PHS 2000B Lab 2 Exercise

Propensity Scores

Monday, January 30, 2023

Propensity Score Methods

We will be using a subset of the California Smokers Cohort (CSC) 2011 to examine the effect of early smoking initiation ($\leq$ 16 years old) on adult depressive symptoms (PHQ-4).

We will consider five confounders as sufficient to have exchangeability of smoking status: highest level of parental education, race, ethnicity, nativity, sex, and age.

Data can be found on Canvas in `dataset.csv`.

Name	Variable	Description
`SMK1AGE`	Exposure	Age of first time smoking (a whole) cigarette: 0 :$\leq$ 16 years, 1 :> 16 years
`DEP`	Outcome	Depression score summed over PROBINTR and PROBDOWN
`SCHOOLLEV`	Confounder	Highest level of schooling for any parent
`HISPN`	Confounder	Hispanic: Yes/No
`RACE`	Confounder	White, Black, Others
`NATVLAND`	Confounder	Nationality: US vs non-US
`AGE`	Confounder	Age groups: 30-39, 40-49, 50-64

We will be exploring three methods using propensity scores: stratification, matching, and weighting.

1. When might you want to use propensity score methods over outcome-based regression modeling?

Propensity score methods are useful in situations in which it is easier to model the exposure than to model the outcome. If we have many confounders, we may not have enough data to model the exposure - outcome relationship, for example, with binary outcomes when we have few events but many covariates can make the model unstable. Propensity scores are also helpful when an outcome is very skewed, the functional form given by covariates may be hard to specify.

2. Please write out the propensity score model for the exposure and propensity score estimator for each individual observation.

$$\text{logit}(Pr(A = 1|C = c)) = \beta_0 + \beta_1 SCHOOLEV + \beta_2 HISPN + \beta_3 NATVLAND + \beta_4 MALE + \beta_5 AGE + \beta_6 RACE$$

$$\hat{S} = \frac{\hat{\beta}_0 + \beta_1 SCHOOLEV + \beta_2 HISPN + \beta_3 NATVLAND + \beta_4 MALE + \beta_5 AGE + \beta_6 RACE}{1 + exp(\beta_0 + \beta_1 SCHOOLEV + \beta_2 HISPN + \beta_3 NATVLAND + \beta_4 MALE + \beta_5 AGE + \beta_6 RACE)}$$

3. Fit the model in R and predict the propensity score for all individuals.

```
# Part 1: Question 3

# Exposure model
m_ps <- glm(SMK1AGE ~ SCHOOLEV + HISPN + NATVLAND + MALE + AGE + RACE,
            family = binomial(link="logit"), data = csc)

# Predicting propensity score
csc$pr_score <- predict(m_ps, type = "response")
```

4. Plot the distribution of propensity scores among the exposed and unexposed and comment on what you see.

There appears to be considerable overlap in the distribution of propensity scores for the exposed and unexposed. Interestingly, it seems like there are a high number of individuals who had high probability of exposure among the unexposed group. This may be an indication that the combination of covariates in our propensity score model are not highly predictive of exposure status.

```
# Part 1: Question 4
# Checking mean overall and by treatment group
csc |>
  group_by(SMK1AGE) |>
  summarise(mean(pr_score))
```

```
## # A tibble: 2 x 2
##   SMK1AGE `mean(pr_score)`
##     <int>            <dbl>
## 1       0            0.612
## 2       1            0.640
```

```
# Plot
ggplot(subset(csc, SMK1AGE == 1), aes(x = pr_score, fill = factor(SMK1AGE))) +
  geom_histogram(aes(y = - after_stat(density))) + # note the negative sign here
  geom_histogram(data = subset(csc, SMK1AGE == 0),
                     aes(x = pr_score, y = after_stat(density), fill = factor(SMK1AGE))) +
  ylab("Density") + xlab("Probability of Exposure") +
  ggtitle("Propensity Scores in Exposed and Unexposed") +
  scale_fill_discrete(name = "Exposure") +
  theme_minimal()
```

