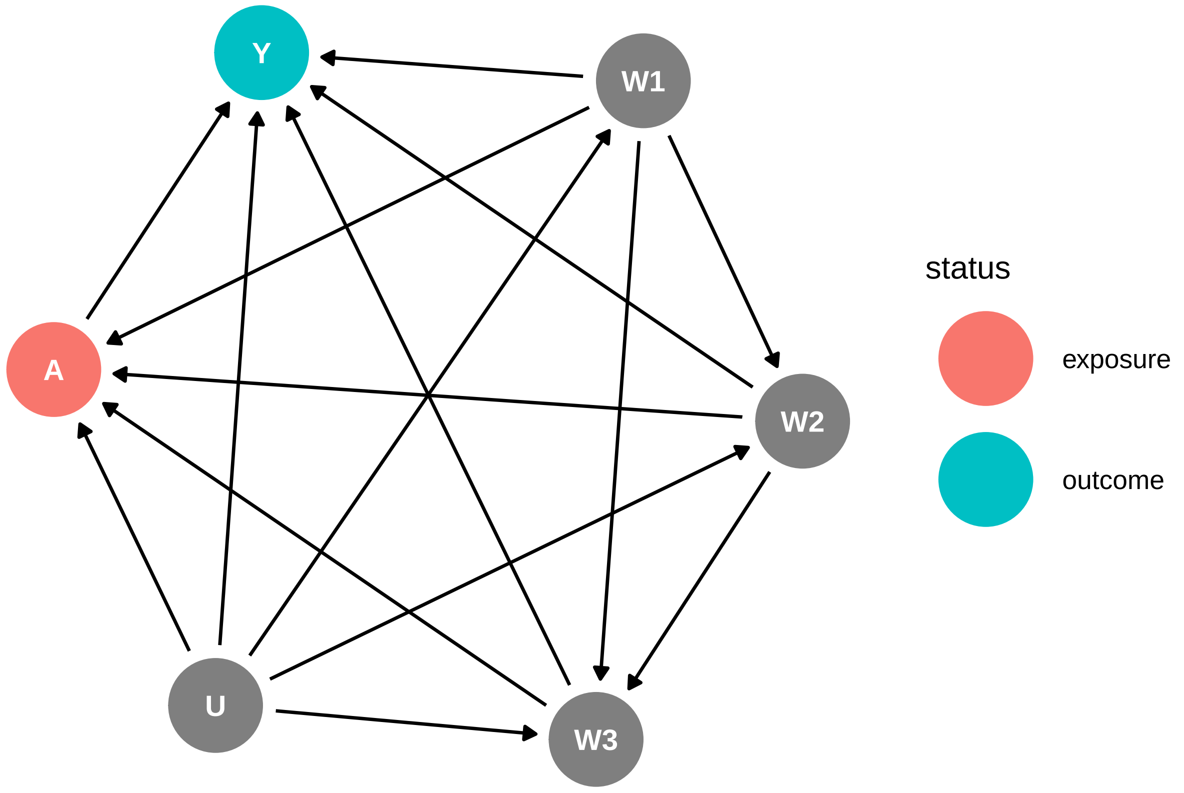
BIOSTAT 683 - final project

```r
library(ggdag)
library(tidyverse)
```

```r
set.seed(123)

sample_dag <- dagify(
  W1 ~ U,
  W2 ~ W1 + U,
  W3 ~ W1 + W2 + U,
  A ~ W1 + W2 + W3 + U,
  Y ~ W1 + W2 + W3 + A + U,
  exposure = "A",
  outcome = "Y"
)

ggdag_status(sample_dag)+
  theme_dag()
```



```r
# getting the data
slpexcov1517 <- read.csv("slpexcov1517.csv")
summary(slpexcov1517)
```

```
##       SEQN           exminwk          targetex          slphrs
##  Min.   : 83732   Min.   :   0.0   Min.   :0.0000   Min.   : 2.000
##  1st Qu.: 88525   1st Qu.:   0.0   1st Qu.:0.0000   1st Qu.: 6.500
##  Median : 93331   Median :  60.0   Median :0.0000   Median : 7.500
##  Mean   : 93322   Mean   : 282.8   Mean   :0.4127   Mean   : 7.379
##  3rd Qu.: 98078   3rd Qu.: 360.0   3rd Qu.:1.0000   3rd Qu.: 8.000
##  Max.   :102956   Max.   :6860.0   Max.   :1.0000   Max.   :14.500
##
##    targetslp           age           raceeth           educ
##  Min.   :0.0000   Min.   :20.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.0000   1st Qu.:31.00   1st Qu.:1.000   1st Qu.:2.000
##  Median :1.0000   Median :43.00   Median :2.000   Median :3.000
##  Mean   :0.7811   Mean   :42.89   Mean   :2.327   Mean   :2.555
##  3rd Qu.:1.0000   3rd Qu.:55.00   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :1.0000   Max.   :64.00   Max.   :4.000   Max.   :4.000
##                                                   NA’s   :1
##     marital        household         income         snoring
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000
##  Median :2.000   Median :3.000   Median :2.000   Median :3.000
##  Mean   :1.641   Mean   :3.393   Mean   :1.945   Mean   :2.679
##  3rd Qu.:2.000   3rd Qu.:5.000   3rd Qu.:2.000   3rd Qu.:4.000
##  Max.   :2.000   Max.   :6.000   Max.   :3.000   Max.   :4.000
##  NA’s   :1                       NA’s   :305     NA’s   :206
##      apnea             bmi            bmicat          waist
##  Min.   :0.0000   Min.   :15.10   Min.   :1.000   Min.   : 62.3
##  1st Qu.:0.0000   1st Qu.:24.80   1st Qu.:2.000   1st Qu.: 89.7
##  Median :0.0000   Median :28.30   Median :3.000   Median : 99.2
##  Mean   :0.3034   Mean   :29.34   Mean   :3.115   Mean   :101.1
##  3rd Qu.:1.0000   3rd Qu.:32.80   3rd Qu.:4.000   3rd Qu.:110.5
##  Max.   :1.0000   Max.   :86.20   Max.   :4.000   Max.   :169.6
##  NA’s   :190      NA’s   :50      NA’s   :50      NA’s   :159
##      smoke           alcohol           phq9          depressed
##  Min.   :0.0000   Min.   :0.000   Min.   : 0.000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:1.000   1st Qu.: 0.000   1st Qu.:0.00000
##  Median :0.0000   Median :1.000   Median : 1.000   Median :0.00000
##  Mean   :0.4876   Mean   :1.182   Mean   : 2.875   Mean   :0.07161
##  3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.: 4.000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :2.000   Max.   :27.000   Max.   :1.00000
##  NA’s   :2        NA’s   :446     NA’s   :301      NA’s   :301
```

```r
# Binary table
binslpexcov1517 <- dplyr::select(slpexcov1517, SEQN, targetex, targetslp)
