930  
## percent\_21yearsandover -0.2195536565  
## percent\_62yearsandover -0.2733620397  
## percent\_65yearsandover -0.0753348807  
## percent\_white 0.0117964060  
## percent\_black .   
## percent\_amindian\_alaskan -0.0003626177  
## percent\_asian -0.0136542886  
## percent\_nativeandother -0.0251290575  
## percent\_other\_nativeandother 0.1304389818  
## percent\_hispanicorlatino 0.1134847261  
## percent\_race\_other -0.0355314482  
## median\_income -0.0044518612  
## mean\_income -0.0165947238  
## employ\_16 -0.5864459024  
## unemploy\_16 -0.6838026513  
## unemploy\_20to64 0.5034295175  
## employ\_20to64 0.2323574951  
## employ\_rename\_20to64 -0.1740740106  
## hsorhigher 0.0460383325  
## bach\_orhigher .   
## less9thgrade -0.0596920131  
## grade9to12 -0.0747070067  
## highschool .   
## somecollege -0.0407177305  
## assoc -0.0442366246  
## bachelors 0.1295308344  
## grad -0.2047536954  
## randn 0.0030265705  
## p2004 0.2017601575

# 3.4 b) OLS propensity score model

BCH\_OLS<-glm(Y ~W + sex + yob + g2000 + g2002 + p2000 + p2002 + cluster + votedav +  
dem + nov + aug + city + hh\_id + totalpopulation\_estimate +  
percent\_male + percent\_female + median\_age + percent\_under5years +  
percent\_5to9years + percent\_10to14years + percent\_15to19years +  
percent\_20to24years + percent\_35to44years +  
percent\_45to54years + percent\_55to59years + percent\_60to64years + percent\_75to84years + percent\_85yearsandolder +  
percent\_18yearsandolder + percent\_21yearsandover + percent\_62yearsandover + percent\_white + percent\_black +  
percent\_amindian\_alaskan + percent\_asian + percent\_nativeandother +  
percent\_other\_nativeandother + percent\_hispanicorlatino +  
percent\_race\_other + mean\_income + employ\_16 +  
unemploy\_16 + unemploy\_20to64 + employ\_20to64 + employ\_rename\_20to64 +  
hsorhigher + bach\_orhigher + less9thgrade + grade9to12 +  
highschool + somecollege + assoc + bachelors + grad + randn +  
p2004, data = data\_new, family = "binomial")

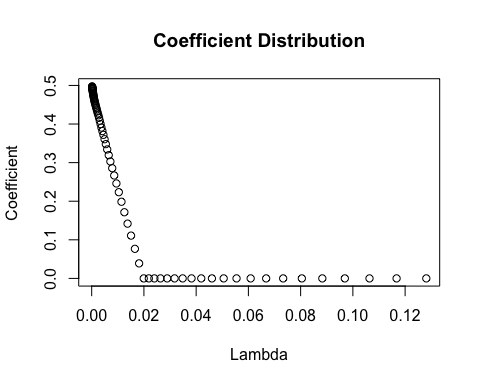
summary(BCH\_OLS)

