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

Jamie Esmond

March 02, 2023

Contents

List of Figures

List of Tables

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 gardens). 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 2018, and the counties rolled out the program city-wide in 2019. You have two datasets: `barrels_rct.csv` with data from the RCT, and `barrels_obs.csv` with observational data from self-selected participants.

These datasets 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 these two datasets to find the causal effect (or average treatment effect (ATE)) of this hypothetical program.

Follow these two examples from class as guides:

- RCTs
- Matching and IPW

As a reference, Figure 1 shows the DAG for the program:

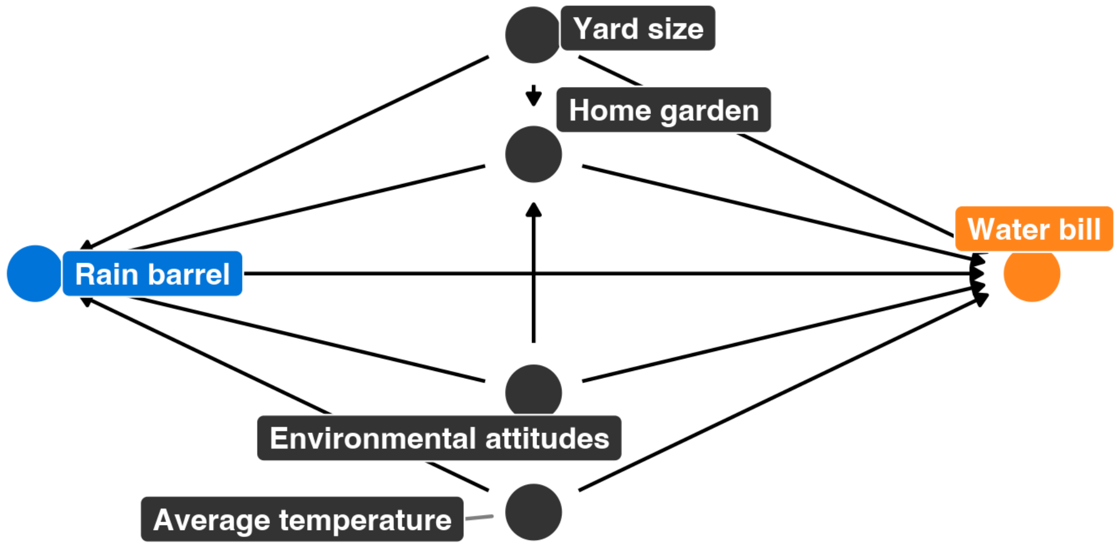

Figure 1: Rain barrel program DAG

```
library(tidyverse)
library(broom)
library(patchwork)
library(MatchIt)
library(kableExtra)
library(modelsummary)
library(huxtable)

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

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

1. Finding causation from a randomized controlled trial

Modified DAG

You can draw a DAG without any arrows to the treatment in an RCT because the random selection for the trial essentially controls for all possible confounding variables. Assuming the selection is truly random, there should be no fundamental difference between the treatment and control groups.

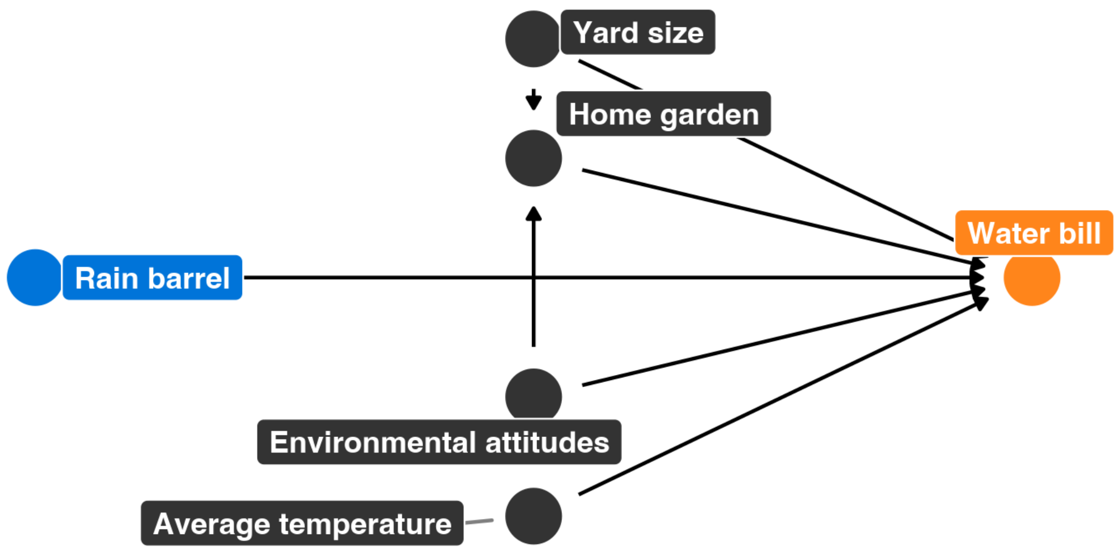

Figure 2: Rain barrel program DAG as an RCT

Check balance

The sample for the RCT contains 493 observations. 272 were assigned to treatment and the remaining 221 where in the control group. Since the two groups have close to the same amount of observations, they should be comparable for analysis. The probability of being in either group is split 45/55, which is close enough to 50/50.

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

barrel	n	prop
No barrel	221	0.448
Barrel	272	0.552

When checking the balance of the pre-treatment characteristics, average yard size, attitudes about the environment, and average temperature are very similar. The only characteristic that had some variation was the presence of a home garden. Those in the control group more likely to have a home garden, but the difference is not statistically significant.

