STA640 Homework 5

Fan Bu

11/3/2020

PART 1

Load the data and required packages first

```r
library(tidyverse)
library(ggplot2)
set.seed(42)

dat = read.delim('data-ps5.txt', sep = ' ')
dat = dat %>% drop_na() # turns out there's no missing data...
#glimpse(dat)
```

Note that since every variable is binary, our estimand is

$$\tau = \mathbb{E}(Y(1,1) - Y(0,0)) = Pr(Y(1,1) = 1) - Pr(Y(0,0) = 1).$$

(a)

Using the formula shown in the lecture slides (here we also need to adjust for the baseline covariate X_1), we have

$$\hat{P}r(Y(1,1) = 1) = \sum_{X_2^{obs}=0,1} \hat{P}r(Y^{obs} \mid W_1 = 1, W_2 = 1, X_2^{obs}, X_1)\hat{P}r(X_2^{obs} \mid W_1 = 1, X_1),$$

and

$$\hat{P}r(Y(0,0) = 1) = \sum_{X_2^{obs}=0,1} \hat{P}r(Y^{obs} \mid W_1 = 0, W_2 = 0, X_2^{obs}, X_1)\hat{P}r(X_2^{obs} \mid W_1 = 0, X_1).$$

We first use R to calculate these two values separately for $X_1 = 0$ and $X_1 = 1$.

```r
# code is kinda convoluted, but I got OCD with tidyverse...

## separate data by X1
dat0 = dat %>% filter(x1==0)
dat1 = dat %>% filter(x1==1)

p_x1 = mean(dat$x1)

## function to calculate the two probs on a dataset
get_p1_p0 <- function(d){
  p1 = d %>% filter(w1==1, w2==1) %>%
  group_by(x2) %>%
  summarize(prob_y1 = mean(y), count_x2 = n()) %>%
  mutate(prob_x2 = count_x2/sum(count_x2)) %>%
  summarize(p1 = sum(prob_y1 * prob_x2)) %>%
  pull()

  p0 = d %>% filter(w1==0, w2==0) %>%
  group_by(x2) %>%
  summarize(prob_y1 = mean(y), count_x2 = n()) %>%
  mutate(prob_x2 = count_x2/sum(count_x2)) %>%
  summarize(p1 = sum(prob_y1 * prob_x2)) %>%
  pull()

  list(p1=p1, p0=p0)
}

# calculate p1 and p0 on X1=0 and X1=0 subset
probs0 = get_p1_p0(dat0)
probs1 = get_p1_p0(dat1)

cat('For X1=0: Pr(Y(1,1)=1) =',probs0$p1, ' Pr(Y(0,0)=1) =', probs0$p0, '\n')
```

```
## For X1=0: Pr(Y(1,1)=1) = 0.2162162  Pr(Y(0,0)=1) = 0.3797678
```

```
cat('For X1=1: Pr(Y(1,1)=1) =',probs1$p1, ' Pr(Y(0,0)=1) =', probs1$p0, '\n')
```

```
## For X1=1: Pr(Y(1,1)=1) = 0.5901639  Pr(Y(0,0)=1) = 0.6271186
```

And then we compute a weighted average of $\hat{\tau}_{(X_1=1)}$ and $\hat{\tau}_{(X_1=0)}$ to get an estimate for τ.

```
(tau = (1-p_x1) * (probs0$p1 - probs0$p0) + p_x1 * (probs1$p1 - probs1$p0))
```

```
## [1] -0.1321556
```

(b)

Do the following things:

1. Specify a model for outcome Y under "randomization":

$$\text{logit}(Pr(Y=1)) \sim W_1 + W_2 + W_1 \times W_2.$$

2. Build propensity score models for time 1 and 2

```
ps1 = glm(w1 ~ x1, data=dat, family = 'binomial')
ps2 = glm(w2 ~ x1 + w1 + x2, data=dat,
          family = 'binomial') # adding interactions doesn't seem helpful
```

and also the unconditional probabilities of treatment assignments at both time points:

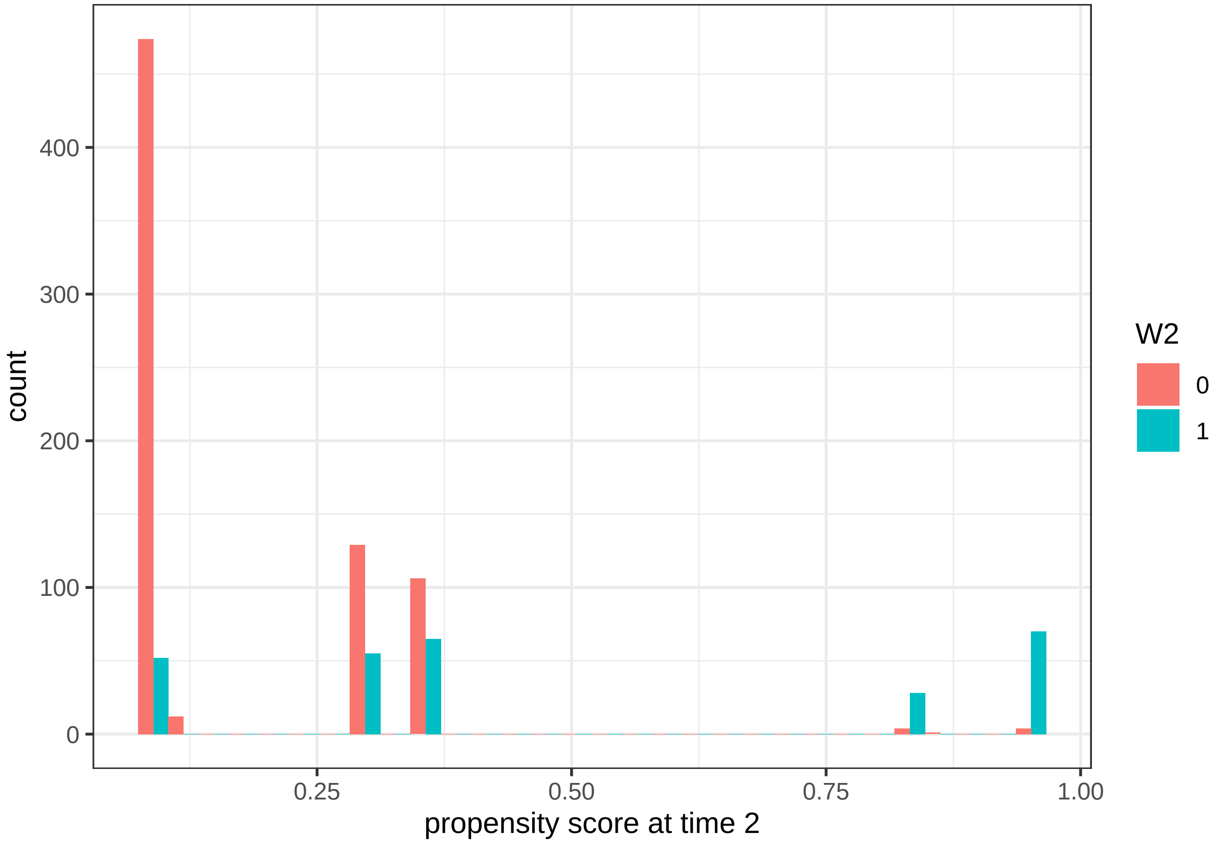
```
up1 = glm(w1 ~ 1, data=dat, family = 'binomial') # a constant prob here
up2 = glm(w2 ~ w1, data=dat, family = 'binomial')
```

3. Estimate the propensity score for all units at each time point and check for overlap.

```
library(broom)

expit <- function(x){ exp(x)/(1+exp(x)) }

# time point 1
e1 = ps1$fitted.values

## check Pr(W1=1) for X1=0 and 1 separately
augment(ps1, newdata = data.frame(x1=c(0,1))) %>%
  mutate(fitted_ps = expit(.fitted)) %>%
  select(x1, fitted_ps)
```

```
## # A tibble: 2 x 2
##      x1 fitted_ps
##   <dbl>     <dbl>
## 1     0    0.0559
## 2     1    0.262
```

```
# time point 2
e2 = ps2$fitted.values

e2_dat = data.frame(PS = e2, W2 = dat$w2)
ggplot(data=e2_dat, aes(x=e2, fill=factor(W2))) +
  geom_histogram(position = 'dodge') +
  labs(x='propensity score at time 2', fill='W2') +
  theme_bw()
```



It seems that at time 1, there is some imbalance, but at time 2 the imbalance is more severe. That being said, all the fitted propensity scores are not very extreme (all within [0.05, 0.96]) so I will not truncate any.

