Problem set 3: RCTs, matching, and inverse probability weighting

David Segovia

10/25/21

Program overview

The metropolitan Atlanta area is interested in helping residents become more environmentally conscious, reduce their water consumption, and save money on their monthly water bills. To do this, Fulton, DeKalb, Gwinnett, Cobb, and Clayton counties have jointly initiated a new program that provides free rain barrels to families who request them. These barrels collect rain water, and the reclaimed water can be used for non-potable purposes (like watering lawns and gardening). Officials hope that families that use the barrels will rely more on rain water and will subsequently use fewer county water resources, thus saving both the families and the counties money.

Being evaluation-minded, the counties hired an evaluator (you!) before rolling out their program. You convinced them to fund and run a randomized controlled trial (RCT) during 2016, and the counties rolled out the program city-wide in 2019. You have the dataset: `barrels_rct.csv` with data from the RCT.

These dataset contain the following variables:

- `id`: A unique ID number for each household
- `water_bill`: The family's average monthly water bill, in dollars
- `barrel`: An indicator variable showing if the family participated in the program
- `barrel_num`: A 0/1 numeric version of `barrel`
- `yard_size`: The size of the family's yard, in square feet
- `home_garden`: An indicator variable showing if the family has a home garden
- `home_garden_num`: A 0/1 numeric version of `home_garden`
- `attitude_env`: The family's self-reported attitude toward the environment, on a scale of 1-10 (10 meaning highest regard for the environment)
- `temperature`: The average outside temperature (these get wildly unrealistic for the Atlanta area; just go with it)

Your goal

Your task in this problem set is to analyze the dataset to find the causal effect (or average treatment effect (ATE)) of this hypothetical program.

```
library(tidyverse)
library(broom)
library(patchwork)
library(MatchIt)


barrels_rct <- read_csv("barrels_rct.csv") %>%
  # This makes it so "No barrel" is the reference category
  mutate(barrel = fct_relevel(barrel, "No barrel"))
```

Finding causation from a randomized controlled trial

First, Check balance

Discuss the sample size for the RCT data and how many people were assigned to treatment/control. Are you happy with this randomization?

```
# Check for balance of numbers in the treatment and control groups

barrels_rct %>%
  count(barrel) %>%
  mutate(prop = n/sum(n))
```

```
## # A tibble: 2 x 3
##   barrel      n  prop
##   <fct>   <int> <dbl>
## 1 No barrel   221 0.449
## 2 Barrel      272 0.552
```

While in a perfect world we would like to assign 50% of the group to the barrel and 50% to the non barrel group, this is not always ideal. ~45% of the sample are not in the program but 55% is in the program.However, both are close enough to be in the program. As long as these groups were chosen by random such as by flipping a coin, this randomization is fine.

Check the balance of the main pre-treatment characteristics. Are you happy with the balance?

```
# You can check the balance of the RCT across different pre-treatment
# characteristics like `home_garden, yard_size, environmental attitudes,
# and average temperature.

barrels_rct %>%
  group_by(barrel) %>%
  summarise(prop_garden = mean(home_garden_num),
            ave_yard = mean(yard_size),
            ave_att = mean(attitude_env),
            ave_tem = mean(temperature))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 5
##   barrel    prop_garden ave_yard ave_att ave_tem
##   <fct>           <dbl>    <dbl>   <dbl>   <dbl>
## 1 No barrel       0.267   21309.    5.52    69.6
## 2 Barrel          0.206   20357.    5.42    69.8
```

The home garden: it seems like there are higher proportions in the people not participating in a program with a garden (0.26) than people in the program (0.20). average yard size: the average yard size is greater for the participants not in the program (~21309.02 square feet) while the yard size is less for participants in the program (20356.86) average environmental attitudes and average temperature both have equal proportions. Average environmental attitudes are, on average, around ~5 for both groups and average outside temperature is ~ 6.

Although the proportion is a bit uneven with the proportion with garden and average yard size, the difference is not huge, so I am happy with the balance.

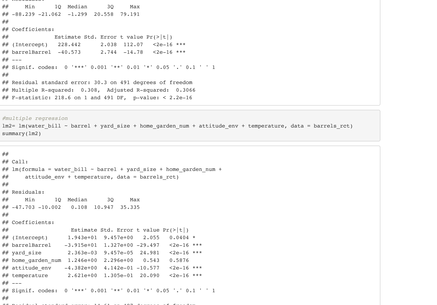
Make Some Graphs to see how balanced it is

```
## You don't need to graph all characteristics, choose one from them;
## Then use two different plots to show the differences between the treatment
## and control groups for the particular characteristic you picked.

plot_diff_garden <- ggplot(barrels_rct, aes(x = barrel, y = home_garden_num, color = barrel)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  guides(color = FALSE) +
  labs(x = NULL, y = "Proportion home garden")

plot_prop_garden <- ggplot(barrels_rct, aes(x = home_garden_num, fill = barrel)) +
  # Using position = "fill" makes the bars range from 0-1 and show the proportion
  geom_bar(position = "fill") +
  labs(x = NULL, y = "Proportion", fill = NULL) +
  scale_fill_manual(values = c("darkblue", "darkred"))

# Show the plots side-by-side
plot_diff_garden + plot_prop_garden
```



Second, Estimate difference

What is the average treatment effect (ATE)?

```
# Find the water bill used by those with barrel and those without, then get the ATE
# What does the number of ATE mean?


barrels_rct %>%
  group_by(barrel) %>%
  summarize(ATE= mean(water_bill))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 2
##   barrel     ATE
##   <fct>    <dbl>
## 1 No barrel  228.
## 2 Barrel     188.
```

The average water bill is $228.44 for no barrel and $187.86 for barrel for an average treatment effect of $40.58. So those who don't participate in the program pay $40.58 more than those who do participate in the program, this is the ATE.

