Biost 544 HW4

John Schoof

12/2/2020

Introduction

The goal of this analysis is to assess the effect of smoking on bone mass density (BMD) in middle aged women. We use the baseline data for a cohort of 3,302 middle aged women from the SWAN observational data. The analysis included 2,118 women with complete data for all covariates and outcome variables. To complete this analysis, I separately assess the effect of smoking on spinal bone mass density and hip bone mass density.

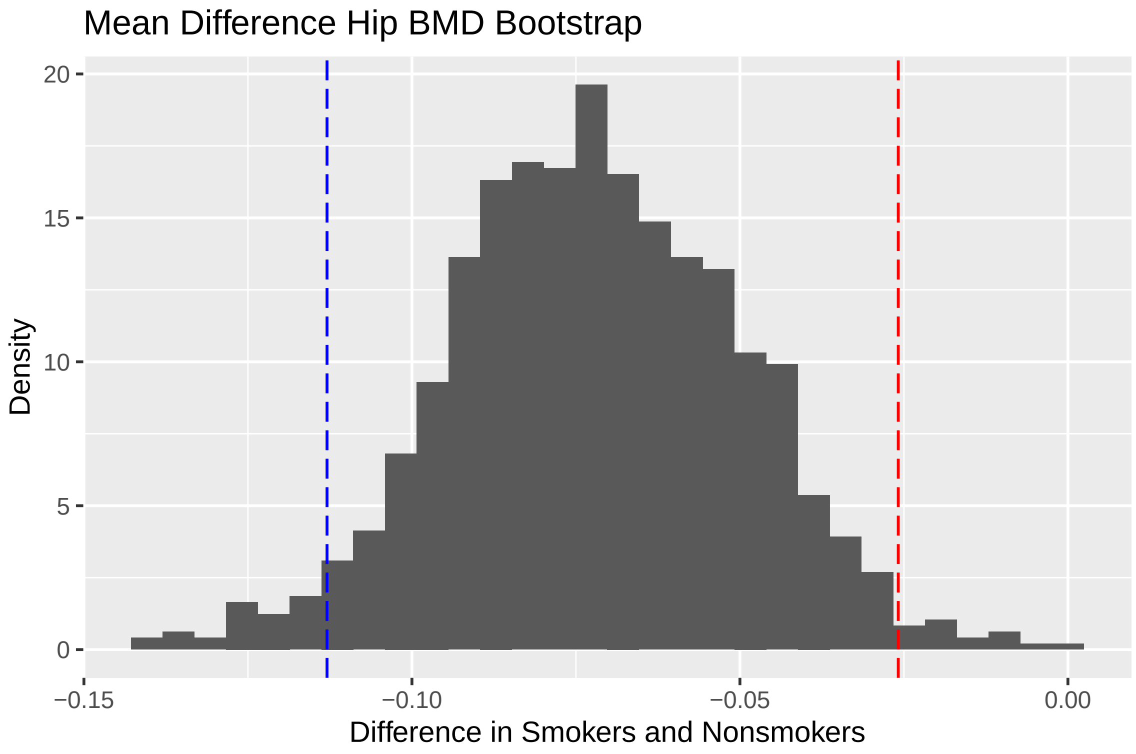
Variables

The exposure variable of interest is a binary measure of having ever been a regular smoker in one's life. I also dichotomized hip and spine BMD at each respective median. Therefore my outcome variables are binary, high/low BMD. I chose to adjust for the following confounding variables: age (continuous), income (categorical), alcohol consumption per week (categorical; 0,1,2,3+). Age is a confounder because older generations are more likely to smoke than younger and older women are more likely to have lower bone mass density. Income is a confounder because those with lower incomes are more likely to smoke and likely more likely to have lower bone mass density. Alcohol consumption is a confounder because those who drink more are more likely to smoke and also more likely to have lower bone mass density.

