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

Answer key - PMAP 8521, Spring 2021

March 1, 2021

Contents

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:

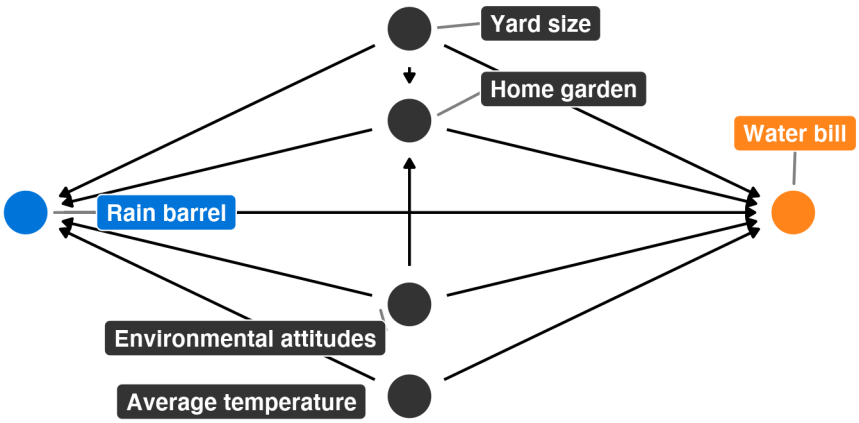


Figure 1: Rain barrel program DAG

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

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 remember from PMAP 8521 that when running an RCT, you can draw the DAG for the program like this (Figure 2). **Why?**

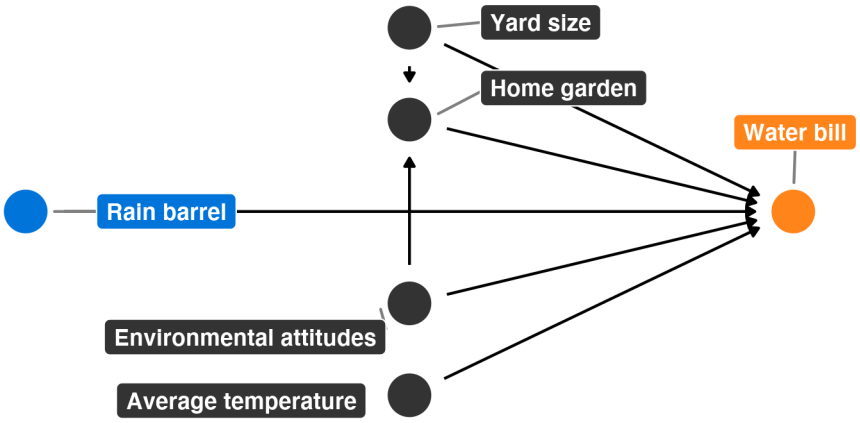


Figure 2: Rain barrel program DAG as an RCT

When you have control over the assignment of treatment status, you get to remove all the arrows pointing into the treatment node in the DAG, since nothing else causes the treatment—it is completely exogenous now. This instantly removes all backdoor confounding relationships between rain barrel use and water bill. Mathematically, we can write this effect using *do* language: $E(\text{Water bill} \mid do(\text{Rain barrel}))$

Check balance

Treatment balance

There were 493 participants in the rain barrel trial. We can check how well the trial was balanced:

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

barrel	n	proportion
No barrel	221	0.45
Barrel	272	0.55

45% of participants were in the program, which is slightly lower than a perfect 50%. We can check if that's a statistically significant difference by using a proportion test in R (this wasn't in the class materials, but it's helpful to know). The null hypothesis in a proportion test is that the proportion of two groups is the same. If the p-value is less than 0.05, we can reject that null and claim that there's a significant difference in proportions; if the p-value is greater the 0.05, we don't have enough evidence to conclude that there's a difference.

```
# table() finds a count of categories in barrels_rct$barrel; we then feed those
# counts into prop.test() to run the actual test
table(barrels_rct$barrel) %>% prop.test() %>% tidy()
```

estimate	statistic	p.value	parameter	conf.low	conf.high	method	a
0.45	5.1	0.02	1	0.4	0.49	1-sample proportions test with continuity correction	t

