PPHA 34600 Program Evaluation

Problem Set 2

5/1/2022

(1) HARRIS are interested in answering the following question: What was the effect of FIONA on profits for the average farmer? To make sure everybody is on the same page, explain to them what the ideal experiment would be for answering this question. Describe the dataset that you'd like to have to carry out this ideal experiment, and use math, words, and the potential outcomes framework to explain what you would estimate and how you would do so. Make sure to be clear about the unit of analysis (ie, what is "i" here?).

Using the potential outcomes framework, we would ideally like to measure the impact of the rainfall-index insurance on the average farmer's profits.

In the ideal experiment, we would be able to randomize farmers into treatment and control groups where treated farmers take up the rainfall insurance, while farmers in the control group do not have insurance. To prevent spillovers from happening, in the ideal experiment, randomization would be conducted at the district level so as to minimize the possibility that farmers that are not offered the insurance find out about the program from their neighbors and get access to the insurance (ie. treatment). In the ideal experiment, there would not be issues of non-compliance either, such that everyone assigned to treatment actually takes up the insurance and vice versa with the control group.

To run this ideal experiment, I would like to have a dataset of farmers by districts with their baseline profits and other baseline observable characteristics so that I could randomize access to the rainfall-index insurance at the district level. I would then construct a balance table by conducting t-tests to make sure the treatment and control groups are balanced in all baseline characteristics and subsequently run the experiment.

Therefore, we could answer HARRIS' question of the effect of the rainfall-index insurance on the profits of the **average farmer** by estimating the **Average Treatment Effect (ATE):**

$$\tau^{ATE} = E[Y_i(1)] - E[Y_i(0)]$$

Where:

- Y is the outcome variable of interest, 2016 profits.
- i is each individual farmer

Under randomization, the **expected** profits for farmers that take up the insurance (ie. $Profits_{farmer}(1)$) is equivalent to the **mean** profits of farmers that take the insurance, and likewise for farmers in the control group. As such, we could estimate ATE by taking the difference in means between both groups:

$$\tau^{ATE} = \overline{Y_i}(D = 1) - \overline{Y_i}(D = 0)$$

Where:

- D refers to treatment status: $D = 1$ for farmers that took up the rainfall insurance and $D = 0$ for farmers that did not take it.
- Y is the outcome variable of interest, 2016 profits.
- i is each individual farmer

Even though I would be randomizing at the district level, the unit of analysis i would be farmer, since HARRIS is interested in the profits of the *average* **farmer**.

(2) HARRIS like what you're suggesting, but think it's answering the wrong question. They aren't going to be able to get every single farmer to participate. They'd instead like to know: What was the effect of FIONA on profits among farmers who took up insurance? Describe in math and words, using the potential outcomes framework, what they'd like to estimate. Explain how this differs from what you described in (1), and describe what component of this estimand you will be fundamentally unable to observe.

Since HARRIS will not be able to get every single farmer to participate in their insurance scheme, this means that HARRIS anticipates an issue with perfect compliance (were they to run an RCT), in which farmers assigned to receive the rainfall insurance will not actually opt in. In this case, we effectively could not estimate ATE.

Instead, what HARRIS is interested in measuring is the the **Average Treatment Effect on the Treated (ATT)** or in other words, the ATE only among those farmers that did take up the insurance. Mathematically, this can be expressed as:

$$\tau^{ATT} = E[\tau_i | D_i = 1] = E[Y_i(1) - Y_i(0) | D_i = 1]$$
$$\tau^{ATT} = E[Y_i(1) | D_i = 1] - E[Y_i(0) | D_i = 1]$$

Where:

- D refers to treatment status: $D = 1$ for farmers that took up the insurance and $D = 0$ for farmers that did not take it.
- Y is the outcome variable of interest, 2016 profits.
- $i = farmer$.

The ATT, therefore is calculated by taking the difference between **(1)** the potential profits for farmers that actually took up the insurance ($D_i = 1$) had they been offered the insurance ($Y_i(1)$) and, **(2)**, the potential profits for farmers that actually took up the insurance ($D_i = 1$) *had they not been offered the insurance* ($Y_i(0)$).

This is different from **ATE** because, instead of measuring the average effect for both the treatment and control groups, the **ATT** only focuses on the average effect of the treated group (those that take up the insurance).

The issue here is the **fundamental problem of causal inference** because we are fundamentally unable to observe the second term $E[Y_i(0) | D_i = 1]$. It is impossible to know what the potential profits of farmers that took up insurance would have been if they never got access to it in the first place. As such, we cannot calculate the ATT.

(3) HARRIS are on board with your explanation. Because FIONA already exists in the real world, they can't run an RCT to study it. However, they do know that not all farmers were offered insurance through FIONA. It turns out that FIONA only impacted certain districts. Non-FIONA districts were not offered any insurance products. Explain what you would recover if you simply compared FIONA farms to non-FIONA farms on average. Describe three concrete examples of why this might be problematic.

In the absence of randomization, if we simply compared FIONA farms to non-FIONA farms, we would be calculating the **naive estimator:**

$$\tau^{N} = \overline{Y}(1) - \overline{Y}(0)$$

Where:

- $N = farms$ since the unit of analysis is FIONA and non-FIONA farms.
- Y is the outcome variable of interest, 2016 profits.

This naive estimator is simply based on **observed outcomes** and it can be very problematic since the treated and untreated farms are very likely to be systematically different in many characteristics, both observable and unobservable, in ways that affect farmer's selection into treatment (opting for insurance or not). As such, if we use this estimator, we would have **selection bias** and would erroneously attribute the impact on profits to FIONA, when it can well be the result of many other characteristics.

To give three concrete examples:

1. It may be the case that due to geographic locations, some districts had a very good rainfall year while other districts didn't. If, for example, the farms that are offered insurance happened to be those farms with a good rainfall year, then those farms do not receive any tangible benefit from opting for insurance against bad rainfall in the year.

Since they had good rainfall that year, they would have more income to invest in fertilizer to further increase their profits. In this case, the increase in profits would come from good rainfall and not because the FIONA scheme was actually effective.

2. Alternatively, it could be the case that in some farms, farmers are culturally more risk adverse. If most of the farmers that are offered the FIONA insurance happened to be risk-adverse, then it is unlikely they will invest in riskier up-front inputs, even if they have access to insurance. Here, we might erroneously conclude that FIONA doesn't work at all, when it might actually have an effect on non risk-adverse farmers.

3. If, for instance, farms that were offered treatment also happened to be located in districts that had an economic boost during the year, then the increase in profits may be the result from that boost rather than from the FIONA scheme, and it may be that doing well economically is a deciding factor on feeling the need to opt for insurance or not.