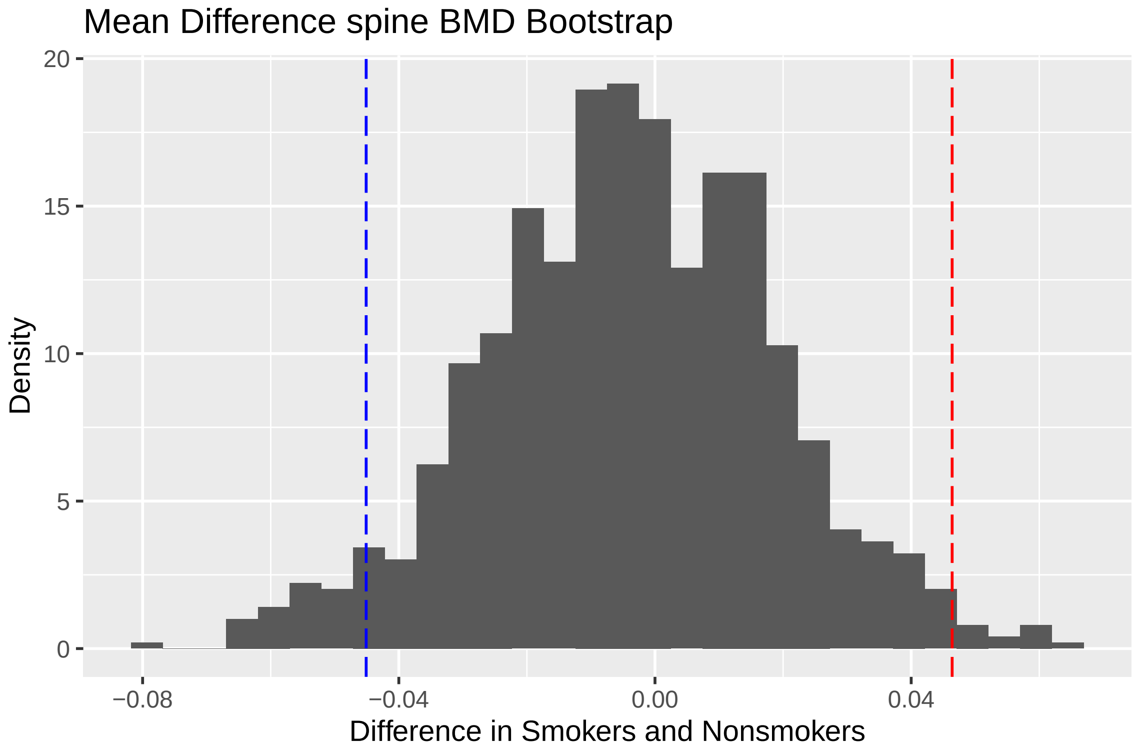
Inverse Probability Weighting Method

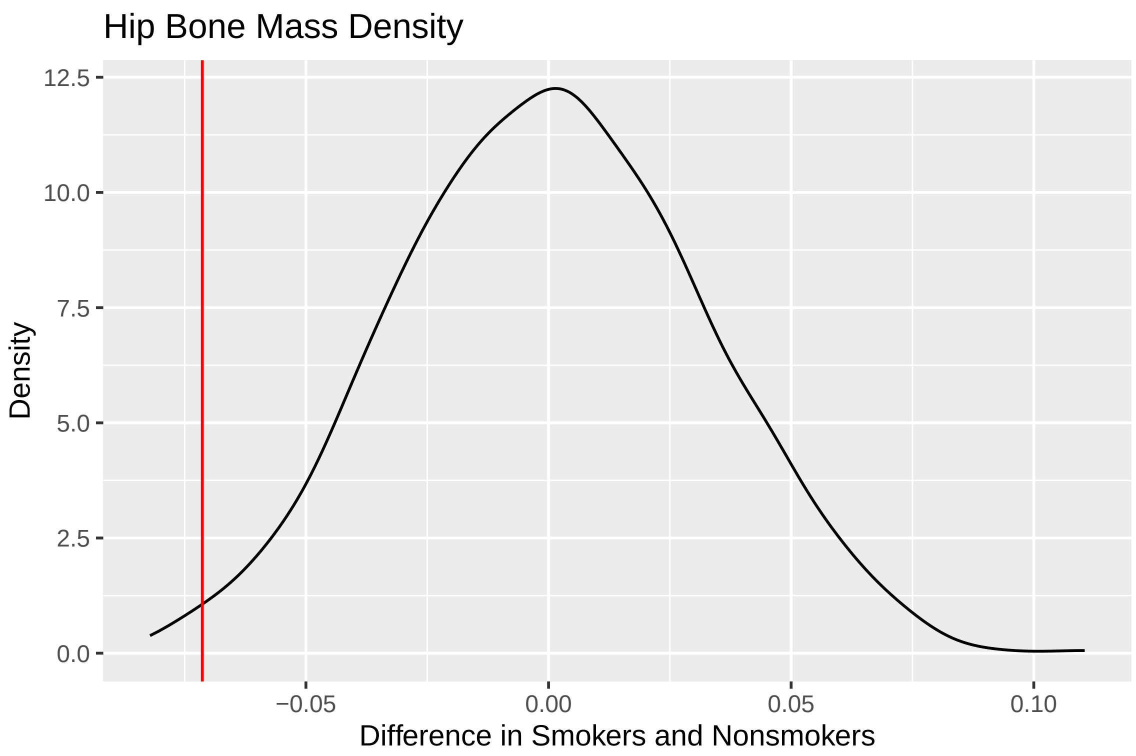
I first use the inverse probability weighting method. I first use a logistic regression model to regress age, income and alcohol on smoke in order to estimate the propensity scores. I then wrote a function to calculate the reweighted difference in mean hip and spine BMD. The estimated difference in mean hip and spine BMDs adjusting for age, income, and alcohol are -.002 and -.07, respectively.
Next, I used the bootstrap method in order to estimate 95% cofidence intervals for each estimate. I write a function to a random sample with replacement and then simulate that 1,000 times to create a sampling distribution. The 2.5th percentile and the 97.5th percentile are the upper and lower bounds, respectively, of the confidence interval. Lastly, we use permutation to create a sampling distribution of outcomes we would expect to see if there were no effect of smoking on BMD. The first figure shows our observed smoking effect on hip BMD lies on the sampling distribution and has a corresponding p-value of 0.008. **We can conclude that there is a significant association between smoking and hip bone mass density.** The second figure shows our observed smoking effect on spine BMD lies on the sampling distribution and has a corresponding p-value of 0.44.