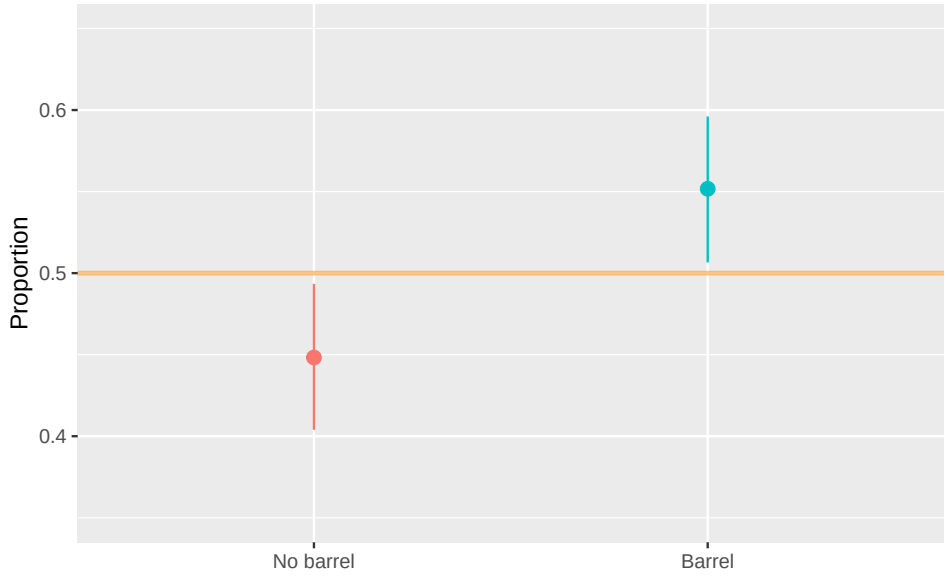
We can also visualize the proportion of groups:

```
# Getting confidence intervals from proportion tests is a little tricky, since
# prop.test() only gives the results for one of the categories (the first number
# that comes out of table()). Technically that's all we need when dealing with
# two numbers---if the proportion of treatment is 75%, the proportion of control
# has to be 25%. But when plotting, it can be helpful to include both groups.
# Here we make two data frames that each have one row and then combine them into
# one. The only difference between the two is that we use rev() on the second
# one to reverse the output of table() so that it runs the test based on the
# control group.
proportion_control <- table(barrels_rct$barrel) %>%
  prop.test() %>%
  tidy() %>%
  mutate(group = "No barrel")

proportion_treatment <- rev(table(barrels_rct$barrel)) %>%
  prop.test() %>%
  tidy() %>%
  mutate(group = "Barrel")

all_proportions <- bind_rows(proportion_control, proportion_treatment) %>%
  mutate(group = fct_inorder(group))

ggplot(all_proportions, aes(x = group, y = estimate, color = group)) +
  geom_hline(yintercept = 0.5, color = "darkorange", alpha = 0.5, size = 1) +
  geom_pointrange(aes(ymin = conf.low, ymax = conf.high)) +
  guides(color = FALSE) +
  coord_cartesian(ylim = c(0.35, 0.65)) +
  labs(x = NULL, y = "Proportion")
```



Based on this graph and the results from `prop.test()`, there is a statistically significant difference between

the two groups ($p = 0.02$), so there are definitely more people in the treatment group than expected. That doesn't necessarily destroy the results, but it's something to keep in mind when reporting the results.