(4) HARRIS hears your concerns, but still wants an estimate of the impacts of FIONA. Given that you're unable to implement your ideal experiment, and you are worried about simple comparisons of FIONA-aided farmers and those without insurance, you'll need to do something a little more sophisticated. Luckily for you and for HARRIS, India makes data on farmers available to the public, in the form of ps2_data.csv. Read the data into R and, as always, make sure everything makes sense. Document and fix any errors.

```
#First I check the unique values for discrete variables, to check for
# consistency in data types.
unique(fiona$farmer_birth_year)
```

```
##  [1] "1968"                   "1964"                   "1960"
##  [4] "1971"                   "1967"                   "1965"
##  [7] "1972"                   "1977"                   "1969"
## [10] "nineteen seventy-three" "1961"                   "1966"
## [13] "1962"                   "1976"                   "1975"
## [16] "1970"                   "1941"                   "1984"
## [19] "1974"                   "1982"                   "1988"
## [22] "1986"                   "1981"                   "1948"
## [25] "1937"                   "1989"                   "1980"
## [28] "1963"                   "1958"                   "1978"
## [31] "1944"                   "1950"                   "1973"
## [34] "1983"                   "1954"                   "1959"
## [37] "1956"                   "nineteen seventy-two"   "1979"
## [40] "1925"                   "1951"                   "1957"
## [43] "1942"                   "1934"                   "1987"
## [46] "1953"                   "1955"                   "1947"
## [49] "1949"                  "1945"                   "1936"
## [52] "1940"                  "1952"                   "1938"
## [55] "1946"                  "1943"                   "1985"
## [58] "1916"                  "1928"                   "1935"
## [61] "1939"                  "1933"                   "1919"
## [64] "1931"                  "1930"                   "1929"
## [67] "1932"
```

```
unique(fiona$crop)
```

```
## [1] "RICE"    "LENTILS" "WHEAT"   "COTTON"
```

```
unique(fiona$district)
```

```
## [1] "KARUR"       "TENKASI"     "MADURAI"     "PUDUKKOTTAI" "THANJAVUR"
## [6] "DINDIGUL"
```

```
unique(fiona$fiona_farmer)
```

```
## [1] 0 1
```

```
unique(fiona$fertilizer_use)
```

```
## [1] 1 0
```

```
# Replace written birth years to numbers in the farmer_birth_year variable
fiona$farmer_birth_year[fiona$farmer_birth_year ==
                          "nineteen seventy-three"]<- 1973
fiona$farmer_birth_year[fiona$farmer_birth_year ==
                          "nineteen seventy-two"] <- 1972

# Convert birth year from string type to integer type
typeof(fiona$farmer_birth_year)
```

```
## [1] "character"
```

```
fiona$farmer_birth_year = strtoi(fiona$farmer_birth_year, base=0L)
typeof(fiona$farmer_birth_year)
```

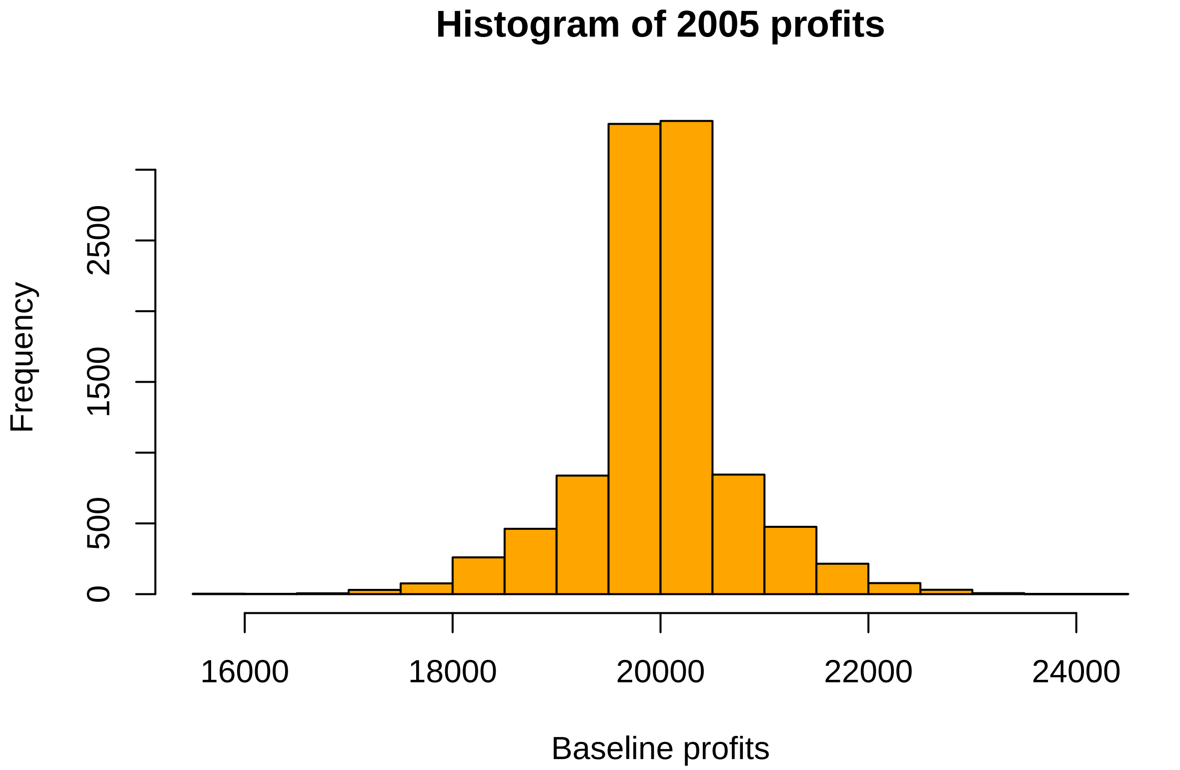
```
## [1] "integer"
```

When looking at the data, there are some observations that need to be cleaned up for consistency. For example, in farmer_birth_year there are observations that are written in words (ie. “nineteen seventy-three”) instead of numbers (ie. 1973). I replaced these observations with numbers and converted the variable from string type to integer.

Otherwise, all other observations are of the same data type. There are no variations in spelling in the case of string variables (ie. district and crop) and luckily, there are no missing values either.

