Big Assignment 4

Governments in many African countries have recently adopted fertilizer subsidies as an attempt to increase agricultural producitivity for small farmers. However there is widespread debate over whether these are effective policies. To generate evidence Carter, Lajaaj and Yang (2021) partnered with the Mozambican government to run a randomized controlled trial, in which farmers were randomly offered a voucher for subsidized fertilizer. (The paper is [Subsidies and the African Green Revolution: Direct Effects and Social Network Spillovers of Randomized Input Subsidies in Mozambique] *American Economic Journal: Applied Economics* 13(2).)

We are going to work with a subset of their data for this problem set to estimate the effects of fertilizer use on maize yields.

Part A:

Suppose the government did not conduct a randomized controlled trial. Instead they surveyed farmers and compared maize yields for those who did and did not use fertilizer. Would this comparison allow you to estimate the effect of fertilizer on maize yields? Explain why or why not using both words and potential outcomes notation. How would the expression for the estimated impact of fertilizer on maize yields differ if fertilizer use were randomized?

No, just observing yield differences between farmers who do and don't use fertilizer would not allow us to credibly estimate the impact of fertilizer on maize yields. We think that farmers choose to use fertilizer for many reasons that may also be correlated with yields, for example, farmers who are more skilled, wealthier, grow hybrid maize varieties, have worse soil quality, etc. are probably more likely to use fertilizer, but these attributes are also likely to affect yields. So we would not be isolating the true effect of fertilizer because we would also pick up these differences by just comparing farmers who do and don't use fertilizer.

In potential outcomes notation, Let T_i equal 1 if farmer i uses fertilizer and 0 otherwise. Let Y^T be yields for farmers who use fertilizer and Y^C be yields for farmers who don't.

What we observe by comparing average yields across fertilizer and non-fertilizer users is
$$E[Y^T|T=1] - E[Y^C|T=0]$$
But what we want is
$$E[Y^T|T=1] - E[Y^T|T=0]$$
i.e. what yields would have been for the group of farmers we observed using fertilizer, had they instead not used fertilizer. Of course, this counterfactual is unobservable. Adding and subtracting $E[Y^T|T=0]$ to the above gives us
$$(E[Y^T|T=1] - E[Y^T|T=0]) + (E[Y^T|T=0] - E[Y^C|T=0])$$
The first part is equal to the true effect while the second part captures selection bias -- people who did use fertilizer would likely have different yields than those who didn't even had they not used fertilizer.

Randomization allows us to recover the true effect of fertilizer by setting this selection bias term to 0. It ensures $E[Y^T|T=0] = E[Y^C|T=0]$ - i.e. farmers who do and don't use fertilizer would have the same expected yields from unfertilized maize because whether they used fertilizer was randomly determined.

Part B:

Now let's analyze the RCT data. The dataset contains observations of nine variables from 390 farm households. The variables are as follows

- $respid$: This is just the unique ID for each farmer
- $lyieldr$: This is the log of maize yields
- $vouch$: This is the treatment variable of interest - equal to 1 if the household was given a voucher to purchase fertilizer (i.e. in the treatment group) and 0 otherwise (i.e in the control group)
- $irrigprev$: This is a dummy equal to 1 if households used irrigation and 0 otherwise
- $pestdprev$: This is a dummy equal to 1 if households used pesticides and 0 otherwise
- $hhhmale$: This is a dummy equal to 1 if the household head is male and 0 if the household head is female
- $hhhage$: This is the age of the household head
- $hhheduc$: This is the number of years of education of the household head
- $hhsize$: This is the number of members in the household

Note that all variables other than maize yield were measured *before* the distribution of vouchers. For irrigation and pesticide use, they were measured the year before the voucher intervention.

a) **Before running any regressions, show how you can obtain the average treatment effect (ATE) of fertilizer vouchers on (log) maize yields. Then write down the regression you could use to estimate the ATE.**

The difference in average log yields between the treatment and control groups is the average treatment effect of fertilizer vouchers. Given that vouchers were randomized, any differences in log yields between the two groups can be attributed to the effect of the vouchers. In math this is just
$$ATE = \overline{\log(yield)}_T - \overline{\log(yield)}_C$$
Equivalently, we could run the regression
$$\log(yield)_i = \beta_0 + \beta_1 vouch_i + u_i$$
where β_1 would equal the difference in log yields -- the ATE.