Pre-treatment characteristic balance

Participants' pre-treatment characteristics appear fairly well balanced. It seems that people in the control group are more likely to have a garden, have a larger yard, and have higher regard for the environment, but these differences aren't statistically significant.

```
barrels_rct %>%
  group_by(barrel) %>%
  summarize(prop_garden = mean(home_garden_num),
            avg_yard_size = mean(yard_size),
            avg_env = mean(attitude_env),
            avg_temp = mean(temperature))
```

barrel	prop_garden	avg_yard_size	avg_env	avg_temp
No barrel	0.27	21309	5.5	70
Barrel	0.21	20357	5.4	70

This is apparent in the following plots. There's no statistically significant difference between home garden use between the two groups ($p = 0.1$):

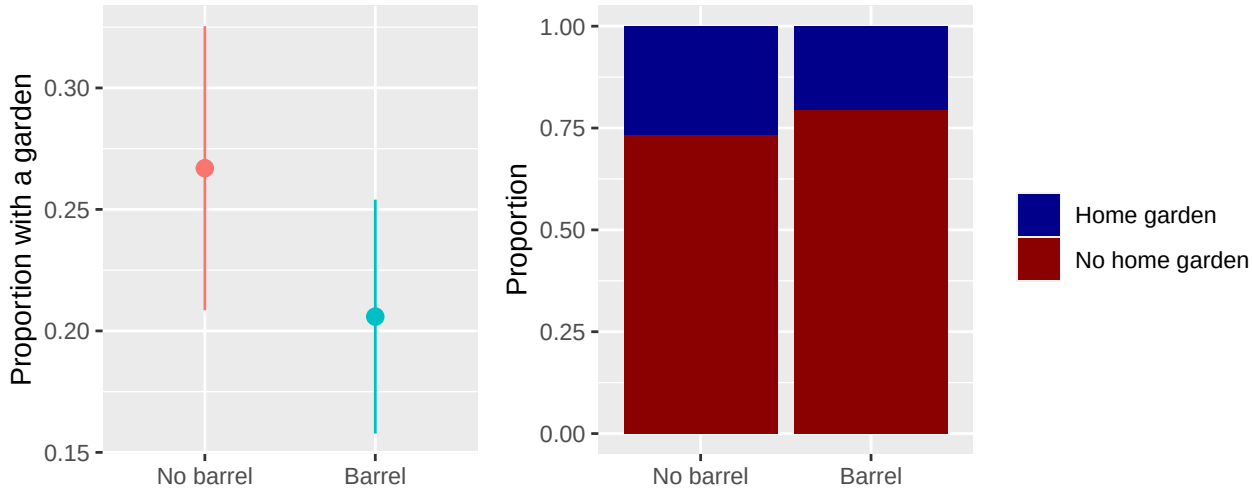
```
t.test(home_garden_num ~ barrel, data = barrels_rct) %>% tidy()
```

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alt
0.06	0.27	0.21	1.6	0.11	451	-0.01	0.14	Welch Two Sample t-test	two

```
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 with a garden")

plot_prop_garden <- 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"))

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



There's also no difference in average yard size (p = 0.217):

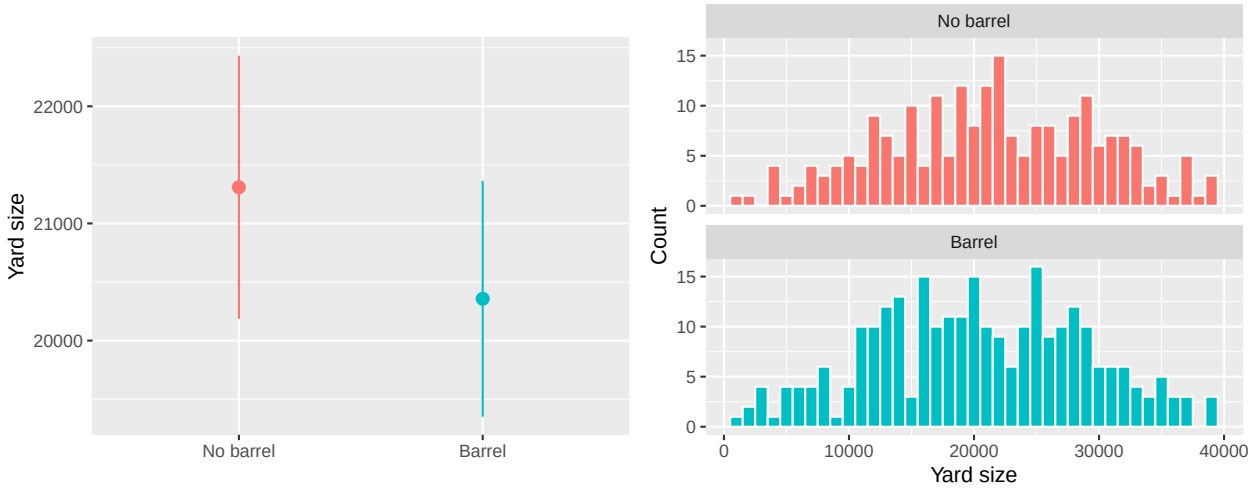
```
t.test(yard_size ~ barrel, data = barrels_rct) %>% tidy()
```

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alt
952	21309	20357	1.2	0.22	469	-560	2464	Welch Two Sample t-test	two

