NSW Worksheet

Math 394, Spring 2023

Josh & Malen

Instructions

Please complete this worksheet in groups of 2 or 3. List the names of the people in your group in the space provided above, and submit your completed worksheet PDF to Gradescope. **Only one person from each group needs to submit a completed worksheet to Gradescope.**

Data

The code below reads in the data:

```
library(tidyverse)
load("/Users/joshuayamamoto/Downloads/nsw_clean.rda")
```

We are using a subset of the the National Supported Work (NSW) Demonstration data that Dehejia and Wahba (1999) used to examine the effectiveness of "propensity score" methods (which we'll study soon). Each row in the data is a person, and the variables are characteristics of that person. The variables are listed below:

Problems

For each problem, put your solution between the bars of red stars in the .Rmd file.

Problem 1. First, we're going to try out some regression estimators on the NSW data.

(a) Calculate and report the unbiased *experimental* difference in means estimate, using the:

- Experimental Treated
- Experimental Control

```
nsw_clean %>%
  filter(group != "3. Non-Experimental Comparison") %>%
  group_by(group) %>%
  summarise(mean = mean(re78)) %>%
  pivot_wider( names_from = "group", values_from = "mean") %>%
  mutate(dim = `1. Experimental Treated` - `2. Experimental Control`) %>%
  select(dim)
```

```
## # A tibble: 1 x 1
##     dim
##   <dbl>
## 1 1794.
```

(b) Calculate and report the biased *non-experimental* difference in means estimate, using the:

- Experimental Treated
- Non-Experimental Comparison

Then, compare this estimate with the estimate from part (a).

```
nsw_clean %>%
  filter(group != "2. Experimental Control") %>%
  group_by(group) %>%
  summarise(mean = mean(re78)) %>%
  pivot_wider( names_from = "group", values_from = "mean") %>%
  mutate(dim = `1. Experimental Treated` - `3. Non-Experimental Comparison`) %>%
  select(dim)
```

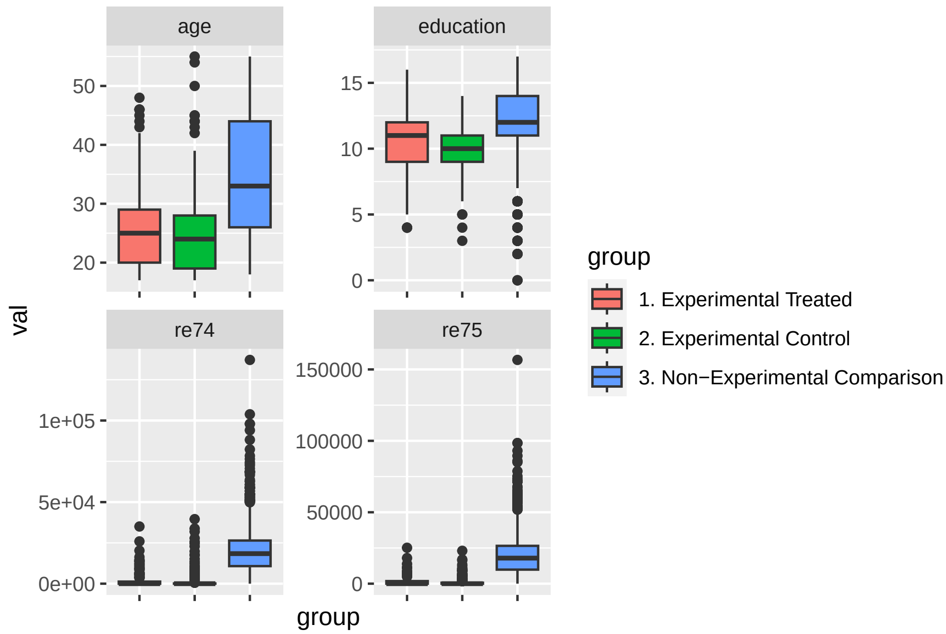
```
## # A tibble: 1 x 1
##       dim
##     <dbl>
## 1 -15205.
```

(c) Create a table/graphic that describes the differences between each of the three groups (per the `group` variable) on the covariates. Comment on any differences you see, and use them to explain why there is a huge difference between the estimators from parts (a) and (b).