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

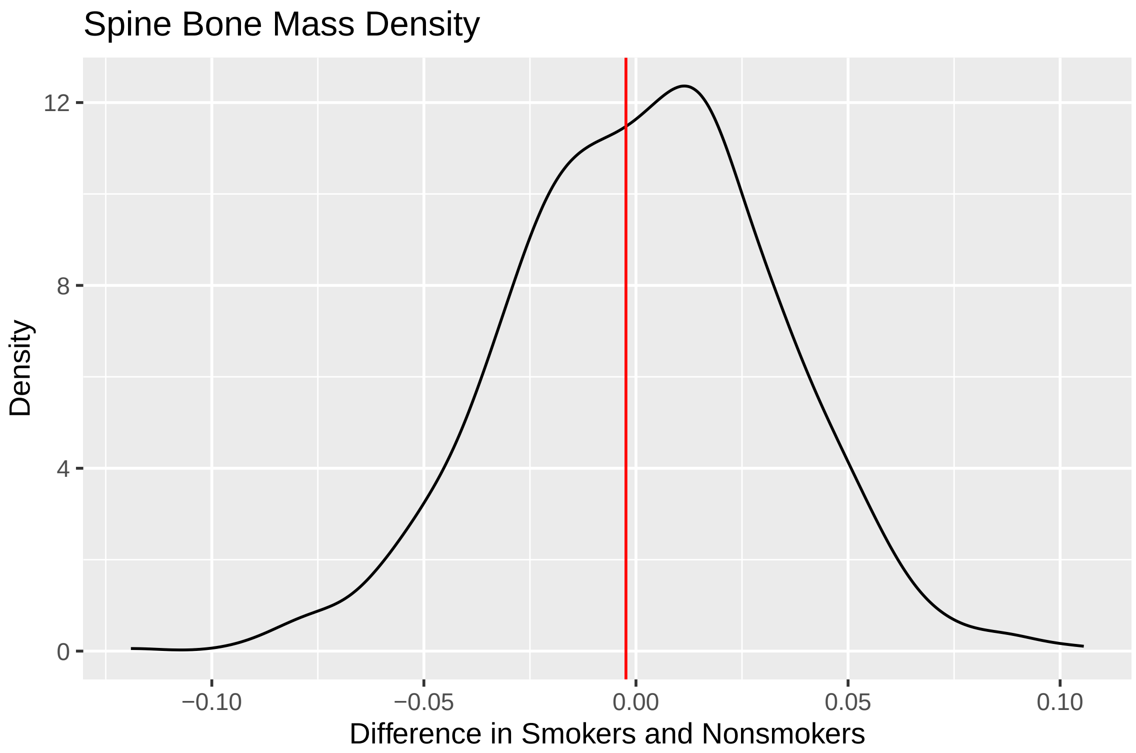


```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




Hip Bone Mass Density
Density
12.5
10.0
7.5
5.0
2.5
0.0
−0.05
0.00
0.05
0.10
Difference in Smokers and Nonsmokers

