Causal Inference Final Project: Effect of Smoking on 10-year Development of Coronary Heart Disease

Bianca Doone, Michael Attah, Graham Casey Gibson, Daniel Saunders, Nutcha Wattanachit

November 25, 2018

Background Story

Coronary heart disease (CHD) is the leading cause of death and serious illness in the United States. The Framingham Heart Study's objective was to identify the common factors or characteristics that contribute to CHD by following its development over time in a large group of participants who had not yet developed overt symptoms of CHD or suffered a heart attack or stroke.

The researchers recruited 5,209 men and women between the ages of 30 and 70 from Framingham, Massachusetts, and began the first round of extensive physical examinations and lifestyle interviews that they would later analyze for common patterns related to CHD development. Over the years, careful monitoring of the Framingham Study population has led to the identification of the major CHD risk factors – high blood pressure, high blood cholesterol, smoking, obesity, diabetes, and physical inactivity. We are interested how the extent of smoking affects the development of CHD, specifically, it is not immediately obvious whether smoking 5 cigarettes per day affects the development of CHD differently than smoking 15 cigarettes per day does.

Causal Roadmap

Step 0: Specify the Scientific Question

What is the effect of smoking on the ten-year development of Coronary Heart Disease?

Target population

The target population is white middle-class men and women aged 30 to 70 in the US.

The sameple in this study is white middle-class men and women aged 30 to 70 (at baseline) in Framingham, Massachusetts. The importance of the major CHD risk factors identified in this group have been shown in other studies to apply almost universally, even though the patterns of distribution may vary. Thus, we are willing to generalize to the target population.

Step 1: Specify a Causal Model

- Endogenous nodes: $X = (W, Z, A, Y)$, where
- W is age, gender, education
- Z is blood pressure (systolic and diastolic), total Cholesterol, prevalence of hypertension, prevalence of stroke, heart rate, BMI, Diabetes prevalence
- A is the number of cigarettes smoked per days
- Y is the ten-year development of coronary heart disease (CHD).

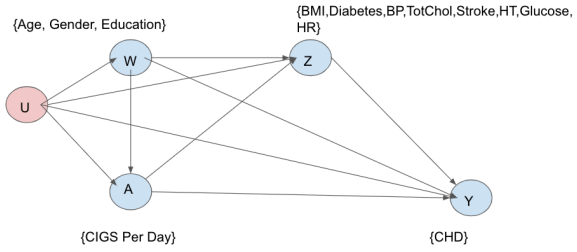


Figure 1: Causal Graph for the SCM

- Exogenous nodes: $U = (U_W, U_Z, U_A, U_Y) \sim \mathbb{P}_U$. We make no assumptions about the distribution $\mathbb{P}_U$.
- Structural equations F:

$$\begin{aligned} W &\leftarrow f_W(U_W) \\ Z &\leftarrow f_Z(W, A, U_Z) \\ A &\leftarrow f_A(W, U_A) \\ Y &\leftarrow f_Y(W, Z, A, U_Y) \end{aligned}$$

There are no exclusion restrictions or assumptions about functional form.

Causal Graph

NEEDS TO BE FIXED!

Step 2: Counterfactuals & Causal Parameter

Causal Parameter

$$\Psi^{*i}(\mathbb{P}^*) = \mathbb{E}^*[Y_i] \quad i \in \{1, 2, 3, 4\}$$

where i represent the bin of cigarettes smoked per day. Y_i denotes the counterfactual outcome (the ten-year development of cardiovascular disease), if possibly contrary to fact, a person's number of cigarettes smoked per day is within i^{th} bin.

Step 3. Specify your observed data and its link to the causal model

The dataset is adapted from Framingham Heart Study. We assume that Gender, Age, Education, and Number of Cigarettes per Day (A) were collected in a questionnaire at baseline. Then, BMI, Diabetes Status, Prevalence of Stroke, Prevalence of Hypertension, Indication of Blood Pressure Medication, Total Cholesterol Level, Blood Pressure, and Heart Rate were all collected after the questionnaire at a doctor's office. Our outcome, Coronary Heart Disease, is collected at a 10-year follow up. Note that this is unlike the original study. We assume our observed data were generated by sampling n from a system described by our structural causal model, so we have $n = 4211$ copies of $O \overset{i.i.d}{\sim} \mathbb{P}_O$. We place no restrictions on the statistical model $\mathcal{M}$, which is thereby non-parametric. BMI was binned using guidelines from the World Health Organization. Total Cholesterol was binned using guidelines from the National Heart, Lung and Blood Institute (NHLBI). Table 1 below shows the counts for each variable in each bin of the exposure, as well as a χ^2-test of independence.

Table 1: Number of Observations in Each Bin

	Level	[0,1)	[1,10)	[10,19)	[20,70]	p
n		2089	471	380	1168	
diabetes (%)	0	2022 (96.8)	463 (98.3)	372 (97.9)	1145 (98.0)	0.079
	1	67 (3.2)	8 (1.7)	8 (2.1)	23 (2.0)	
prevalentStroke (%)	0	2071 (99.1)	469 (99.6)	377 (99.2)	1166 (99.8)	0.095
	1	18 (0.9)	2 (0.4)	3 (0.8)	2 (0.2)	
prevalentHyp (%)	0	1337 (64.0)	338 (71.8)	293 (77.1)	860 (73.6)	<0.001
	1	752 (36.0)	133 (28.2)	87 (22.9)	308 (26.4)	
age (%)	[32, 42)	363 (17.4)	117 (24.8)	110 (28.9)	309 (26.5)	<0.001
	[42, 49)	464 (22.2)	137 (29.1)	125 (32.9)	401 (34.3)	
	[49, 56)	527 (25.2)	97 (20.6)	71 (18.7)	254 (21.7)	
	[56, 70]	735 (35.2)	120 (25.5)	74 (19.5)	204 (17.5)	
education (%)	1	915 (43.8)	188 (39.9)	137 (36.1)	470 (40.2)	0.003
	2	574 (27.5)	150 (31.8)	121 (31.8)	399 (34.2)	
	3	367 (17.6)	76 (16.1)	71 (18.7)	170 (14.6)	
	4	233 (11.2)	57 (12.1)	51 (13.4)	129 (11.0)	
BP (%)	0	1139 (54.5)	267 (56.7)	211 (55.5)	701 (60.0)	0.025
	1	950 (45.5)	204 (43.3)	169 (44.5)	467 (40.0)	
totChol (%)	[0, 200)	386 (18.6)	108 (23.4)	84 (22.3)	227 (19.7)	0.004
	[200, 240)	723 (34.9)	151 (32.7)	157 (41.8)	401 (34.9)	
	[240, 600]	962 (46.5)	203 (43.9)	135 (35.9)	522 (45.4)	
gender (%)	0	1400 (67.0)	346 (73.5)	222 (58.4)	386 (33.0)	<0.001
	1	689 (33.0)	125 (26.5)	158 (41.6)	782 (67.0)	
bmi (%)	[0, 18.5)	19 (0.9)	13 (2.8)	7 (1.8)	17 (1.5)	<0.001
	[18.5, 25)	785 (37.8)	245 (52.2)	225 (59.2)	568 (48.8)	
	[25, 30)	940 (45.3)	170 (36.2)	117 (30.8)	467 (40.1)	
	[30, 56.8]	333 (16.0)	41 (8.7)	31 (8.2)	112 (9.6)	
heartRate (%)	[0, 60)	122 (5.8)	27 (5.7)	19 (5.0)	28 (2.4)	<0.001
	[60, 143]	1967 (94.2)	444 (94.3)	360 (95.0)	1140 (97.6)	
CHD (%)	0	1784 (85.4)	420 (89.2)	320 (84.2)	958 (82.0)	0.002
	1	305 (14.6)	51 (10.8)	60 (15.8)	210 (18.0)	