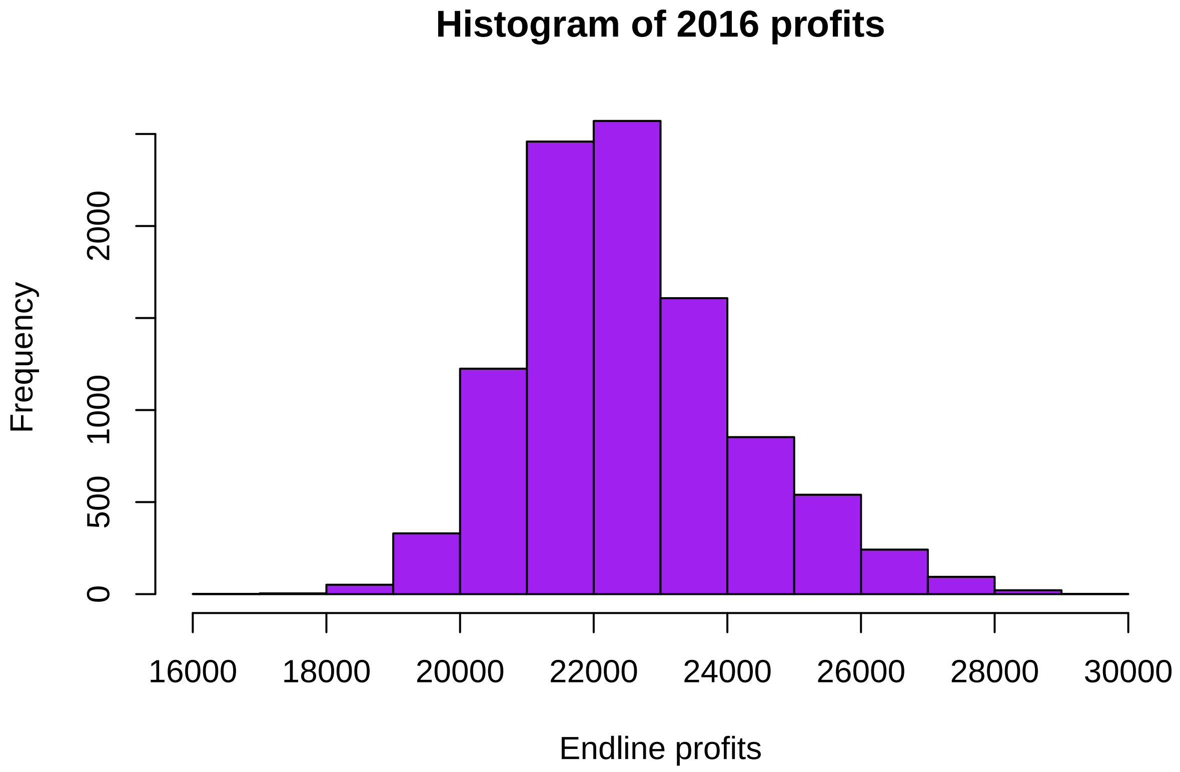
Then, I plot histograms of the outcome variable of interest, profits, at baseline and endline to detect if there are any visible outliers in the data.

```
hist(fiona$profits_2005,
     main = "Histogram of 2005 profits",
     xlab="Baseline profits",
     col="orange")
```



Histogram of 2005 profits

```r
hist(fiona$profits_2016,
     main = "Histogram of 2016 profits",
     xlab="Endline profits",
     col="purple")
```



Histogram of 2016 profits

While the baseline 2005 profits follows a normal distribution, compared to a right skewed distribution in the case of endline 2016 profits, there are no visible outliers in the data that need to be dropped.

Use the variables contained in the dataset to describe, using math and words, two (related) potential approaches to estimating the effect of FIONA on profits. Make sure to be clear about your unit of analysis, and be explicit about how these designs apply to FIONA (ie, describe things in terms of "profits," not just "outcome"). Hint: HARRIS wants you to describe two selection-on-observables designs.

Given the variables available in the dataset and the absence of randomization, we can estimate ATE under Selection on Observables (SOO) designs.

1. REGRESSION ADJUSTMENT: In this approach, we adjust the regular ATE regression by including controls in the form of additional covariates X_i's.

$$Y_i = \alpha + \tau D_i + \gamma X_i + v_i$$

Where:

- D_i refers to treatment status $Di = 1$ for farms that took up FIONA insurance and $D_i = 0$ for those that didn't.
- X_i represents control variables, such as: crop dummies, farmer_birth_year, etc.
- Y_i is the outcome variable, 2016 profits
- $i = farmer$

Since there was no randomization in the FIONA scheme, $E[\varepsilon \mid D_i] \neq 0$. Instead, we have **conditional independence** $E[\varepsilon \mid D_i, X_i] = 0$, where the expected error term *conditional* on taking up the insurance AND covariates (eg. crop dummies, farmer_birth_year) is 0. This means that, assuming this assumption holds as well as the functional form above, we need to include additional covariates in our regression so as to recover τ^{ATE}.

2. MATCHING: In this second approach, we match untreated farmers with treated farmers with identical covariates (eg. crop dummies, farmer_birth_year). Then, the difference in 2016 profits between treated and untreated farmers is the average treatment estimate $\hat{\tau}$.

This requires dividing the data into *cells* by the different covariates (eg. farmer_birth_year, crops). For each cell, we need to calculate the mean profits for the treated $\overline{Y}_T$ and the mean profits of the untreated $\overline{Y}_U$, and then take the differences between both $\overline{Y}_T - \overline{Y}_U$ for each covariate. Finally, $\hat{\tau}^{ATE}$ is a weighted average of these differences. We can also calculate ATT and ATN by using weights for the treated and untreated, respectively.

