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

Data Exploration

```r
fhs = read.csv('framingham.csv', header = T)
head(fhs)
```

```
##   male age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 1    1  39         4             0          0      0               0
## 2    0  46         2             0          0      0               0
## 3    1  48         1             1         20      0               0
## 4    0  61         3             1         30      0               0
## 5    0  46         3             1         23      0               0
## 6    0  43         2             0          0      0               0
##   prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose
## 1            0        0     195 106.0    70 26.97        80      77
## 2            0        0     250 121.0    81 28.73        95      76
## 3            0        0     245 127.5    80 25.34        75      70
## 4            1        0     225 150.0    95 28.58        65     103
## 5            0        0     285 130.0    84 23.10        85      85
## 6            1        0     228 180.0   110 30.30        77      99
##   TenYearCHD
## 1          0
## 2          0
## 3          0
## 4          1
## 5          0
## 6          0
```

```r
summary(fhs)
```

```
##       male             age          education     currentSmoker
##  Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :49.00   Median :2.000   Median :0.0000
##  Mean   :0.4292   Mean   :49.58   Mean   :1.979   Mean   :0.4941
##  3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000
##                                   NA's   :105
##    cigsPerDay         BPMeds         prevalentStroke     prevalentHyp
##  Min.   : 0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median : 0.000   Median :0.00000   Median :0.000000   Median :0.0000
##  Mean   : 9.006   Mean   :0.02962   Mean   :0.005896   Mean   :0.3106
##  3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :70.000   Max.   :1.00000   Max.   :1.000000   Max.   :1.0000
```

```
##  NA's   :29       NA's   :53
##     diabetes          totChol          sysBP           diaBP
##  Min.   :0.00000   Min.   :107.0   Min.   : 83.5   Min.   : 48.0
##  1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.0
##  Median :0.00000   Median :234.0   Median :128.0   Median : 82.0
##  Mean   :0.02571   Mean   :236.7   Mean   :132.4   Mean   : 82.9
##  3rd Qu.:0.00000   3rd Qu.:263.0   3rd Qu.:144.0   3rd Qu.: 90.0
##  Max.   :1.00000   Max.   :696.0   Max.   :295.0   Max.   :142.5
##                    NA's   :50
##       BMI          heartRate          glucose         TenYearCHD
##  Min.   :15.54   Min.   : 44.00   Min.   : 40.00   Min.   :0.0000
##  1st Qu.:23.07   1st Qu.: 68.00   1st Qu.: 71.00   1st Qu.:0.0000
##  Median :25.40   Median : 75.00   Median : 78.00   Median :0.0000
##  Mean   :25.80   Mean   : 75.88   Mean   : 81.96   Mean   :0.1519
##  3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00   3rd Qu.:0.0000
##  Max.   :56.80   Max.   :143.00   Max.   :394.00   Max.   :1.0000
##  NA's   :19      NA's   :1        NA's   :388
```

```r
table(fhs$TenYearCHD)
```

```
##
##    0    1
## 3596  644
```

```r
table(fhs$currentSmoker)
```

```
##
##    0    1
## 2145 2095
```

```r
table(fhs$cigsPerDay)
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14
## 2145   67   18  100    9  121   18   12   11  130  143    5    3    3    2
##   15   16   17   18   19   20   23   25   29   30   35   38   40   43   45
##  210    3    7    8    2  734    6   55    1  218   22    1   80   56    3
##   50   60   70
##    6   11    1
```

```r
dim(fhs)
```

```
## [1] 4240   16
```

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 62 in the US.

The sameple in this study is white middle-class men and women aged 30 to 62 (at baseline) in Framingham, Massachusetts. We are willing to generalize to the target population because it is reasonable to assume that

SES and risk factors among the sample and the target population are adequately homogenous.

Step 1: Specify a Causal Model

- Endogenous nodes: $X = (W1, W2, A, Y)$, where $W1$ is a group of covariates, W2 is another group of covariates, A is smoking status, and Y is the ten-year development of cardiovascular disease.
- Exogenous nodes: $U = (U_{W1}, U_{W2}, U_A, U_Y) \sim \mathbb{P}_U$. We make no assumptions about the distribution $\mathbb{P}_U$.
- Structural equations F:

$$\begin{aligned} W1 &\leftarrow f_{W1}(U_{W1}) \\ W2 &\leftarrow f_{W2}(W1, U_{W2}) \\ A &\leftarrow f_A(W1, W2, U_A) \\ Y &\leftarrow f_Y(W1, W2, A, U_Y) \end{aligned}$$

There are no exclusion restrictions or assumptions about functional form.