Step 4. Identifiability

Since we made no independence assumptions on our exogenous background factors, we will need to make additional independence assumptions for identifiability. For the target causal parameter in the SCM $\mathcal{M}^*$ to be identified from the observed data distribution, we need to make a randomization and a positivity assumption.

1) Randomization Assumption:

Conditional on W, the counterfactual outcome is independent of the observed treatment:

$$Y \perp A|W$$

$$U_A \perp U_Y, U_A \perp U_W, U_A \perp U_Z$$

In this augmented/working SCM ($\mathcal{M}^{**}$), the unmeasured background factor of A (cigarettes smoked per day) is independent of the unmeasured background factor of Y (10yr CHD), the unmeasured background

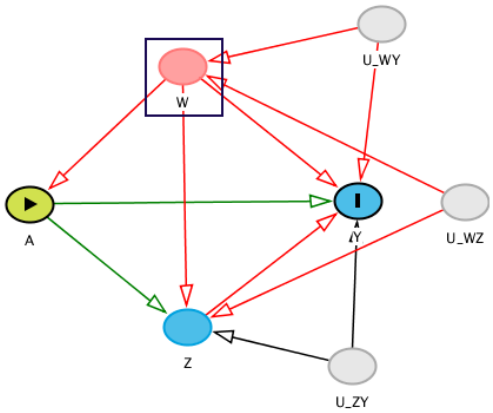


Figure 2: Causal Graph for the SCM

factor of W (baseline age, gender, education, diabetes, BMI), and the unmeasured background factor of Z (prevalence of stroke, hypertension, blood pressure, blood pressure medication, heart rate).

Since W,Z, and Y include SES and biological factors that affect human health, we aviod assuming independence between their unmeasured background factors. Thus, we consider it is more plausible to make the indepedence assumptions listed. We do not the mediator Z to avoid opening a backdoor path. Under M^*∗, the backdoor criterion holds conditional on W. Additional data on factors that affect heath status and determinants could help with identifiability, but those factors are not well-understood.

2) Positivity Assumption:

There must be a positive probability of each treatment condition within each possible strata of W. We need a positive probability of cigarettes smoked per day for each strata of W.

$$min_{i \in A} \mathbb{P}_0(A = i | W = w) > 0$$
$$\text{for all } w \text{ for which} \mathbb{P}_0(W = w) \geq 0$$

where i denote the index of a bin of A.

THIS NEEDS TO BE FIXED!!!!

We are concerned about a positivity assumption violation for a bin with high numbers of cigarettes smoked per day since certain strata of W might not smoke at all, and binning could make particular stratas have low probabilities of smoking certain numbers of cigarettes per day. We can informally check for a positivity assumption violation from tables of A given a strata of W. One table from a strata of W is shown below:

From the table above, we can see that we have 0 female whose BMI is in the range [18.5, 25), age is in the range of [42, 49), education level is missing, who does not have Diabetes, who does not smoke cigarettes. Thus, we have a sparsity issue.

Table 2: Table for BMI = [18.5, 25), Gender = 0, Diabetes = 0, A = 0

Strata	Education: 1	Education: 2	Education: 3	Education: 4	Education: NA
Age: [32, 42)	26	60	32	12	1
Age: [42, 49)	40	42	41	18	0
Age: [49, 56)	43	52	42	17	5
Age: [56, 70]	75	41	36	16	2

NEED TO FIX ABOVE!!!!

We also calculated the predicted probability of each bin of A (the number of cigarettes smoked per day) given a strata of W:

	c.0.285367089748959..0.0612950549329788..0.0592083666852972..	c.0.427146494355085..0.0743610835769698..0.06809
i=1	0.2853671	
i=2	0.0612951	
i=3	0.0592084	
i=4	0.0765399	

From the table, we can see that the probabiliy of falling into the last bin of A, which is ≥ 20 cigarettes smoked per day, is very close to zero, indicating a practical violation of positivity assumption. Theoretically, randomizing the number of cigarettes smoked per day could help with identifiability, but it is not feasible.

Step 5. Statistical Model and Estimand

The target parameter of $\mathbb{P}_0$, which equals the causal parameter in the augmented causal model $\mathcal{M}^{**}$ is given by the G-Computation formula:

$$\begin{aligned}\Psi_0(\mathbb{P}_0^i) &= \mathbb{E}_0[\mathbb{E}_0[Y|A=i,W=w]] \\ &= \sum_w \mathbb{E}_0[Y|A=i,W=w] * \mathbb{P}_0(W=w)\end{aligned}$$

Step 6. Estimation

Conditional Mean outcome

```
library(mgcv)
```

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
## This is mgcv 1.8-20. For overview type 'help("mgcv-package")'.
```