(5) Produce a balance table which displays the differences between FIONA and non-FIONA farmers on observable characteristics. Interpret this table. Does this table make you feel better or worse about your concerns in (3)?

```r
# Convert crop and district to numerical categorical values
fiona$crop_cat = as.numeric(as.factor(fiona$crop))
fiona$district_cat = as.numeric(as.factor(fiona$district))

# Convert crop variable to dummies for each crop
fiona$rice = ifelse(fiona$crop == 'RICE', 1, 0)
fiona$lentils = ifelse(fiona$crop == 'LENTILS', 1, 0)
fiona$wheat = ifelse(fiona$crop == 'WHEAT', 1, 0)
fiona$cotton = ifelse(fiona$crop == 'COTTON', 1, 0)

# Convert district variable to dummies for each district
```

```
fiona$karur = ifelse(fiona$district == 'KARUR', 1, 0)
fiona$tenkasi = ifelse(fiona$district == 'TENKASI', 1, 0)
fiona$pudukkottai = ifelse(fiona$district == 'PUDUKKOTTAI', 1, 0)
fiona$thanjavur = ifelse(fiona$district == 'THANJAVUR', 1, 0)
fiona$dindigul = ifelse(fiona$district == 'DINDIGUL', 1, 0)
fiona$madurai = ifelse(fiona$district == 'MADURAI', 1, 0)
```

```
# Separate the dataset into treated and untreated groups:
treated <- filter(fiona, fiona_farmer == 1)
untreated <- filter(fiona, fiona_farmer == 0)

# Segment treated and untreated by each available baseline characteristic and
# run a t-test to check if they are balanced:

# (1) Pre-treatment profits (profits_2005)
profits_t <- treated %>% pull(profits_2005)
profits_u <- untreated %>% pull(profits_2005)
mt1 = mean(treated$profits_2005)
mu1 = mean(untreated$profits_2005)
sdt1 = sd(treated$profits_2005)
sdu1 = sd(untreated$profits_2005)
nt1 = nrow(treated)
nu1 = nrow(untreated)
t_test <-t.test(profits_t, profits_u)
p1 = t_test$p.value

# (2) Farmer birth year (farmer_birth_year)
fby_t <- treated %>% pull(farmer_birth_year)
fby_u <- untreated %>% pull(farmer_birth_year)
mt2 = mean(treated$farmer_birth_year)
mu2 = mean(untreated$farmer_birth_year)
sdt2 = sd(treated$farmer_birth_year)
sdu2 = sd(untreated$farmer_birth_year)
nt2 = nrow(treated)
nu2 = nrow(untreated)
t_test <-t.test(fby_t, fby_u)
p2 = t_test$p.value

# (3) All crops(crop_cat)
all_crops_t <- treated %>% pull(crop_cat)
all_crops_u <- untreated %>% pull(crop_cat)
mt3 = mean(treated$crop_cat)
mu3 = mean(untreated$crop_cat)
sdt3 = sd(treated$crop_cat)
sdu3 = sd(untreated$crop_cat)
nt3 = nrow(treated)
nu3 = nrow(untreated)
t_test <-t.test(all_crops_t, all_crops_u)
p3 = t_test$p.value

# (4) Rice
rice_t <- treated %>% pull(rice)
rice_u <- untreated %>% pull(rice)
mt4 = mean(treated$rice)
```

```r
mu4 = mean(untreated$rice)
sdt4 = sd(treated$rice)
sdu4 = sd(untreated$rice)
nt4 = nrow(subset(treated, rice=="1"))
nu4 = nrow(subset(untreated, rice=="1"))
t_test <-t.test(rice_t, rice_u)
p4 = t_test$p.value

# (5) Lentils
lentils_t <- treated %>% pull(lentils)
lentils_u <- untreated %>% pull(lentils)
mt5 = mean(treated$lentils)
mu5 = mean(untreated$lentils)
sdt5 = sd(treated$lentils)
sdu5 = sd(untreated$lentils)
nt5 = nrow(subset(treated, lentils=="1"))
nu5 = nrow(subset(untreated, lentils=="1"))
t_test <-t.test(lentils_t, lentils_u)
p5 = t_test$p.value

# (6) Wheat
wheat_t <- treated %>% pull(wheat)
wheat_u <- untreated %>% pull(wheat)
mt6 = mean(treated$wheat)
mu6 = mean(untreated$wheat)
sdt6 = sd(treated$wheat)
sdu6 = sd(untreated$wheat)
nt6 = nrow(subset(treated, wheat=="1"))
nu6 = nrow(subset(untreated, wheat=="1"))
t_test <-t.test(wheat_t, wheat_u)
p6 = t_test$p.value

# (7) Cotton
cotton_t <- treated %>% pull(cotton)
cotton_u <- untreated %>% pull(cotton)
mt7 = mean(treated$cotton)
mu7 = mean(untreated$cotton)
sdt7 = sd(treated$cotton)
sdu7 = sd(untreated$cotton)
nt7 = nrow(subset(treated, cotton=="1"))
nu7 = nrow(subset(untreated, cotton=="1"))
t_test <-t.test(cotton_t, cotton_u)
p7 = t_test$p.value

# (8) All districts (district_cat)
all_dist_t <- treated %>% pull(district_cat)
all_dist_u <- untreated %>% pull(district_cat)
mt8 = mean(treated$district_cat)
mu8 = mean(untreated$district_cat)
sdt8 = sd(treated$district_cat)
sdu8 = sd(untreated$district_cat)
nt8 = nrow(treated)
nu8 = nrow(untreated)
```

```
t_test <-t.test(all_dist_t, all_dist_u)
p8 = t_test$p.value

# (9) Karur (karur)
karur_t <- treated %>% pull(karur)
karur_u <- untreated %>% pull(karur)
mt9 = mean(treated$karur)
mu9 = mean(untreated$karur)
sdt9 = sd(treated$karur)
sdu9 = sd(untreated$karur)
nt9 = nrow(subset(treated, karur=="1"))
nu9 = nrow(subset(untreated, karur=="1"))
t_test <-t.test(karur_t, karur_u)
p9 = t_test$p.value

# (10) Tenkasi (tenkasi)
tenkasi_t <- treated %>% pull(tenkasi)
tenkasi_u <- untreated %>% pull(tenkasi)
mt10 = mean(treated$tenkasi)
mu10 = mean(untreated$tenkasi)
sdt10 = sd(treated$tenkasi)
sdu10 = sd(untreated$tenkasi)
nt10 = nrow(subset(treated, tenkasi=="1"))
nu10 = nrow(subset(untreated, tenkasi=="1"))
t_test <-t.test(tenkasi_t, tenkasi_u)
p10 = t_test$p.value

# (11) Pudukkottai (pudukkottai)
pudukkottai_t <- treated %>% pull(pudukkottai)
pudukkottai_u <- untreated %>% pull(pudukkottai)
mt11 = mean(treated$pudukkottai)
mu11 = mean(untreated$pudukkottai)
sdt11 = sd(treated$pudukkottai)
sdu11 = sd(untreated$pudukkottai)
nt11 = nrow(subset(treated, pudukkottai=="1"))
nu11 = nrow(subset(untreated, pudukkottai=="1"))
t_test <-t.test(pudukkottai_t, pudukkottai_u)
p11 = t_test$p.value

# (12) Thanjavur (thanjavur)
thanjavur_t <- treated %>% pull(thanjavur)
thanjavur_u <- untreated %>% pull(thanjavur)
mt12 = mean(treated$thanjavur)
mu12 = mean(untreated$thanjavur)
sdt12 = sd(treated$thanjavur)
sdu12 = sd(untreated$thanjavur)
nt12 = nrow(subset(treated, thanjavur=="1"))
nu12 = nrow(subset(untreated, thanjavur=="1"))
p12 = NaN

# (13) Dindigul (dindigul)
dindigul_t <- treated %>% pull(dindigul)
dindigul_u <- untreated %>% pull(dindigul)
```

```r
mt13 = mean(treated$dindigul)
mu13 = mean(untreated$dindigul)
sdt13 = sd(treated$dindigul)
sdu13 = sd(untreated$dindigul)
nt13 = nrow(subset(treated, dindigul=="1"))
nu13 = nrow(subset(untreated, dindigul=="1"))
t_test <-t.test(dindigul_t, dindigul_u)
p13 = t_test$p.value

# (13) Madurai (madurai)
madurai_t <- treated %>% pull(madurai)
madurai_u <- untreated %>% pull(madurai)
mt14 = mean(treated$madurai)
mu14 = mean(untreated$madurai)
sdt14 = sd(treated$madurai)
sdu14 = sd(untreated$madurai)
nt14 = nrow(subset(treated, madurai=="1"))
nu14 = nrow(subset(untreated, madurai=="1"))
t_test <-t.test(madurai_t, madurai_u)
p14 = t_test$p.value


# Create data frame to store computed values
baseline = c("Profits 2005", "Farmer Birth Year", "Crop", "Rice", "Lentils",
             "Wheat", "Cotton", "District", "Karur", "Tenkasi", "Pudukkottai",
             "Thanjavur", "Dindigul", "Madurai")
mean_treated = c(mt1, mt2, mt3, mt4, mt5, mt6, mt7, mt8, mt9, mt10, mt11,
                 mt12, mt13, mt14)
sd_treated= c(sdt1, sdt2, sdt3, sdt4, sdt5, sdt6, sdt7, sdt8, sdt9, sdt10,
              sdt11, sdt12, sdt13, sdt14)
n_treated = c(nt1, nt2, nt3, nt4, nt5, nt6, nt7, nt8, nt9, nt10, nt11, nt12,
              nt13, nt14 )
mean_untreated = c(mu1, mu2, mu3, mu4, mu5, mu6, mu7, mu8, mu9, mu10, mu11,
                   mu12, mu13, mu14)
sd_untreated = c(sdu1, sdu2, sdu3, sdu4, sdu5, sdu6, sdu7, sdu8, sdu9, sdu10,
                 sdu11, sdu12, sdu13, sdu14)
n_untreated = c(nu1, nu2, nu3, nu4, nu5, nu6, nu7, nu8, nu9, nu10, nu11, nu12,
                nu13, nu14)
diff_means = c(mt1-mu1, mt2-mu2, mt3-mu3, mt4-mu4, mt5-mu5, mt6-mu6, mt7-mu7,
               mt8-mu8, mt9-mu9, mt10-mu10, mt11-mu11, mt12 - mu12, mt13-mu13,
               mt14-mu14)
p_value = c(p1, p2, p3, p4, p5, p6, p7, p8, p9, p10, p11, p12, p13, p14)
df = data.frame(baseline, mean_treated, mean_untreated, diff_means, sd_treated,
                sd_untreated, n_treated, n_untreated, p_value)
```