Step 2: Counterfactuals & Causal Parameter

Causal Parameter

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

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. Here we let $\mathbb{W} = \{W_1, W_2\}$

G-Computation

$$\Psi_O(\mathbb{P}_O^i) = \mathbb{E}_o[\mathbb{E}_o[Y|A = a \text{ in bin } i, \mathbb{W}]]$$

$$\Psi_n(\mathbb{P}_n^i) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_n(Y|A = a \text{ in bin } i, \mathbb{W})$$

where j indexes 1 to (TBD) observations.

IPTW

$$IPTW^i = \frac{1}{n}\sum_{j}^{n} Y \frac{\mathbb{I}(A \in i)}{P(A \in i|\mathbb{W})}$$

Targeted Maximum Likelihood Estimate (TMLE)

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

- Describe your observed data and its link to the causal model you have specified. If you feel that in reality the link between your causal model and the observed data is more complex than we have learned in class (n i.i.d. copies of random variable O), explain why. But for this project, stick with the simple link we have learned in class.

The dataset is adapted from Framingham Heart Study. All covariate data is assumed to be collected at baseline, and then a 10-year follow up on CHD (unlike the study).

- Be sure to include a basic descriptive table of your data that provides information on the outcome, exposure, and covariate distributions. (i.e. a classic "Table 1" in the applied public health and medical literature.) Feel free to ask for guidance if you are not sure what this should look like.

Table 1: Descriptive Table

Variable Name	Covariate	Description
male	Gender	binary: male = 1 female = 0
age	Age	ordinal: 32-38, 39-40, 41-43, 44-45, 46-48 49-51, 52-54, 55-57, 58-61, 62-70
education	Education level	ordinal: 1 = some high school, 2 = high school/GED 3 = some college/vocational school, 4 = college
currentSmoker	Current Smoking Status	binary: 1 = Yes 0 = No
cigsPerDay	Number of cigarettes per day	ordinal: 0, 1-19, 20-70
BPMeds	Indicator of blood pressure medication	binary: 1 = Yes 0 = No
prevalentStroke	Prevalence of Stroke	binary: 1 = Yes 0 = No
prevalentHyp	Prevalence of Hypertension	binary: 1 = Yes 0 = No
diabetes	Prevalence of Diabetes	binary: 1 = Yes 0 = No
totChol	Total Cholesterol Level	ordinal: 0-79, 80-89, 90-600
sysBP	Systolic Blood Pressure	ordinal: 0-119, 120-139, 140-295
diaBP	Diastolic Blood Pressure	ordinal: 0-79, 80-89, 90-142.5
BMI	Body Mass Index	ordinal: 0-18.4, 18.5-24.9, 25-29.9, 30-56.8
heartRate	Heart Rate	ordinal: 0-59, 60 -143
glucose	Glucose Level	ordinal: 0-77, 78-394
TenYearCHD	Ten Year Follow-Up Prevalence	binary: 1 = Yes (had CHD) 0 = No (do not have CHD)

Remove NA's (need something smarter later)

Step 4. Identifiability

Is your target causal parameter identified under your initial causal model? If not, under what additional assumptions would it be identified? How plausible are these for your particular problem? Are there additional data or changes to your study design that would improve their plausibility?