summary(binslpexcov1517)
```

```
##       SEQN           targetex        targetslp
##  Min.   : 83732   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 88525   1st Qu.:0.0000   1st Qu.:1.0000
##  Median : 93331   Median :0.0000   Median :1.0000
##  Mean   : 93322   Mean   :0.4127   Mean   :0.7811
##  3rd Qu.: 98078   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :102956   Max.   :1.0000   Max.   :1.0000
```

```
#Contingency tables
binslpexcov1517 %>% select(-SEQN) %>% table()
```

```
##         targetslp
## targetex    0    1
##        0  544 1683
##        1  286 1279
```

```
#          targetslp
# targetex    0    1
#        0  544 1683
#        1  286 1279

slpexcov1517 %>%
  dplyr::select(targetex, targetslp) %>%
  group_by(targetex, targetslp) %>%
  summarise(n = n())
```

```
## `summarise()` has grouped output by 'targetex'. You can override using the `.groups` argument.

## # A tibble: 4 x 3
## # Groups:   targetex [2]
##   targetex targetslp     n
##      <int>     <int> <int>
## 1        0         0   544
## 2        0         1  1683
## 3        1         0   286
## 4        1         1  1279
```

```
# Fit a logistic model to the data without confounders and look at results
glm.slpex = glm(targetslp ~ targetex, family = binomial(link = "logit"), data = slpexcov1517)
summary(glm.slpex)
```

```
##
## Call:
## glm(formula = targetslp ~ targetex, family = binomial(link = "logit"),
##     data = slpexcov1517)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8437   0.6353   0.6353   0.7484   0.7484
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.12938    0.04932  22.899  < 2e-16 ***
## targetex     0.36846    0.08192   4.498 6.87e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3985.3  on 3791  degrees of freedom
## Residual deviance: 3964.7  on 3790  degrees of freedom
## AIC: 3968.7
##
## Number of Fisher Scoring iterations: 4
```

```
exp(cbind(OR = coef(glm.slpex), confint(glm.slpex)))
```

```
## Waiting for profiling to be done...
```

```
##                   OR    2.5 %   97.5 %
## (Intercept) 3.093750 2.810878 3.410527
## targetex    1.445504 1.232041 1.698740
```

```
# Fit data with covariates and see how effects change
glm.slpexcov = glm(targetslp ~ targetex + age + factor(raceeth) + factor(educ) + factor(marital) +
                     factor(household) + factor(income) + factor(snoring) + factor(apnea) + bmi +
                     waist + factor(smoke) + factor(alcohol) + factor(depressed),
                   family = binomial(link = "logit"), data = slpexcov1517)
summary(glm.slpexcov)
```

```
##
## Call:
## glm(formula = targetslp ~ targetex + age + factor(raceeth) +
##     factor(educ) + factor(marital) + factor(household) + factor(income) +
##     factor(snoring) + factor(apnea) + bmi + waist + factor(smoke) +
##     factor(alcohol) + factor(depressed), family = binomial(link = "logit"),
##     data = slpexcov1517)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3600   0.4280   0.6067   0.7344   1.3582
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.688443   0.453792   3.721 0.000199 ***
