HAD7002 HW1
Due May 7

Question 1

(a) Based on the Table 2.2 data below, calculate the crude and standardized risk differences. Compare and interpret the results. Here we do non-model based standardization, using empirical proportions as inputs in the standardization formula. Point estimates are sufficient, we will cover calculation of confidence intervals and other uncertainty estimates later.

- of interests are crude and standardized risk difference, whose definitions are:

$$RD_{\text{crude}} = P(Y=1|A=1) - P(Y=1|A=0)$$

$$RD_{\text{standardized}} = \sum_{l=0}^{1} P(Y=1|A=1,L=l)P(L=l) - \sum_{l=0}^{1} P(Y=1|A=0,L=l)P(L=l)$$

- the data gives the following
  - $P(L=0) = \frac{8}{20} = 0.4$ and $P(L=1) = \frac{12}{20} = 0.6$
  - $P(Y=1|A=1,L=0) = \frac{1}{4}$ and $P(Y=1|A=1,L=1) = \frac{6}{9}$
  - $P(Y=1|A=0,L=0) = \frac{1}{4}$ and $P(Y=1|A=0,L=1) = \frac{2}{3}$
- compute $\hat{RD}_{\text{crude}}$ and $\hat{RD}_{\text{standardized}}$ from the data.

$$\hat{RD}_{\text{crude}} = P(Y=1|A=1) - P(Y=1|A=0) = \frac{7}{13} - \frac{3}{7} = 0.11$$

$$\hat{RD}_{\text{standardized}} = = (\frac{1}{4}0.4 + \frac{6}{9}0.6) - (\frac{1}{4}0.4 + \frac{2}{3}0.6) = 0.5 - 0.5 = 0$$

- compare crude and standardized risk differences
  From the calculation, we know that $\hat{RD}_{\text{crude}} = 0.11$ and $\hat{RD}_{\text{standardized}} = 0$. Before standardization, the crude RD estiamte is 0.11, indicating the risk of $Y = 1$ in $A = 1$ group is greater than that in $A = 0$ group, however, after standardization, the risk of $Y = 1$ in both gorups are the same. This may indicate that L is a potential confounder.

(b) Fit an appropriate regression model for the outcome, and use it to calculale risk difference using model-based standardization. In the model, you can include an interaction term between the exposure and covariate, or leave it out. Try both models, and see if it makes a difference. With a saturated model, you should be able to verify the result in (a) exactly.

- step 1: fit outcome model using actual data. we specify the model as:

$$\text{Model 1: } logit(P(Y=1|A,L)) = \alpha + \beta_1 A + \beta_2 L$$

```
Q1_m1_fit <- glm(Y ~ A+L,
                family=binomial(link=logit),
                data = Q1_dat)
```

- step 2: get the predicted values using `predict` function call:

```
## the data/had everyone in the treated and in the
   untreated
Q1_pred_dat <- data.frame(L = c(Q1_dat$L,Q1_dat$L) ,
                          A = c(rep(1,nrow(Q1_dat)),
                             rep(0,nrow(Q1_dat))))
## calculate the predicted value E(Y|A,L)
model1_pred= predict(Q1_m1_fit, Q1_pred_dat, type = "
   response") %>%
data.frame()

## calculate double expectation E(E(Y|A,L)) = E(
   predicted vals from model )
 model1_pred_dat <- cbind(Q1_pred_dat,model1_pred =
    model1_pred$.)

model1_pred_dat %>% unique()
# L A model1_pred
# 0 1   0.2500000
# 1 1   0.6666667
# 0 0   0.2500000
# 1 0   0.6666667
```

- step 3: calculate the risk difference as:

$$E(E[Y \mid A=1,L]) - E(E[Y \mid A=0,L])$$
$$= \frac{1}{n}\sum_{i=1}^{n}\Big\{E[Y_i \mid A_i=1,L_i;\hat{\alpha},\hat{\beta}_1\,\hat{\beta}_2] - E[Y_i \mid A_i=0,L_i;\hat{\alpha},\hat{\beta}_1\,\hat{\beta}_2]\Big\}$$

$$\hat{RD}_{\text{standardized}} = \frac{1}{20}\{(8\times 0.25+12\times 0.67)-(8\times 0.25+12\times 0.67)\} = 0.5-0.5 = 0$$

```
model1_pred_dat %>%
    group_by(A) %>%
    summarise(mean_diff =mean(model1_pred))
# A tibble: 2    2
#       A mean_diff
#   <dbl>     <dbl>
#      0     0.500
#      1     0.500
```

similarly, we calculated the standardized risk difference using a saturated outcome model (i.e., with A*L interaction). the model is specified as :

$$\text{Model 2: } logit(P(Y=1|A,L)) = \alpha + \beta_1 A + \beta_2 L + \beta_3 AL$$