Table 1: Balance Table

Variable	Mean Treated	Mean Untreated	Diff Means	SD Treated	SD Untreated	N Treated	N Untreated	p-value
Profits 2005	20001.713	19989.878	11.835	1036.465	616.214	2500	7500	0.589
Farmer Birth Year	1968.820	1968.932	-0.112	5.042	7.715	2500	7500	0.406
Crop	2.958	3.000	-0.042	0.827	0.775	2500	7500	0.024
Rice	0.379	0.400	-0.021	0.485	0.490	947	3000	0.059
Lentils	0.300	0.300	0.000	0.458	0.458	750	2250	1.000
Wheat	0.300	0.300	0.000	0.458	0.458	750	2250	1.000
Cotton	0.021	0.000	0.021	0.144	0.000	53	0	0.000
District	6.000	3.000	3.000	0.000	1.414	2500	7500	0.000
Karur	0.000	0.200	-0.200	0.000	0.400	0	1500	0.000
Tenkasi	0.000	0.200	-0.200	0.000	0.400	0	1500	0.000
Pudukkottai	0.000	0.200	-0.200	0.000	0.400	0	1500	0.000
Thanjavur	1.000	0.000	1.000	0.000	0.000	2500	0	NaN
Dindigul	0.000	0.200	-0.200	0.000	0.400	0	1500	0.000
Madurai	0.000	0.200	-0.200	0.000	0.400	0	1500	0.000

The available observables at baseline in this dataset are the following:

1. Profits 2005
2. Farmer birth year
3. Crops
    - Rice
    - Lentils
    - Wheat
    - Cotton
4. District
    - Karur
    - Tenkasi
    - Pudukkottai
    - Thanjavur
    - Dindigul
    - Madurai

In the case of crops and district, it is necessary to convert them into numerical categorical variables in order to run formal t-tests and check for balance (eg. cotton, lentils, rice and wheat are converted to 1, 2, 3, 4). I additionally convert them to dummy variables with the purpose of checking for balance for each type of crop and district (eg. 1 if rice, 0 otherwise).

As a result, I obtained the following balance table:

```
# Create Balance Table
kbl(df, caption = "Balance Table", booktabs = T, digits = 3,col.names =
      c("Variable", "Mean Treated","Mean Untreated", "Diff Means", "SD Treated",
        "SD Untreated", "N Treated", "N Untreated",
        "p-value"))%>% kable_styling(latex_options = "scale_down")
```

The balance table reports the means, standard deviations, difference in means and the p-value from running a Welch Two-Sample t-test, which tests for the following hypotheses:

$$H_0 : mean \text{ treatment} - mean \text{ control} = 0$$

$$H_A : mean \text{ treatment} - mean \text{ control} \neq 0$$

There are several observations that can be made from the balance table:

- The **baseline profits** are **balanced** between treated and untreated groups. Since the p-value $= 0.589 > 0.05$, we fail to reject the null that the true difference in means is 0. The difference in means is relatively small, -11.835 between treated and untreated. We can observe the standard deviation in profits is very high in both groups, particularly for the untreated group: 1036.465.
- The **farmer birth year** is also **balanced** between both groups. The p-value $= 0.406 > 0.05$, such that we fail to reject the null at the 5% level. The difference in means is very small at 0.112.
- If we analyze the **crops** variable, in general terms, we find that it is **unbalanced**. The p-value is small at $0.024 < 0.05$, such that we reject the null in favor of the alternative at the 5% significance level. The difference in means, however, is very small, at 0.042. To understand better, we can check for balance for each type of crop:
  - If we specifically look at **rice**, the p-value $0.059 > 0.05$ such that we fail to reject the null, therefore the variable is statistically **balanced**. However, we can see that the number of rice crops in the treated group, 3000, is more than double that of the treated group, 947.
  - As to **lentils**, it is completely **balanced** since the p-value is at it's highest possible value of $1 > 0.05$, such that we we fail to reject the null. There are 2250 lentil crops in the treated group compared to 750 in the untreated group and the difference in means is 0.
  - For **wheat**, this characteristic is also completely **balanced** with a p-value of $1 > 0.05$. The statistics are exactly the same as those for lentils.
  - For **cotton**, the characteristic is completely **unbalanced**, where the p-value of $0 < 0.05$, such that we reject the null in favor of the alternative. This is visible as there are 0 cotton crops in the treatment group, compared to 53 in the untreated group, so there is clearly no balance at all. It is because of cotton that the variable crop as a whole is unbalanced, even though the difference in means for crop is very, very small.
- Looking at **district**, this variable is completely **unbalanced**, where the p-value of $0 < 0.05$. This is to be expected since the FIONA program only impacted certain districts. By design, each district is entirely offered treatment or not, so we can expected every district to be unbalanced:
  - All of the districts **Karur, Tenkasi, Pudukkottai, Dindigul and Madurai** are completely **unbalanced**. In the case of **Thanjavur**, it is not possible for R to run the t-test because all the data is constant. However, for the same reason as the 5 other districts, it is also unbalanced given that all observations in the district correspond to the treated group.

To better understand the disparity in the sample size for the treated and the untreated, I made the following table to summarize the offer of insurance, which was done at the district level:

```r
ndt1 = nrow(subset(treated, district_cat=="1"))
ndu1 = nrow(subset(untreated, district_cat=="1"))
ndt2 = nrow(subset(treated, district_cat=="2"))
ndu2 = nrow(subset(untreated, district_cat=="2"))
ndt3 = nrow(subset(treated, district_cat=="3"))
ndu3 = nrow(subset(untreated, district_cat=="3"))
ndt4 = nrow(subset(treated, district_cat=="4"))
ndu4 = nrow(subset(untreated, district_cat=="4"))
ndt5 = nrow(subset(treated, district_cat=="5"))
ndu5 = nrow(subset(untreated, district_cat=="5"))
ndt6 = nrow(subset(treated, district_cat=="6"))
ndu6 = nrow(subset(untreated, district_cat=="6"))
tot_ndt = 1500 * 5

district = c("Dindigul", "Karur", "Madurai", "Pudukkottai", "Tenkasi",
             "Thanjavur", "Total")
n_treated = c(ndt1, ndt2, ndt3, ndt4, ndt5, ndt6, tot_ndt)
n_untreated = c(ndu1, ndu2, ndu3, ndu4, ndu5, ndu6, ndu6)
df2 = data.frame(district, n_treated, n_untreated)

kbl(df2, caption = "Treated and Untreated Observations by District",
    booktabs = T, digits = 3,col.names =
      c("District", "N Treated", "N Untreated"))%>% kable_styling(
        latex_options = c("striped","hold_position"), position = "center")
```

Table 2: Treated and Untreated Observations by District

District	N Treated	N Untreated
Dindigul	0	1500
Karur	0	1500
Madurai	0	1500
Pudukkottai	0	1500
Tenkasi	0	1500
Thanjavur	2500	0
Total	7500	0

As we can observe, 5 out of 6 districts reported in the dataset were not offered the insurance, with 1500 farmers each. On the other hand, the district of Thanjavur was the only one to be offered the insurance, with a total of 2500 farmers. In total, 7500 farmers were not offered the insurance compared to only 2500 that were (ie. onlt 25% of the sample was offered the insurance).