## targetex            0.372773   0.105710   3.526 0.000421 ***
## age                -0.007625   0.004471  -1.705 0.088118 .
## factor(raceeth)2   -0.163198   0.139485  -1.170 0.241998
## factor(raceeth)3   -0.778182   0.133716  -5.820  5.9e-09 ***
## factor(raceeth)4   -0.250682   0.152796  -1.641 0.100873
## factor(educ)2      -0.256649   0.143162  -1.793 0.073018 .
## factor(educ)3      -0.060024   0.146147  -0.411 0.681286
## factor(educ)4       0.470659   0.176235   2.671 0.007571 **
## factor(marital)2    0.033478   0.121958   0.275 0.783695
## factor(household)2 -0.109821   0.196421  -0.559 0.576087
## factor(household)3 -0.095248   0.207804  -0.458 0.646696
## factor(household)4 -0.263762   0.213943  -1.233 0.217627
## factor(household)5 -0.366306   0.223922  -1.636 0.101868
## factor(household)6 -0.249541   0.229755  -1.086 0.277428   
## factor(income)2     0.024217   0.115202   0.210 0.833499   
## factor(income)3    -0.202953   0.150150  -1.352 0.176482   
## factor(snoring)2    0.063241   0.149085   0.424 0.671426   
## factor(snoring)3   -0.106954   0.155668  -0.687 0.492040   
## factor(snoring)4   -0.177607   0.147607  -1.203 0.228883   
## factor(apnea)1     -0.063786   0.108132  -0.590 0.555262   
## bmi                -0.054857   0.024035  -2.282 0.022464 * 
## waist               0.019071   0.009613   1.984 0.047261 * 
## factor(smoke)1     -0.002851   0.102511  -0.028 0.977812   
## factor(alcohol)1    0.109962   0.130380   0.843 0.399007   
## factor(alcohol)2    0.008028   0.131784   0.061 0.951428   
## factor(depressed)1 -0.537253   0.178907  -3.003 0.002674 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2894.6  on 2768  degrees of freedom
## Residual deviance: 2767.2  on 2742  degrees of freedom
##   (1023 observations deleted due to missingness)
## AIC: 2821.2
## 
## Number of Fisher Scoring iterations: 4
```

```
exp(cbind(OR = coef(glm.slpexcov), confint(glm.slpexcov)))
```

```
## Waiting for profiling to be done...
```

```
##                           OR     2.5 %     97.5 %
## (Intercept)        5.4110489 2.2290591 13.2134288
## targetex           1.4517548 1.1812176  1.7880032
## age                0.9924043 0.9837344  1.0011336
## factor(raceeth)2   0.8494228 0.6462309  1.1168344
## factor(raceeth)3   0.4592402 0.3530287  0.5964448
## factor(raceeth)4   0.7782697 0.5773499  1.0514048
## factor(educ)2      0.7736399 0.5836186  1.0232825
## factor(educ)3      0.9417421 0.7063489  1.2530444
## factor(educ)4      1.6010486 1.1344335  2.2646448
## factor(marital)2   1.0340451 0.8129502  1.3115780
## factor(household)2 0.8959945 0.6071855  1.3125144
## factor(household)3 0.9091472 0.6029635  1.3628389
## factor(household)4 0.7681564 0.5033569  1.1653424
## factor(household)5 0.6932909 0.4457925  1.0732542
## factor(household)6 0.7791585 0.4956179  1.2208539
## factor(income)2    1.0245129 0.8162507  1.2824760
## factor(income)3    0.8163170 0.6080652  1.0957938
## factor(snoring)2   1.0652832 0.7948516  1.4265405
## factor(snoring)3   0.8985669 0.6620063  1.2191784
## factor(snoring)4   0.8372717 0.6259125  1.1167359
## factor(apnea)1     0.9382055 0.7596079  1.1607789
## bmi                0.9466202 0.9030462  0.9923266
## waist              1.0192541 1.0002701  1.0386996
## factor(smoke)1     0.9971530 0.8156805  1.2192572
## factor(alcohol)1   1.1162354 0.8633103  1.4396635
## factor(alcohol)2   1.0080598 0.7773772  1.3034999
## factor(depressed)1 0.5843512 0.4132842  0.8343249
```

```
# model (variables) selection
#step(glm.slpexcov, scope = ~ targetex + age + factor(raceeth) + factor(educ) + factor(marital) +
                     #factor(household) + factor(income) + factor(snoring) + #factor(apnea) + bmi +
                     #waist + factor(smoke) + factor(alcohol) + #factor(depressed), direction = "backwa
# missing values are a problem to model selection
# data would need to be clean before



# Removed covariates that have many NAs or seem unimportant
glm.slpexcov2 = glm(targetslp ~ targetex + age + factor(raceeth) + factor(educ) + factor(marital) +
                      bmi + waist + factor(depressed),
                   family = binomial(link = "logit"), data = slpexcov1517)
summary(glm.slpexcov2)
```

```