```
barrels_rct %>%
  group_by(barrel) %>%
  summarise(avg_yard = mean(yard_size),
            prop_garden = mean(home_garden_num),
            avg_attitude = mean(attitude_env),
            avg_temp = mean(temperature))
```

barrel	avg_yard	prop_garden	avg_attitude	avg_temp
No barrel	2.13e+04	0.267	5.52	69.6
Barrel	2.04e+04	0.206	5.42	69.8

```
plot_diff_yard <- ggplot(barrels_rct, aes(x = barrel, y = yard_size, color = barrel)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  guides(color = "none") +
  labs(x = NULL, y = "Proportion with Home Garden")

plot_hist_yard <- ggplot(barrels_rct, aes(x = yard_size, fill = barrel)) +
  geom_histogram(binwidth = 1000, color = "white") +
  guides(fill = "none") +
  labs(x = "Yard Size", y = "Count") +
  facet_wrap(vars(barrel), ncol = 1)

plot_diff_yard + plot_hist_yard
```

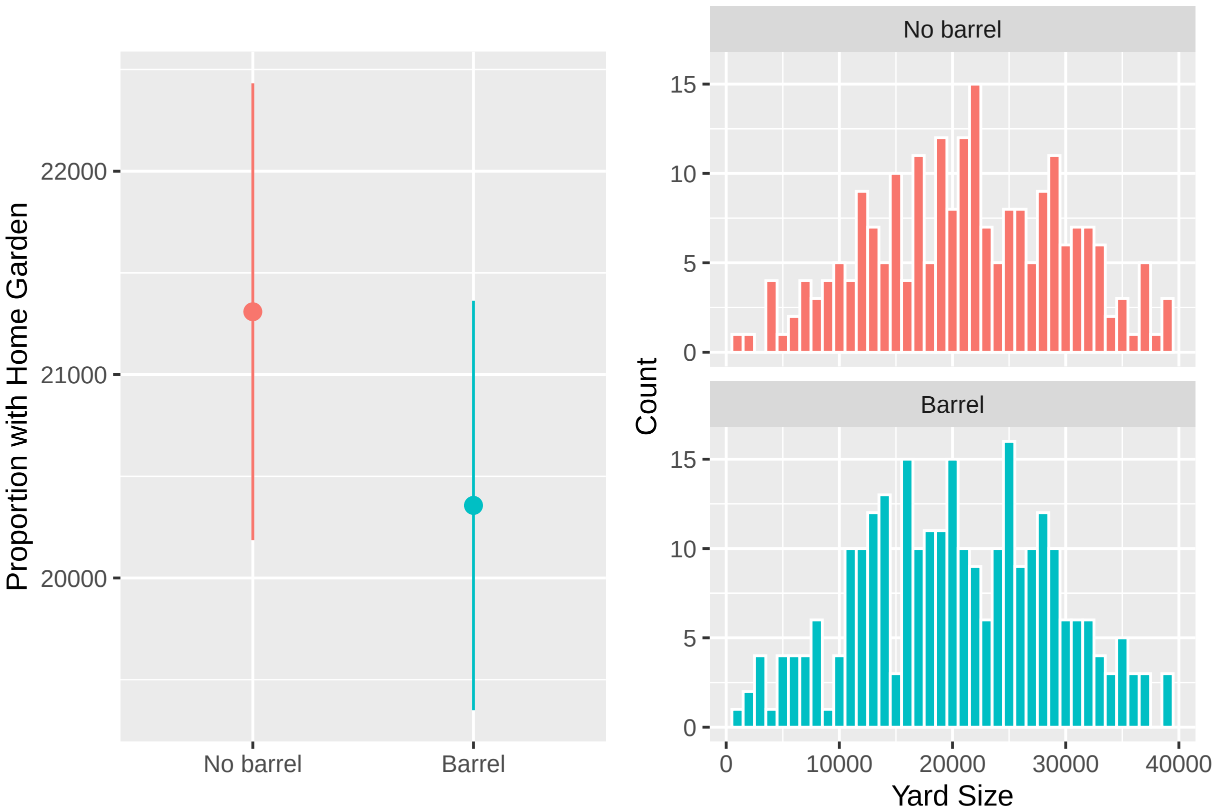

Figure 3: Balance of Pre-treatment Yard Size