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

Stratification

1. Create five strata based on the quintiles of the overall propensity score distribution.

```
# Part 2: Question 1

csc$ps_strat <- quantcut(csc$pr_score, q = 5)

table(csc$ps_strat)
```

```
##
## [0.281,0.574] (0.574,0.635] (0.635,0.671] (0.671,0.699] (0.699,0.753]
##           174           131           154           158           144
```

```
csc <- csc |>
  mutate(strata = case_when(ps_strat=="[0.281,0.574]" ~ 1,
                            ps_strat=="(0.574,0.635]" ~ 2,
                            ps_strat=="(0.635,0.671]" ~ 3,
                            ps_strat=="(0.671,0.699]" ~ 4,
                            ps_strat=="(0.699,0.753]" ~ 5))
```

2. Check the covariate distribution within each of the strata.
3. Estimate the causal effect within each strata of the study population.

```r
# OPTION 1:
# Estimate the means
means <- aggregate(x = csc$DEP,
          by = list(csc$SMK1AGE, csc$ps_strat),
          FUN = mean)

print(c(means[2,3]-means[1,3],
        means[4,3]-means[3,3],
        means[6,3]-means[5,3],
        means[8,3]-means[7,3],
        means[10,3]-means[9,3]))
```

```
## [1] -0.05461294  0.41797060 -0.12755102  0.06267970 -0.13653846
```

```r
# OPTION 2:
# create a vector to contain the stratum-specific estimates
coef_est <- rep(NA, 5)
# and their standard errors
se_est <- rep(NA, 5)

# run a loop through all the strata
for(i in 1:5){
  # estimate the model in each stratum
  mod <- lm(DEP ~ SMK1AGE,
            data = csc,
            subset = strata == i)
  # save the coefficients and standard errors
  coef_est[i] <- coef(mod)[2]
  se_est[i] <- summary(mod)$coefficients[2,2]
}

# Combine and print coefficients and standard errors
rbind(coef_est, se_est)
```

```
##                 [,1]      [,2]       [,3]      [,4]       [,5]
## coef_est -0.05461294 0.4179706 -0.1275510 0.0626797 -0.1365385
## se_est    0.26217679 0.3175990  0.3193608 0.3362703  0.2938098
```

```r
# OPTION 3
csc_nested <- csc |>
  nest_by(strata) |>
  mutate(model = list(lm(DEP ~ SMK1AGE, data = data)),
         effect = list(broom::tidy(model))) |>
  select(strata, effect) |>
  tidyr::unnest(cols=c(effect)) |>
  filter(term == "SMK1AGE")
```

4. Using these estimates, calculate the total average treatment effect and the average treatment effect among the treated and interpret.

The **total average treatment effect** is 0.0207, indicating that the average difference in depression score is 0.0207 comparing had everyone started smoking at $\leq$ 16 years old vs. had everyone started smoking at $>$ 16 years old. The **average treatment among the treated** is 0.0125, indicating that among those

who started smoking at $\leq$ 16 years old, the average difference in depression score is 0.0207 comparing had everyone started smoking at $\leq$ 16 years old vs. had everyone started smoking at $>$ 16 years old.

```
# Total average treatment effect
# Obtain proportions of individuals in each strata
prop.table(table(csc$strata))
```

```
##
##         1         2         3         4         5
## 0.2286465 0.1721419 0.2023653 0.2076216 0.1892247
```

```
(-0.0546*0.229) + (0.418*0.172) + (-0.128*0.202) + (0.0627*0.208) + (-0.137*0.189)
```

```
## [1] 0.0206852
```

```
# Average treatment effect among treated
# Obtain proportions of individuals in each strata AMONG TREATED
csc |>
  filter(SMK1AGE==1) |>
  group_by(strata) |>
  summarise(count = n()) |>
  mutate(prop = count/sum(count))
```

```
## # A tibble: 5 x 3
##   strata count  prop
##    <dbl> <int> <dbl>
## 1      1    92 0.192
## 2      2    74 0.154
## 3      3    98 0.205
## 4      4   111 0.232
## 5      5   104 0.217
```

```
(-0.0546*0.192) + (0.418*0.154) + (-0.128*0.205) + (0.0627*0.232) + (-0.137*0.217)
```

```
## [1] 0.0124662
```

```
# Average treatment effect among untreated
# Obtain proportions of individuals in each strata AMONG TREATED
csc |>
  filter(SMK1AGE==0) |>
  group_by(strata) |>
  summarise(count = n()) |>
  mutate(prop = count/sum(count))
```

```
## # A tibble: 5 x 3
##   strata count  prop
##    <dbl> <int> <dbl>
## 1      1    82 0.291
## 2      2    57 0.202
## 3      3    56 0.199
## 4      4    47 0.167
## 5      5    40 0.142
```

```
(-0.0546*0.291) + (0.418*0.202) + (-0.128*0.199) + (0.0627*0.167) + (-0.137*0.142)
```

```
## [1] 0.0340923
```