```
## [1] 0.008
```


Spine Bone Mass Density
Density
12
8
4
0
−0.10
−0.05
0.00
0.05
0.10
Difference in Smokers and Nonsmokers

```
## [1] 0.443
```

Standardization Method

Using the standardization method the estimate the difference in mean hip BMD between smokers and nonsmokers is 0 and the estimate for the difference in mean spine BMD is -0.001.

Appendix: R Code

```r
knitr::opts_chunk$set(echo = TRUE, fig.width=6, fig.height=4)
knitr::opts_knit$set(root.dir = ("/Users/johnschoof/Documents/Analyses/Data Science"))
options(digits = 3) ## Formats output to 3 digits
library(ggplot2)
library(haven)
library(tidyverse)
library(readr)
library(data.table)
library(knitr)
library(glmnet)
library(foreign)

## read in SWAN data
swan <- read_dta("datasets/swan_data.dta")


swan.use <- swan %>% select(SPBMDT0, HPBMDT0, SMOKERE0, AGE0,
                            INCOME0, ALCHWK0, HORMPIL0)

swan.use <- swan.use %>% mutate(hip = HPBMDT0,
                                spine = SPBMDT0,
                                smoke = SMOKERE0,
                                age = AGE0,
                                income = INCOME0,
                                alcohol = ALCHWK0) %>%
                        select(hip, spine, smoke, age,
                               income, alcohol)


swan.use <- swan.use %>% filter(income >= 0,
                                smoke >= 0)
swan.use$smoke[swan.use$smoke==1] <- 0
swan.use$smoke[swan.use$smoke==2] <- 1
data <- swan.use %>% na.omit()

## Dichotomize BMD vars
data <- data %>% mutate(hip.bin = (hip < median(hip)),
                        spine.bin = (spine < median(spine)))

(med.hip <- median(median(data$hip)))
(med.spine <- median(median(data$spine)))


# propensity of being included in sample aka propensity of getting treatment
# thats why the glm model is regressed on smoking
# propensity score captures the adjustment of confounders
# as opposed to regression model adjusting for these confounders
# we're treating whether someone is a smoker or non-smoker as a function of these other confounding var
# propensity score tries to represent randomization if it were a randomized trial

propen.model <- glm(smoke ~ age + income + alcohol, family = binomial(), data = data)

propensities <- predict(propen.model, data = data, type = "response")

trunc.prop <- propensities %>% pmax(0.05) %>% pmin(0.95)
# the denominators can become zero which would give us unstable results
# a bit ad hoc


##
calc_weighted_outcome <- function(outcome, label, props){
  weights <- rep(0, length(outcome))

  representative.propen <- mean(label)
  actual.propen <- props

  treat.ind <- which(label == 1)
  weights[treat.ind] <- representative.propen/actual.propen[treat.ind]
  weights[-treat.ind]<- (1 - representative.propen)/(1- actual.propen[-treat.ind])

  weighted.outcome <- outcome*weights

  return(weighted.outcome)
}

calc_stat_weighted <- function(weighted.outcome, label){
  return(mean(weighted.outcome[label == 1]) - mean(weighted.outcome[label == 0]))
}
## instead of using the original outcome we are reweighting the outcome to account for the fact that sm

wt.outcome.spine <- calc_weighted_outcome(data$spine.bin, data$smoke, trunc.prop)

(mean.diff.spine <- calc_stat_weighted(wt.outcome.spine, data$smoke))
## instead of using the original outcome we are reweighting the outcome to account for the fact that sm

wt.outcome.hip <- calc_weighted_outcome(data$hip.bin, data$smoke, trunc.prop)

(mean.diff.hip <- calc_stat_weighted(wt.outcome.hip, data$smoke))
# resampling the data based on same size with replacement
# the rest is same as before
# "pretending" that our sample is entire population
#

do_one <- function(dat){
  resample.inds <- sample(1:nrow(dat), replace = TRUE)
  resample.dat <- dat[resample.inds,]

  propen.model <- glm(smoke ~ age + income + alcohol,
                      family = binomial(), data = resample.dat)
  propensities <- predict(propen.model, data = resample.dat, type = "response")
  trunc.prop <- propensities %>% pmax(0.05) %>% pmin(0.95)

  wt.outcome.hip.boot <- calc_weighted_outcome(resample.dat$hip.bin,
                                                   resample.dat$smoke,
                                                   trunc.prop)
  wt.outcome.spine.boot <- calc_weighted_outcome(resample.dat$spine.bin,
                                                   resample.dat$smoke,
                                                   trunc.prop)
  mean.diff.hip.boot <- calc_stat_weighted(wt.outcome.hip.boot, resample.dat$smoke)
  mean.diff.spine.boot <- calc_stat_weighted(wt.outcome.spine.boot, resample.dat$smoke)



  return(c(mean.diff.hip.boot, mean.diff.spine.boot))
}
set.seed(2)
boot.dist <- replicate(1e3,
                    do_one(data))

boot.data <- data.frame(t(boot.dist))
colnames(boot.data) <- c("mean.diff.hip.boot", "mean.diff.spine.boot")