---

non-experimental is higher in everything and often more variable.

```
library(ggridges)
nsw_clean %>%
  select(-c(black, hispanic, married, nodegree)) %>%
  pivot_longer(cols = age:re75, names_to = "var", values_to = "val") %>%
  ggplot(aes(y = val, x = group, fill = group)) +
  geom_boxplot() +
  facet_wrap(~var, scales = "free") +
  theme(
    axis.text.x = element_blank()
  )
```



non-experimental is less black, less hispanice, more married, less degree-ed.

```
nsw_clean %>%
  select(c(black, hispanic, married, nodegree, group)) %>%
  group_by(group) %>%
  summarise(across(black:nodegree, mean))
```

```
## # A tibble: 3 x 5
##   group                         black hispanic married nodegree
##   <chr>                         <dbl>    <dbl>   <dbl>    <dbl>
## 1 1. Experimental Treated       0.843   0.0595   0.189    0.708
## 2 2. Experimental Control       0.827   0.108    0.154    0.835
## 3 3. Non-Experimental Comparison 0.251   0.0325   0.866    0.305
```

(d) Using the:

- Experimental Treated
- Non-Experimental Comparison

calculate and report a $\hat{\tau}_{\text{ols}}$ estimator, its robust standard error, and a 95% confidence interval. Is your estimate close to estimate from (a)? Does your confidence interval contain the estimate from (a)? **Don't feel you need to restrict yourself to the untransformed** X – feel free to add non-linear terms (e.g., square or interaction terms)!

```
library(sandwich)
model_data <- nsw_clean %>%
  filter(group != "2. Experimental Control") %>%
```

```r
  mutate(D = ifelse(group == "1. Experimental Treated", 1, 0))



ols <- lm(sqrt(re78) ~ D +  1/(1+exp(-model_data$re75/1000000)) + education + age, model_data)

coef(ols)
```

```
## (Intercept)           D   education         age
##  42.1502467 -54.9603501   6.5288116   0.3134582
```

```r
library(sandwich)
V <- vcovHC(ols, 'HC0')
se <- sqrt(V["D", "D"])
ols$coefficients['D'] + se * c(qnorm(.025), qnorm(.975))
```

```
## [1] -62.75669 -47.16401
```

(e) Repeat part (d) using a $\hat{\tau}_{\text{Lin}}$ estimator. **Again, don't feel you need to restrict yourself to the untransformed X!**

(f) Now calculate another regression estimator of your choice (e.g., a more complex $\hat{\tau}_{\text{ols}}$ or $\hat{\tau}_{\text{Lin}}$, KRLS, random forest, etc.)! **You do not need to calculate a standard error or confidence interval.** As a reminder, you should be using the:

- Experimental Treated
- Non-Experimental Comparison

and your estimator should be of the form:

$$\widehat{\text{ATT}} = \frac{1}{n_1}\sum_{i:D_i=1} Y_i - \frac{1}{n_1}\sum_{i:D_i=1}\hat{f}_0(X_i) \quad \text{or} \quad \widehat{\text{ATT}} = \frac{1}{n_1}\sum_{i:D_i=1}\hat{f}_1(X_i) - \frac{1}{n_1}\sum_{i:D_i=1}\hat{f}_0(X_i)$$

where $\hat{f}_d$ is a regression model for $f_d(X) = E[Y(d) \mid X]$. *Note: It might take a while to run a machine-learning method (e.g., KRLS) on the full dataset. For the purposes of this exercise, it might make sense to either (i) randomly sample a subset of the full dataset, (ii) code up your answer in class and run your code when you get home, or (iii) choose a different method.*

(g) Which of the estimators from parts (d)-(f) was closest to the experimental difference in means from part (a)?

Problem 2. Now we're going to use the weights in the `kbalw` variable. Note that:

- `kbalw=1` for the Experimental Treated
- `kbalw=NA` for the Experimental Control
- `sum(kbalw)=2490` for the Non-Experimental Comparison

(a) Regress `re78` (Y) on `treated` (D) in a *weighted* least squares regression (use the `weights` option in the `lm()` function) and report the estimated coefficient for D, using the:

- Experimental Treated
- Non-Experimental Comparison

How close is the resulting estimate to the unbiased experimental estimate from Problem 1a? How does this estimate compare to the estimators you tried out in Problem 1? (Was this result annoying at all?)

```r
ols <- lm(re78 ~ D +  1/(1+exp(-model_data$re75/1000000)) + education + age, weights = kbalw, model_dat

coef(ols)["D"]
```

```
##        D
## 1677.296
```

(b) By performing the regression in part (a), you solved the following problem:

$$(\hat{\alpha}_{\text{wdim}}, \hat{\tau}_{\text{wdim}}) = \underset{\alpha,\tau}{\text{argmin}} \sum_{i=1}^{n} w_i \Big(Y_i - (\alpha + \tau D_i)\Big)^2$$

and reported $\hat{\tau}_{\text{wdim}}$ as an estimator. How does this optimization problem differ from the usual least squares problem? What changes do the weights imply?