The $\hat{RD}_{\text{standardized}}$ is 0.5 -0.5 =0. So we conclude that model 1 and model 2 did not make a difference.

```
## the model
Q1_m2_fit <- glm(Y ~ A*L,
                 family=binomial(link=logit),data = Q1_
                    dat)

## the data
Q1_pred_dat <- data.frame(L = c(Q1_dat$L,Q1_dat$L) ,
                          A = c(rep(1,nrow(Q1_dat)),

                                rep(0,nrow(Q1_dat)) ))
## calculate the predicted value E(Y|A,L)
model2_pred= predict(Q1_m2_fit, Q1_pred_dat, type = "
   response") %>%
   data.frame()

## calculate double expectation E(E(Y|A,L))
model2_pred_dat <- cbind(Q1_pred_dat,model2_pred = model2
   _pred$.)

model2_pred_dat %>% unique()
# L A model2_pred
# 0 1   0.2500000
# 1 1   0.6666667
# 0 0   0.2500000
# 1 0   0.6666667

model2_pred_dat %>%
   group_by(A) %>%
   summarise(mean_diff =mean(model2_pred))
# A tibble: 2    2
#       A mean_diff
#   <dbl>     <dbl>
#      0      0.500
#      1      0.500
```

Question 2 In the NHEFS data, we are interested in smoking cessation between 1971 (baseline) visit and 1982 (follow-up) visit as the exposure, and weight change between 1971 and 1982 as the outcome. Everyone in this dataset was a smoker at baseline. As potential confounders, we consider age, sex, race, weight at baseline,

college education or more (dichotomized), and daily number of cigarettes smoked at baseline

(a) Calculate the crude and standardized risk difference between quitting and not quitting. Compare and interpret the results. As you have to adjust for multiple covariates, some of these continuous, you have to use model-based standardization. Please explain and justify your modeling choices.

- crude mean difference
  the crude mean difference is defined as:

$$RD_{\text{crude}} = E(wt82\_71|qsmk = 1) - E(wt82\_71|qsmk = 0)$$

  The estiamte of RD_{crude} from the data is

$$\hat{RD}_{\text{crude}} = 4.53 - 1.98 = 2.55$$

```
Q2_crude_RD <- Q2_dat %>% group_by(qsmk) %>%
    summarise(mean_weight_change = mean(wt82_71))
Q2_crude_RD
# A tibble: 2    2
#   qsmk mean_weight_change
#  <dbl>               <dbl>
#     0                1.98
#     1                4.53
# 4.53 - 1.98
# [1] 2.55
```

- standardized mean difference is calculated using model-based standardization.
  - step 1: fit the outcome model. The model is specified as:

$$E(wt82\_71|qsmk, age, sex, race, wt71, education, smokeintensity)$$
$$= \beta_0 + \beta_1 qsmk + \beta_2 age + \beta_3 sex + \beta_4 race + \beta_5 wt71 + \beta_6 education + \beta_7 smokeintensit$$

```
## 1. dataset
data(nhefs_complete)
Q2_dat <- nhefs_complete %>%
    select(qsmk,wt82_71,age,sex,race,wt71,
    education,smokeintensity) %>%
    mutate(education = as.factor(ifelse(education
       ==5,1,0)))

### 2. glm
xvariables <- select(Q2_dat,-c(wt82_71))
xvariables <- colnames(xvariables)

```

```
formula <- as.formula(paste0("wt82_71 ~ ",paste(
   xvariables,collapse = "+")))

Q2_model <- glm(
    formula,
    family = gaussian(link="identity"),
    data = Q2_dat)
```

– step 2: The predicted weight change:

```
## 3. prediciton
Q2_pred_dat <-  rbind(
    select(Q2_dat,-c(qsmk,wt82_71)),
    select(Q2_dat,-c(qsmk,wt82_71))) %>%
    mutate(qsmk= c(rep(0,nrow(Q2_dat)) ,
                   rep(1,nrow(Q2_dat))))

Q2_pred= predict(Q2_model, Q2_pred_dat) %>%
  data.frame()

weight_pred_dat <- cbind(Q2_pred_dat,
                         weight_gain_pred = Q2_
                            pred$.)
```