```r
plot_diff_barrels_rct <- 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 = "none") +
  labs(x = NULL, y = "Proportion with Home Garden")

plot_prop_barrels_rct <- ggplot(barrels_rct, aes(x = barrel, fill = home_garden)) +
  geom_bar(position = "fill") +
  labs(x = NULL, y = "Proportion", fill = NULL) +
  scale_fill_manual(values = c("darkblue", "darkred")) +
  theme(legend.position = "bottom")


plot_diff_barrels_rct + plot_prop_barrels_rct
```

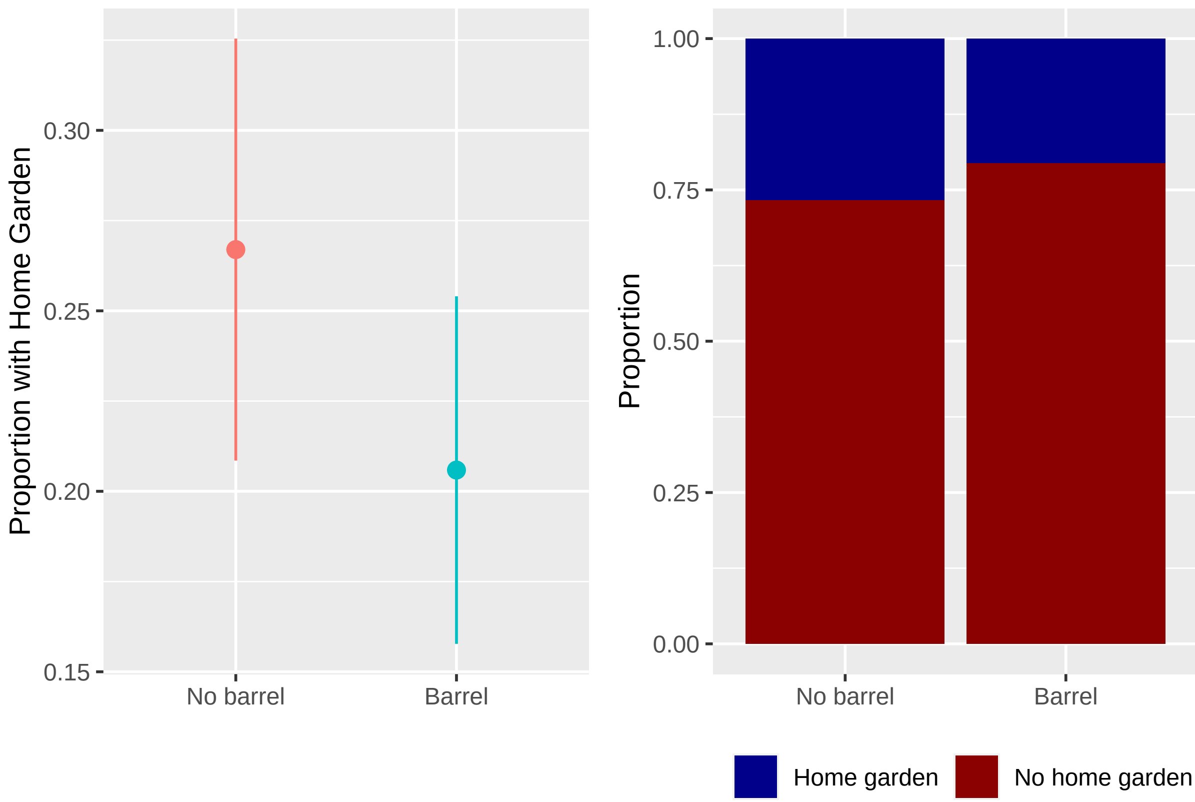

Figure 4: Balance of Pre-treatment Home Garden