The optimization differs from usual least squares in that we include the weights in order to account for the distributions of covariates within the non-experimental comparison group.

(c) Within each of the groups (per the `group` variable), create a table that reports the means of **at least five** covariates *and* **at least three** non-linear transformations of the covariates (e.g., square terms, interaction terms, log transformations, etc.). Then, add to the table the *weighted* mean of the same covariates and non-linear transformations among the Non-Experimental Comparison group (the `weighted.mean()` function might be useful):

$$\frac{1}{n_0} \sum_{i:D_i=0} w_i X_i$$

What do you notice?

The weights are making the untransformed covariate means equal across the groups

```
nsw_clean %>%
  select(age, re75, group) %>%
  group_by(group) %>%
  summarise(across(age:re75, .fns = list(mean = mean, mean_sqrt = ~ mean(sqrt(.x)), mean_cubed = ~ mean
```

```
## # A tibble: 3 x 7
##   group                           age_m~1 age_m~2 age_m~3 re75_~4 re75_~5 re75_~6
##   <chr>                             <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 1. Experimental Treated            25.8    5.04  21555.   1532.    22.0 1.76e11
## 2 2. Experimental Control            25.1    4.96  19934.   1267.    17.3 1.43e11
## 3 3. Non-Experimental Comparison     34.9    5.84  54102.  19063.   125.  2.10e13
## # ... with abbreviated variable names 1: age_mean, 2: age_mean_sqrt,
## #   3: age_mean_cubed, 4: re75_mean, 5: re75_mean_sqrt, 6: re75_mean_cubed
```

```
nsw_clean %>%
  select(age, re75, group, kbalw) %>%
  mutate(age = age*kbalw, re75 = re75*kbalw) %>%
  group_by(group) %>%
  summarise(across(age:re75, .fns = list(mean = mean, mean_sqrt = ~ mean(sqrt(.x)), mean_cubed = ~ mean
```

```
## # A tibble: 3 x 7
##   group                          age_m~1 age_m~2 age_m~3 re75_~4 re75_~5 re75_m~6
##   <chr>                            <dbl>   <dbl>   <dbl>   <dbl>   <dbl>    <dbl>
## 1 1. Experimental Treated           25.8   5.04   2.16e4   1532.   22.0   1.76e11
## 2 2. Experimental Control           NA     NA     NA          NA    NA     NA
## 3 3. Non-Experimental Comparis~     25.8   0.756  2.94e8   1532.    5.25  1.30e14
## # ... with abbreviated variable names 1: age_mean, 2: age_mean_sqrt,
## #   3: age_mean_cubed, 4: re75_mean, 5: re75_mean_sqrt, 6: re75_mean_cubed
```

(d) Now we'll try a weighted sampling exercise. Use the following code to sample values of `re75` from the Non-Experimental Comparison group with replacement, and proportional to the weights in `kbalw`:

```
set.seed(394)

# Making subset of data with only non-experimental comparison group
nec <- nsw_clean[nsw_clean$group=="3. Non-Experimental Comparison", ]

# Weighted sampling
wnec_re75 <- sample(nec$re75, size=1e5, replace=T, prob=nec$kbalw/sum(nec$kbalw))
```

Using histograms, compare the distributions of `re75` among the:

- Experimental Treated
- Non-Experimental Comparison
- The `wnec_re75` vector created above

What do you notice?