In terms of the numbers of farmers offered insurance, there is a lot of unbalance, which makes it challenging to compare these two groups when we have relatively little observations for those farmers that were offered insurance.

From the histograms in part (3), I got a positive sense that the data does not have any drastic outliers and that there are no missing values. However, this analysis makes me feel worse, because most of the baseline variables (accounting for dummies) are unbalanced. The district variable will not be very useful since entire districts received treatment or did not. Within the crop variable, 3 out 4 crops are balanced, with the exception of cotton, which makes it difficult to analyze the effect of the insurance scheme for this type of crop, if there is no counterfactual to compare it to.

(6) HARRIS are interested in your approach in (4), but would like to know a bit more about how much they should believe your proposal. Describe the assumptions required for these designs to be valid in math and in words. To the extent possible, assess the validity of these assumptions using the provided data.

When it comes to SOO designs, two main assumptions are required to hold:

1. Conditional independence: conditional on observables, treatment assignment is independent of potential outcomes, which means that all differences between $D_{farmer} = 1$ and $D_{farmer} = 0$ are assumed to be *as good as random*. This is a **very strong assumption** and can be mathematically expressed as:

$$(Y_i(1), Y_i(0)) \perp D_i \mid X_i$$

Where:

- D_i refers to treatment status $Di = 1$ for farms that took up FIONA insurance and $D_i = 0$ for those that didn't.

- Y_i is the outcome variable, 2016 profits
- X_i represents control variables, such as: crop dummies, farmer_birth_year, etc.
- $i = farmer$

This means that once we control for X_i's (eg. baseline profits, farmer birth year, crop), opting for insurance D_i is *as good as random* and we have essentially eliminated selection bias.

2. Common support - The probability that $D_i = 1$ for all levels of X_i is between 0 and 1:

$$0 < Pr(D_i = 1 \mid X_i = x) < 1$$

Where:

- D_i refers to treatment status $Di = 1$ for farms that took up FIONA insurance and $D_i = 0$ for those that didn't.
- X_i represents control variables, such as: crop dummies, farmer_birth_year, etc.
- $i = farmer$

This means that for each covariate, there are *both treated and untreated units* (ie. farmers that took up the insurance and farmers that did not).

Additional to the conditional independence and the common support assumptions, there are two more assumptions that need to hold when employing the regression adjustment approach:

Regression adjustment additional assumptions:

- The mean covariates for the treated are **close** to the mean covariates for the untreated, ie: $\overline{X}_T$ close to $\overline{X}_U$. If the difference between the covariates is large, $\mid \overline{X}_T - \overline{X}_U \mid$, then the $\hat{\tau}$ estimate will be biased.
- The true relationship is assumed to be of the following **functional form**:

$$Y_i = \alpha + \tau D_i + \gamma X_i$$

Where:

- Y_i is the outcome variable, 2016 profits
- D_i refers to treatment status $Di = 1$ for farms that took up FIONA insurance and $D_i = 0$ for those that didn't.
- X_i represents control variables, such as: crop dummies, farmer_birth_year, etc.
- $i = farmer$

In the case of **Matching**, since treated and untreated units are matched *exactly*, it's a more flexible estimator in which the functional form does not matter and only the first two SOO assumptions described above are required.

ASSESSING THE VALIDITY OF THESE ASUMPTIONS:

- **(1) Conditional Independence Assumption**
  - It is not possible to test this assumption. However, it is very, very unlikely for this assumption to hold as we would essentially have to account for all possible covariates X_i's that are correlated with taking up insurance. This is never actually possible because, for one, we can never account for unobservables, and second, due to limitations in the data we can only control for *some* covariates. In this dataset, we have few covariates: profits_2005, farmer birth year, crop and district. There are many more covariates that could be inducing selection bias for which we have no data. For example: size of farmer household, education or training in agriculture, additional jobs/income, savings, etc.
- **(2) Common Support Assumption**
  - From the balance table, we found that there were 0 farmers that produce cotton and took up the insurance. All 53 farmers that produce cotton were not treated. Likewise, by design, offer

of insurance was done at the district level, so for each district, all farmers are entirely treated or entirely untreated.

```
crop = c("Rice", "Lentils", "Wheat", "Cotton")
n_treated = c(nt4, nt5, nt6, nt7)
n_untreated = c(nu4, nu5, nu6, nu7)
df3 = data.frame(crop, n_treated, n_untreated)

kbl(df3, caption = "Treated and Untreated Observations by Crop",
    booktabs = T, digits = 3,col.names =
      c("Crop", "N Treated", "N Untreated"))%>% kable_styling(
        latex_options = c("striped","hold_position"), position = "center")
```

Table 3: Treated and Untreated Observations by Crop

Crop	N Treated	N Untreated
Rice	947	3000
Lentils	750	2250
Wheat	750	2250
Cotton	53	0

```
kbl(df2, caption = "Treated and Untreated Observations by District",
    booktabs = T, digits = 3,col.names =
      c("District", "N Treated", "N Untreated"))%>% kable_styling(
        latex_options = c("striped","hold_position"), position = "center")
```

Table 4: Treated and Untreated Observations by District

District	N Treated	N Untreated
Dindigul	0	1500
Karur	0	1500
Madurai	0	1500
Pudukkottai	0	1500
Tenkasi	0	1500
Thanjavur	2500	0
Total	7500	0

- Therefore, the common support assumption **doesn't hold for cotton and district**, since there are not treated units and untreated for *every level* of X. If we look at the numbers, there are a lot more untreated farmers

(7500) than treated farmers (2500). We can expect a lot of farmers to be unmatched due to lack of overlap.