```
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 = FALSE) +
  labs(x = NULL, y = "Yard size")

plot_hist_yard <- ggplot(barrels_rct, aes(x = yard_size, fill = barrel)) +
  geom_histogram(binwidth = 1000, color = "white") +
  guides(fill = FALSE) +
  labs(x = "Yard size", y = "Count") +
  facet_wrap(vars(barrel), ncol = 1)

plot_diff_yard + plot_hist_yard
```



There's also no difference in attitudes towards the environment (p = 0.584):

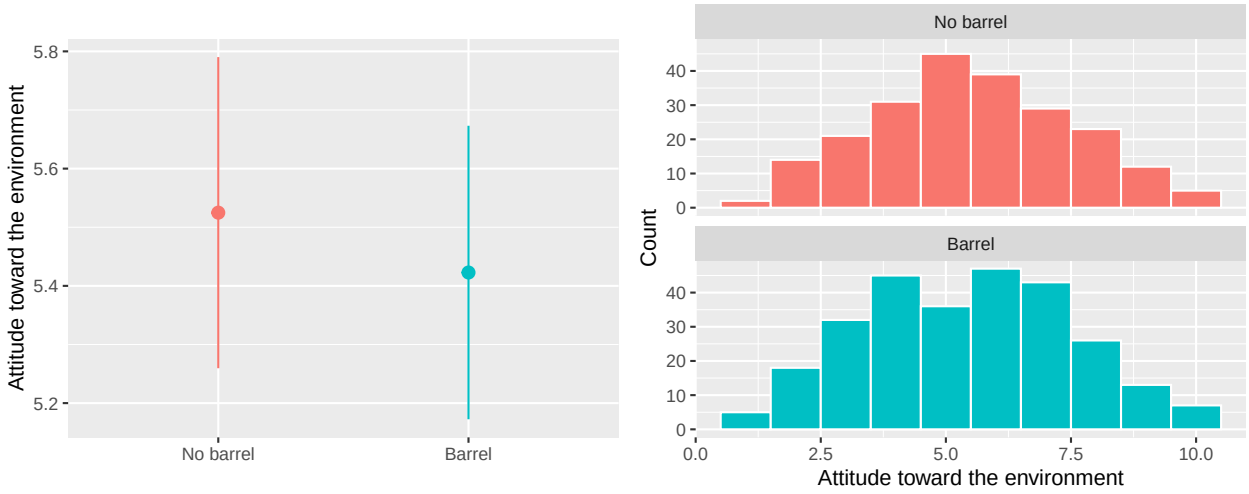
```
t.test(attitude_env ~ barrel, data = barrels_rct) %>% tidy()
```

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alt
0.1	5.5	5.4	0.55	0.58	478	-0.26	0.47	Welch Two Sample t-test	two