```r
plot_diff_attitude <- ggplot(barrels_rct, aes(x = barrel, y = attitude_env, color = barrel)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  guides(color = "none") +
  labs(x = NULL, y = "Proportion with Home Garden")

plot_hist_attitude <- ggplot(barrels_rct, aes(x = attitude_env, fill = barrel)) +
  geom_histogram(binwidth = 1, color = "white") +
  guides(fill = "none") +
```

```
  labs(x = "Yard Size", y = "Count") +
  facet_wrap(vars(barrel), ncol = 1)

plot_diff_attitude + plot_hist_attitude
```

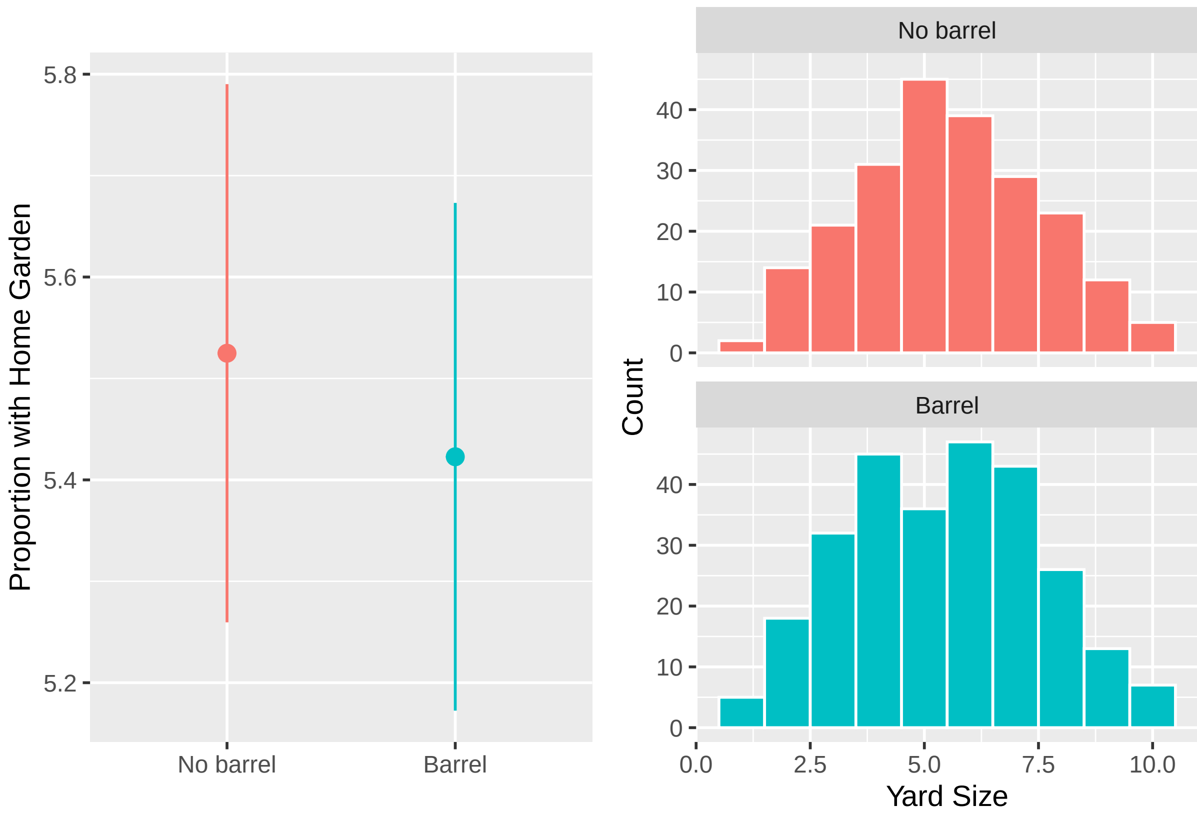

Figure 5: Balance of Pre-treatment Attitude about the Environment