- Also, in the case of **farmer birth year**, it will be difficult to find exact matches without running into the **curse of dimensionality** since there are a total of 65 different farmer birth years. As we can see below, there are many birth years for which there are very few farmers (eg. 7 where there is only 1 farmer), so that it will not be possible to find an exact match at every single level.
- Thus, the common support assumption **does not hold** for birth year either.

```
fiona %>%
  group_by(farmer_birth_year) %>%count() %>% arrange(n)
```

```
## # A tibble: 65 x 2
## # Groups:   farmer_birth_year [65]
##    farmer_birth_year     n
##                <int> <int>
##  1              1916     1
##  2              1919     1
##  3              1928     1
##  4              1929     1
##  5              1930     1
##  6              1931     1
##  7              1933     1
##  8              1932     2
##  9              1925     3
## 10              1935     3
## # ... with 55 more rows
```

- With the exact matching estimator, it will **not be possible** to find exact matches for 2005 profits because it is a continuous variable and we will run into the **curse of dimensionality**.
- In conclusion, the following table summarizes that the common support assumption **only holds** for three covariates: **rice, wheat and lentils.**

```
var = c("Profits 2005", "Farmer Birth Year", "Rice", "Lentils",
             "Wheat", "Cotton", "Karur", "Tenkasi", "Pudukkottai",
             "Thanjavur", "Dindigul", "Madurai")
holds = c("No", "No", "Yes", "Yes", "Yes", "No", "No", "No", "No", "No", "No",
          "No")
df4 = data.frame(var, holds)
kbl(df4, caption = "Common Support Assumption by Variable",
    booktabs = T, digits = 3,col.names =
      c("Variable", "Assumption Holds"))%>% kable_styling(
        latex_options = c("striped","hold_position"), position = "center")
```

Table 5: Common Support Assumption by Variable

Variable	Assumption Holds
Profits 2005	No
Farmer Birth Year	No
Rice	Yes
Lentils	Yes
Wheat	Yes
Cotton	No
Karur	No
Tenkasi	No
Pudukkottai	No
Thanjavur	No
Dindigul	No
Madurai	No

- **(3) Closeness Assumption**
  - For the $\overline{X}_T$ close to $\overline{X}_U$ regression adjustment assumption, we can refer back to the balance table in (5). The covariates that are statistically balanced are: 2005 profits, farmer birth year, rice, lentils and wheat. For these 5 variables, the difference in means between treated and untreated is very small.

Discuss whether you think you will be able to obtain a credible estimate of the answer to the questions described in (1) and (2) based on the data, and use concrete examples to explain why or why not.

Since this is not an RCT where $E[\varepsilon_i|D_i] = 0$ holds, we cannot establish causality in any of these relationships. I predict that it will not be possible to obtain a very credible estimate with neither SOO design, in particular because of the conditional independence assumption being violated. As previously explained, we are missing key covariates that very likely are correlated with the treatment variable, such as the number of dependents or children. Farmers that have more dependents to sustain economically, might be more willing to enroll in insurance to avoid losing profits. Likewise, if there are long-term diseases in the family that require stable profits.

Additionally, there is the issue of unobservables such as farmers being risk adverse or not, or having distrust of insurance schemes. SOO designs are unable to control for unobservables and cannot provide reliable estimates in such situations, so it's likely that the estimates will not perform significantly better than the naive estimator.

(7) Use a regression-based approach to estimate the effect of FIONA on farmer profits. Describe which variables you chose to include in your regression, and explain why you chose these. Did you leave any variables out? If yes, explain why. Interpret your results. What are the strengths and weaknesses of this approach? How do your results differ from what you find if you instead use the naive estimator?

The data we have available from FIONA represents farmers that self-selected into treatment (opting-in for the rain insurance). For this reason, I chose to include as many covariates available in the dataset for which **both the common support and closeness assumption holds**. The goal is to minimize the risk of violating the conditional independence assumption by excluding covariates that are potentially correlated with farmer selection into the insurance scheme, given that those covariates satisfy the required SOO assumptions.

These are the variables I included:

- **fiona_farmer** since this is out treatment variable, this is the coefficient that will estimate our ATE
- **lentils**, **rice** and **wheat**, because some types of crop may be more rainfall-dependent, making it a variable that farmers consider when self-selecting into treatment. I.

I excluded all of the following variables:

- **farmer_birth_year**, while it satisfy the closeness assumption, it does not satisfy common support
- **cotton** because it does not satisfy common support nor closeness asumption
- **all district dummies** because they don't satisfy common support nor the closeness assumption
- **fertilizer use**, this is a post-treatment variable that should *never be included* as doing so would introduce bias into the model.
- **profits_2005**, satisfies closeness but does not satisfy common support. Also, 2005 profits are likely very different to profits in 2014 (if anything due to inflation) when the study actually started. It is not relevant to include this variable because it's unlikely that the profits of a farmer in 2005 is correlated with his/her decision to opt into insurance 9 years later.

```r
reg <- lm(profits_2016 ~ fiona_farmer + lentils + rice + wheat,  data = fiona)
summary(reg)
```

```
##
## Call:
## lm(formula = profits_2016 ~ fiona_farmer + lentils + rice + wheat,
##     data = fiona)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5224.7  -838.5     1.3   842.0  5299.9
```

```
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  21420.61     179.01 119.665 < 0.0000000000000002 ***
## fiona_farmer  2375.95      29.91  79.425 < 0.0000000000000002 ***
## lentils        625.85     179.45   3.488             0.000489 ***
## rice           301.64     179.12   1.684             0.092207 .
## wheat          711.07     179.45   3.963            0.0000747 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1285 on 9995 degrees of freedom
## Multiple R-squared:  0.3972, Adjusted R-squared:  0.397
## F-statistic:  1647 on 4 and 9995 DF,  p-value: < 0.00000000000000022
```

From the regression, we obtain the following results:

- Holding everything else constant, the difference in endline 2016 profits between a farmer that opted for FIONA insurance and a farmer that didn't, is estimated to be approximately 2375.95 Indian rupees. This ATE estimate is highly statistically significant at all conventional levels of significance, such that we reject the null in favor of the alternative that ATE is significantly different from 0.
- Ceteris paribus, rice, lentils and wheat are statistically significant at the 5% level. Rice is not statistically significant at the 5% level.
- The coefficient on the three crops is positive, which means that growing each of these crops, all else constant, is associated with an increase in profits, in particular wheat.

Running the naive estimator

If, instead we run the naive estimator, we obtain the following results:

```
# Naive estimator
naive <- lm(profits_2016 ~ fiona_farmer,  data = fiona)
summary(naive)
```

```
##
## Call:
## lm(formula = profits_2016 ~ fiona_farmer, data = fiona)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5438.4  -843.4     6.0   840.3  5142.6
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  21942.34      14.99 1463.95 <0.0000000000000002 ***
## fiona_farmer  2369.56      29.98   79.05 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1298 on 9998 degrees of freedom
## Multiple R-squared:  0.3846, Adjusted R-squared:  0.3845
## F-statistic:  6248 on 1 and 9998 DF,  p-value: < 0.00000000000000022
```

- Ceteris paribus, the difference in endline 2016 profits between a farmer that opted for FIONA insurance and a farmer that didn't, is estimated to be approximately 2369.56 Indian rupees. This ATE estimate is highly statistically significant at all conventional levels of significance. This is a minimal difference of

-6.39 rupees from the estimate obtained in the adjusted regression.