b) **Read in the data from `ps4_data.dta` (remember you need to use the `read_dta` function from the `haven` package to read `.dta` files. You will also probably want to load `tidyverse`). Then, estimate the regression you wrote down using `lm()` and test whether the ATE is statistically significant at a 95\% confidence level.**

In [1]:
```
library(haven)
library(tidyverse)

df <- read_dta('ps4_data.dta')

print(paste0('ATE=',mean(subset(df, vouch==1)$lyieldr)-mean(subset(df, vouch==0)$lyieldr)))
summary(lm(lyieldr~vouch, data=df))
```

```
── Attaching packages ─────────────────────────────── tidyverse 1.3.1 ──

✔ ggplot2 3.3.5     ✔ purrr   0.3.4
✔ tibble  3.1.3     ✔ dplyr   1.0.7
✔ tidyr   1.1.3     ✔ stringr 1.4.0
✔ readr   2.0.1     ✔ forcats 0.5.1

── Conflicts ────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

```
[1] "ATE=0.208840550692112"

Call:
lm(formula = lyieldr ~ vouch, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0641 -0.6338  0.0916  0.6229  4.1123 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.30702    0.07638  82.575   <2e-16 ***
vouch        0.20884    0.11044   1.891   0.0594 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.091 on 389 degrees of freedom
Multiple R-squared:  0.009108,	Adjusted R-squared:  0.006561 
F-statistic: 3.576 on 1 and 389 DF,  p-value: 0.05938
```

Vouchers appear to have increased yields by almost 21\%, but with $p = .0594$ this is not quite significant at the 5\% level (but is at 10\%). Therefore, we fail to reject the null hypothesis that the voucher program had no effect on yields at the 5\% level.

Part C:

a) Now let's check to make sure our sample is balanced across treatment and control. For each of the four household demographics variables, conduct a t-test against the null that they are on average equal between treatment and control.

Coding Hints: The command for a t-test is `t.test()`. One way to test whether coviarate X is correlated with treatment status T, you can put `X~T` inside `t.test()`, just like you would for `lm()`. If you just want to print the t-stat or the p-value rather than the entire test output, you can call `$statistic` or `$p.value`, respectively.

You can repeat this separately for each covariate, but if you want to be extra fancy you can use the `lapply()` function. `lapply()` is super useful for applying a function repeatedly over different variables. The way you would use it here is `lapply(X, function (x) FUN)` where `X` is the data frame of the variables you are interested in testing and `FUN` is the function you want to apply. One of the arguments of this function should be (small) `x`, which serves as a stand-in for the column of (big) `X` that you want to apply the function over. Finally, if you want to show the output as a single data frame (instead of a list), you can wrap all of this with `as.data.frame()`. With all of this information, you could theoretically produce a nice-ish table of p-values for the four tests you run using a single line of code. You are not required or expected to use this method, but it might come in handy if you find yourself having to work with larger datasets in the future.

In [16]:
```
as.data.frame(lapply(df[grep('^hh', names(df))], function(x) t.test(x ~ vouch, data=df)$p.value))
t.test(hhhmale ~ vouch, data=df)
t.test(hhsize ~ vouch, data=df)$estimate
```

A data.frame: 1 × 4

hhhmale	hhheduc	hhhage	hhsize
<dbl>	<dbl>	<dbl>	<dbl>
0.7262042	0.5804594	0.8129967	0.04358971