```
plot_diff_temp <- ggplot(barrels_rct, aes(x = barrel, y = temperature, color = barrel)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  guides(color = "none") +
  labs(x = NULL, y = "Proportion with Home Garden")

plot_hist_temp <- ggplot(barrels_rct, aes(x = temperature, fill = barrel)) +
  geom_histogram(binwidth = 5, color = "white") +
  guides(fill = "none") +
  labs(x = "Yard Size", y = "Count") +
  facet_wrap(vars(barrel), ncol = 1)

plot_diff_temp + plot_hist_temp
```

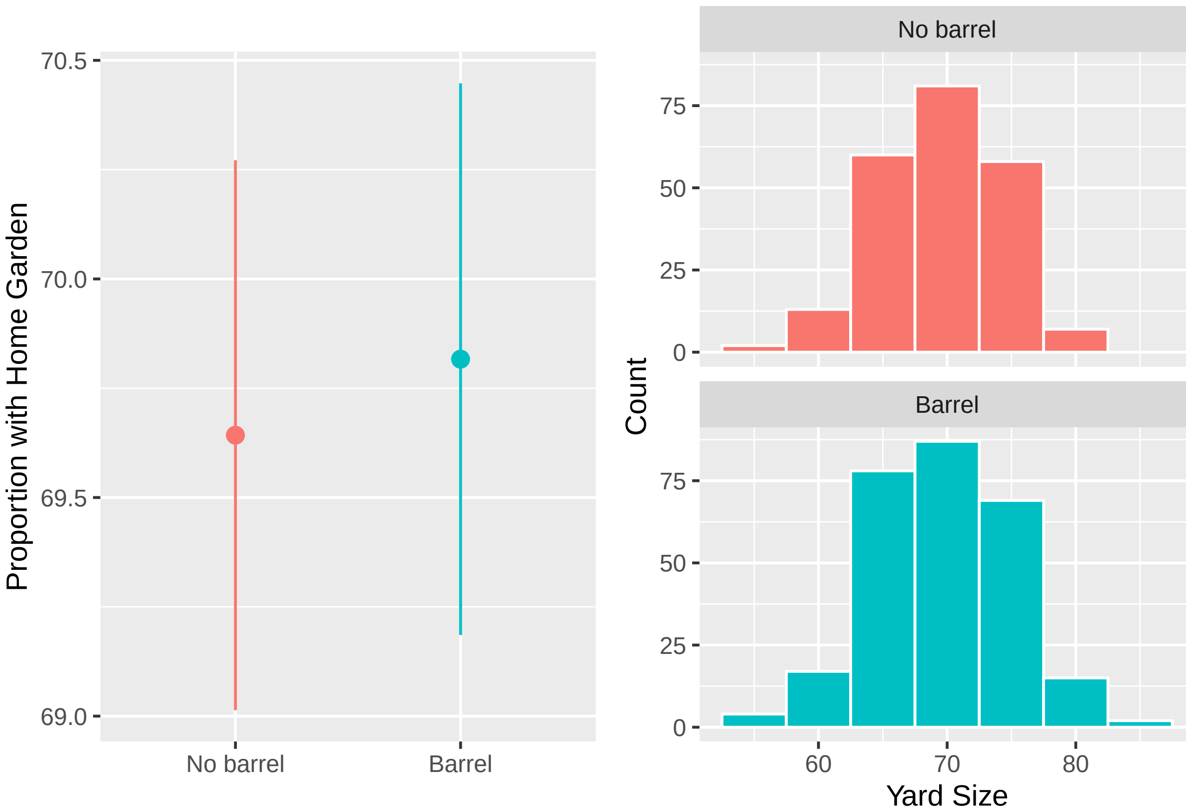

Figure 6: Balance of Pre-treatment Temperature

Estimate difference

According to the data, the difference in the water bill charges between the treatment and control group is $40.57. This result is statistically significant. However, there may be credibility issues, considering there are other unobserved variables which could influence the price of the water bill, such as number of people in the household.

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

barrel	avg_bill
No barrel	228
Barrel	188

```
model_rct <- lm(water_bill ~ barrel, data = barrels_rct)

modelsummary(model_rct,
             coef_rename = c(barrelBarrel = "Barrel"),
             output = "kableExtra",
             statistic = "conf.int",
             title = "RCT Results") %>%
    row_spec(c(1,3,5,7,9,11), background = "#f7fabe")
```

