Propensity Score Weighting and Generalized Boosted Regression in R

Shenyang Guo and Peter Sun

September, 2021

# Load Packages

library(tidyverse)  
library(haven)  
library(sjlabelled)  
library(lmtest)  
library(sandwich)  
library(gbm)  
select <- dplyr::select  
knitr::opts\_chunk$set(dpi = 300,fig.width = 7)

# Propensity Score Weighting

## Load Data

d <- read\_dta("data/chpt5\_2\_original.dta") %>%  
 haven::zap\_formats() %>%  
 sjlabelled::remove\_all\_labels() %>%  
 as\_tibble()

## Create ATE and ATT Weights

d.weights <- d %>%  
 mutate(ate\_w = ifelse(kuse == 0, 1/(1-ps), 1/ps),  
 att\_w = ifelse(kuse == 0, ps/(1-ps), 1))

## PSW With ATE Weights

Use the weights argument in lm() to run OLS regression using weights and lmtest::coeftest() to control for clustering effects.

m3 <- lm(lwss97 ~ kuse + male + black + age97 + pcged97 + mratio96,   
 data = d.weights, weights = ate\_w)   
lmtest::coeftest(m3, vcov. = vcovCL(m3, cluster = d.weights$pcg\_id))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 84.19992 4.83032 17.4315 < 2.2e-16 \*\*\*  
## kuse -5.16399 1.42438 -3.6254 0.0003031 \*\*\*  
## male -1.62201 1.09186 -1.4855 0.1377180   
## black -2.49898 1.34670 -1.8556 0.0638009 .   
## age97 0.73868 0.18075 4.0867 4.727e-05 \*\*\*  
## pcged97 0.99264 0.35596 2.7886 0.0053938 \*\*   
## mratio96 1.13856 0.32220 3.5337 0.0004286 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## PSW With ATT Weights

m4 <- lm(lwss97 ~ kuse + male + black + age97 + pcged97 + mratio96,   
 data = d.weights, weights = att\_w)  
lmtest::coeftest(m4, vcov. = vcovCL(m4, cluster = d.weights$pcg\_id))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 85.29467 5.05331 16.8790 < 2.2e-16 \*\*\*  
## kuse -4.62058 1.41182 -3.2728 0.001102 \*\*   
## male -1.58995 1.14705 -1.3861 0.166020   
## black -2.74605 1.41605 -1.9392 0.052756 .   
## age97 0.61577 0.20001 3.0787 0.002136 \*\*   
## pcged97 0.92698 0.36718 2.5246 0.011738 \*   
## mratio96 1.26018 0.33556 3.7555 0.000183 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Sample Imbalance Check Code

Use logistic regression for dummy covariates and OLS regression for continuous covariates. The full code can be found in Section 7.3.1 of the PSA-R code.

c1 <- glm(male ~ kuse, family = quasibinomial,   
 data = d.weights, weights = ate\_w)  
lmtest::coeftest(c1, vcov. = vcovCL(c1, cluster = d.weights$pcg\_id))

