Problem Set 1

Applied Methods

Andie Creel

2023-01-25

Question One: Randomization

This analysis will use the Dehijia and Wahba sample from the Lalonde dataset of the NSW experiment. The dataset is lalonde_nsw.csv. The outcome variable is re78 (real earnings in 1978). The treatment indicator is treat. The remaining variables are potential covariates. Assume for the purposes of this problem set that treat is completely randomly assigned.

a. Calculate the average treatment effect of the policy $E(\tau_i)$ using a simple difference in means.

Under random assignment, we know that $\hat{\tau}_{ATE} = \sum_i \frac{Y_{i,1}}{D_i} - \sum_i \frac{Y_{i,0}}{1-D_i}$.

```
# --------------------------------------------------------------
# read in data
# --------------------------------------------------------------
# myData <-
# vroom('https://raw.githubusercontent.com/paulgp/applied-methods-phd/main/homework/Homework1/lalonde_n
# vroom_write(myData, file = 'lalonde_nsw.csv')

myData <- vroom("lalonde_nsw.csv")

# --------------------------------------------------------------
# calculating tau
# --------------------------------------------------------------
# calculating number of treated and control
calc_ATE <- function(df) {
    D1 <- sum(df$treat)
    D0 <- sum(1 - df$treat)

    # sum of outcome for treated
    Y1_df <- df %>%
        filter(treat == 1) %>%
        select(re78)

    Y1 <- sum(Y1_df$re78)
    rm(Y1_df)

    # sum of outcome for control
    Y0_df <- df %>%
        filter(treat == 0) %>%
        select(re78)

    Y0 <- sum(Y0_df$re78)
    rm(Y0_df)

    # calculating tau hat
    t_hat <- Y1/D1 - Y0/D0
    return(t_hat)
}

t_hat <- calc_ATE(myData)
```

I find that $\hat{\tau}$ is 1794.34.

b. Calculate the average treatment effect on the treated of the policy $E(\tau_i|treat = 1)$. How does it compare to part a?

Because treatment is randomly assigned (and we assume there are no violators to their treatment status), $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$. Therefore the average treatment on the treated with be the same as the average treatment effect and the answer does not differ from part a.

c. Test the null of $\tau_i = 0$ for all i using a randomization test. *N.B.* Hold fixed the number of treated and control (e.g. assume the treatment count would be held fixed) and permute the labels randomly 1000 times – you do not need to fully do every permutation (there would be too many). Report the quantile that your estimate from the previous question falls.

```
# ------------------------------------------------------------------------------
# functions
# ------------------------------------------------------------------------------
set.seed(123)

# get a sample ATE
get_sample_ate <- function() {
    # step one: randomize assignment of treatment and
    # control
    mySample <- myData %>%
        mutate(treat = sample(treat, replace = FALSE))

    # step two: calculate t_hat
    calc_ATE(mySample)
}


# ------------------------------------------------------------------------------
# Create a distribution and find p value
# ------------------------------------------------------------------------------

# create a distribution
myDist <- replicate(1000, get_sample_ate())

# indicate if sample's tau is bigger than estimated
```

```r
# tau
myDist_df <- data.frame(sample_tau = myDist) %>%
    mutate(perc = if_else(sample_tau > t_hat, 1, 0))

# calc p value
pval <- sum(myDist_df$perc)/1000
```

I find that the p-value is 0.001.

d. Run a regression using robust standard errors (you may use canned software) of the outcome on the treatment dummy, and compare the p-values from this test to the previous answer.

```r
myReg <- lm(data = myData, formula = re78 ~ treat)

# robust standard errors
myRobustSE <- sqrt(diag(vcovHC(myReg)))

# treatment effect
tau_reg <- summary(myReg)$coefficients[2, 1]

# p-value for null 0 avg treatment effect in reg.1, vs
# 2 sided alternative
reg_p_val <- unname(round(2 * (1 - pnorm(abs(tau_reg/myRobustSE[2]))),
    4))
```

I find a p-value in part c that is less than .01, as does the canned regression that finds a p-value of 0.0076. The average treatment effects are equivalent, as the regression finds the treatment effect to be 1794.3423819.