```
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 = "none") +
  labs(x = NULL, y = "Water Bill")
```

Table 1: RCT Results

	(1)
(Intercept)	228.442
	[224.437, 232.447]
Barrel	−40.573
	[−45.966, −35.181]
Num.Obs.	493
R2	0.308
R2 Adj.	0.307
AIC	4766.6
BIC	4779.2
Log.Lik.	−2380.295
RMSE	30.24

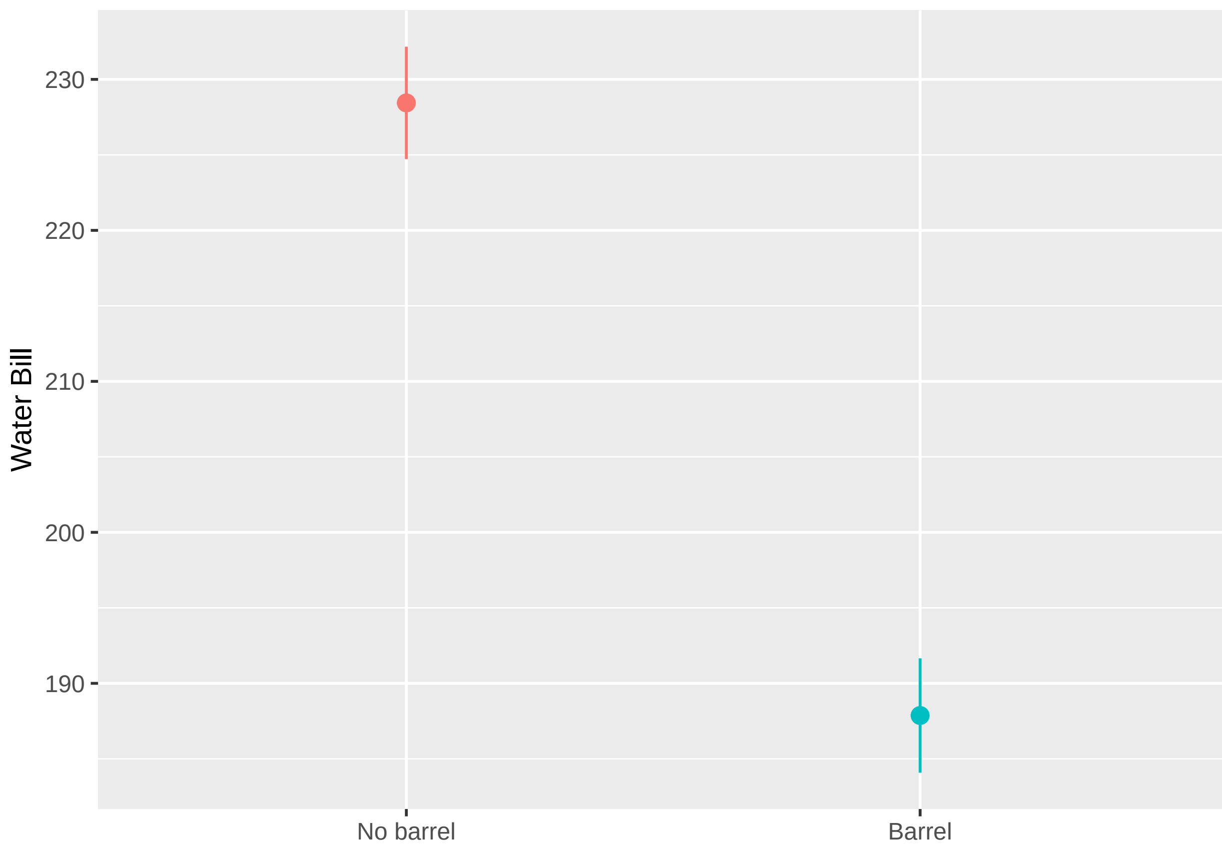

Figure 7: Barrel Water Bill Difference

2. Finding causation from observational data

Naive difference in means

The average difference in water bills for those in the program and those not in the program from the observational data is that those in the program had water bills that were $29.86 lower than those not in the program. This result is not terribly credible because there are several confounding variables which could influence the treatment and the outcome.