Step 5. Statistical Model and Estimand

Step 6. Estimation

Conditional Mean outcome

```
library(mgcv)
```

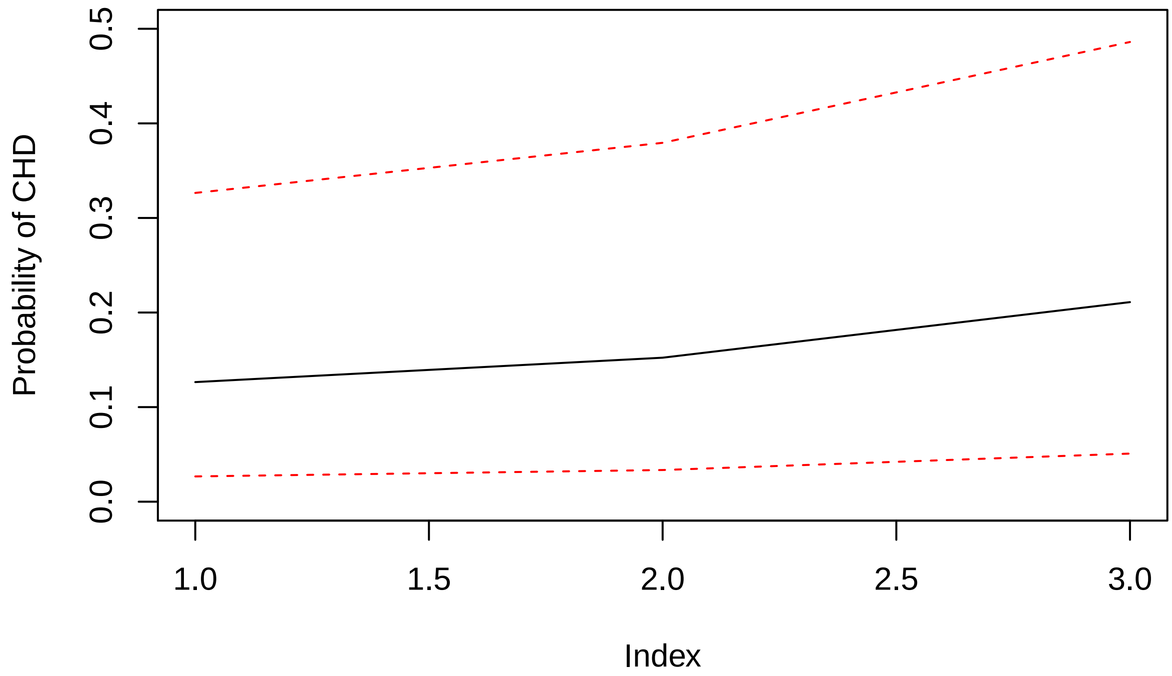
```
## Loading required package: nlme

## This is mgcv 1.8-23. For overview type 'help("mgcv-package")'.
```

```
glm_fit <- glm( CHD ~ cigsPerDay + education + age + diabetes + bmi , data = fhs_binned, family = "binor
summary(glm_fit)
```

```
##
## Call:
## glm(formula = CHD ~ cigsPerDay + education + age + diabetes +
##     bmi, family = "binomial", data = fhs_binned)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4081  -0.6352  -0.4508  -0.3031   2.7066
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.75474    0.12971 -13.528  < 2e-16 ***
## cigsPerDay.L  0.47268    0.08036   5.882 4.05e-09 ***
## cigsPerDay.Q  0.08321    0.10230   0.813 0.416017
## education2   -0.20096    0.12031  -1.670 0.094845 .
## education3   -0.24260    0.14561  -1.666 0.095693 .
## education4   -0.04079    0.16165  -0.252 0.800787
## age.L         2.28373    0.20094  11.365  < 2e-16 ***
## age.Q        -0.02872    0.18721  -0.153 0.878060
## age.C        -0.05312    0.18450  -0.288 0.773423
## age^4         0.10738    0.18026   0.596 0.551365
## age^5         0.05985    0.18796   0.318 0.750153
## age^6        -0.09165    0.17491  -0.524 0.600307
## age^7        -0.25058    0.16574  -1.512 0.130554
## age^8         0.13756    0.17141   0.803 0.422245
## age^9         0.14049    0.15857   0.886 0.375613
## diabetes      0.79005    0.22583   3.498 0.000468 ***
## bmi.L         0.21194    0.29804   0.711 0.477016
## bmi.Q         0.20046    0.22738   0.882 0.377988
## bmi.C        -0.14459    0.12156  -1.189 0.234267
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3121.2  on 3657  degrees of freedom
## Residual deviance: 2855.0  on 3639  degrees of freedom
## AIC: 2893
##
## Number of Fisher Scoring iterations: 5
```

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


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] <- mean(predict(glm_fit, newdata=intervene_on_bin(i), type='response'))
  average_treatment_effect_ci[i,] <- quantile(predict(glm_fit, newdata=intervene_on_bin(i), type='respon