Matching

1. Look at the number of people in the exposed and unexposed. Which group has more people? What implications does this have for matching and the effect that we will identify?

There are more people in the exposed group than unexposed. Since there are more of the exposed, we will want to match to the unexposed group, as this ensures that every unexposed has a match. Consequently, we will be looking at the average treatment effect among the unexposed.

```
# Number of people in the exposed and unexposed groups
table(csc$SMK1AGE)
```

```
##
##   0   1
## 282 479
```

2. Fit the nearest-neighbor matching, using the same exposure model. What are the dimensions of the matched dataset compared to the original? How many people are in the exposed and unexposed groups in the matched dataset?

The original dataset has 761 rows and 13 columns, while the matched dataset has 564 rows and 16 columns. In the original dataset, there were 282 unexposed individuals and 479 exposed individuals, while in the matched dataset there were 282 unexposed and 282 exposed individuals. This is because we specified we were matching to the unexposed, as there were more unexposed than exposed. We dropped those exposed individuals for whom there were no nearest neighbor match among the unexposed.

```
# 1:1 nearest-neighbor matching
# set the seed immediately before running so we all get the same matches
set.seed(1234)

nearest_match <- matchit(SMK1AGE ~ as.factor(SCHOOLEV) + as.factor(HISPN) + as.factor(NATVLAND) + as.fa
                         method = "nearest", distance = "glm", link="logit", discard = "treated",
                         data = csc)
```

```
## Warning: Fewer control units than treated units; not all treated units will get
## a match.
```

```
#nearest_match <- matchit(SMK1AGE ~ as.factor(SCHOOLEV) + as.factor(HISPN) + #as.factor(NATVLAND) + as.
#                          method = "nearest", distance = "glm", link="logit", discard = #"both",
#                          data = csc)
#
# Note if we change the code so that we have discard = "both" then we get a message that says "not all
# This change would mean that our estimate of the causal effect would be the ATE among those with commo



# Output the data for further analysis
nearest_data <- match.data(nearest_match)

# Dimensions of datasets
```

```r
dim(csc)
```

```
## [1] 761  13
```

```r
dim(nearest_data)
```

```
## [1] 564  16
```

```r
# Number in matched subset
table(csc$SMK1AGE)
```

```
##
##   0   1
## 282 479
```