##   
## Call:  
## glm(formula = Y ~ W + sex + yob + g2000 + g2002 + p2000 + p2002 +   
## cluster + votedav + dem + nov + aug + city + hh\_id + totalpopulation\_estimate +   
## percent\_male + percent\_female + median\_age + percent\_under5years +   
## percent\_5to9years + percent\_10to14years + percent\_15to19years +   
## percent\_20to24years + percent\_35to44years + percent\_45to54years +   
## percent\_55to59years + percent\_60to64years + percent\_75to84years +   
## percent\_85yearsandolder + percent\_18yearsandolder + percent\_21yearsandover +   
## percent\_62yearsandover + percent\_white + percent\_black +   
## percent\_amindian\_alaskan + percent\_asian + percent\_nativeandother +   
## percent\_other\_nativeandother + percent\_hispanicorlatino +   
## percent\_race\_other + mean\_income + employ\_16 + unemploy\_16 +   
## unemploy\_20to64 + employ\_20to64 + employ\_rename\_20to64 +   
## hsorhigher + bach\_orhigher + less9thgrade + grade9to12 +   
## highschool + somecollege + assoc + bachelors + grad + randn +   
## p2004, family = "binomial", data = data\_new)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0321 -0.8725 -0.6324 1.1344 2.5526   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.264535 1.848336 -0.143 0.886195   
## W 0.505142 0.053940 9.365 < 2e-16 \*\*\*  
## sex -0.024371 0.011282 -2.160 0.030759 \*   
## yob 0.003819 0.012657 0.302 0.762861   
## g2000 -0.056909 0.012615 -4.511 6.45e-06 \*\*\*  
## g2002 -0.021335 0.018139 -1.176 0.239506   
## p2000 0.094582 0.012206 7.749 9.27e-15 \*\*\*  
## p2002 0.163103 0.014159 11.519 < 2e-16 \*\*\*  
## cluster -40.987884 112.317606 -0.365 0.715165   
## votedav 2.994533 10.330492 0.290 0.771913   
## dem -0.034610 0.011770 -2.941 0.003277 \*\*   
## nov 0.210600 0.021728 9.693 < 2e-16 \*\*\*  
## aug 0.372145 0.020267 18.362 < 2e-16 \*\*\*  
## city 0.041038 0.013754 2.984 0.002848 \*\*   
## hh\_id 40.889429 112.317469 0.364 0.715819   
## totalpopulation\_estimate -0.073495 0.023890 -3.076 0.002096 \*\*   
## percent\_male -0.006551 0.019062 -0.344 0.731101   
## percent\_female NA NA NA NA   
## median\_age -0.183360 0.088331 -2.076 0.037911 \*   
## percent\_under5years 0.080948 0.060463 1.339 0.180635   
## percent\_5to9years 0.202737 0.059414 3.412 0.000644 \*\*\*  
## percent\_10to14years 0.178530 0.058553 3.049 0.002296 \*\*   
## percent\_15to19years 0.176881 0.062674 2.822 0.004769 \*\*   
## percent\_20to24years 0.101945 0.044518 2.290 0.022024 \*   
## percent\_35to44years -0.025011 0.026822 -0.932 0.351083   
## percent\_45to54years 0.019744 0.034766 0.568 0.570100   
## percent\_55to59years 0.025021 0.026862 0.931 0.351628   
## percent\_60to64years -0.008876 0.024309 -0.365 0.715006   
## percent\_75to84years 0.103819 0.030598 3.393 0.000691 \*\*\*  
## percent\_85yearsandolder 0.006536 0.020988 0.311 0.755490   
## percent\_18yearsandolder 0.666990 0.082951 8.041 8.93e-16 \*\*\*  
## percent\_21yearsandover -0.166516 0.114700 -1.452 0.146569   
## percent\_62yearsandover -0.265233 0.083631 -3.171 0.001517 \*\*   
## percent\_white 0.028989 0.132859 0.218 0.827278   
## percent\_black 0.019320 0.096556 0.200 0.841410   
## percent\_amindian\_alaskan 0.001274 0.025992 0.049 0.960894   
## percent\_asian -0.021127 0.067913 -0.311 0.755730   
## percent\_nativeandother -0.021633 0.012990 -1.665 0.095837 .   
## percent\_other\_nativeandother 0.133522 0.019585 6.817 9.27e-12 \*\*\*  
## percent\_hispanicorlatino 0.125133 0.033706 3.712 0.000205 \*\*\*  
## percent\_race\_other -0.059415 0.048439 -1.227 0.219982   
## mean\_income -0.042785 0.070810 -0.604 0.545698   
## employ\_16 -0.731728 0.097016 -7.542 4.61e-14 \*\*\*  
## unemploy\_16 -0.765294 0.067747 -11.296 < 2e-16 \*\*\*  
## unemploy\_20to64 0.582879 0.065057 8.959 < 2e-16 \*\*\*  
## employ\_20to64 0.340376 0.082267 4.137 3.51e-05 \*\*\*  
## employ\_rename\_20to64 -0.179916 0.026846 -6.702 2.06e-11 \*\*\*  
## hsorhigher -0.088534 1.228056 -0.072 0.942528   
## bach\_orhigher -17.854252 4.219231 -4.232 2.32e-05 \*\*\*  
## less9thgrade -1.191184 0.427307 -2.788 0.005309 \*\*   
## grade9to12 -1.926730 0.698323 -2.759 0.005796 \*\*   
## highschool -5.498824 1.955657 -2.812 0.004927 \*\*   
## somecollege -2.255679 0.787149 -2.866 0.004162 \*\*   
## assoc -1.195139 0.409988 -2.915 0.003556 \*\*   
## bachelors 4.927977 1.947864 2.530 0.011408 \*   
## grad 4.429497 1.876451 2.361 0.018247 \*   
## randn 0.005159 0.011175 0.462 0.644333   
## p2004 0.204296 0.014458 14.130 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 54192 on 42267 degrees of freedom  
## Residual deviance: 46776 on 42211 degrees of freedom  
## AIC: 46890  
##   
## Number of Fisher Scoring iterations: 15