```
##
## Attaching package: 'mgcv'
```

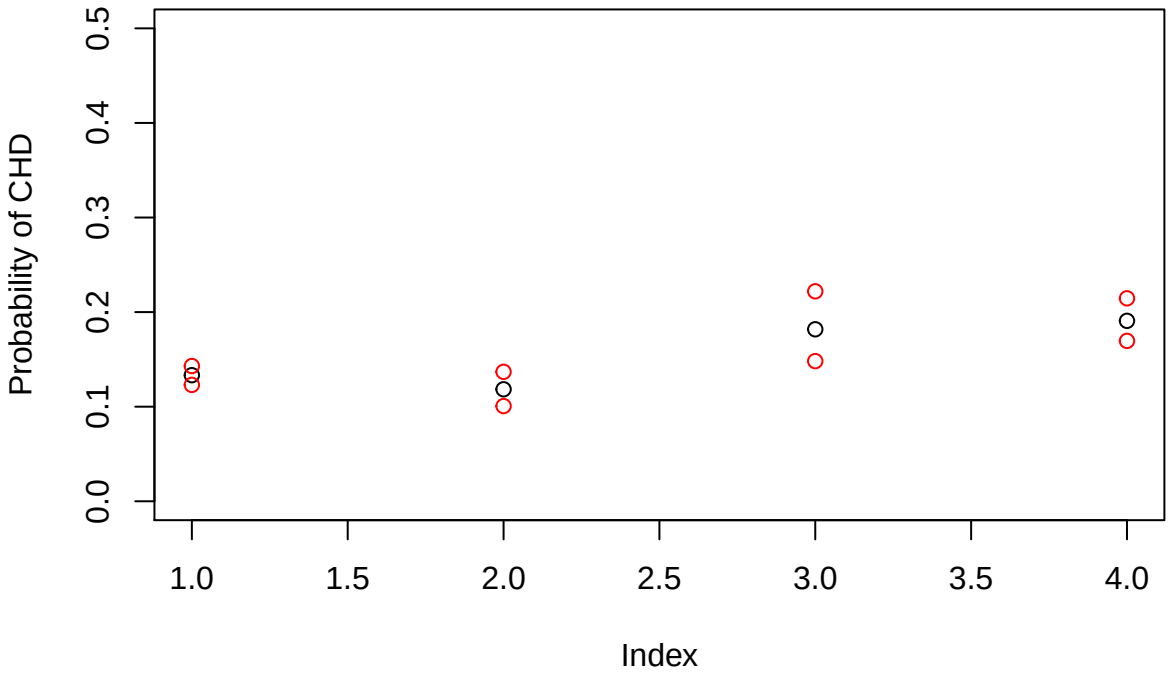
```
## The following object is masked from 'package:nnet':
##
##     multinom
```

```
n_bootstrap_samples <-10
conMean_bootstrap <- matrix(NA, nrow=n_bootstrap_samples,ncol=4)
```

```
for(i in 1:n_bootstrap_samples){
  set.seed(i)
  fhs_binned_sample <- fhs_binned[sample(1:nrow(fhs_binned),nrow(fhs_binned),replace=TRUE),]

  intervene_on_bin <- function(i){
  fhs_binned_i <- fhs_binned_sample
  fhs_binned_i$cigsPerDay <-levels(fhs_binned_sample$cigsPerDay)[i]
  return (fhs_binned_i)
}



## SATURED REGRESSION MODEL FOR NPMLE
glm_fit <- glm( CHD ~ cigsPerDay*education*age*gender, data = fhs_binned_sample, family = "binomial")

for (j in 1:length(levels(fhs_binned$cigsPerDay))){
  conMean_bootstrap[i,j] <- mean(predict(glm_fit, newdata=intervene_on_bin(j), type='response'))
}
}
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

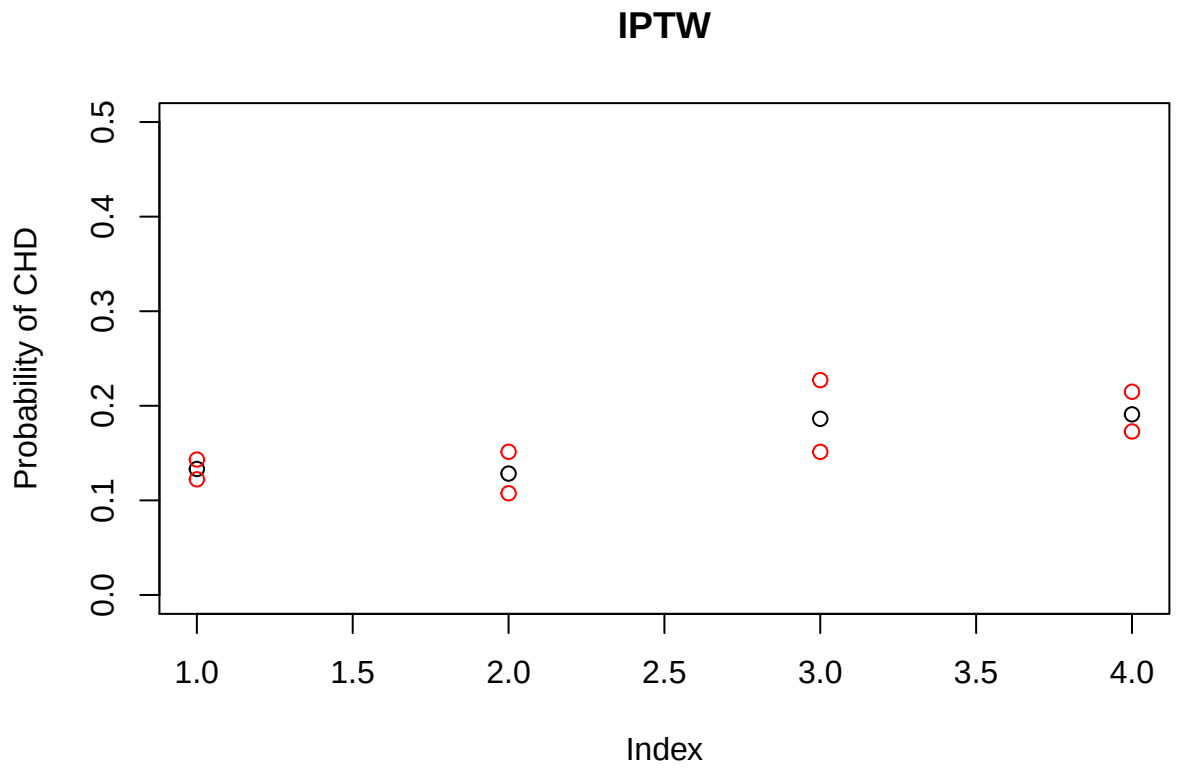
```
average_treatment_effect <- c()
average_treatment_effect_ci <- matrix(NA,nrow=length(levels(fhs_binned_sample$cigsPerDay)),ncol=2)
for (i in 1:length(levels(fhs_binned_sample$cigsPerDay))){
  average_treatment_effect[i] <- colMeans(conMean_bootstrap)[i]
  average_treatment_effect_ci[i,] <- quantile(conMean_bootstrap[,i],probs=c(.025,.975))

}

plot(average_treatment_effect,ylab="Probability of CHD",ylim=c(0,.5),
     main = "Simple Substitution")
points(average_treatment_effect_ci[,1],col='red',lty=2)
points(average_treatment_effect_ci[,2],col='red',lty=2)
```