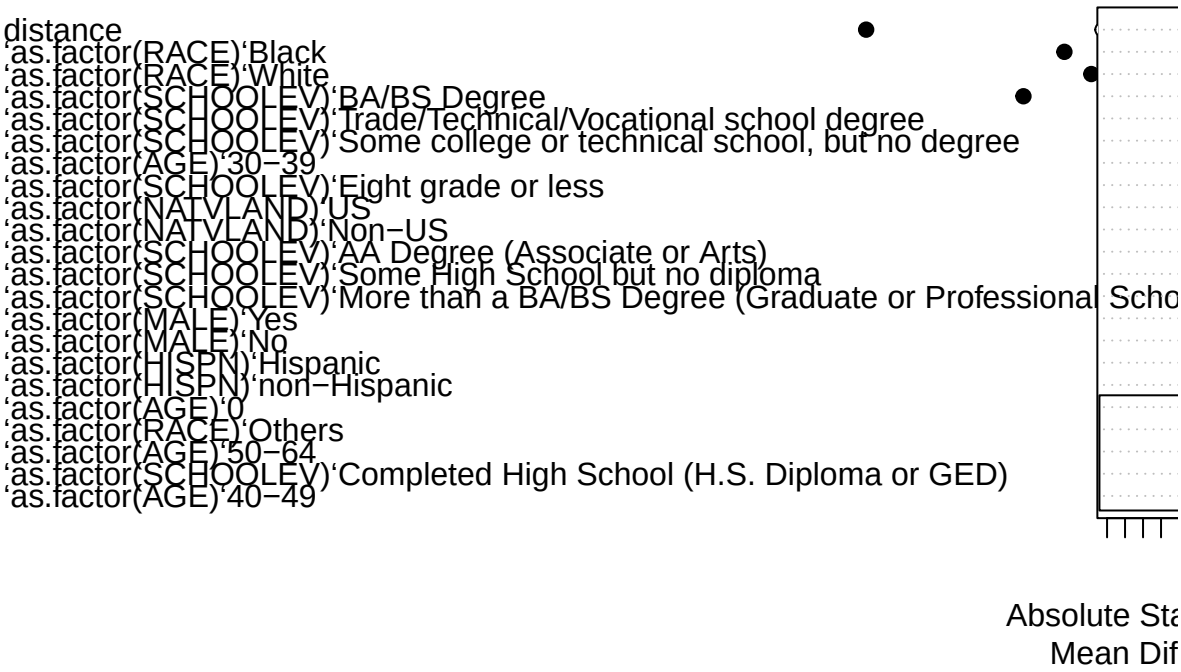
```r
table(nearest_data$SMK1AGE)
```

```
##
##   0   1
## 282 282
```

3. Check the covariate balance in the matched subset. Are there any imbalances of concern? What is one way to address these potential imbalances?

From the plot, we find there are many imbalances, as indicated by SMD ≥ 0.1 for many covariates. This may indicate that the matching was unsuccessful. We could use a doubly robust method, i.e., combine matching with another strategy, to address potential 'residual confounding'.

```r
# Check for covariate balance in matched subset
plot(summary(nearest_match), var.order = "unmatched")
```




```
# Code to look at the table
#summary(nearest_match)
```

4. Estimate the causal effect and interpret.

This is the average treatment effect among the unexposed, since we matched to the unexposed group. Among those who were unexposed, the difference in the average depression score is -0.067 had everyone been exposed compared to had everyone been unexposed.

```
# Causal effect
lm(DEP ~ SMK1AGE, data = nearest_data)
```

```
##
## Call:
## lm(formula = DEP ~ SMK1AGE, data = nearest_data)
##
## Coefficients:
## (Intercept)      SMK1AGE
##     1.40780     -0.06738
```

Weighting

1. What are some differences between stabilized and unstabilized weights?

Unstabilized weights will create a pseudo-population with size $N *$ (number of exposure categories), while stabilized weights create a pseudo-population with size N. Unstabilized weights are subject to extreme

weights, whereas the numerator of stabilized weights helps reduce the opportunity for extreme weights.

2. Calculate unstabilized weights and check the distribution. Are there any extreme weights? What is one option for dealing with extreme weights?