– step 3: estimated mean difference and corresponding 95% CI

$$\hat{RD}_{\text{standardized}} = 5.12 - 1.78 = 3.34$$

The corresponding 95% CI is: [2.35,4.27]

```
weight_pred_dat %>%
   group_by(qsmk) %>%
   summarise(mean_weight_gain =mean(weight_gain_
      pred))
# A tibble: 2    2
#   qsmk mean_weight_gain
#  <dbl>            <dbl>
#     0             1.78
#     1             5.12
# 5.12 - 1.78
# [1] 3.34
```

```
input_dat <- Q2_dat
set.seed(1017)
boot_standard_means <- lapply(1:1000, function(x){
  boot_id <- sample(c(1:nrow(input_dat)), size=
     nrow(input_dat), replace=TRUE)
  boot_dat <- input_dat[boot_id,]
  standard_mean_f(dat = boot_dat)[[2]]
```

```
})
quantile(boot_standard_means %>% unlist(),
    c(0.025,0.975))
# 2.5%   97.5%
# 2.34975 4.27000
```

(b) Using an appropriate stratified standardization formula,and appropriate model(s), investigate whether there was effect modification by sex.
structural model can be used to model if sex is a potential effect modifier. The potential outcome structural model is specified as:

$$E(Y_{qsmk}|sex, \boldsymbol{\theta}) = \theta_0 + \theta_1 \times qsmk + \theta_2 sex \times qsmk + \theta_3 sex$$

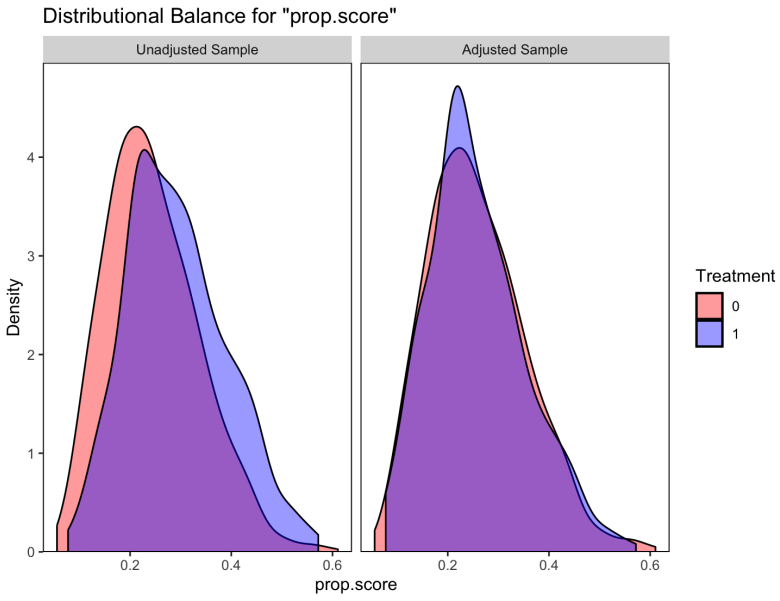
where $\theta_1 + \theta_2$ is the causal effect of qsmk on the averaged weight change in the male group, and θ_1 is the causal effect of qsmk on average weight change in the female group. If the effect if modified by sex, $\theta_1 + \theta_2 \neq \theta_1$, that is, $\theta_2 \neq 0$. we fit the model using a pseudo-population that is created using inverse probability weighting technique.