Question Two: Propensity Scores

This analysis will use the dataset from Problem 1 as well as the PSID dataset from Dehijia and Wahba, lalonde-psid.csv. These datasets have identical variables. The new dataset is a sample of observations from the Panel Survey of Income Dynamics that can be used as alternative control observations. Importantly, these observations were not in the initial randomization.

a. Using age, education, hispanic, black, married, nodegree, RE74 and RE75, construct a propensity score using the treated group in lalonde-nsw.csv and the control sample of lalonde-psid.csv. Use a logit regression model to do so (you may use a canned routine to run the regression). Report the average p-score for the treated and control samples, and plot the propensity score densities for the treatment and control groups.

Step one: create new data set with treated from nsw and control for psid

Step two: calculate propensity scores using logit

$\pi(\mathbf{X_i}) = Pr(D_i = 1|\mathbf{X_i})$

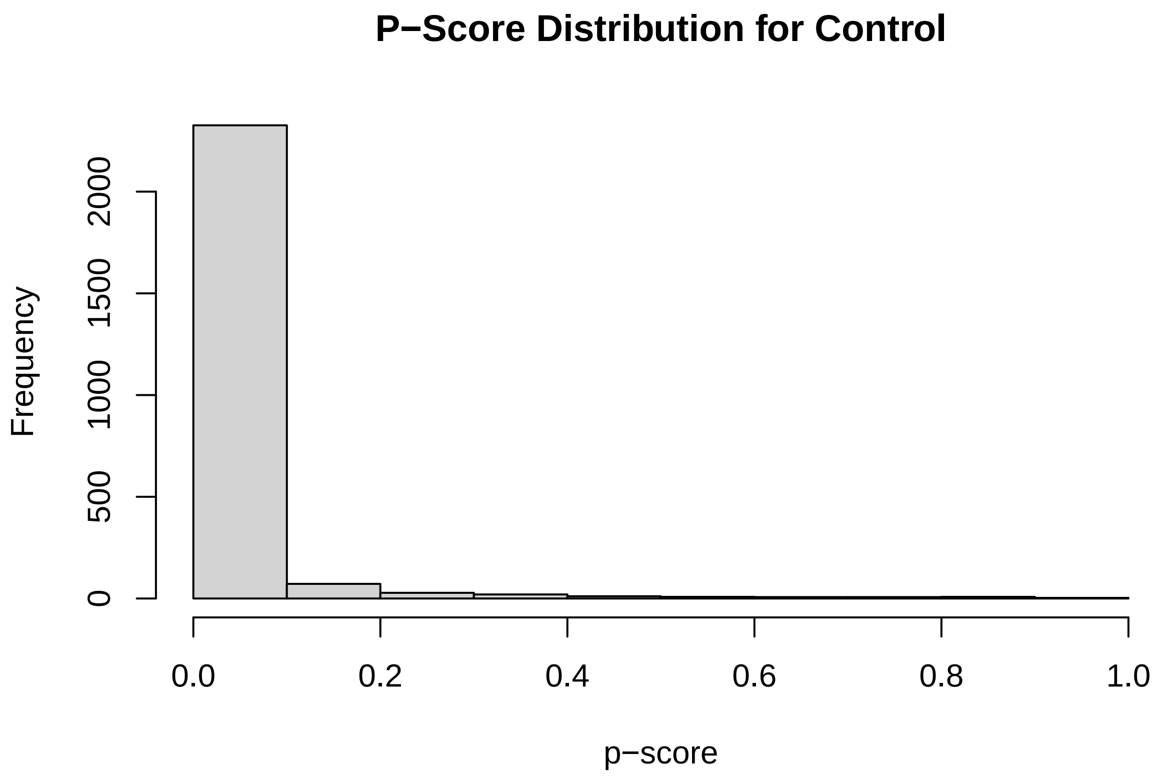
Step three: Subset by treated and control, calculate average propensity score for two groups.

```r
# mySample <-
# vroom('https://raw.githubusercontent.com/paulgp/applied-methods-phd/main/homework/Homework1/lalonde_p
# vroom_write(mySample, 'lalonde_psid.csv')

# ----------------------------------------------------------------------------
# read in data, make new data set
# ----------------------------------------------------------------------------
myData <- vroom("lalonde_nsw.csv") %>%
    filter(treat == 1)


mySample <- vroom("lalonde_psid.csv") %>%
    filter(treat == 0)

myData_2 <- bind_rows(myData, mySample)
rm(myData, mySample)

myData_2 <- myData_2 %>%
    select(treat, age, education, hispanic, black, married,
        nodegree, re74, re75, re78)


# ----------------------------------------------------------------------------
# calculate pscores
# ----------------------------------------------------------------------------

# logit
myLogit <- glm(treat ~ age + education + hispanic + black +
    married + nodegree + re74 + re75, data = myData_2, family = "binomial")
```