```
barrels_obs %>%
  group_by(barrel) %>%
  summarize(number = n(),
            avg = mean(water_bill))
```

barrel	number	avg
No barrel	736	225
Barrel	505	195

```
model_bill <- lm(water_bill ~ barrel, data = barrels_obs)

modelsummary(model_bill,
             coef_rename = c(barrelBarrel = "Barrel"),
             output = "kableExtra",
             statistic = "conf.int",
             title = "Naive difference in means") %>%
    row_spec(c(1,3,5,7,9,11), background = "#f7fabe")
```

Table 2: Naive difference in means

	(1)
(Intercept)	224.800
	[222.705, 226.895]
Barrel	−29.860
	[−33.144, −26.576]
Num.Obs.	1241
R2	0.204
R2 Adj.	0.204
AIC	11 881.0
BIC	11 896.3
Log.Lik.	−5937.482
RMSE	28.95

Adjustment with Mahalanobis nearest-neighbor matching

```
matched_data <- matchit(barrel_num ~ yard_size + home_garden_num + attitude_env + temperature,
                        data = barrels_obs,
                        method = "nearest",
                        distance = "mahalanobis",
                        replace = TRUE)
summary(matched_data)
```

```
##
## Call:
## matchit(formula = barrel_num ~ yard_size + home_garden_num +
##     attitude_env + temperature, data = barrels_obs, method = "nearest",
##     distance = "mahalanobis", replace = TRUE)
##
## Summary of Balance for All Data:
##                 Means Treated Means Control Std. Mean Diff. Var. Ratio
## yard_size          21173.5505    20166.7283          0.1192     1.0373
## home_garden_num        0.2574        0.1848          0.1661          .
## attitude_env           5.5426        5.1196          0.2078     1.0604
## temperature           71.7804       68.6484          0.6243     1.1025
##                 eCDF Mean eCDF Max
## yard_size          0.0368   0.0706
## home_garden_num    0.0726   0.0726
## attitude_env       0.0423   0.1058
## temperature        0.1284   0.2692
##
## Summary of Balance for Matched Data:
##                 Means Treated Means Control Std. Mean Diff. Var. Ratio
## yard_size          21173.5505    21082.9505          0.0107     1.0600
## home_garden_num        0.2574        0.2574          0.0000          .
## attitude_env           5.5426        5.5347          0.0039     1.0316
## temperature           71.7804       71.5305          0.0498     1.1064
##                 eCDF Mean eCDF Max Std. Pair Dist.
## yard_size          0.0098   0.0337          0.1454
## home_garden_num    0.0000   0.0000          0.0000
## attitude_env       0.0059   0.0178          0.0486
## temperature        0.0114   0.0475          0.1692
##
## Sample Sizes:
##               Control Treated
## All            736.       505
## Matched (ESS)  208.52     505
## Matched        300.       505
## Unmatched      436.         0
## Discarded        0.         0
```

```
matched_data_for_real <- match.data(matched_data)
```

```
model_match <- lm(water_bill ~ barrel,
                  data = matched_data_for_real)
```

Table 3: Matched Results

	(1)
(Intercept)	230.901
	[227.605, 234.198]
Barrel	−35.961
	[−40.123, −31.800]
Num.Obs.	805
R2	0.264
R2 Adj.	0.263
AIC	7714.6
BIC	7728.6
Log.Lik.	−3854.281
RMSE	29.05

Table 4: Matched & Weighted Results

	(1)
(Intercept)	234.371
	[231.077, 237.664]
Barrel	−39.431
	[−43.589, −35.273]
Num.Obs.	805
R2	0.301
R2 Adj.	0.301
AIC	7758.9
BIC	7772.9
Log.Lik.	−3876.430
RMSE	29.13

```r
modelsummary(model_match,
             coef_rename = c(barrelBarrel = "Barrel"),
             output = "kableExtra",
             statistic = "conf.int",
             title = "Matched Results") %>%
    row_spec(c(1,3,5,7,9,11), background = "#f7fabe")
```

```r
model_match_wts <- lm(water_bill ~ barrel,
                      data = matched_data_for_real,
                      weights = weights)

modelsummary(model_match_wts,
             coef_rename = c(barrelBarrel = "Barrel"),
             output = "kableExtra",
             statistic = "conf.int",
             title = "Matched & Weighted Results") %>%
    row_spec(c(1,3,5,7,9,11), background = "#f7fabe")
```

Adjustment with inverse probability weighting