##
## Call:
## glm(formula = targetslp ~ targetex + age + factor(raceeth) +
##     factor(educ) + factor(marital) + bmi + waist + factor(depressed),
##     family = binomial(link = "logit"), data = slpexcov1517)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2416   0.4576   0.6234   0.7374   1.2798
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.367665   0.346292   3.949 7.83e-05 ***
## targetex             0.328647   0.093760   3.505 0.000456 ***
## age                 -0.008757   0.003680  -2.379 0.017341 *
## factor(raceeth)2    -0.155546   0.122161  -1.273 0.202915
## factor(raceeth)3    -0.704612   0.118080  -5.967 2.41e-09 ***
## factor(raceeth)4    -0.208653   0.136504  -1.529 0.126376
## factor(educ)2       -0.142141   0.124364  -1.143 0.253060
## factor(educ)3       -0.049859   0.124047  -0.402 0.687730
## factor(educ)4        0.447969   0.144199   3.107 0.001892 **
## factor(marital)2    -0.015536   0.092465  -0.168 0.866570
## bmi                 -0.081716   0.021169  -3.860 0.000113 ***
## waist                0.027740   0.008479   3.272 0.001069 **
## factor(depressed)1 -0.475600   0.151744  -3.134 0.001723 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3603.6  on 3405  degrees of freedom
## Residual deviance: 3472.4  on 3393  degrees of freedom
##   (386 observations deleted due to missingness)
## AIC: 3498.4
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
exp(cbind(OR = coef(glm.slpexcov2), confint(glm.slpexcov2)))
```

```
## Waiting for profiling to be done...
```

```
##                           OR     2.5 %    97.5 %
## (Intercept)        3.9261707 1.9934057 7.7501641
## targetex           1.3890867 1.1567584 1.6707580
## age                0.9912809 0.9841461 0.9984517
## factor(raceeth)2   0.8559473 0.6735931 1.0875914
## factor(raceeth)3   0.4943005 0.3918898 0.6226879
## factor(raceeth)4   0.8116768 0.6217147 1.0620424
## factor(educ)2      0.8674986 0.6793976 1.1064715
## factor(educ)3      0.9513636 0.7454649 1.2125541
## factor(educ)4      1.5651299 1.1807391 2.0786851
## factor(marital)2   0.9845844 0.8207332 1.1794098
## bmi                0.9215338 0.8840135 0.9605409
## waist              1.0281279 1.0112292 1.0454153
## factor(depressed)1 0.6215118 0.4631588 0.8402080
```

```
head(slpexcov1517)
```

```
##    SEQN exminwk targetex slphrs targetslp age raceeth educ marital household
## 1 83732     180        1    5.5         0  62       1    4       2         2
## 2 83733       0        0    8.0         1  53       1    2       1         1
## 3 83741     240        1    6.5         1  22       3    3       1         3
## 4 83744       0        0    4.0         0  56       3    2       1         1
## 5 83747     840        1   10.0         1  46       1    4       2         2
## 6 83750     120        0    8.0         1  45       4    1       1         5
##   income snoring apnea  bmi bmicat waist smoke alcohol phq9 depressed
## 1      2       3    NA 27.8      3 101.1     1       1    1         0
## 2      1       2     0 30.8      4 107.9     1       2    2         0
## 3      2       1     0 28.0      3  86.6     1       2    1         0
## 4      1      NA     1 33.6      4 116.0     0      NA    0         0
## 5      1       2     0 27.6      3 104.3     1       1    2         0
## 6      2       1     0 24.1      2  90.1     1       2    0         0
```

```
ObsData <- subset(slpexcov1517,
                  select = c(targetslp,
                  targetex, age,
                  raceeth, educ,
                  marital, bmi,
                  waist,depressed))

summary(ObsData)
```

```
##    targetslp         targetex           age           raceeth
##  Min.   :0.0000   Min.   :0.0000   Min.   :20.00   Min.   :1.000
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:31.00   1st Qu.:1.000
##  Median :1.0000   Median :0.0000   Median :43.00   Median :2.000
##  Mean   :0.7811   Mean   :0.4127   Mean   :42.89   Mean   :2.327
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:55.00   3rd Qu.:3.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :64.00   Max.   :4.000
##
##       educ          marital           bmi             waist
##  Min.   :1.000   Min.   :1.000   Min.   :15.10   Min.   : 62.3
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:24.80   1st Qu.: 89.7