```
	Welch Two Sample t-test

data:  hhhmale by vouch
t = 0.35044, df = 383.15, p-value = 0.7262
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -0.05753045  0.08248589
sample estimates:
mean in group 0 mean in group 1 
      0.8627451       0.8502674 
```

mean in group 0: 6.66606304984467 **mean in group 1:** 7.08251317299624

Here we use the `lapply` function to run t-tests for each variable in the dataset whose name starts with 'hh' (see `?grep` for how this works), saves the p-values, and presents them in a dataframe/table. For comparison, we show the full output for the t-test for the first variable, $hhhmale$, and the means by treatment status for household size.

b) Do you conclude that the sample is well-balanced? Are you surprised or concerned by any of the results? If so, what could you do to address your concerns?

The table above shows the p-value from each of the four hypothesis tests. We see no evidence that household head attributes (sex, years of education, or age) differ between treatment and control (here we *don't* want to be able to reject the null that they are identical), but we see that household sizes are significantly (at 95\% confidence) different in the treatment group. Looking at the means by group, treatment households are slightly (but significantly) bigger on average.

This is a little surprising but not a cause for concern. Since we assume that we'll falsely reject the null 5\% of the time for each hypothesis test we run at a 95\% confidence level, we'd expect to get one false rejection from four tests about $1 - 0.95^4 \approx 18\%$ of the time. Remember that randomization only guarantees that treatment and control groups will be balanced in expectation. We are just dealing with one sample of 390 households, so some variables may still be correlated with treatment by chance. To address any concern about this correlation, we can always control for household size in our regression to purge our treatment effect estimates of any correlation that arose due to chance.

Part D

a) What happens if you control for additional covariates (household demographics and previous use of irrigation and pesticides) in this regression? What (if any) advantages are there to controlling for these variables? What (if any) disadvantages are there?

Since the provision of fertilizer vouchers was randomized, it should not be correlated with any other covariates. As a result, including additional covariates in the regression should not affect our estimate of the average treatment effect of the voucher program. Of course, we saw that household size is correlated with voucher receipt, so it's possible that our estimated treatment effect will change somewhat due to including this variable as a control.

The advantages for controlling for variables are 1) addressing the possible affects of sample imbalances (which can happen by chance as we saw in part C) and 2) increasing precision of our estimates (reducing standard errors). If the other variables we've added explain a lot of the variation in log yields and are (for the most part) uncorrelated with $vouch$, this will reduce our standard errors (by reducing the residual variation in u). There are no real disadvantages to adding these variables as long as they cannot be affected by treatment (if they can be affected by treatment, we might have a 'bad controls' problem), but some people say that adding covariates in a context where treatment was randomized is unncessary because the simple regression in part B is "good enough".

b) **Run the regression controlling for all six of these variables. Interpret how your results change, if at all.**

In [17]:
```
summary(lm(lyieldr~vouch+hhhmale+hhheduc+hhhage+hhsize+irrigprev+pestdprev, data=df))
```

```
Call:
lm(formula = lyieldr ~ vouch + hhhmale + hhheduc + hhhage + hhsize + 
    irrigprev + pestdprev, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0468 -0.6732  0.0744  0.6050  4.1302 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.413447   0.350036  18.322   <2e-16 ***
vouch        0.224879   0.110090   2.043   0.0418 *  
hhhmale      0.287254   0.161247   1.781   0.0756 .  
hhheduc      0.005474   0.021105   0.259   0.7955    
hhhage      -0.007948   0.004525  -1.756   0.0798 .  
hhsize      -0.007872   0.027174  -0.290   0.7722    
irrigprev    0.419174   0.269942   1.553   0.1213    
pestdprev    0.321620   0.238140   1.351   0.1776    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.08 on 383 degrees of freedom
Multiple R-squared:  0.04362,	Adjusted R-squared:  0.02614 
F-statistic: 2.495 on 7 and 383 DF,  p-value: 0.0162
```

We see that the coefficient on $vouch$ has increased slightly (but is not statistically distinguishable from the one in part B). However, we note that it has just become statistically significant at the 5\% level - we can now reject the null that the vouchers had no effect on maize yields. This is a benefit of the increased precision from including covariates.

Part E

You think households might differ in their ability to correctly apply fertilizer in order to increase their maize yields. You hypothesize that households that use pesticides might have more knowledge about appropriate input use, so might experience larger increases in maize yields if they receive a fertilizer voucher. Propose, implement, and interpret a test of whether the effect of vouchers on log maize yields is different for households that used pesticides the prior year, building on the model from Part D.

We would specify the following model:

$$log(yield) = \beta_0 + \beta_1 vouch + \beta_2 vouch * pestdprev + \beta_3 hhhmale + \beta_4 hhheduc + \beta_5 hhhage + \beta_6 hhsize + \beta_7 irrigprev + \beta_8 pestdprev$$

In this model, β_2 is the differential effect of the voucher treatment on log maize yield for households that used pesticide the year before. The null hypothesis we want to test is $H_0 : \beta_2 = 0$ against the alternative $H_1 : \beta_2 \neq 0$.

In [24]:
```
summary(lm(lyieldr~vouch+vouch:pestdprev+hhhmale+hhheduc+hhhage+hhsize+irrigprev+pestdprev, data=df))
```