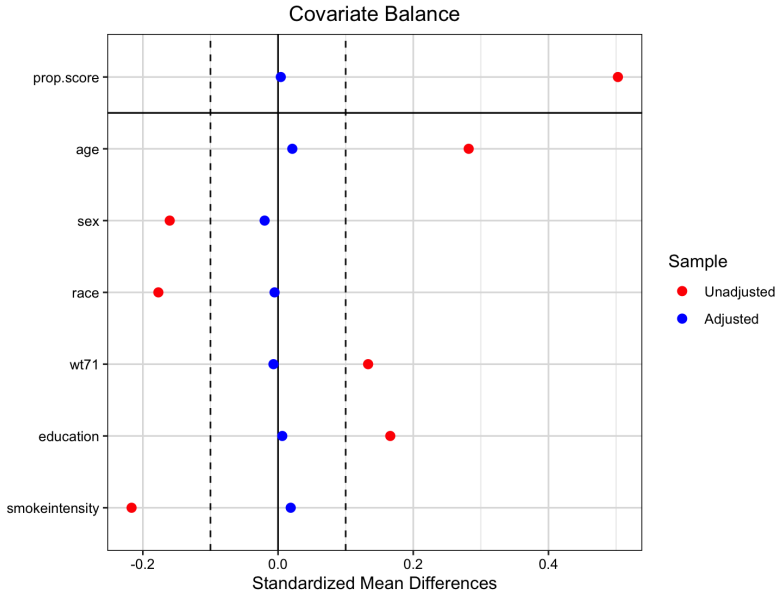
- create pseudo-population using IPTW. (weight is stabilized)

```
covaraites <- select(Q2_dat,-c(qsmk,wt82_71))
baselines <- colnames(covaraites)
ps_formula <- as.formula(paste0("qsmk ~ ",
                                paste(baselines,
                                   collapse = "+")))
IPTW <- weightit(ps_formula,
                 data = Q2_dat,
                 method = "glm", #using the default
                    logistic regression;
                 stabilize = TRUE)

# summary(IPTW)
```

- check the covariance balance: we see overlap of the propensity scores between two groups, and the standardized mean difference of the covariates between treatment groups are around zero.

```
library(cobalt)
bal.plot(IPTW,
         which="both",
         type = "density",
         colors = c("red","blue"))

love.plot(IPTW,
          binary = "std",
          grid = TRUE,
          thresholds = c(m = .1),
          colors = c("red","blue"))
```





- fit the potential outcome model using this weighted population. (by adding weight in the `svyglm` function call. From the output, we see the coefficient estimate of the interaction term $qsmk : sex1$ is -0.54, and the corresponding P-value is greater than a significance level of 0.05, so there is no statistical evidence for the hypothesis that sex is an effect modifier.

```
dweight <-svydesign(id=~1, ## no cluster
                    weights=~IPTW$weights, data=Q2_
                       dat)
Q2_fit_iptw <- svyglm(wt82_71 ~ qsmk*sex,
```

```
                        family = gaussian(link="
                           identity"),
                        design = dweight)
summary(Q2_fit_iptw)

# Call:
# svyglm(formula = wt82_71 ~ qsmk * sex, design =
   dweight, family # = gaussian(link = "identity"))

# Survey design:
# svydesign(id = ~1, weights = ~IPTW$weights, data =
   Q2_dat)

# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  1.81944    0.30434   5.978 2.79e-09 **
   *
# qsmk         3.63440    0.72962   4.981 7.02e-07 **
   *
# sex1        -0.03827    0.44263  -0.086    0.931
# qsmk:sex1   -0.54263    1.05814  -0.513    0.608
# ---
# Signif. codes:  0    ***    0.001    **    0.01
   *    0.05    .    0.1          1

# (Dispersion parameter for gaussian family taken to
   be 61.22353)

# Number of Fisher Scoring iterations: 2
```

Question 3 For the following questions please identify all paths between A and Y and write out the set of variables that satisfy the backdoor criterion.

(a) Consider the causal DAG below to write out the paths and the required set.