The minimum weight is 1.3 while the maximum weight is 4, so there do not appear to be extreme weights. We can trim the weights so that the most extreme weights. Another option is to use stabilized weights, if appropriate.

```
# Create weights
csc$weight <- (csc$SMK1AGE*(1/csc$pr_score) +
                         (1-csc$SMK1AGE)*(1/(1-csc$pr_score)))

# this is short hand of calculating, using the fact that csc$SMK1AGE*(1/csc$pr_score)==0 for untreated
# and (1-csc$SMK1AGE)*(1/csc$pr_score)==0 for treated group SKK1AGE==1

# Check that sum of weights is sample size*2 (i.e. has mean 2)
sum(csc$weight)
```

```
## [1] 1521.289
```

```
mean(csc$weight)
```

```
## [1] 1.999066
```

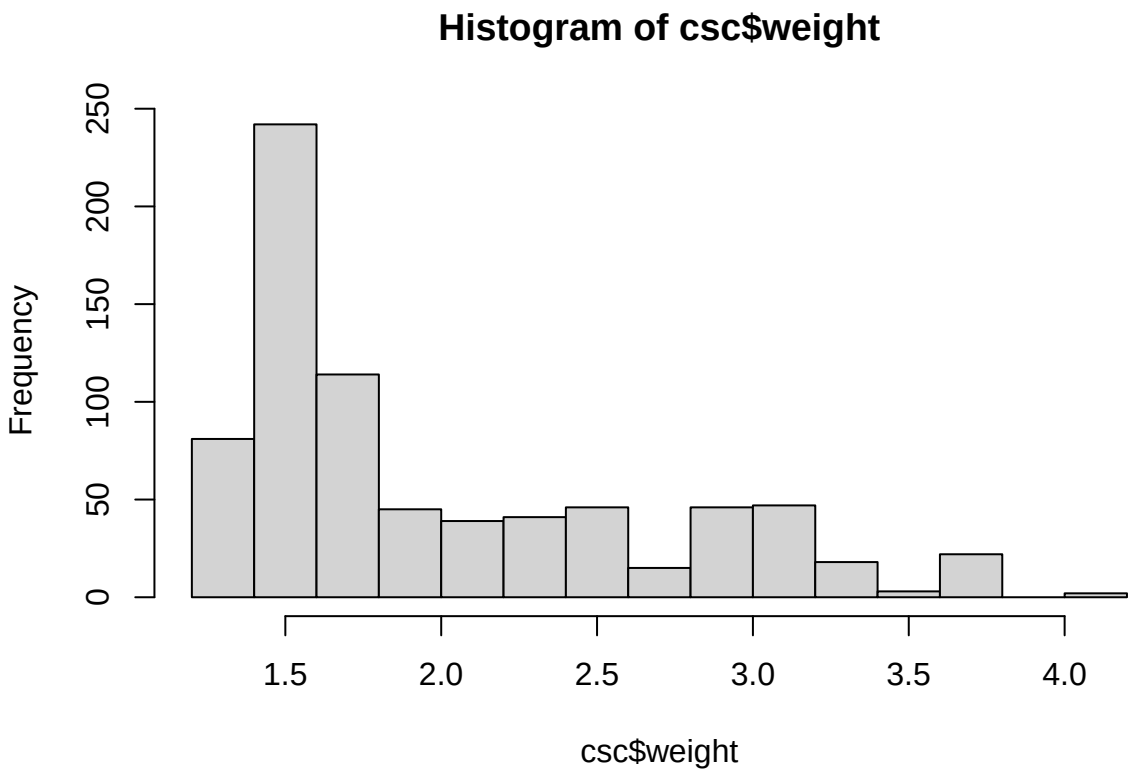
```
nrow(csc)*2
```

```
## [1] 1522
```

```
# Check the distribution of the weights
summary(csc$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.329   1.485   1.712   1.999   2.430   4.041
```

```
hist(csc$weight)
```



3. Estimate the causal effect and interpret. Is this effect marginal or conditional?

The average difference in depression score is 0.0336 comparing had everyone started smoking at $\leq$ 16 years old vs. had everyone started smoking at $>$ 16 years old. This is a marginal effect among the total population.

```
# Fit regression in weighted population with robust standard errors
weighted_model <- geeglm(DEP ~ SMK1AGE , data = csc,
                         weights = weight, family = gaussian, id = psraid, corstr = "ind")

coef(weighted_model)
```

```
## (Intercept)     SMK1AGE
##  1.38956121  0.03357467
```