##  Median :3.000   Median :2.000   Median :28.30   Median : 99.2
##  Mean   :2.555   Mean   :1.641   Mean   :29.34   Mean   :101.1
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:32.80   3rd Qu.:110.5
##  Max.   :4.000   Max.   :2.000   Max.   :86.20   Max.   :169.6
##  NA’s   :1       NA’s   :1       NA’s   :50      NA’s   :159
##    depressed
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.07161
##  3rd Qu.:0.00000
##  Max.   :1.00000
##  NA’s   :301
```

```
# 301 NA for DEPRESSED (7.94%), we should do multiple imputation
# 159 NA for WAIST (4.19%)

ObsData <- na.exclude(ObsData)

ObsData <- ObsData %>% mutate(A = targetex,
                              Y = targetslp) %>%
  select(-targetex, -targetslp)

head(ObsData)
```

```
##   age raceeth educ marital  bmi waist depressed A Y
## 1  62       1    4       2 27.8 101.1         0 1 0
## 2  53       1    2       1 30.8 107.9         0 0 1
## 3  22       3    3       1 28.0  86.6         0 1 1
## 4  56       3    2       1 33.6 116.0         0 0 0
## 5  46       1    4       2 27.6 104.3         0 1 1
## 6  45       4    1       1 24.1  90.1         0 0 1
```

```
# Removed covariates that have many NAs or seem unimportant
glm.slpexcov2 = glm(targetslp ~ targetex + age + factor(raceeth) + factor(educ) + factor(marital) +
                      bmi + waist + factor(depressed),
                   family = binomial(link = "logit"), data = slpexcov1517)
summary(glm.slpexcov2)
```

```
##
## Call:
## glm(formula = targetslp ~ targetex + age + factor(raceeth) +
##     factor(educ) + factor(marital) + bmi + waist + factor(depressed),
##     family = binomial(link = "logit"), data = slpexcov1517)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2416   0.4576   0.6234   0.7374   1.2798
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.367665   0.346292   3.949 7.83e-05 ***
## targetex            0.328647   0.093760   3.505 0.000456 ***
## age                -0.008757   0.003680  -2.379 0.017341 *
## factor(raceeth)2   -0.155546   0.122161  -1.273 0.202915
## factor(raceeth)3   -0.704612   0.118080  -5.967 2.41e-09 ***
## factor(raceeth)4   -0.208653   0.136504  -1.529 0.126376
## factor(educ)2      -0.142141   0.124364  -1.143 0.253060
## factor(educ)3      -0.049859   0.124047  -0.402 0.687730
## factor(educ)4       0.447969   0.144199   3.107 0.001892 **
## factor(marital)2   -0.015536   0.092465  -0.168 0.866570
## bmi                -0.081716   0.021169  -3.860 0.000113 ***
## waist               0.027740   0.008479   3.272 0.001069 **
## factor(depressed)1 -0.475600   0.151744  -3.134 0.001723 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3603.6  on 3405  degrees of freedom
## Residual deviance: 3472.4  on 3393  degrees of freedom
##   (386 observations deleted due to missingness)
## AIC: 3498.4
##
## Number of Fisher Scoring iterations: 4
```

```
exp(cbind(OR = coef(glm.slpexcov2), confint(glm.slpexcov2)))
```

```
## Waiting for profiling to be done...
```

```
##                           OR     2.5 %    97.5 %
## (Intercept)        3.9261707 1.9934057 7.7501641
## targetex           1.3890867 1.1567584 1.6707580
## age                0.9912809 0.9841461 0.9984517
## factor(raceeth)2   0.8559473 0.6735931 1.0875914
## factor(raceeth)3   0.4943005 0.3918898 0.6226879
## factor(raceeth)4   0.8116768 0.6217147 1.0620424
## factor(educ)2      0.8674986 0.6793976 1.1064715
## factor(educ)3      0.9513636 0.7454649 1.2125541
## factor(educ)4      1.5651299 1.1807391 2.0786851
## factor(marital)2   0.9845844 0.8207332 1.1794098
## bmi                0.9215338 0.8840135 0.9605409
## waist              1.0281279 1.0112292 1.0454153
## factor(depressed)1 0.6215118 0.4631588 0.8402080
```

```
#set.seed(252)
```

```r
library(SuperLearner)
SL.library <- c("SL.mean", "SL.glm", "SL.step.interaction")

# using SuperLearner

X <- subset(ObsData, select = -Y)

X1 <- X0 <- X

X1$A <- 1
X0$A <- 0

SL.outcome <- SuperLearner(Y = ObsData$Y,
                           X = X,
                           SL.library = SL.library,
                           family = "binomial")
SL.outcome
```

```