4. Calculate stabilized weights for all units.

```
sw = data.frame(e1 = e1, e2 = e2, up1 = up1$fitted.values,
                up2 = up2$fitted.values,
                w1 = dat$w1, w2 = dat$w2) %>%
  mutate(ipw1_inv = ifelse(w1==1, e1, 1-e1),
         ipw2_inv = ifelse(w2==1, e2, 1-e2)) %>%
  mutate(sw = (up1 * up2)/(ipw1_inv*ipw2_inv))
```

5. Fit the weighted outcome model (specified in 1).

```r
sms.mod = glm(y ~ w1 * w2, data=dat, weights = sw$sw,
              family = 'binomial')
```

And finally, obtain an estimate for τ using the fitted model:

```r
# (0,0) and (1,1)
contra = data.frame(w1=c(0,1), w2=c(0,1))
# calculate the estimated Pr(Y(1,1)=1) - Pr(Y(1,1)=1)
cat('Estimate for tau:\n')
```

```
## Estimate for tau:
```

```r
augment(sms.mod, newdata = contra) %>%
  mutate(prob = expit(.fitted)) %>%
  summarise(tau = diff(prob)) %>%
  select(tau) %>%
  pull()
```

```
## [1] -0.1393842
```

(c)

Do the following things (according to the steps in Section 3 of Keil et al., 2018).

1. Specify the joint model for (x_t, w_t, y) $(t = 1, 2)$ for the target population:

- a model for X_2 (intermediate outcome), determined by $p^{(x_2)}(x1, w1) = Pr(X_2 \mid X_1 = x_1, W_1 = w_1)$ (4 probabilities).
- a model for Y (outcome), determined by $p^{(y)}(x1, w1, x2, w2) = Pr(Y \mid X_1 = x_1, W_1 = w_1, X_2 = x_2, W_2 = w_2)$ (16 probabilities).

2. Specify the priors: $Unif(0, 1) = Beta(1, 1)$ for each probability in the above.
3. Sample from the target population via $p(X_1)$; here we simply use the empirical estimate for $Pr(X_1 = 1)$ in our sample data, which is 0.248.

```r
p_x1 = mean(dat$x1)
```

4. Set the treatment sequences (that we care about); they are $(W_1, W_2) = (1, 1)$ and $(W_1, W_2) = (0, 0)$.
5. Draw from the posterior distribution of parameters; note here we can utilize the Beta-Binomial conjugacy, and so with independent $Unif(0, 1)$ priors, the posterior distribution of the probability (of getting a 1) in each cell is essentially $Beta(1 + \text{num. of 1s}, 1 + \text{num. of 0s})$.

```r
# number of samples
S = 5000

# get X1's from p_x1
X1s = rbernoulli(S, p=p_x1) %>% as.numeric()

# the posteriors for X_2 probs
post_X2 = dat %>% count(x1,w1,x2) %>%
  mutate(post = n+1)

# posterior samples for X_2 probs (for w1=0 and w1=1)
get_probs_X2 <- function(x1_vec, w){
  res = numeric(length(x1_vec))
  n0 = sum(x1_vec==0)
  n1 = sum(x1_vec==1)
  a0 = post_X2 %>% filter(x1==0, w1==w, x2==1) %>%
    select(post) %>% pull()
  b0 = post_X2 %>% filter(x1==0, w1==w, x2==0) %>%
    select(post) %>% pull()
  #cat(a0, b0, '\n')
  res[x1_vec==0] = rbeta(n0, a0, b0)

  a1 = post_X2 %>% filter(x1==1, w1==w, x2==1) %>%
    select(post) %>% pull()
  b1 = post_X2 %>% filter(x1==1, w1==w, x2==0) %>%
    select(post) %>% pull()
  res[x1_vec==1] = rbeta(n1, a1, b1)
  #cat(a1, b1, '\n')
  res
}

probs_X2 = list('0' = get_probs_X2(X1s, 0), '1' = get_probs_X2(X1s, 1))


# the posteriors for Y probs
post_Y = dat %>% count(x1,w1,x2,w2,y) %>%
  mutate(post = n+1)
```