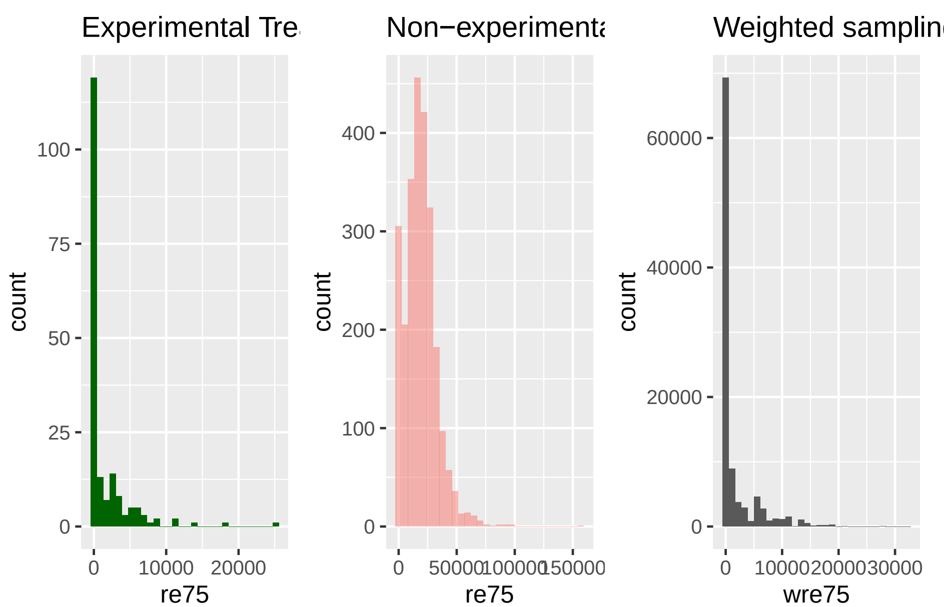
The weights are making the distribution of the continuous covariate in the non experimental comparison group look similar to the distribution of that covariate in the experimental treated group.

```
library(patchwork)
p1 <- nsw_clean %>%
  filter(group  == "1. Experimental Treated") %>%
  ggplot(aes(x = re75)) +
  geom_histogram(show.legend = F,  fill = "darkgreen") +
  labs(title = "Experimental Treated")

p2 <- nsw_clean %>%
  filter(group  == "3. Non-Experimental Comparison") %>%
  ggplot(aes(x = re75, fill = group)) +
  geom_histogram(alpha = 0.5, show.legend = F) +
  labs(title = "Non-experimental Comparison")

p3 <- tibble(
  wre75 = wnec_re75
) %>%
  ggplot(aes(x = wre75)) +
  geom_histogram() +
  labs(title = "Weighted sampling")

p1 + p2 + p3
```



(e) Repeat part (d) using a different *continuous* covariate.

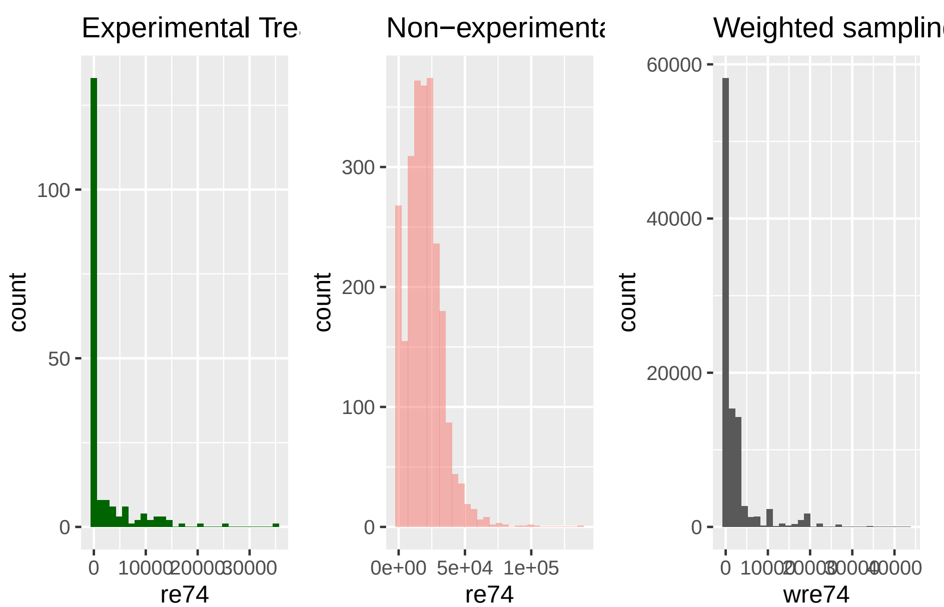
```r
wnec_re74 <- sample(nec$re74, size=1e5, replace=T, prob=nec$kbalw/sum(nec$kbalw))

p1 <- nsw_clean %>%
  filter(group  == "1. Experimental Treated") %>%
  ggplot(aes(x = re74)) +
  geom_histogram(show.legend = F,  fill = "darkgreen") +
  labs(title = "Experimental Treated")

p2 <- nsw_clean %>%
  filter(group  == "3. Non-Experimental Comparison") %>%
  ggplot(aes(x = re74, fill = group)) +
  geom_histogram(alpha = 0.5, show.legend = F) +
  labs(title = "Non-experimental Comparison")

p3 <- tibble(
  wre74 = wnec_re74
) %>%
  ggplot(aes(x = wre74)) +
  geom_histogram() +
  labs(title = "Weighted sampling")

p1 + p2 + p3
```



(f) As best you can, explain in words what you think the weights in `kbalw` are "doing"?

It appears that the kbalw weights work to make the distributions of the covariates in the non-experimental comparison group more closely align with the distribution of covariates in the experimental treated group. This allows us to recover the difference in means with greater accuracy by correcting for bias in the covariates.