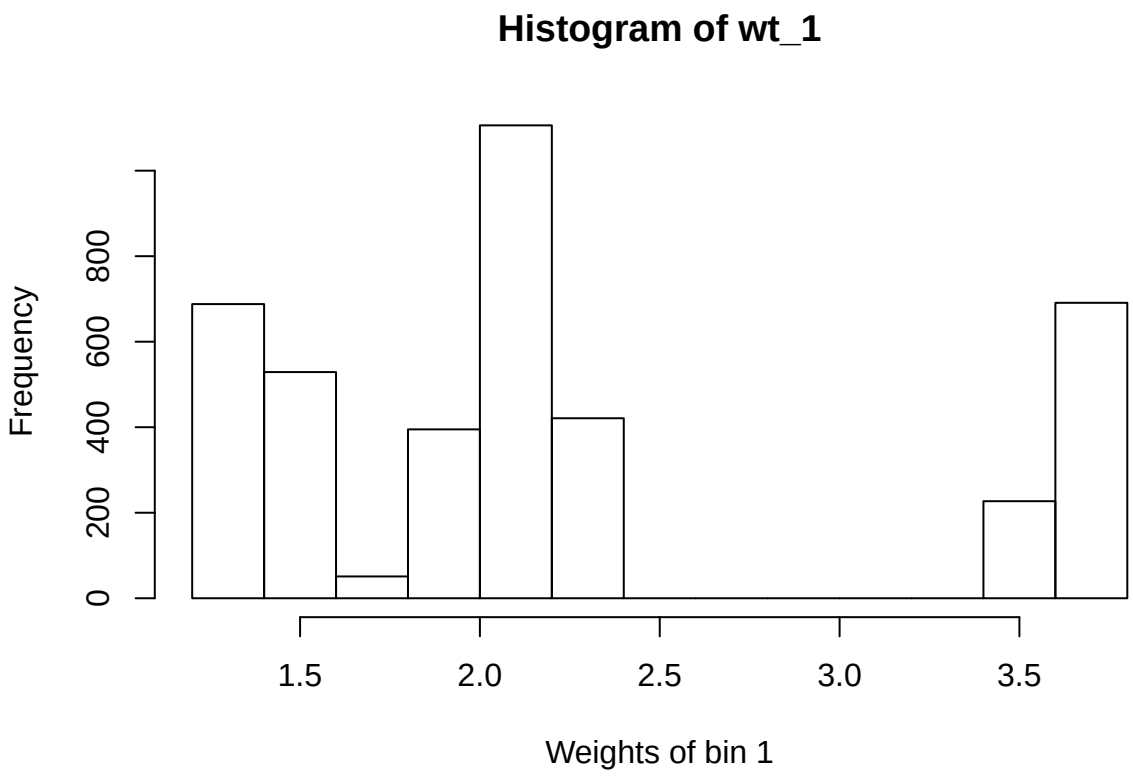
IPTW

```
n_bootstrap_samples <-10
iptw_bootstrap <- matrix(NA, nrow=n_bootstrap_samples,ncol=4)

for (i in 1:n_bootstrap_samples){
  set.seed(i)
  fhs_binned_sample <- fhs_binned[sample(1:nrow(fhs_binned),nrow(fhs_binned),replace=TRUE),]

### BIN 1
glm_fit_iptw_bin_1 <- glm( cigsPerDay_bin_1 ~   education + age + gender, data = fhs_binned_sample, fami
prob.1W <- predict(glm_fit_iptw_bin_1, type= "response")
wt_1<- 1/prob.1W
summary(wt_1)
IPTW_bin_1<- mean( wt_1*as.numeric(fhs_binned_sample$cigsPerDay_bin_1==1)*as.numeric(fhs_binned_sample$


### BIN 2
glm_fit_iptw_bin_2 <- glm( cigsPerDay_bin_2 ~   education + age + gender, data = fhs_binned_sample, fami
prob.1W <- predict(glm_fit_iptw_bin_2, type= "response")
wt_2<- 1/prob.1W
summary(wt_2)
IPTW_bin_2<- mean( wt_2*as.numeric(fhs_binned_sample$cigsPerDay_bin_2==1)*as.numeric(fhs_binned_sample$

### BIN 3
glm_fit_iptw_bin_3 <- glm( cigsPerDay_bin_3 ~  education + age + gender, data = fhs_binned_sample, famil
prob.1W <- predict(glm_fit_iptw_bin_3, type= "response")
wt_3<- 1/prob.1W
```