6. Draw the posterior predictive samples and get posterior samples for $\tau = Pr(Y(1,1) = 1) - Pr(Y(0,0) = 1)$.

```r
# sample p00 and p11 for Y
# given the X1 and X2 samples drawn

# draw X2 samples first
# under W1 = 0 and W1 = 1
X2s = list('0' = rbernoulli(S, p=probs_X2$`0`),
           '1' = rbernoulli(S, p=probs_X2$`1`))

# then draw probs of Y=1 under W=(0,0) and W=(1,1), conditioned on X1 and X2
get_probs_Y <- function(X1_vec, X2_vec, w_1, w_2){
  res = numeric(S)

  for(x_1 in c(0,1)){
    for(x_2 in c(0,1)){
      n_this = sum(X1_vec == x_1 & X2_vec == x_2)
      if(n_this > 0){
        a_this = post_Y %>%
        filter(x1==x_1, w1==w_1, x2==x_2, w2==w_2, y==1) %>%
        select(post) %>% pull()
        # if no observed data in cell, set it to prior
        if(length(a_this)==0){ a_this = 1}
        b_this = post_Y %>%
          filter(x1==x_1, w1==w_1, x2==x_2, w2==w_2, y==0) %>%
          select(post) %>% pull()
        # again, if no observed data in cell, set it to prior
        if(length(b_this)==0){ b_this = 1}

        res[X1_vec == x_1 & X2_vec == x_2] = rbeta(n_this, a_this, b_this)
      }
    }
  }
  res
}

probY_00 = get_probs_Y(X1s, X2s$`0`, 0,0)
probY_11 = get_probs_Y(X1s, X2s$`1`, 1,1)
```

Here we report posterior mean as well as a 95% credible interval for τ:

```
Bayes_tau = probY_11 - probY_00
cat('Posterior mean:', mean(Bayes_tau), '\n95% CI:', quantile(Bayes_tau,c(.025, .975)))
```

```
## Posterior mean: -0.1156828
## 95% CI: -0.3353844 0.1145443
```

(d)

In the context of this problem, the joint model of all variables can be factorized as

$$\begin{aligned}&p(X_1, W_1, X_2(0), X_2(1), W_2, Y(0,0), Y(0,1), Y(1,0), Y(1,0))\\=&p(Y(0,0), Y(0,1), Y(1,0), Y(1,0) \mid X_1, W_1, X_2(0), X_2(1), W_2)\\&\times p(W_2 \mid X_1, W_1, X_2(0), X_2(1))p(X_2(0), X_2(1) \mid X_1, W_1)p(W_1 \mid X_1)p(X_1).\end{aligned}$$

Then I need to specify 5 models in total:

1. model for $Y(0,0), Y(0,1), Y(1,0), Y(1,0)$ (final potential outcomes) given X_1, W_1, X_2, W_2
2. model for W_2 (2nd treament assignment) given X_1, W_1, X_2
3. model for $X_2(0), X_2(1)$ (potential intermediate outcomes) given X_1, X_1
4. model for W_1 (1st treament assignment) given X_1
5. model for X_1 (the target population distribution)

Here all of them can be specified as (marginal) binary outcome models, and again I can adopt independent $unif(0,1)$ priors for all the cell probabilities. Estimation for τ can be done by drawing from the posterior distributions of the parameters.

PART 2

For any treatment sequence (a_1, a_2, a_3), we have

$$\begin{aligned}&Pr(Y(a_1, a_2, a_3) = 1)\\=&\sum_{(X_1^{obs}, X_2^{obs}) \in \mathcal{X}} Pr(Y^{obs} = 1 \mid W_1 = a_1, X_1^{obs}, W_2 = a_2, X_2^{obs}, W_3 = a_3)\\&\qquad\times Pr(X_1^{obs} \mid W_1 = a_1)\\&\qquad\times Pr(X_2^{obs} \mid W_1 = a_1, X_1^{obs}, W_2 = a_2),\end{aligned}$$

where $\mathcal{X} = \{(0,0), (0,1), (1,0), (1,1)\}$ is the set of all possible combinations of (X_1^{obs}, X_2^{obs}).

Then to estimate $\tau = \mathbb{E}(Y(1,1,1) - Y(0,0,0))$ we will first estimate $Pr(Y(1,1,1) = 1)$ and $Pr(Y(0,0,0) = 1)$ and then take the difference.

Note that for treatments $(1,1,1)$, we only have $(X_1^{obs}, X_2^{obs}) = (1,1)$, so

$$Pr(Y(1,1,1) = 1) = 60\% \times 100\% \times 100\% = 60\%.$$

Then for treatments $(0, 0, 0)$, we need to sum over all four combinations:

$$\begin{aligned} & Pr(Y(0,0,0) = 1) \\ =& 0 + 40\% \times 50\% \times 50\% + 40\% \times 50\% \times 50\% + 60\% \times 50\% \times 50\% \\ =& 35\%. \end{aligned}$$

Therefore our estimate for the causal effect is

$$\hat{\tau} = 60\% - 35\% = 25\%.$$