}

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



IPTW

```
### Create pairwise binary variables for each bin
fhs_binned$cigsPerDay_bin_1 <- ifelse(fhs_binned$cigsPerDay == "[0, 0.9)",1,0)
fhs_binned$cigsPerDay_bin_2 <- ifelse(fhs_binned$cigsPerDay == "[0.9, 20)",1,0)
fhs_binned$cigsPerDay_bin_3 <- ifelse(fhs_binned$cigsPerDay == "[20, 70]",1,0)

### BIN 2
glm_fit_iptw_bin_1 <- glm( cigsPerDay_bin_1 ~   education + age + diabetes + bmi , data = fhs_binned, fa
prob.1W <- predict(glm_fit_iptw_bin_1, type= "response")
wt_1<- 1/prob.1W
summary(wt_1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.174   1.624   1.973   2.094   2.470   3.962
```

```
IPTW_bin_1<- mean( wt_1*as.numeric(fhs_binned$cigsPerDay_bin_1==1)*as.numeric(fhs_binned$CHD==1))

### BIN 2
glm_fit_iptw_bin_2 <- glm( cigsPerDay_bin_2 ~   education + age + diabetes + bmi , data = fhs_binned, fa
```

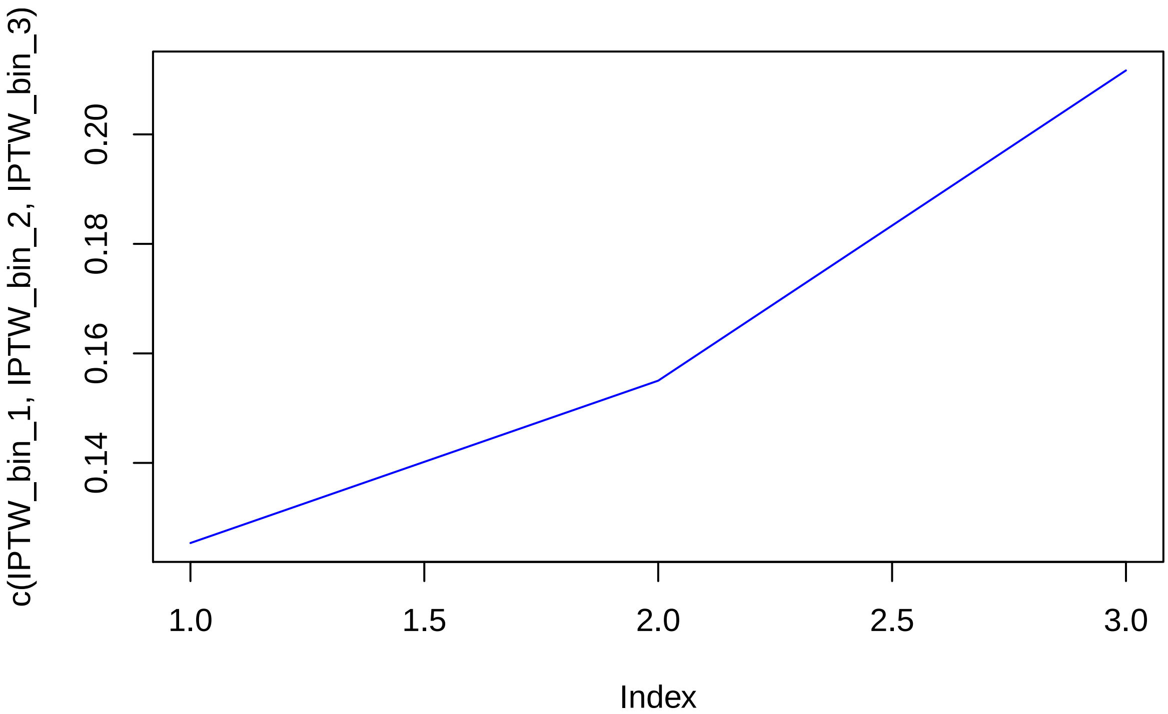
```r
prob.1W <- predict(glm_fit_iptw_bin_2, type= "response")
wt_2<- 1/prob.1W
summary(wt_2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.476   3.895   5.272   5.445   6.456  12.287
```

```r
IPTW_bin_2<- mean( wt_2*as.numeric(fhs_binned$cigsPerDay_bin_2==1)*as.numeric(fhs_binned$CHD==1))

### BIN 3
glm_fit_iptw_bin_3 <- glm( cigsPerDay_bin_3 ~   education + age + diabetes + bmi , data = fhs_binned, fa
prob.1W <- predict(glm_fit_iptw_bin_3, type= "response")
wt_3<- 1/prob.1W
summary(wt_3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.340   2.818   3.327   3.927   4.520  13.164
```

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

plot(c(IPTW_bin_1,IPTW_bin_2,IPTW_bin_3),type='l',col='blue')
```



Superlearner/TMLE

BIN 3

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

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