What is the effect of the program on participants' water bills? How credible is this result?

```
# Based on the regression model result, explain the causal effect and make a conclusion.
barrels_rct
```

```
## # A tibble: 493 x 9
##       id water_bill barrel barrel_num yard_size home_garden home_garden_num
##    <dbl>      <dbl> <fct>       <dbl>     <dbl> <dbl>                 <dbl>
##  1     1        209. Barrel          1     25811 No home ga…               0
##  2     2        238. Barrel          1     39479 Home garden               1
##  3     3        164. Barrel          1     13297 No home ga…               0
##  4     4        226  Barrel          1     28259 No home ga…               0
##  5     5        232. No ba…          0     21479 No home ga…               0
##  6     6        279. No ba…          0     28986 Home garden               1
##  7     7        224. No ba…          0      7041 No home ga…               0
##  8     8        186. Barrel          1     29434 No home ga…               0
##  9     9        199. Barrel          1     24779 Home garden               1
## 10    10        172. Barrel          1     29741 Home garden               1
## # … with 483 more rows, and 2 more variables: attitude_env <dbl>,
## #   temperature <dbl>
```

```
#single regression
lm1= lm(water_bill ~ barrel, data = barrels_rct)
summary(lm1)
```

```
##
## Call:
## lm(formula = water_bill ~ barrel, data = barrels_rct)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.229 -21.062  -1.299  20.558  79.191
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   228.442      2.038  112.07   <2e-16 ***
## barrelBarrel  -40.573      2.744  -14.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.3 on 491 degrees of freedom
## Multiple R-squared:  0.308,  Adjusted R-squared:  0.3066
## F-statistic: 218.6 on 1 and 491 DF,  p-value: < 2.2e-16
```

```
#multiple regression
lm2= lm(water_bill ~ barrel + yard_size + home_garden_num + attitude_env + temperature, data = barrels_rct)
summary(lm2)
```

```
##
## Call:
## lm(formula = water_bill ~ barrel + yard_size + home_garden_num +
##     attitude_env + temperature, data = barrels_rct)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.703 -10.002   0.108  10.947  35.235
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.943e+01  9.457e+00   2.055   0.0404 *
## barrelBarrel -3.915e+01  1.123e+00 -29.497   <2e-16 ***
## yard_size     2.363e-03  9.457e-05  24.981   <2e-16 ***
## home_garden_num  1.246e+00  2.294e+00   0.543   0.5876
## attitude_env    -4.392e+00  4.142e-01 -10.577   <2e-16 ***
## temperature   2.421e+00  1.305e-01  20.090   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.61 on 487 degrees of freedom
## Multiple R-squared:  0.8465, Adjusted R-squared:  0.8388
## F-statistic: 513.2 on 5 and 487 DF,  p-value: < 2.2e-16
```

A simple *linear regression* shows that those in the program that receive free rain barrels pay $40.57 less than those not in the program on their water bills. The intercept coefficient also equals $228.44, which means the average monthly bill for the base group, which is for the participants not in the program. These two coefficients are statistically significant at the p<.001*** level.

When holding yard size, home garden, environmental attitudes, and outside temperature constant, the coefficient goes down to $39.15 but still shows a significant effect of the program on reducing their water bills. This is statistically significant at the p<.001*** level.The R-squared value also increases to 84.05%, which means that 84.05% of the variation in monthly bills can be explained by this model.

Plot the causal effect

```
# Use geom_pointrange to visualize the effect

barrel= barrels_rct %>%
  filter(barrel== "Barrel")
summary(barrel$water_bill)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   59.43  164.38  186.09  187.87  209.31  267.06
```

```
  #Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  #59.43  164.38  186.09  187.87  209.31  267.06

nobarrel= barrels_rct %>%
  filter(barrel== "No barrel")
summary(nobarrel$water_bill)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   156.7   208.3   229.1   228.4   247.1   293.3
```

```
  #Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  #156.7   208.3   229.1   228.4   247.1   293.3

ggplot(barrels_rct, aes(x = barrel, y = water_bill, color = barrel)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  guides(color = FALSE) +
  labs(x = NULL, y = "Monthly water bills")
```

Those in the program pay less on their water bills, with a range of $99.63-$267.06 with an average of $187.87

Those not in the program pay more on their water bills with a range of $154.7-$293.30 with an average of $228.40 on their monthly water bills.

At last, knit your file and submit it to your blackboard.