# Estimate Propensity Scores Using Generalized Boosted Regression

## Load Data

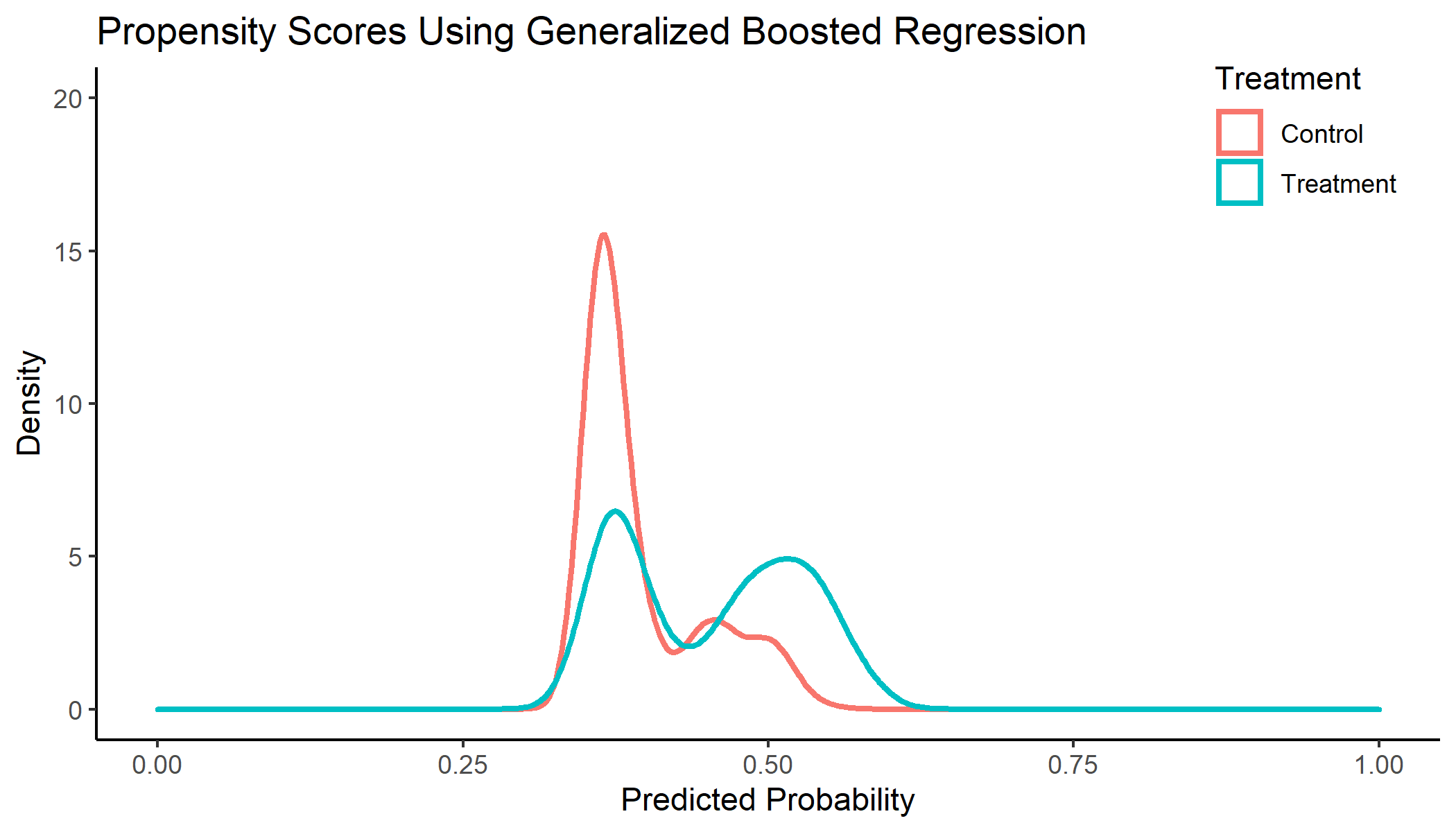
d2 <- read\_dta("data/g3aca1.dta") %>%  
 haven::zap\_formats() %>%  
 sjlabelled::remove\_all\_labels() %>%  
 as\_tibble() %>%  
 select(intbl, ageyc, fmale, blck, whit, hisp, pcedu, ipovl, pcemft, fthr,  
 dicsagg2, dicsint2, dccereg2, dccscom2, dccpros2, draggr2) %>%  
 drop\_na()

## Generate Propensity Scores

f5 <- as.formula(paste("intbl ~ ", paste(names(select(d2, -intbl)),   
 collapse = " + ")))  
set.seed(1000)  
m5 <- gbm::gbm(formula = f5,  
 data = d2,  
 distribution = "bernoulli",  
 n.trees = 1000,  
 train.fraction = 0.8,  
 interaction.depth = 4,  
 shrinkage = 0.0005)  
psb <- gbm::predict.gbm(m5, data = d2, type = "response") # Create ps

## Plot and Summarize Results

d2 %>%  
 mutate(psb = psb,  
 intbl = factor(intbl, labels = c("Control", "Treatment"))) %>%  
 ggplot(aes(x = psb, color = intbl)) + theme\_classic() +  
 geom\_density(size = 1) + xlim(0, 1) + ylim(0, 20) +  
 labs(x = "Predicted Probability", y = "Density",  
 title = "Propensity Scores Using Generalized Boosted Regression", color = "Treatment") +  
 theme(legend.position = c(0.9, 0.9))



summary(psb)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.3480 0.3660 0.3869 0.4229 0.4836 0.5883