##
## Call:
## SuperLearner(Y = ObsData$Y, X = X, family = "binomial", SL.library = SL.library)
##
##
##
##                              Risk      Coef
## SL.mean_All             0.1726745 0.1482501
## SL.glm_All              0.1695816 0.8517499
## SL.step.interaction_All 0.1705521 0.0000000
```

```r
# expected outcome, given exposure and covariates
expY.givenAW <- predict(SL.outcome, newdata = X)$pred
expY.given1W <- predict(SL.outcome, newdata = X1)$pred
expY.given0W <- predict(SL.outcome, newdata = X0)$pred

# observing the data
head(data.frame(A = ObsData$A,
                expY.givenAW,
                expY.given1W,
                expY.given0W))
```

```
##   A expY.givenAW expY.given1W expY.given0W
## 1 1    0.8542428    0.8542428    0.8185140
## 2 0    0.7664982    0.8119719    0.7664982
## 3 1    0.7829755    0.7829755    0.7318740
## 4 0    0.7175422    0.7707651    0.7175422
## 5 1    0.8792139    0.8792139    0.8501365
## 6 0    0.7028350    0.7581054    0.7028350
```

```r
tail(data.frame(A = ObsData$A,
                expY.givenAW,
                expY.given1W,
                expY.given0W))
```

```
##      A expY.givenAW expY.given1W expY.given0W
## 3401 0    0.6676843    0.7273043    0.6676843
## 3402 0    0.7617877    0.8080682    0.7617877
## 3403 0    0.7292783    0.7807732    0.7292783
## 3404 0    0.8046007    0.8430841    0.8046007
## 3405 0    0.7207159    0.7734797    0.7207159
## 3406 0    0.8010043    0.8401823    0.8010043
```

```
# simple substitution
PsiHat.SS <- mean(expY.given1W - expY.given0W)
PsiHat.SS
```

```
## [1] 0.04584497
```

```
# -----
# Estimating propensity score with SuperLearner
# --------

X <- subset(ObsData, select = - c(A, Y))
#X

SL.exposure <- SuperLearner(Y = ObsData$Y,
                            X = X,
                            SL.library = SL.library,
                            family = "binomial")
SL.exposure
```

```
##
## Call:
## SuperLearner(Y = ObsData$Y, X = X, family = "binomial", SL.library = SL.library)
##
##
##
##                              Risk      Coef
## SL.mean_All             0.1725644 0.1273358
## SL.glm_All              0.1697644 0.8726642
## SL.step.interaction_All 0.1706331 0.0000000
```

```
# generating probability of exposure given baseline covariates
probA1.givenW <- SL.exposure$SL.predict

# above is equivalent to :
check <- predict(SL.exposure, newdata = X)$pred
head(check)
```

```
##           [,1]
## [1,] 0.8345637
## [2,] 0.7806627
## [3,] 0.7709730
## [4,] 0.7301527
## [5,] 0.8649784
## [6,] 0.7180601
```

```
sum(probA1.givenW != check)
```

```
## [1] 0
```

```
# this should be zero

probA0.givenW <- 1 - probA1.givenW

# summary table
summary(data.frame(probA1.givenW, probA0.givenW))
```

```
##  probA1.givenW    probA0.givenW
##  Min.   :0.4657   Min.   :0.1177
##  1st Qu.:0.7479   1st Qu.:0.1835
##  Median :0.7849   Median :0.2151
##  Mean   :0.7783   Mean   :0.2217
##  3rd Qu.:0.8165   3rd Qu.:0.2521
##  Max.   :0.8823   Max.   :0.5343
```

```
# creating the clever covariate H(A,W) for each observation

H.AW <- as.numeric(ObsData$A==1)/probA1.givenW - as.numeric(ObsData$A==0)/probA0.givenW

H.1W <- 1/probA1.givenW
H.0W <- -1/probA0.givenW

head(data.frame(A = ObsData$A,
                H.AW,
                H.1W,
                H.0W))
```

```
##   A      H.AW     H.1W      H.0W
## 1 1  1.198231 1.198231 -6.044621
## 2 0 -4.559188 1.280963 -4.559188
## 3 1  1.297062 1.297062 -4.366297
## 4 0 -3.705799 1.369577 -3.705799
## 5 1  1.156098 1.156098 -7.406220
## 6 0 -3.546856 1.392641 -3.546856
```

```
# IPTW estimator of G-computation formula
PsiHat.IPTW <- mean(H.AW * ObsData$Y)
PsiHat.IPTW
```

```
## [1] -1.631941
```

```
# update estimator of E_0(Y|A,W)
logitUpdate <- glm(ObsData$Y ~ -1 + offset(qlogis(expY.givenAW)) + H.AW,
                                          family = "binomial")

epsilon <- logitUpdate$coef
epsilon
```

```
##        H.AW
## 0.004393361
```

```r
# targeted estimates
expY.givenAW.star <- plogis(qlogis(expY.givenAW) + epsilon * H.AW)
expY.given1W.star <- plogis(qlogis(expY.given1W) + epsilon * H.1W)
expY.given0W.star <- plogis(qlogis(expY.given0W) + epsilon * H.0W)


# tlooking at epsilon with another regression update
coef(glm(ObsData$Y ~ -1 + offset(qlogis(expY.givenAW.star)) + H.AW,
                                          family = "binomial"))
```

```
##          H.AW
## -3.454763e-17
```