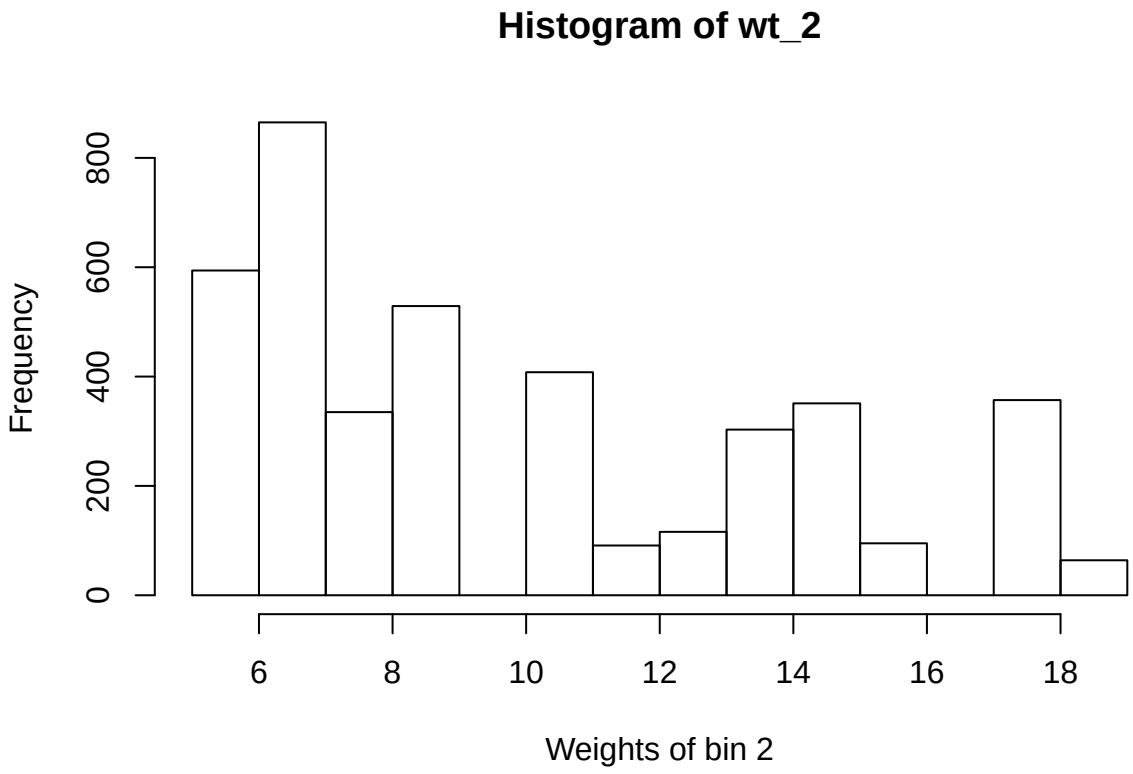
```r
summary(wt_3)
IPTW_bin_3<- mean( wt_3*as.numeric(fhs_binned_sample$cigsPerDay_bin_3==1)*as.numeric(fhs_binned_sample$


### BIN 4
glm_fit_iptw_bin_4 <- glm( cigsPerDay_bin_4 ~  education + age + gender, data = fhs_binned_sample, famil
prob.1W <- predict(glm_fit_iptw_bin_4, type= "response")
wt_4<- 1/prob.1W
summary(wt_4)
IPTW_bin_4<- mean( wt_4*as.numeric(fhs_binned_sample$cigsPerDay_bin_4==1)*as.numeric(fhs_binned_sample$

iptw_bootstrap[i,] = c(IPTW_bin_1, IPTW_bin_2, IPTW_bin_3, IPTW_bin_4)
}

average_treatment_effect <- c()
average_treatment_effect_ci <- matrix(NA,nrow=length(levels(fhs_binned_sample$cigsPerDay)),ncol=2)
for (i in 1:length(levels(fhs_binned_sample$cigsPerDay))){
  average_treatment_effect[i] <- colMeans(iptw_bootstrap)[i]
  average_treatment_effect_ci[i,] <- quantile(iptw_bootstrap[,i],probs=c(.025,.975))

}

plot(average_treatment_effect,ylab="Probability of CHD",ylim=c(0,.5),
     main = "IPTW")
points(average_treatment_effect_ci[,1],col='red',lty=2)
points(average_treatment_effect_ci[,2],col='red',lty=2)
```



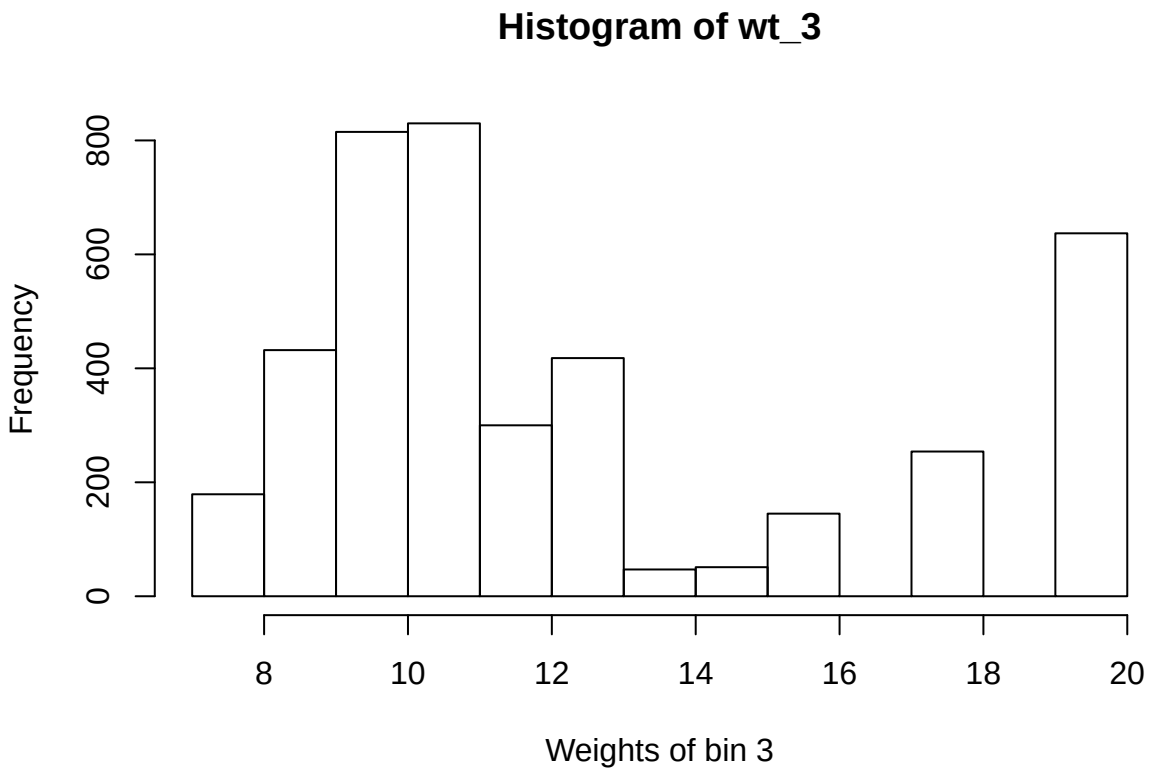

```r
hist(wt_1,xlab="Weights of bin 1")
```