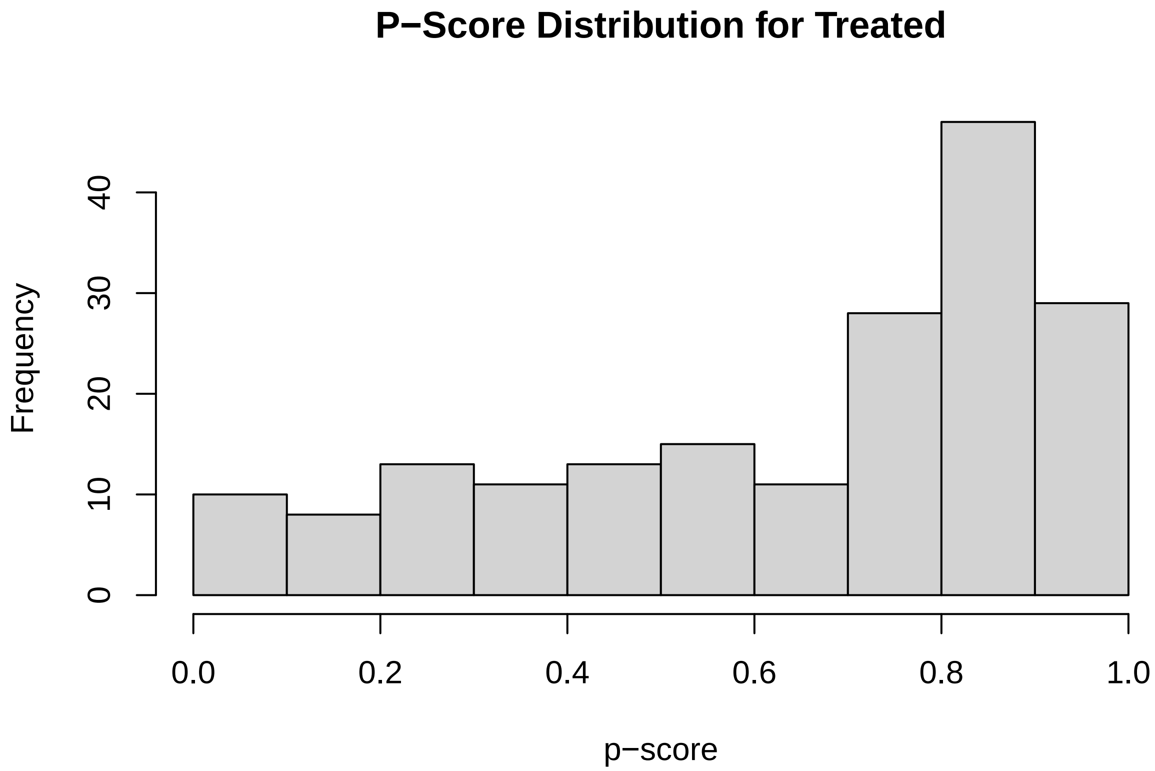
```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
# predict treatment given model (LOGIT RETURNS A LOG
# ODDS)
myData_2$log_odds <- predict(myLogit, newdata = myData_2)

# Transform to probability
myData_2$odds <- exp(myData_2$log_odds)
myData_2$p_score <- myData_2$odds/(1 + myData_2$odds)

# ----------------------------------------------------------------------------
# avg pscore for treated and control
# ----------------------------------------------------------------------------

# treated
treat_df <- myData_2 %>%
    filter(treat == 1)

p_score_treat <- mean(treat_df$p_score)

# control
control_df <- myData_2 %>%
    filter(treat == 0)
```

```
p_score_control <- mean(control_df$p_score)

# ----------------------------------------------------------------------------
# plots
# ----------------------------------------------------------------------------

hist(control_df$p_score, main = "P-Score Distribution for Control",
    xlab = "p-score")
```



```
hist(treat_df$p_score, main = "P-Score Distribution for Treated",
    xlab = "p-score")
```