```
Call:
lm(formula = lyieldr ~ vouch + vouch:pestdprev + hhhmale + hhheduc + 
    hhhage + hhsize + irrigprev + pestdprev, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0375 -0.6508  0.0630  0.5944  4.1168 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      6.406996   0.350358  18.287   <2e-16 ***
vouch            0.205602   0.113251   1.815   0.0702 .  
hhhmale          0.291712   0.161459   1.807   0.0716 .  
hhheduc          0.006095   0.021134   0.288   0.7732    
hhhage          -0.007758   0.004535  -1.711   0.0880 .  
hhsize          -0.007809   0.027190  -0.287   0.7741    
irrigprev        0.414364   0.270185   1.534   0.1259    
pestdprev        0.175087   0.311020   0.563   0.5738    
vouch:pestdprev  0.354424   0.483472   0.733   0.4640    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.081 on 382 degrees of freedom
Multiple R-squared:  0.04496,	Adjusted R-squared:  0.02496 
F-statistic: 2.248 on 8 and 382 DF,  p-value: 0.02351
```

We observe that the coefficient on $vouch$ is similar to the previous model. Since $vouch$ now represents the average treatment effect among households that did not use pesticide the year before, this suggests there are not many households that used pesticide. The coefficient estimate on $vouch : pestdprev$ is positive and larger in magnitude, but is not statistically significant. Specifically, we have $p = 0.464$, meaning we fail to reject the null that there is no differential effect of the voucher for households that used pesticide the year before at any reasonable significance level.

Part F

Of course, not everyone who was offered a voucher ended up redeeming it and using fertilizer. So what we've estimated so far for the average treatment effect of *receiving a fertilizer subsidy voucher on yields* can be interpreted as an **intent to treat** estimate for the effect of *fertilizer application on yields*.

Suppose 75\% of people who were offered vouchers ended up using fertilizer, compared to 10\% of people in the control group. Write down a Treatment on the Treated estimator and calculate it using your result from Part D (you don't have to run any additional regressions for this part). Under what circumstances would this be equal to the average treatment effect of using fertilizer? Do you think these hold in this context?

In [18]:
```
# numerator: ITT estimate
# denominator: share of compliers
(summary(lm(lyieldr~vouch+hhhmale+hhheduc+hhhage+hhsize+irrigprev+pestdprev,
          data=df))$coefficients[2,1])/(.75-.1)
```

0.345967654299587

The ToT is simply the ITT divided by the share of compliers (.65). We estimate that fertilizer application increases average maize yields 32.1%.

Since we know that the control group is a good counterfactual for the treatment group, we can believe that 10\% of the households in the treatment group would have used fertilizer anyways (always takers) and that 25\% of households would not have used fertlizier even had they been given subsidy vouchers (never takers). Thus, we can infer that 65\% of households use fertilizer if and only if they receive vouchers -- these are the only people for whom fertilizer *use* is changing (compliers). 65% of our ITT estimate is thus the actual effect of the treatment on the treated, while the other 35% is the null effect coming from people whose behavior isn't changed by the intervention. In math, $ITT = 0.65 * TOT + 0.35 * 0$.

So the ToT is just the ITT weighted by the inverse of the share of compliers, or in math:
$$ToT = \frac{ITT}{\%\text{ compliers}} = \frac{\overline{\log(yield)}_T - \overline{\log(yield)}_C}{\overline{use\_fert}_T - \overline{use\_fert}_C} = \frac{.2249}{.65} = 0.346$$

This is only equal to the average treatment effect of using fertilzier under two assumptions.

1. Receiving the vouchers only affects yields through fertilizer use and
2. The effects of fertilzier on yields are the same for compliers as they are for everyone else.

There are a few different reasons 1) might not hold. People who receive fertilizer vouchers would likely alter the amounts of other inputs they use (to the extent that labor, pesticides and seeds are complements/substitutes for fertilizer). For example, hybrid seeds are much more responsive to fertilizer but generally higher yielding. If the vouchers cause people to use the money they saved on fertilizer on other inputs, then the ToT would also capture the effects of these other inputs on yields.

Assumption 2) is very unlikely to hold. We think that people have different returns to fertilizer (e.g. based on skill, landholdings and soil quality) and it is quite likely that the people who stand most to gain from fertilizer use are more likely willing to pay the unsubsidized price. Conversely, people who only use fertilizer if they receive a subsidy but are not willing to pay the market price may simply be doing so because fertilizer is not very effective on their plots. So in this case $TOT < ATE$. You could also tell a story that if farmers are credit constrained and there are decreasing marginal returns to fertilizer, then the subsidy disproportionately affects credit constrained farmers that are using less fertilizer and thus have higher returns, so $TOT > ATE$. In either case, it's unlikely that these two effects would be exactly the same. Who the "compliers" are is always important for contextualizing effects!