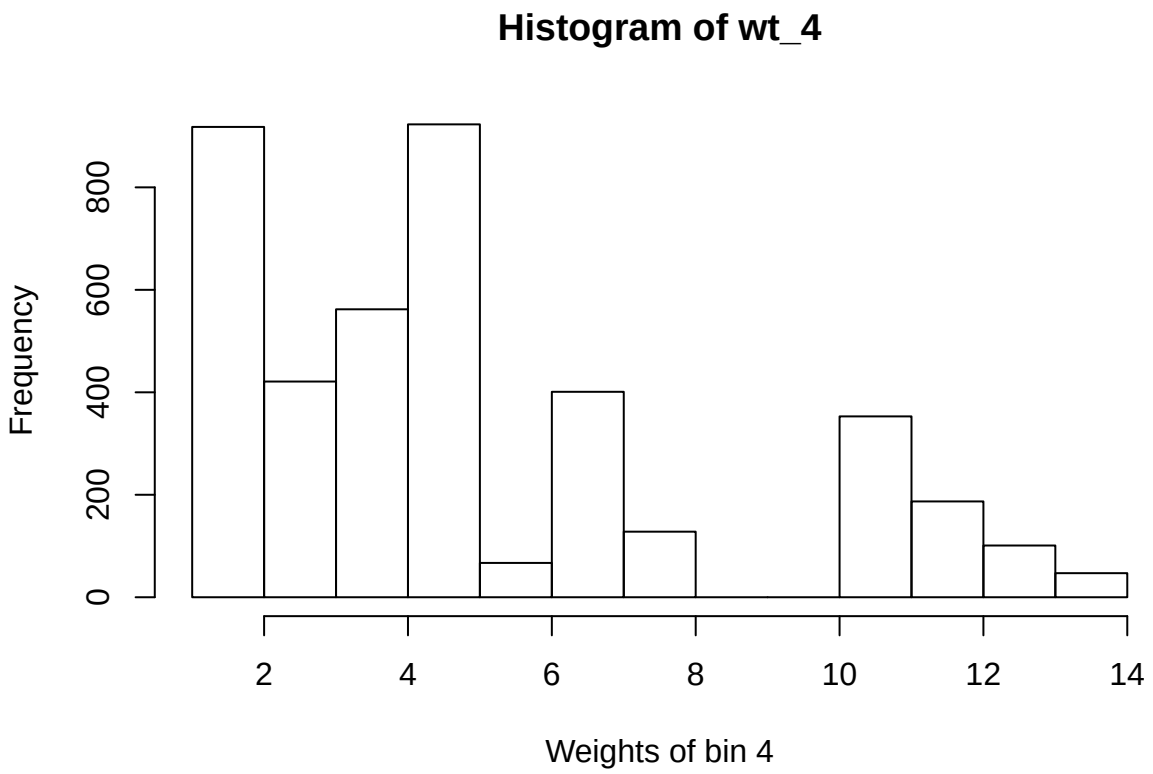

```
hist(wt_2,xlab="Weights of bin 2")
```




```
hist(wt_3,xlab="Weights of bin 3")
```




```
hist(wt_4,xlab="Weights of bin 4")
```



Superlearner/TMLE

```
library('SuperLearner')
```

```
## Warning: package 'SuperLearner' was built under R version 3.4.4
```

```
## Loading required package: nnls
```

```
## Super Learner
```

```
## Version: 2.0-24
```

```
## Package created on 2018-08-10
```

```
SL.library<- c("SL.glmnet","SL.randomForest","SL.nnet","SL.earth","SL.bayesglm")

n_bootstrap_samples <-1
tmle_bootstrap_bin_1 <- c()

for (i in 1:n_bootstrap_samples){
  set.seed(i)
  ### BIN1 TMLE
  fhs_binned_sample <- fhs_binned[sample(1:nrow(fhs_binned),nrow(fhs_binned),replace=TRUE),]

  X_minus_bin_1<- subset(fhs_binned_sample, select= c("cigsPerDay_bin_1", "education", "age","gender") )
  X_minus_bin_1$age <- as.numeric(X_minus_bin_1$age)
  X_minus_bin_1_all_bin_1 <- X_minus_bin_1
  X_minus_bin_1_all_bin_1$cigsPerDay_bin_1 <- 1

  ##CONVERT FACTORS TO NUMERIC FOR SUPERLEARNER


  SL.outcome<- SuperLearner(Y=as.numeric(fhs_binned_sample$CHD==1), X=X_minus_bin_1, SL.library=SL.libra


  expY.givenA1 <-  predict(SL.outcome, newdata=X_minus_bin_1_all_bin_1)$pred
  SL.exposure<- SuperLearner(Y=as.numeric(fhs_binned$cigsPerDay_bin_1==1), X=subset(X_minus_bin_1, selec
  probA1.givenW<- SL.exposure$SL.predict
  H.AW<- as.numeric(fhs_binned$cigsPerDay_bin_1==1)/probA1.givenW
  logitUpdate<- glm(fhs_binned$CHD ~ -1 +offset(qlogis(expY.givenA1)) + H.AW, family='binomial')
  epsilon<- logitUpdate$coef
  expY.givenAW.star<- plogis(qlogis(expY.givenA1)+ epsilon*H.AW)
  PsiHat.TMLE_bin_1<- mean(expY.givenAW.star)#- expY.givenOW.star)
  tmle_bootstrap_bin_1[i] <- PsiHat.TMLE_bin_1
}
```

```
## Loading required package: arm
```

```
## Warning: package 'arm' was built under R version 3.4.4
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Loading required package: Matrix
## Loading required package: lme4
## Warning: package 'lme4' was built under R version 3.4.4
##
## Attaching package: 'lme4'
## The following object is masked from 'package:nlme':
##
##     lmList
##
## arm (Version 1.10-1, built: 2018-4-12)
## Working directory is /Users/gcgibson/causal_project/finalproj
## Loading required package: glmnet
## Loading required package: foreach
## Loaded glmnet 2.0-12
## Loading required package: randomForest
## Warning: package 'randomForest' was built under R version 3.4.4
## randomForest 4.6-14
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
## The following object is masked from 'package:dplyr':
##
##     combine
## Loading required package: earth
## Warning: package 'earth' was built under R version 3.4.4
## Loading required package: plotmo
## Warning: package 'plotmo' was built under R version 3.4.4
## Loading required package: plotrix
##
## Attaching package: 'plotrix'
## The following object is masked from 'package:arm':
##
##     rescale
## Loading required package: TeachingDemos
```

```
  #### BIN 2 TMLE

tmle_bootstrap_bin_2 <- c()
for (i in 1:n_bootstrap_samples){
  set.seed(i)
  X_minus_bin_2<- subset(fhs_binned_sample, select= c("cigsPerDay_bin_2", "education", "age","gender") )
  X_minus_bin_2$age <- as.numeric(X_minus_bin_2$age)
  X_minus_bin_2_all_bin_2 <- X_minus_bin_2
```