```
model_barrel <- glm(barrel ~ yard_size + home_garden + attitude_env + temperature,
                    data = barrels_obs,
                    family = binomial(link = "logit"))

barrel_probabilities <- augment_columns(model_barrel,
                                         barrels_obs,
                                         type.predict = "response") %>%
  rename(propensity = .fitted)

barrel_probabilities %>%
  select(id, barrel, yard_size, home_garden, attitude_env, temperature, propensity) %>%
  head()
```

id	barrel	yard_size	home_garden	attitude_env	temperature	propensity
1	Barrel	2.58e+04	No home garden	3	81.1	0.705
2	No barrel	3.95e+04	Home garden	7	65.1	0.354
3	Barrel	1.33e+04	No home garden	7	75.9	0.629
4	No barrel	2.83e+04	No home garden	2	70.6	0.339
5	Barrel	2.15e+04	No home garden	5	66.7	0.292
6	Barrel	2.89e+04	Home garden	8	75.3	0.686

```
barrel_ipw <- barrel_probabilities %>%
  mutate(ipw = (barrel_num / propensity) + ((1 - barrel_num) / (1 - propensity)))

barrel_ipw %>%
  select(id, barrel, yard_size, home_garden, attitude_env, temperature, propensity, ipw) %>%
  head()
```

id	barrel	yard_size	home_garden	attitude_env	temperature	propensity	ipw
1	Barrel	2.58e+04	No home garden	3	81.1	0.705	1.42
2	No barrel	3.95e+04	Home garden	7	65.1	0.354	1.55
3	Barrel	1.33e+04	No home garden	7	75.9	0.629	1.59
4	No barrel	2.83e+04	No home garden	2	70.6	0.339	1.51
5	Barrel	2.15e+04	No home garden	5	66.7	0.292	3.43
6	Barrel	2.89e+04	Home garden	8	75.3	0.686	1.46

Table 5: IPW Results

	(1)
(Intercept)	228.214
	[225.882, 230.546]
Barrel	−39.050
	[−42.326, −35.775]
Num.Obs.	1241
R2	0.306
R2 Adj.	0.306
AIC	12 016.2
BIC	12 031.6
Log.Lik.	−6005.090
RMSE	29.30

```r
model_ipw <- lm(water_bill ~ barrel,
                data = barrel_ipw,
                weights = ipw)

modelsummary(model_ipw,
             coef_rename = c(barrelBarrel = "Barrel"),
             output = "kableExtra",
             statistic = "conf.int",
             title = "IPW Results") %>%
    row_spec(c(1,3,5,7,9,11), background = "#f7fabe")
```

Table 6: RCT and Observational Results

	RCT	Naive correlation	Matched	Matched & Weighted	IPW
(Intercept)	228.44	224.80	230.90	234.37	228.21
Barrel	−40.57	−29.86	−35.96	−39.43	−39.05
Num.Obs.	493	1241	805	805	1241
R2	0.308	0.204	0.264	0.301	0.306
R2 Adj.	0.307	0.204	0.263	0.301	0.306
AIC	4766.6	11 881.0	7714.6	7758.9	12 016.2
BIC	4779.2	11 896.3	7728.6	7772.9	12 031.6
Log.Lik.	−2380.295	−5937.482	−3854.281	−3876.430	−6005.090
RMSE	30.24	28.95	29.05	29.13	29.30

3. Comparing results

```
modelsummary(list("RCT" = model_rct,
                  "Naive correlation" = model_bill,
                  "Matched" = model_match,
                  "Matched & Weighted" = model_match_wts,
                  "IPW" = model_ipw),
             coef_rename = c(barrelBarrel = "Barrel"),
             output = "kableExtra",
             statistic = NULL,
             title = "RCT and Observational Results",
             fmt =  2) %>%
    row_spec(c(1,3,5,7,9), background = "#f7fabe")
```

The matched and weighed estimate and the IPW estimate are the closest to the RCT estimate, and they are all very similar. The consistency across methods indicates a higher degree of accuracy. The observational ATEs are sufficient to prove program effect when they are matched and weighted or the IPW method was used. The naive correlation and simply matched estimates are less credible. The matched and weighed estimate, the IPW estimate, and the RCT estimate are all very close to each other, especially both observational estimates. Since there is not a major difference in these estimates, and different methods produce a similar result, indicating consistency and credibility. This program could be rolled out throughout Georgia, but some characteristics should be considered. Locations with different sized yards, gardens, political views, and temperature may respond differently to the program. There may also be other unobserved characteristics that could impact the effectiveness of the program.