I get an average p-score the treatment of 0.64 and 0.03 for the control.

b. Using your p-score estimates, calculate the IPW and SIPW estimate for control and treated mean of the outcome, and the average treatment effect. Contrast these estimates to the control mean of the outcome from the NSW sample, and the treatment effect from last week's problem set.

Notes to self: If we assume strong ignorability, $Y_i(0), Y_i(1) \perp\!\!\!\perp D_i|\mathbf{X_i}$, then we know $(Y_i(0), Y_i(1) \perp\!\!\!\perp D_i|\pi(\mathbf{X_i})$. This allows us to solve a high dimensional problem with a scalar value.

IPW

Horvitz-Thompson estimator (aka inverse probability estimator):

$$E[\tau_{IPW}] = E[\frac{Y_i D_i}{\pi(X_i)} - \frac{Y_i(1-D_i)}{1-\pi(X_i)}]$$

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_i \frac{Y_i D_i}{\pi_i} - \frac{Y_i(1-D_i)}{1-\pi_i}$$

```
# ----------------------------------------------------------------------------
# function outcome is re78
# ----------------------------------------------------------------------------

getHC <- function(df) {
    hc <- mean(df$re78 * df$treat/df$p_score - df$re78 *
        (1 - df$treat)/(1 - df$p_score))
}

tau_hat_HC <- getHC(myData_2)
```

I estimate the IPW is 7859.29.

SIPW

$$\hat{\tau}_{SIPW} = \frac{n^{-1}\sum_i \frac{Y_i D_i}{\hat{\pi}_i}}{n^{-1}\sum_i \frac{D_i}{\hat{\pi}_i}} - \frac{n^{-1}\sum_i \frac{Y_i(1-D_i))}{(1-\hat{\pi}_i)}}{n^{-1}\sum_i \frac{1-D_i}{1-\hat{\pi}_i}}$$

```
# ----------------------------------------------------------------------------
# function
# ----------------------------------------------------------------------------

getSPIW <- function(df) {
    # t_hat_spiw = term_1/temp_2 - term_3/term_4
    term_1 <- mean((df$re78 * df$treat)/df$p_score)
    term_2 <- mean(df$treat/df$p_score)

    term_3 <- mean((df$re78 * (1 - df$treat))/(1 - df$p_score))
    term_4 <- mean((1 - df$treat)/(1 - df$p_score))

    tau_hat_spiw <- term_1/term_2 - term_3/term_4

}

tau_hat_spiw <- getSPIW(myData_2)
```

I estimate the SIPW is -1.008635×10^4.

ATE

```
ate_b <- calc_ATE(myData_2)
```

I estimate the ATE is -1.520478×10^4.