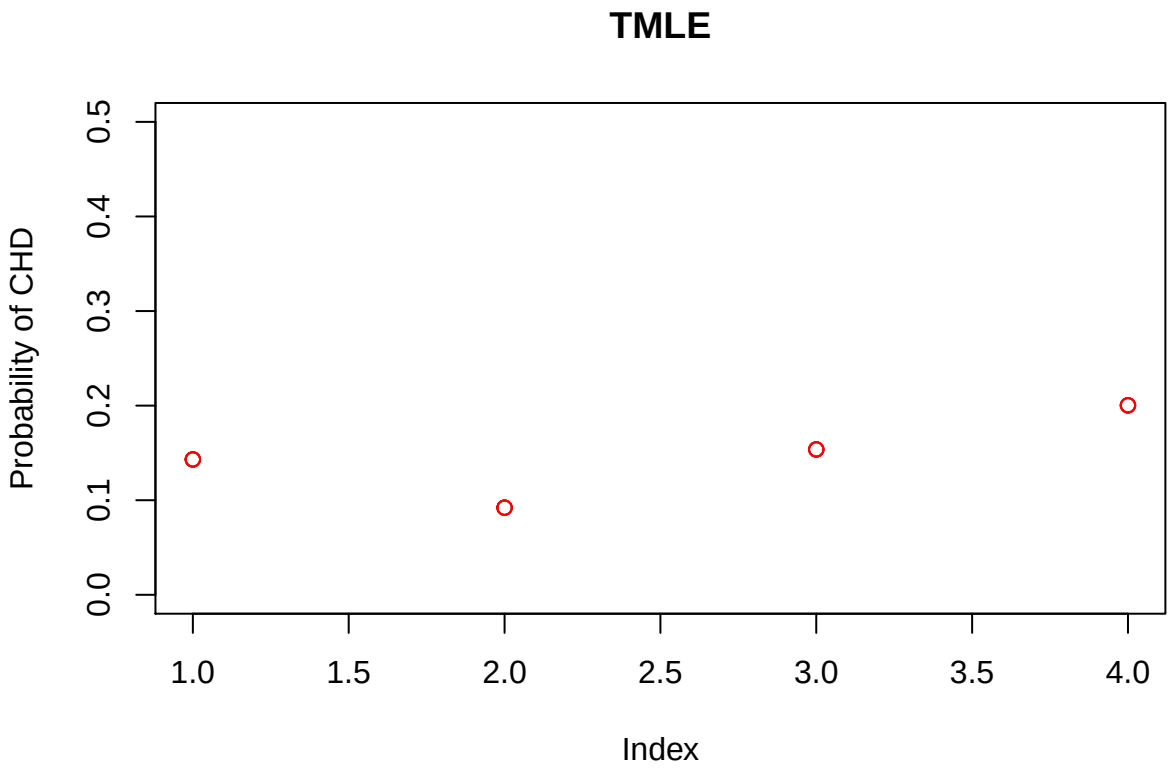
```
  X_minus_bin_2_all_bin_2$cigsPerDay_bin_2 <- 1

  SL.outcome<- SuperLearner(Y=as.numeric(fhs_binned_sample$CHD==1), X=X_minus_bin_2, SL.library=SL.libra
  expY.givenA1 <-  predict(SL.outcome, newdata=X_minus_bin_2_all_bin_2)$pred
  SL.exposure<- SuperLearner(Y=as.numeric(fhs_binned$cigsPerDay_bin_2==1), X=subset(X_minus_bin_2, selec
  probA1.givenW<- SL.exposure$SL.predict
  H.AW<- as.numeric(fhs_binned$cigsPerDay_bin_2==1)/probA1.givenW
  logitUpdate<- glm(fhs_binned$CHD ~ -1 +offset(qlogis(expY.givenA1)) + H.AW, family='binomial')
  epsilon<- logitUpdate$coef
  expY.givenAW.star<- plogis(qlogis(expY.givenA1)+ epsilon*H.AW)
  PsiHat.TMLE_bin_2<- mean(expY.givenAW.star)#- expY.givenOW.star)
  tmle_bootstrap_bin_2[i] <- PsiHat.TMLE_bin_2
}
```

```
tmle_bootstrap_bin_3 <- c()

for (i in 1:n_bootstrap_samples){
  #### BIN 3 TMLE

  X_minus_bin_3<- subset(fhs_binned_sample, select= c("cigsPerDay_bin_3", "education", "age","gender") )
  X_minus_bin_3$age <- as.numeric(X_minus_bin_3$age)

  X_minus_bin_3_all_bin_3 <- X_minus_bin_3
  X_minus_bin_3_all_bin_3$cigsPerDay_bin_3 <- 1
  SL.outcome<- SuperLearner(Y=as.numeric(fhs_binned_sample$CHD==1), X=X_minus_bin_3, SL.library=SL.libra
  expY.givenA1 <-  predict(SL.outcome, newdata=X_minus_bin_3_all_bin_3)$pred
  SL.exposure<- SuperLearner(Y=as.numeric(fhs_binned$cigsPerDay_bin_3==1), X=subset(X_minus_bin_3, selec
  probA1.givenW<- SL.exposure$SL.predict
  H.AW<- as.numeric(fhs_binned$cigsPerDay_bin_3==1)/probA1.givenW
  logitUpdate<- glm(fhs_binned$CHD ~ -1 +offset(qlogis(expY.givenA1)) + H.AW, family='binomial')
  epsilon<- logitUpdate$coef
  expY.givenAW.star<- plogis(qlogis(expY.givenA1)+ epsilon*H.AW)
  PsiHat.TMLE_bin_3 <- mean(expY.givenAW.star)
  tmle_bootstrap_bin_3[i] <- PsiHat.TMLE_bin_3
}
```

```
  #### BIN 4  TMLE
tmle_bootstrap_bin_4 <- c()
for (i in 1:n_bootstrap_samples){

  X_minus_bin_4<- subset(fhs_binned_sample, select= c("cigsPerDay_bin_4", "education", "age","gender") )
  X_minus_bin_4$age <- as.numeric(X_minus_bin_4$age)

  X_minus_bin_4_all_bin_4 <- X_minus_bin_4
  X_minus_bin_4_all_bin_4$cigsPerDay_bin_4 <- 1
  SL.outcome<- SuperLearner(Y=as.numeric(fhs_binned_sample$CHD==1), X=X_minus_bin_4, SL.library=SL.libra
  expY.givenA1 <-  predict(SL.outcome, newdata=X_minus_bin_4_all_bin_4)$pred
  SL.exposure<- SuperLearner(Y=as.numeric(fhs_binned$cigsPerDay_bin_4==1), X=subset(X_minus_bin_4, selec
  probA1.givenW<- SL.exposure$SL.predict
  H.AW<- as.numeric(fhs_binned$cigsPerDay_bin_4==1)/probA1.givenW
  logitUpdate<- glm(fhs_binned$CHD ~ -1 +offset(qlogis(expY.givenA1)) + H.AW, family='binomial')
  epsilon<- logitUpdate$coef
  expY.givenAW.star<- plogis(qlogis(expY.givenA1)+ epsilon*H.AW)
  PsiHat.TMLE_bin_4 <- mean(expY.givenAW.star)
```

```
  tmle_bootstrap_bin_4[i] <- PsiHat.TMLE_bin_4
}
```

```
tmle_bootstrap <- cbind(tmle_bootstrap_bin_1,tmle_bootstrap_bin_2,tmle_bootstrap_bin_3,tmle_bootstrap_bi

average_treatment_effect <- c()
average_treatment_effect_ci <- matrix(NA,nrow=length(levels(fhs_binned$cigsPerDay)),ncol=2)
for (i in 1:length(levels(fhs_binned$cigsPerDay))){
  average_treatment_effect[i] <- colMeans(tmle_bootstrap)[i]
  average_treatment_effect_ci[i,] <- quantile(tmle_bootstrap[,i],probs=c(.025,.975))

}

plot(average_treatment_effect,ylab="Probability of CHD",ylim=c(0,.5),
     main = "TMLE")
points(average_treatment_effect_ci[,1],col='red',lty=2)
points(average_treatment_effect_ci[,2],col='red',lty=2)
```