# 3.5 a) How ATE coef changes with regularization

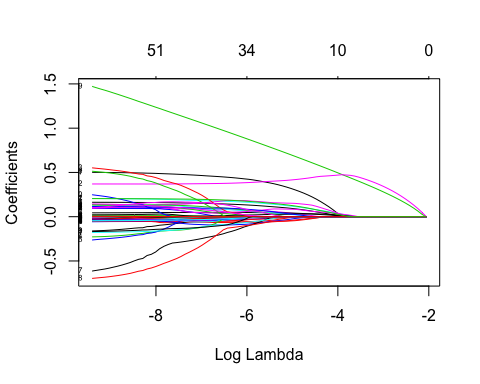
# single‐equation LASSO of Y on X and W.plot how the coefficient on the treatment indicator changes with lambda.

temp = coef(ate\_lasso\_cv, s = ate\_lasso\_cv$lambda)  
coef\_W = as.matrix(temp)  
plot(ate\_lasso\_cv$lambda, coef\_W[2,], xlab = "Lambda", ylab = "Coefficient", main = "Coefficient Distribution")



# 3.5 optional plot the coefficients on some other covariates as lambda changes

plot(ate\_lasso\_cv$glmnet.fit, "lambda", label = TRUE)



## No.4 Double machine learning WHERE TO SHOW RESULTS???? WHAT TO COMPARE ??

residual\_yx = Y - predict(BCH\_outcome\_lambda, newx=as.matrix(data\_new[, -(1:2)]),   
 s = 0.0023, type = "response")  
residual\_wx = W - predict(BCH\_treatment\_lambda, newx=as.matrix(data\_new[, -(1:2)]),   
 s = 0.0023, type = "response")  
dml\_est = lm(residual\_yx ~ residual\_wx) # Residual on residual regression  
coef(dml\_est)['residual\_wx']

## residual\_wx   
## 0.09886502

summary(dml\_est)

##   
## Call:  
## lm(formula = residual\_yx ~ residual\_wx)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.7897 -0.3220 -0.1956 0.4980 0.9490   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.254e-08 2.118e-03 0.000 1   
## residual\_wx 9.887e-02 1.058e-02 9.345 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4355 on 42266 degrees of freedom  
## Multiple R-squared: 0.002062, Adjusted R-squared: 0.002038   
## F-statistic: 87.33 on 1 and 42266 DF, p-value: < 2.2e-16

# Q4b) Residual balancing

X <- data\_new[, -c(1, 2)]  
W <- data\_new[, 2]  
Y <- data\_new[, 1]  
  
#tau.hat <- residualBalance.ate(X, Y, W, estimate.se = TRUE, optimizer = "pogs")  
#print(paste("point estimate:", round(tau.hat[1], 2)))  
#print(paste0("95% CI for tau: (", round(tau.hat[1] - 1.96 \* tau.hat[2], 2), ", ", round(tau.hat[1] + 1.96 \* tau.hat[2], 2), ")"))