```r
# interpretation??
## clever covariate not changing same as lab??

PsiHat.TMLE <- mean(expY.given1W.star - expY.given0W.star)
PsiHat.TMLE # 0.05065004
```

```
## [1] 0.05039597
```

```r
# comparing the estimates
c(PsiHat.SS, PsiHat.IPTW, PsiHat.TMLE)
```

```
## [1]  0.04584497 -1.63194126  0.05039597
```

```r
# ltme package
set.seed(123)

library(ltmle)

ltmle.SL <- ltmle(data = ObsData,
                  Anodes = "A",
                  Ynodes = "Y",
                  abar = list(1,0),
                  SL.library = SL.library,
                  estimate.time = F)
summary(ltmle.SL)
```

```
## Estimator:  tmle
## Call:
## ltmle(data = ObsData, Anodes = "A", Ynodes = "Y", abar = list(1,
##     0), SL.library = SL.library, estimate.time = F)
##
## Treatment Estimate:
##    Parameter Estimate:  0.80595
##     Estimated Std Err:  0.012208
##               p-value:  <2e-16
##     95% Conf Interval: (0.78202, 0.82988)
##
## Control Estimate:
##    Parameter Estimate:  0.75417
##     Estimated Std Err:  0.010213
##               p-value:  <2e-16
##     95% Conf Interval: (0.73415, 0.77418)
##
## Additive Treatment Effect:
##    Parameter Estimate:  0.051785
##     Estimated Std Err:  0.015795
##               p-value:  0.0010438
##     95% Conf Interval: (0.020826, 0.082743)
##
## Relative Risk:
##    Parameter Estimate:  1.0687
##   Est Std Err log(RR):  0.020156
##               p-value:  0.00098485
##     95% Conf Interval: (1.0273, 1.1117)
##
## Odds Ratio:
##    Parameter Estimate:  1.3539
##   Est Std Err log(OR):  0.094905
##               p-value:  0.0014119
##     95% Conf Interval: (1.1241, 1.6306)
```

```
# ltmle package provides estimates and inference for (under identifiability assumptions):

# 1. expected outcome under the exposure (treatment estimate) = 0.80595
# 2. expected outcome under no exposure (control estimate) = 0.75417
# 3. additive treatment effect = 0.051785  (THIS!)
```

```
# call ltmle with main terms parametric regression for both E(U|A,M) & P(A=1|W)

ltmle.parametric <- ltmle(data = ObsData,
                          Anodes = "A",
                          Ynodes = "Y",
                          abar = list(1,0),
                          Qform = c(Y = "Q.kplus1 ~ A + age +
                          factor(raceeth) + factor(educ) + factor(marital) +
                          bmi + waist + factor(depressed)"),
                          gform = "A ~ age + factor(raceeth) +
                          factor(educ) + factor(marital) +
                          bmi + waist + factor(depressed)",
                          estimate.time = F)
summary(ltmle.parametric)
```

```
## Estimator:  tmle
## Call:
## ltmle(data = ObsData, Anodes = "A", Ynodes = "Y", Qform = c(Y = "Q.kplus1 ~ A + age +\n
##     gform = "A ~ age + factor(raceeth) + \n                          factor(educ) + factor(marital) +
##     abar = list(1, 0), estimate.time = F)
##
```

```
## Treatment Estimate:
##    Parameter Estimate:  0.80235
##     Estimated Std Err:  0.012523
##               p-value:  <2e-16
##     95% Conf Interval: (0.7778, 0.82689)
##
## Control Estimate:
##    Parameter Estimate:  0.75366
##     Estimated Std Err:  0.010083
##               p-value:  <2e-16
##     95% Conf Interval: (0.73389, 0.77342)
##
## Additive Treatment Effect:
##    Parameter Estimate:  0.048693
##     Estimated Std Err:  0.015972
##               p-value:  0.002299
##     95% Conf Interval: (0.017388, 0.079998)
##
## Relative Risk:
##    Parameter Estimate:  1.0646
##   Est Std Err log(RR):  0.020421
##               p-value:  0.0021703
##     95% Conf Interval: (1.0228, 1.1081)
##
## Odds Ratio:
##    Parameter Estimate:  1.3269
##   Est Std Err log(OR):  0.09524
##               p-value:  0.0029807
##     95% Conf Interval: (1.1009, 1.5992)
```

```r
# call ltmle with unadjusted

ObsData <- data.frame(U=1, ObsData)

ltmle.unadj <- ltmle(data = ObsData,
                     Anodes = "A",
                     Ynodes = "Y",