Step 7. Result Interpretation

What is the statistical interpretation of your analyses? Discuss differences (or lack thereof) in the estimates provided by the different estimators. What is the causal interpretation of your results and how plausible is it? What are key limitations of your analysis? How might these results (if at all) inform policy, understanding, and/or the design of future studies?

```
mean(as.numeric(fhs_binned[fhs_binned$cigsPerDay == "[10, 20)" & fhs_binned$age == "[32, 42)",]$CHD)-1)
```

```
## [1] 0.05454545
```

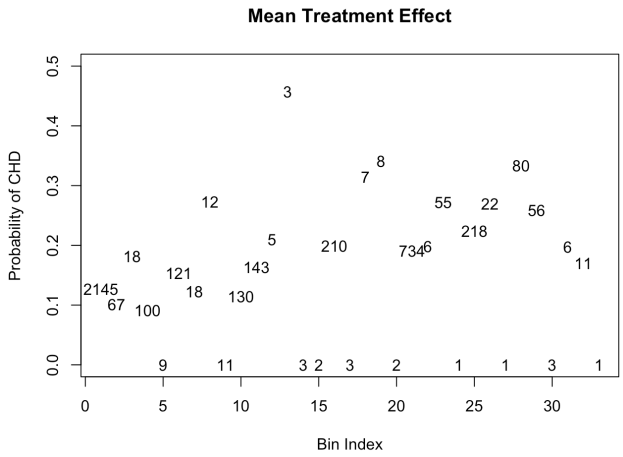


Figure 3: Expected Outcome

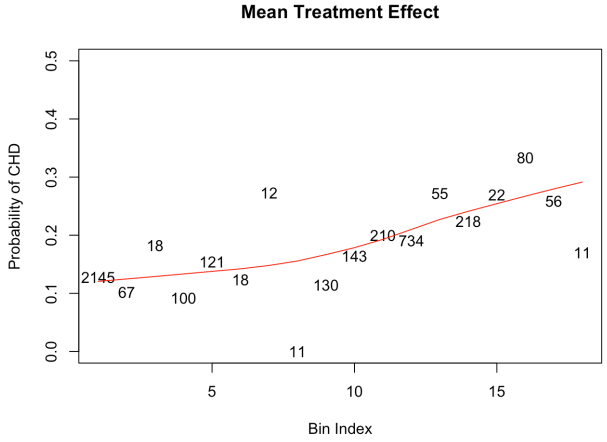


Figure 4: Expected Outcome

```
mean(as.numeric(fhs_binned[fhs_binned$cigsPerDay == "[0, 1)" & fhs_binned$age == "[32, 42)",]$CHD)-1)
```

```
## [1] 0.02479339
```

```
mean(as.numeric(fhs_binned[fhs_binned$cigsPerDay == "[1, 10)" & fhs_binned$age == "[56, 70]",]$CHD)-1)
```

```
## [1] 0.15
```

```
mean(as.numeric(fhs_binned[fhs_binned$cigsPerDay == "[10, 20)" & fhs_binned$age == "[56, 70]",]$CHD)-1)
```

```
## [1] 0.3378378
```

Statistical:

The probability of developing CHD in 10 years increases as theexposure increases among subjects with similar baseline covariatesForcigsPerDay [1,10), the exposure appears to have a“protective effect”; this may be due to reporting bias / practicalpositivity violationsFor more than 20 cigarettes, the exposure effect starts to plateau,approaching a constant effect in risk of CHDTMLE estimates more conservative than conditional

mean & IPTW

Counterfactual:

The estimated effect of the exposure is the counterfactual probability of CHD in 10 years under the exposure category"cigsPerDay",A_i $i \in \{0, 1, 2, 3\}$), with the assumptions thatW(age, sex, BMI, education, and diabetes) satisfy the backdoorcriterion + positivity.Plausibility:As expected, we observed a general increase in estimated risk ofCHD as the (binned) exposure (cigsPerDay) is increased.Less plausible is the estimated reduction in risk of CHD for $cigsPerDay \in [0, 10)$; this may be due to reporting bias, or may infact be evidence for a "protective effect" of smoking on risk of CHD.

Team Member Contribution

Michael Attah

- Causal graph
- Causal question
- Estimation...
- Interpretation?
- etc

Bianca Doone

- Bootstrap confidence interval
- Data Retrieval
- Causal graph
- Causal question
- Causal Parameter
- Specify observed data and link to SCM
- Descriptive table
- Put together Table 1

Casey Graham

- Categorize covariates
- Causal Parameter
- Conditional mean outcome
- IPTW
- Superlearner/TMLE
- Causal graph
- Causal question
- Statistical Estimand: G-comp
- etc

Daniel Saunders

- Data exploration
- Estimation...
- Interpretation?
- Causal graph
- Causal question

- etc

Nutcha Wattanachit

- Background story
- Descriptive table
- Causal graph
- Causal question
- Causal Parameter
- Specify observed data and link to SCM
- Identifiability: Randomization assumption
- Identifiability: Positivity assumption
- Statistical Estimand: G-comp

References

Boston University & the National Heart, Lung, & Blood Institute, Framingham Heart Study. 5 Dec. 2018: https://www.framinghamheartstudy.org