## how does your interpretation of these CIs change

distance.U.L <- quantile(boot.data$mean.diff.hip.boot, c(0.025,0.975)) - mean.diff.hip
CI.hip <- mean.diff.hip - distance.U.L[2:1]

distance.U.L <- quantile(boot.data$mean.diff.spine.boot, c(0.025,0.975)) - mean.diff.spine
CI.spine <- mean.diff.spine - distance.U.L[2:1]

ggplot(data = boot.data, aes(x = mean.diff.hip.boot, y=..density..)) +
  geom_histogram()+
  geom_vline(xintercept = CI.hip[1], color = "blue", linetype = "longdash") +
  geom_vline(xintercept = CI.hip[2], color = "red", linetype = "longdash") +
  labs(x = "Difference in Smokers and Nonsmokers",
                   y = "Density",
                   title = "Mean Difference Hip BMD Bootstrap")

ggplot(data = boot.data, aes(x = mean.diff.spine.boot, y=..density..)) +
  geom_histogram()+
  geom_vline(xintercept = CI.spine[1], color = "blue", linetype = "longdash") +
  geom_vline(xintercept = CI.spine[2], color = "red", linetype = "longdash") +
  labs(x = "Difference in Smokers and Nonsmokers",
                   y = "Density",
                   title = "Mean Difference spine BMD Bootstrap")
do.one.propen <- function(outcome, propen){
  n <- length(outcome)
  label <- rbinom(n,1,propen)

  weights <- rep(0,n)
  representative <- mean(label)
  actual <- propen
  ind.t <- which(label == 1)
  weights[ind.t] <- (representative/actual)[ind.t]
  weights[-ind.t] <- ((1-representative)/(1-actual))[-ind.t]

  return(mean((weights*outcome)[ind.t]) - mean((weights*outcome)[-ind.t]))
}

set.seed(2)
rerandomized.diffs.hip <-
  replicate(1e3, do.one.propen(data$hip.bin, trunc.prop))
rerandomized.diffs.spine <-
  replicate(1e3, do.one.propen(data$spine.bin, trunc.prop))
ggplot(data.frame(diffs = rerandomized.diffs.hip), aes(x = diffs, y = ..density..)) +
  geom_density() +
  geom_vline(xintercept = mean.diff.hip, color = "red") +
  labs(x = "Difference in Smokers and Nonsmokers",
                   y = "Density",
                   title = "Hip Bone Mass Density")

mean(rerandomized.diffs.hip < mean.diff.hip)


ggplot(data.frame(diffs = rerandomized.diffs.spine), aes(x = diffs, y = ..density..)) +
  geom_density() +
  geom_vline(xintercept = mean.diff.spine, color = "red") +
  labs(x = "Difference in Smokers and Nonsmokers",
                   y = "Density",
                   title = "Spine Bone Mass Density")

mean(rerandomized.diffs.spine < mean.diff.spine)

outcome.reg.hip <- glm(hip.bin ~ smoke + age + income + alcohol,
                          family=binomial, data=data)

hip.smoker <- data %>% mutate(smoke == 1)
hip.nonsmoker <- data%>% mutate(smoke == 0)

(standardized.est.hip <- mean(  predict(outcome.reg.hip,
                                   hip.smoker,
                                   type = "response") -
                             predict(outcome.reg.hip,
                                     hip.nonsmoker,
                                     type = "response")))


outcome.reg.spine <- glm(spine.bin ~ smoke + age + income + alcohol,
                          family=binomial, data=data)

spine.smoker <- data %>% mutate(smoke = 1)
spine.nonsmoker <- data%>% mutate(smoke = 0)

(standardized.est.spine <- mean(  predict(outcome.reg.spine,
                                   spine.smoker,
                                   type = "response") -
                             predict(outcome.reg.spine,
                                     spine.nonsmoker,
                                     type = "response")))
```