```
plot_diff_env <- 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 = FALSE) +
  labs(x = NULL, y = "Attitude toward the environment")

plot_hist_env <- ggplot(barrels_rct, aes(x = attitude_env, fill = barrel)) +
  geom_histogram(binwidth = 1, color = "white") +
  guides(fill = FALSE) +
  labs(x = "Attitude toward the environment", y = "Count") +
  facet_wrap(vars(barrel), ncol = 1)

plot_diff_env + plot_hist_env
```



And finally, there's no difference in average temperature (p = 0.702):

```
t.test(temperature ~ barrel, data = barrels_rct) %>% tidy()
```

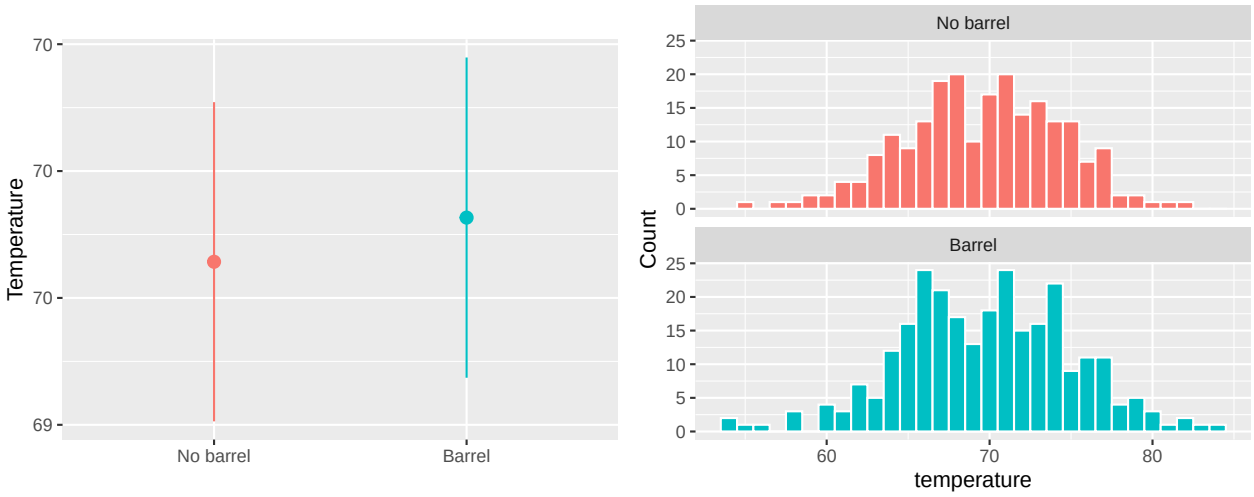
estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alt
-0.17	70	70	-0.38	0.7	486	-1.1	0.72	Welch Two Sample t-test	two

```
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 = FALSE) +
  labs(x = NULL, y = "Temperature")

plot_hist_temp <- ggplot(barrels_rct, aes(x = temperature, fill = barrel)) +
  geom_histogram(binwidth = 1, color = "white") +
  guides(fill = FALSE) +
```

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

plot_diff_temp + plot_hist_temp
```



Estimate difference

We calculate the causal effect of the program by finding the difference in the average water bill in the treatment and control groups. We do this with a simple regression model:

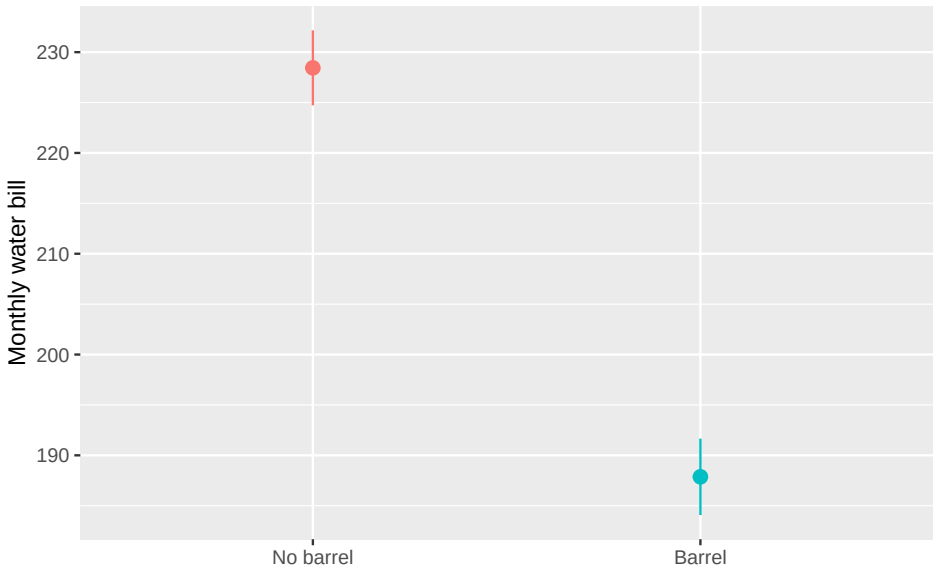
```
model_rct <- lm(water_bill ~ barrel, data = barrels_rct)
tidy(model_rct)
```

term	estimate	std.error	statistic	p.value
(Intercept)	228	2.0	112	0
barrelBarrel	-41	2.7	-15	0

The rain barrel program *causes* water bills to be $40 lower, on average, and the finding is statistically significant ($p < 0.001$). I would find this result fairly credible (even though treatment wasn't perfectly balanced).

Here's what that effect looks like:

```
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 bill")
```


Monthly water bill
230
220
210
200
190
No barrel
Barrel

2. Finding causation from observational data

We can use the observational rain barrel data to try to find a similar causal effect. There are more rows in this data (n = 1241), but participants self-selected into the program, which means we have to deal with counfounders and backdoors.

Naive difference in means

If we calculate a naive difference in means, we find that those who used barrels save $30 on their water bill. This is wrong, though, because of self-selection.

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

barrel	number	avg_bill
No barrel	736	225
Barrel	505	195

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

term	estimate	std.error	statistic	p.value
(Intercept)	225	1.1	211	0
barrelBarrel	-30	1.7	-18	0

Adjustment with Mahalanobis nearest-neighbor matching

Because we know that home garden use, yard size, attitudes toward the environment, and temperatures cause rain barrel use, we'll try to find observations with similar values of these confounders who both used and didn't use rain barrels, thus creating well-matched pseudo treatment and control groups. We do this first with one-to-many Mahalanobis nearest-neighbor matching.

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

```
##
## Call:
## matchit(formula = barrel_num ~ yard_size + attitude_env + home_garden_num +
##     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 eCDF Mean eCDF Max
## yard_size            21173.55      20166.73            0.12        1.0     0.037    0.071
## attitude_env             5.54          5.12            0.21        1.1     0.042    0.106
## home_garden_num          0.26          0.18            0.17          .     0.073    0.073
## temperature             71.78         68.65            0.62        1.1     0.128    0.269
##
##
## Summary of Balance for Matched Data:
##                 Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max Std. Pair
## yard_size            21173.55      21080.76           0.011        1.0     0.009    0.030
## attitude_env             5.54          5.54           0.003        1.0     0.007    0.020
## home_garden_num          0.26          0.26           0.000          .     0.000    0.000
## temperature             71.78         71.59           0.038        1.1     0.010    0.042
##
## Percent Balance Improvement:
##                 Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
## yard_size                   91      -14.6        75       58
## attitude_env                99       36.7        83       81
## home_garden_num            100          .       100      100
## temperature                 94       -5.3        92       85
##
## Sample Sizes:
##               Control Treated
## All               736     505
## Matched (ESS)     209     505
## Matched           300     505
## Unmatched         436       0
## Discarded           0       0
```

All 505 barrel users were paired with 301 non-barrel users, and we discard 435 non-barrel users that don't match well. We use this matched data in a new regression model:

```
barrels_matched <- match.data(matched)

model_matched <- lm(water_bill ~ barrel, data = barrels_matched)
tidy(model_matched)
```

term	estimate	std.error	statistic	p.value
(Intercept)	231	1.7	137	0
barrelBarrel	-36	2.1	-17	0

After matching, the causal effect of the rain barrel program is now $35, which is larger than the naive ATE (and closer to the "true" RCT ATE of $40).