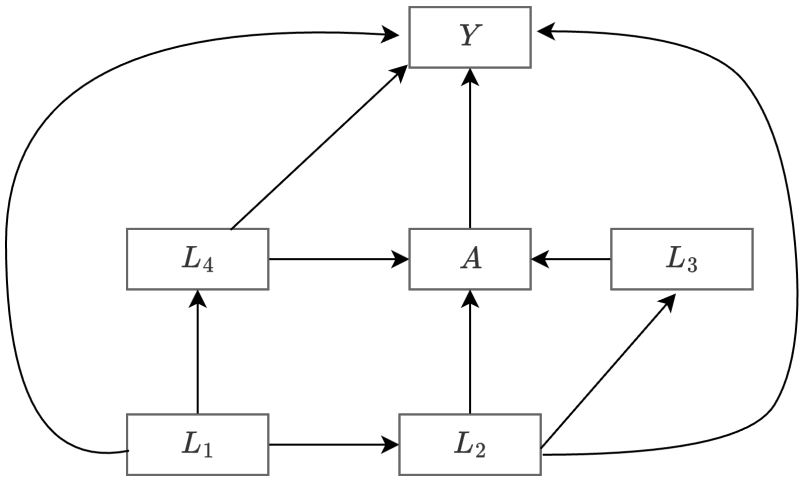
- There are 8 paths, namely:
  1. $A \rightarrow Y$
  2. $A \leftarrow L_2 \rightarrow Y$
  3. $A \leftarrow L_2 \leftarrow L_1 \rightarrow L_3 \rightarrow L_4 \rightarrow Y$
  4. $A \leftarrow L_2 \leftarrow L_1 \rightarrow L_4 \rightarrow Y$
  5. $A \leftarrow L_3 \rightarrow L_4 \rightarrow Y$
  6. $A \leftarrow L_3 \leftarrow L_1 \rightarrow L_2 \rightarrow Y$
  7. $A \leftarrow L_3 \leftarrow L_1 \rightarrow L_4 \rightarrow Y$
  8. $A \leftarrow L_3 \rightarrow L_4 \leftarrow L_1 \rightarrow L_2 \rightarrow Y$

- sets that satisfy the backdoor criterion

set 1: $\{L_2, L_3\}$
set 2: $\{L_2, L_4\}$
set 3: $\{L_1, L_2, L_3\}$
set 4: $\{L_1, L_2, L_4\}$
set 5: $\{L_2, L_3, L_4\}$
set 6: $\{L_1, L_2, L_3, L_4\}$

(b) Draw a causal DAG corresponding to the R code of simex3.r below. Identify paths between A and Y and the required covariates set.

- causal DAG



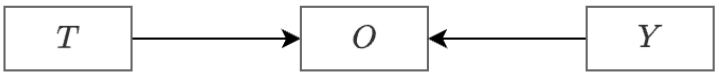
- There are 10 paths in total
  1. $A \rightarrow Y$
  2. $A \leftarrow L_2 \rightarrow Y$
  3. $A \leftarrow L_2 \leftarrow L_1 \rightarrow Y$
  4. $A \leftarrow L_2 \leftarrow L_1 \rightarrow L_4 \rightarrow Y$
  5. $A \leftarrow L_3 \leftarrow L_2 \rightarrow Y$
  6. $A \leftarrow L_3 \leftarrow L_2 \leftarrow L_1 \rightarrow Y$
  7. $A \leftarrow L_3 \leftarrow L_2 \leftarrow L_1 \rightarrow L_4 \rightarrow Y$
  8. $A \leftarrow L_4 \rightarrow Y$
  9. $A \leftarrow L_4 \leftarrow L_1 \rightarrow Y$
  10. $A \leftarrow L_4 \leftarrow L_1 \rightarrow L_2 \rightarrow Y$
- sets that satisfy the backdoor criterion

set 1: $\{L_2, L_4\}$
set 2: $\{L_2, L_3, L_4\}$
set 3: $\{L_1, L_2, L_4\}$

set 4: $\{L_1, L_2, L_3, L_4\}$

Question 4 A hypothetical SARS-CoV-2 Vaccine Trial randomized participants to a vaccine (T = 1) or placebo (T = 0). A safety study followed the participants for one year to determine the occurrence of adverse events (Y = 1) or not (Y = 0). Unfortunately, as is common for studies with a long follow-up period, several participants dropped out of the study early, so that their value for Y was missing. Let O = 1 indicate that Y was observed. Suppose participants are randomized to placebo and participants experiencing adverse events tended to drop out of the study early. Suppose also that, unknown to the investigators, the vaccine did not cause adverse events.

(a) Draw the causal DAG for these variables.



(b) What type of bias is presented in this example?

Selection bias (specifically, differential loss to follow up or informative censoring). If only participants with O = 1 are selected for analysis/conditional on O, the blocked backdoor path $T \rightarrow O \leftarrow Y$ by the collider is opened .

(c) Explain why assessing the safety of the vaccine by estimating the association between T and Y for those with O = 1 is a biased analysis.

There is only one blocked path from the treatment T to the outcome Y, $T \rightarrow O \leftarrow Y$, it is blocked by the common effect (collider). However, if conditioning on the common effect O, the blocked backdoor path is opened, we would observe association between T and Y, which is a result of selection bias.