```
## Super Learner
```

```
## Version: 2.0-23
```

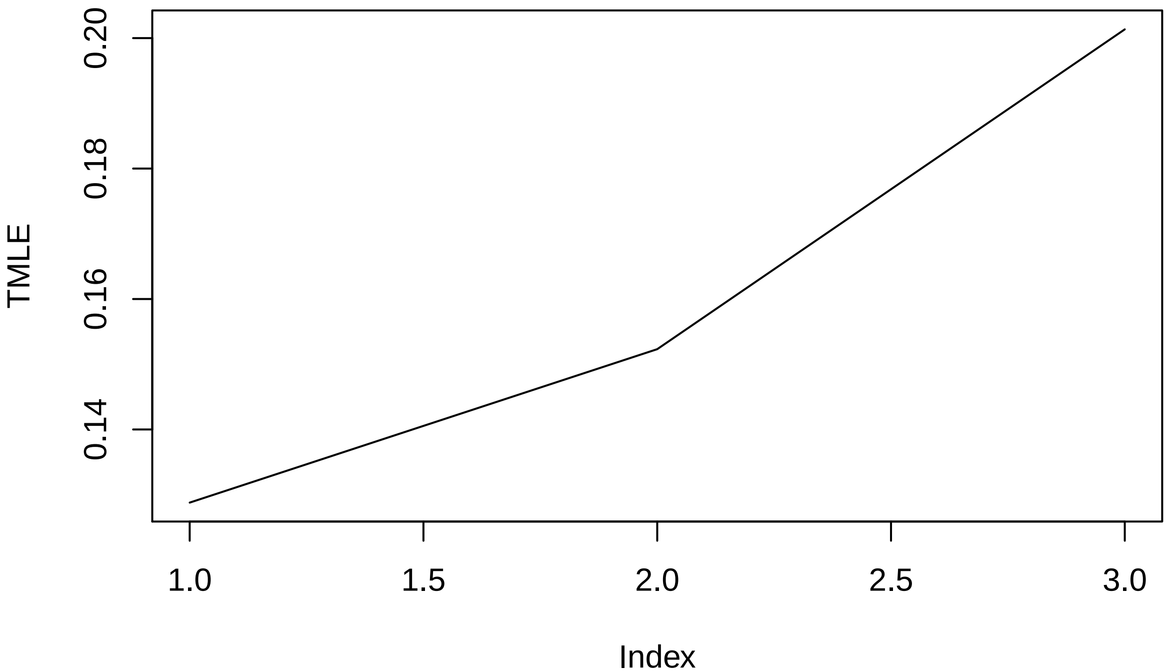
```
## Package created on 2018-03-09
```

```r
SL.library<- c("SL.glmnet")


### Bin1 TMLE

X_minus_bin_1<- subset(fhs_binned, select= c("cigsPerDay_bin_1", "education", "age","diabetes","bmi") )
X_minus_bin_1_all_bin_1 <- X_minus_bin_1
X_minus_bin_1_all_bin_1$cigsPerDay_bin_1 <- 1
SL.outcome<- SuperLearner(Y=as.numeric(fhs_binned$CHD==1), X=X_minus_bin_1, SL.library=SL.library, famil
```

```
## Loading required package: glmnet

## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-16
```

```r
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, select=
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.given0W.star)


X_minus_bin_2<- subset(fhs_binned, select= c("cigsPerDay_bin_2", "education", "age","diabetes","bmi") )
X_minus_bin_2_all_bin_2 <- X_minus_bin_2
X_minus_bin_2_all_bin_2$cigsPerDay_bin_2 <- 1
SL.outcome<- SuperLearner(Y=as.numeric(fhs_binned$CHD==1), X=X_minus_bin_2, SL.library=SL.library, famil
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, select=
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.given0W.star)
#### BIN 3 TMLE

X_minus_bin_3<- subset(fhs_binned, select= c("cigsPerDay_bin_3", "education", "age","diabetes","bmi") )
X_minus_bin_3_all_bin_3 <- X_minus_bin_3
X_minus_bin_3_all_bin_3$cigsPerDay_bin_3 <- 1
SL.outcome<- SuperLearner(Y=as.numeric(fhs_binned$CHD==1), X=X_minus_bin_3, SL.library=SL.library, famil
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, select=
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)#- expY.given0W.star)
```

```r
plot(c(PsiHat.TMLE_bin_1,PsiHat.TMLE_bin_2,PsiHat.TMLE_bin_3),type='l',ylab="TMLE")
```



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?