We can use the weights from `matchit()` to fix the imbalance in weighting that happens from reusing matched observations, thus increasing accuracy:

```
model_matched_weighted <- lm(water_bill ~ barrel,
                             data = barrels_matched, weights = weights)
tidy(model_matched_weighted)
```

term	estimate	std.error	statistic	p.value
(Intercept)	235	1.7	139	0
barrelBarrel	-40	2.1	-19	0

Now the ATE is $39, which is surprisingly close to the RCT ATE!

Adjustment with inverse probability weighting

We also use inverse probability weighting to adjust for the confounding backdoors. We use logistic regression to model the choice to use a barrel based on yard size, attitudes toward the environment, use of a home garden, and tempurature, and then use the predicted probabilities from this model as weights to calculate the ATE of the barrel program on water bills.

```
# Generate propensity scores
wants_barrel_model <- glm(barrel ~ yard_size + attitude_env + home_garden + temperature,
                          data = barrels_obs, family = binomial(link = "logit"))

barrel_propensities <- augment_columns(wants_barrel_model, barrels_obs,
                                       type.predict = "response") %>%
  rename(p_barrel = .fitted)

# Calculate ATE weights
barrels_ipw <- barrel_propensities %>%
  mutate(ipw = (barrel_num / p_barrel) + ((1 - barrel_num) / (1 - p_barrel))) %>%
  # Truncate high weights at 10
  mutate(ipw = ifelse(ipw > 10, 10, ipw))

model_ipw <- lm(water_bill ~ barrel,
                data = barrels_ipw, weights = ipw)
tidy(model_ipw)
```

term	estimate	std.error	statistic	p.value
(Intercept)	228	1.2	193	0
barrelBarrel	-39	1.7	-23	0

After using inverse probability weighting, the observational ATE for the rain barrel program is $38.70, which is again much closer to $40 than the naive estimate.

3. Comparing results

Here are all the ATEs we found:

```
modelsummary(list("RCT" = model_rct, "Naive" = model_naive,
                  "Matching, unweighted" = model_matched,
                  "Matching, weighted" = model_matched_weighted,
                  "IPW, truncated" = model_ipw)) %>%
  kableExtra::row_spec(3, bold = TRUE, color = "white", background = "orange")
```

	RCT	Naive	Matching, unweighted	Matching, weighted	IPW, truncated
(Intercept)	228.442	224.800	230.775	234.595	228.214
	(2.038)	(1.068)	(1.687)	(1.683)	(1.184)
barrelBarrel	**-40.573**	**-29.860**	**-35.835**	**-39.655**	**-38.671**
	(2.744)	(1.674)	(2.129)	(2.125)	(1.667)
Num.Obs.	493	1241	805	805	1241
R2	0.308	0.204	0.261	0.302	0.303
R2 Adj.	0.307	0.204	0.260	0.302	0.302
AIC	4766.6	11881.0	7721.5	7764.7	12007.1
BIC	4779.2	11896.3	7735.6	7778.7	12022.5
Log.Lik.	-2380.295	-5937.486	-3857.766	-3879.332	-6000.544
F	218.573	318.158	283.226	348.164	537.960

If this were a real program with a real RCT, I'd believe the RCT ATE the most, since it doesn't suffer from selection bias and unobserved confounding. If I could only rely on observational data, and I didn't know what the actual true value is, I'd use both the weighted Mahalanobis and IPW models, since they each take care of DAG backdoors. Each of these ATEs is statistically significant and fairly substantive—saving $40 a month on my water bill would be great! Given these findings, it might be worth rolling this program out statewide.