                     abar = list(1,0),
                     Qform = c(Y = "Q.kplus1 ~ A"),
                     gform = "A ~ U",
                     estimate.time = F)
summary(ltmle.unadj)
```

```
## Estimator:  tmle
## Call:
## ltmle(data = ObsData, Anodes = "A", Ynodes = "Y", Qform = c(Y = "Q.kplus1 ~ A"),
##     gform = "A ~ U", abar = list(1, 0), estimate.time = F)
##
## Treatment Estimate:
##    Parameter Estimate:  0.81594
##     Estimated Std Err:  0.010293
##               p-value:  <2e-16
##     95% Conf Interval: (0.79576, 0.83611)
##
```

```
## Control Estimate:
##    Parameter Estimate:  0.75151
##     Estimated Std Err:  0.0096934
##               p-value:  <2e-16
##     95% Conf Interval: (0.73251, 0.77051)
##
## Additive Treatment Effect:
##    Parameter Estimate:  0.064429
##     Estimated Std Err:  0.014139
##               p-value:  5.1918e-06
##     95% Conf Interval: (0.036717, 0.09214)
##
## Relative Risk:
##    Parameter Estimate:  1.0857
##   Est Std Err log(RR):  0.018042
##               p-value:  5.1365e-06
##     95% Conf Interval: (1.048, 1.1248)
##
## Odds Ratio:
##    Parameter Estimate:  1.4658
##   Est Std Err log(OR):  0.085974
##               p-value:  8.6786e-06
##     95% Conf Interval: (1.2385, 1.7348)
```

```r
# --
# explore double robustness
# --

ltmle.DR <- ltmle(data = ObsData,
                  Anodes = "A",
                  Ynodes = "Y",
                  abar = list(1,0),
                  SL.library = SL.library,
                  gform = "A ~ U",
                  estimate.time = F)
summary(ltmle.DR)
```

```
## Estimator:  tmle
## Call:
## ltmle(data = ObsData, Anodes = "A", Ynodes = "Y", gform = "A ~ U",
##     abar = list(1, 0), SL.library = SL.library, estimate.time = F)
##
## Treatment Estimate:
##    Parameter Estimate:  0.81095
##     Estimated Std Err:  0.010329
##               p-value:  <2e-16
##     95% Conf Interval: (0.79071, 0.83119)
##
## Control Estimate:
##    Parameter Estimate:  0.75584
##     Estimated Std Err:  0.0096574
##               p-value:  <2e-16
##     95% Conf Interval: (0.73691, 0.77477)
##
```

```
## Additive Treatment Effect:
##    Parameter Estimate:  0.05511
##     Estimated Std Err:  0.014072
##               p-value:  8.9943e-05
##     95% Conf Interval: (0.027529, 0.082691)
##
## Relative Risk:
##    Parameter Estimate:  1.0729
##   Est Std Err log(RR):  0.017947
##               p-value:  8.808e-05
##     95% Conf Interval: (1.0358, 1.1113)
##
## Odds Ratio:
##    Parameter Estimate:  1.3857
##   Est Std Err log(OR):  0.084977
##               p-value:  0.00012376
##     95% Conf Interval: (1.1731, 1.6368)
```

```
# Additive treatment effect =  0.055591

# P.S: an estimator is consistent if the point estimates converge (in probability)
# to the estimand as sample size n tend to infinity

# Our sample size is 3406 (way more than the 1000 of lab5).
# Not sure if there’s a need to increase sample sizes.
```

```
# Alternative TMLE implementations

# calculate 2-dimensional clever covariate

H.1W <- as.numeric(ObsData$A==1)/probA1.givenW
H.0W <- as.numeric(ObsData$A==0)/probA0.givenW

# target

logitUpdate<- glm(ObsData$Y~ -1 + offset(qlogis(expY.givenAW)) +
+ H.0W + H.1W, family="binomial")

eps<-logitUpdate$coef
eps
```

```
##         H.0W        H.1W
## -0.003523222  0.023298560
```

```
# obtain the targeted estimates
expY.givenAW.star <- plogis(qlogis(expY.givenAW) + eps[’H.0W’]*H.0W + eps[’H.1W’]*H.1W)

expY.given1W.star <- plogis( qlogis(expY.given1W) + eps[’H.1W’]/probA1.givenW )

expY.given0W.star <- plogis(qlogis(expY.given0W) + eps[’H.0W’]/probA0.givenW )

TMLE2 <- data.frame(cbind(
psi1 = mean(expY.given1W.star),
```

```r
psi0 = mean(expY.given0W.star),
diff = mean(expY.given1W.star) - mean(expY.given0W.star),
ratio = mean(expY.given1W.star) /mean(expY.given0W.star)
))

TMLE2
```

```
##        psi1      psi0       diff    ratio
## 1 0.8101392 0.7566914 0.05344783 1.070634
```

```r
# diff = 0.05339405 (yeah!!!)
```