All of these estimates are extremely different from the treatment effect estimated in question one.

c. Compare the ATE in the previous question to the treatment effect estimated using a linear regression using the PSID and NSW treatment sample, with age, education, hispanic, black, married, nodegree, RE74 and RE75 as controls.

```
myReg_1 <- lm(data = myData_2, formula = re78 ~ treat +
    age + education + hispanic + black + married + nodegree +
    re74 + re75)
summary(myReg_1)
```

```
##
## Call:
## lm(formula = re78 ~ treat + age + education + hispanic + black +
##     married + nodegree + re74 + re75, data = myData_2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -64870  -4302   -435   3786 110412
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -129.74267 1688.51705  -0.077   0.9388
## treat        751.94636  915.25722   0.822   0.4114
## age          -83.56559   20.81380  -4.015 6.11e-05 ***
## education    592.61019  103.30278   5.737 1.07e-08 ***
## hispanic    2163.28116 1092.29036   1.981   0.0478 *
## black       -570.92798  495.17772  -1.153   0.2490
## married     1240.51951  586.25390   2.116   0.0344 *
## nodegree     590.46694  646.78416   0.913   0.3614
## re74           0.27812    0.02792   9.960  < 2e-16 ***
## re75           0.56809    0.02756  20.613  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10070 on 2665 degrees of freedom
## Multiple R-squared:  0.5864, Adjusted R-squared:  0.585
## F-statistic: 419.8 on 9 and 2665 DF,  p-value: < 2.2e-16
```

I estimate an ATE of 751.95 when estimating using an OLS regression with controls. This is a very different treatment effect than the HC estimator and SIPW (which are very different from one another as well).

d. Now revisit your estimates from part a and b, and following Crump et al. (2009), discard all units with estimated propensities outside the range of [0.1, 0.9]. Reestimate the IPW and SIPW estimator of the ATE from part b using this trimmed sample.

```r
# trim
myData_3 <- myData_2 %>%
    filter(p_score < 0.9) %>%
    filter(p_score > 0.1)

# HC
tau_hat_HC_trim <- getHC(myData_3)

# SPIW
tau_hat_spiw_trim <- getSPIW(myData_3)
```

I find that $\hat{\tau}_{HC_{trim}}$ is 1060.38 and $\hat{\tau}_{SPIW_{trim}}$ is 783.86. These are more similar to the treatment effect estimated by the regression.

e. Finally, calculate the IPW and SIPW estimates for the ATE using this trimmed sample for Black and non-Black individuals. Compare this estimate to the ATE for Black and non-Black individuals using the full randomized sample.

IPW and SPIW for Black people (trimmed)

```
# Black
myData_black <- myData_3 %>%
    filter(black == 1)

# HC
tau_hat_HC_black <- getHC(myData_black)

# SPIW
tau_hat_spiw_black <- getSPIW(myData_black)
```

I find that $\hat{\tau}_{HC_{black}}$ is -44.02 and $\hat{\tau}_{SPIW_{black}}$ is 589.52.

IPW and SPIW for non-Black people (trimmed)

```
# Black
myData_non <- myData_3 %>%
    filter(black == 0)

# HC
tau_hat_HC_non <- getHC(myData_non)

# SPIW
tau_hat_spiw_non <- getSPIW(myData_non)
```

I find that $\hat{\tau}_{HC_{non}}$ is 5424.52 and $\hat{\tau}_{SPIW_{non}}$ is 1249.18.

ATE for Black and non-Black people (full sample)

```
myBlack_full <- myData_2 %>%
    filter(black == 1)

myNon_full <- myData_2 %>%
    filter(black == 0)

ate_black <- calc_ATE(myBlack_full)
ate_non <- calc_ATE(myNon_full)
```

I find that $\hat{\tau}_{ATE_{Black}}$ is -9733.24 and $\hat{\tau}_{ATE_{non-Black}}$ is -1.596082×10^4.

For Black people, the IPW and SIPW treatement affects vary greatly from one another as well as the ATE. This is true for non-Black people as well.