- The R^2 of the naive estimator model is of a relative size at 0.38, meaning a lot of the change in 2016 profits is not being captured by the model.
- In the absence of randomization, we cannot determine causality from this model and the estimated ATE coefficient is not reliable, as the likelihood of selection bias is very high (ie. farmers that self-selected into the insurance are very likely to be systematically different from those who did not)

ASSESSMENT OF THE ADJUSTED REGRESSION APPROACH:

1. Strengths

- Compared to the naive estimator, the adjusted regression is barely different. The R^2 of 0.40 is slightly higher than the 0.38 obtained in the naive estimator, which means that adding the covariates allowed us to control for a little more of the variation in endline profits.

2. Weaknesses

- It is very, very likely that the conditional independence assumption does not hold in this case. We can infer this from the R^2, because there is much in the change in 2016 profits that is unexplained by the model. This is without a doubt because the selection bias is not eliminated simply by adding a few crop controls.
- Although we added controls to the regression, they are either not enough or not the right controls. For this approach to perform better, we would need to include more, if not all, the covariates that are strongly correlated with the selection into the FIONA insurance.
- We also cannot add more covariates available in the dataset like 2005 profits, farmer birth year, cotton and district variables because these do not fulfill the required SOO assumptions. Given the high selection bias, however, it is unlikely that including those controls would change the result by much.
- It is very possible that there are unobservables that drive selection into the insurance, for example, farmers that are more risk-adverse might be more willing to choose insurance and see it is a benefit. The adjusted regression approach cannot account for this issue and is therefore **biased**.
- Since this is not an RCT where $E[\varepsilon_i|D_i] = 0$ holds, and the conditional independence assumption very likely is violated, we cannot establish causality in any of these relationships.
- In all, the adjusted regression performs the same as the naive estimator.

(8) Use an exact matching approach to estimate the effect of FIONA on farmer profits. What variables should you include in the matching procedure? Begin by estimating the answer to the question in (1). Then, estimate the answer to the question in (2). Are these meaningfully different? Would you have expected these results to be the same? Why or why not? What are the strengths and weaknesses of this approach? How do your results differ from what you find if you instead use the naive estimator? From what you found in (8)? Did you run into the Curse of Dimensionality with this analysis? If yes, describe how it affected your approach. If not, describe how the Curse could have generated problems in this setting.

There are 4 steps to calculate Exact Matching:

1. Divide the data into cells uniquely identified by the covariates.

I am going to use the three crop dummies for which there are observations both in the treated and untreated groups: rice, lentil and wheat.

All other variables that do not satisfy the common support assumption are excluded (ie.for district there are no observations in both groups and profits_2005 is a continuous variable which would lead to the curse of dimensionality to automatically kick in with no exact matches found and the method automatically failing).

```
# STEP 1 - divide data into cells
fiona_match <- matchit(fiona_farmer ~ lentils + rice + wheat, method = "exact",
                       data = fiona)
df_match <- match.data(fiona_match)
nrow(df_match)
```

```
## [1] 9947
```

```
length(unique(df_match$subclass))
```

```
## [1] 3
```

From this first step, we obtain **9447 matched observations**. This is a very high number of matches considering the total number of observations in the data is 10000. In total, we have **3 cells** since only 3 crop dummies are being utilized for exact matching.

```
# To check if matching worked, we do a balance table.

# Separate the dataset into treated and untreated groups:
treated <- filter(df_match, fiona_farmer == 1)
untreated <- filter(df_match, fiona_farmer == 0)

# Run t-tests and calculate means:

# (1) Rice
rice_t <- treated %>% pull(rice)
rice_u <- untreated %>% pull(rice)
mt1 = mean(treated$rice)
mu1 = mean(untreated$rice)
t_test <-t.test(rice_t, rice_u)
p1 = t_test$p.value

# (3) Lentils
lentils_t <- treated %>% pull(lentils)
lentils_u <- untreated %>% pull(lentils)
mt2 = mean(treated$lentils)
mu2 = mean(untreated$lentils)
t_test <-t.test(lentils_t, lentils_u)
p2 = t_test$p.value

# (4) Wheat
wheat_t <- treated %>% pull(wheat)
wheat_u <- untreated %>% pull(wheat)
mt3 = mean(treated$wheat)
mu3 = mean(untreated$wheat)
t_test <-t.test(wheat_t, wheat_u)
p3 = t_test$p.value

# Create data frame to store computed values
baseline = c("Rice", "Lentils", "Wheat")
mean_treated = c(mt1, mt2, mt3)
mean_untreated = c(mu1, mu2, mu3)
p_value = c(p1, p2, p3)
diff_means = c(mt1-mu1, mt2-mu2, mt3-mu3)
df = data.frame(baseline, mean_treated, mean_untreated, diff_means, p_value)
```

```r
# Post-Matching Balance Table
kbl(df, caption = "Post-Matching Balance Table", booktabs = T, digits = 3,col.names =
      c("Variable", "Mean Treated",
        "Mean Untreated", "Diff Means",
        "p-value"))%>% kable_styling(latex_options =
                                   c("stripped", "hold_position"),
                                 position = "center")
```

Table 6: Post-Matching Balance Table

Variable	Mean Treated	Mean Untreated	Diff Means	p-value
Rice	0.387	0.4	-0.013	0.253
Lentils	0.306	0.3	0.006	0.544
Wheat	0.306	0.3	0.006	0.544

From this new balance table, post-matching, we can observe that matching was successful because all p-values > 0.05, we fail to reject the null that the true difference in means is 0. This tells us that all three covariates selected for matching are balanced.

2. Next, calculate $\overline{Y}_T$ and $\overline{Y}_U$ *for each cell.*

```r
# STEP 2

# Filter the matched observations by treated and untreated
matched_treated <- filter(df_match, fiona_farmer == 1)
matched_untreated <- filter(df_match, fiona_farmer == 0)

# Group by cell (subclass) and calculate the mean of outcome variable
# (ie.profits_2016)
means_treated <- matched_treated %>%
  group_by(subclass) %>%
  dplyr::summarize(Mean = mean(profits_2016))

means_untreated <- matched_untreated %>%
  group_by(subclass) %>%
  dplyr::summarize(Mean = mean(profits_2016))

# Convert lists to a dataframe with the means of treated and untreated
means <- cbind(data.frame(means_treated), data.frame(means_untreated))
colnames(means) <- c('cell_t', 'treated_mean','cell_u', 'untreated_mean')
means <- means[ -c(3) ]

head(means)
```

```
##   cell_t treated_mean untreated_mean
## 1      1     23443.41       21928.95
## 2      2     24729.22       21944.19
## 3      3     25027.61       21958.36
```

3. Calculate the difference in means $\overline{Y}_T - \overline{Y}_U$ for each cell.

```r
#Step 3, calculate the difference in means for each cell
means$diff_means <- means$treated_mean - means$untreated_mean
head(means)
```

```
##   cell_t treated_mean untreated_mean diff_means
## 1      1     23443.41       21928.95   1514.468
## 2      2     24729.22       21944.19   2785.022
## 3      3     25027.61       21958.36   3069.251
```

4. Estimate $\hat{\tau}$ as a weighted average.

- First I calculate this for (1), which is the τ^{ATE}:

```
# STEP 4 (1) WEIGHTED ATE

# Store the sample size. For ATE, weights are based on ALL observations from
# the full sample:
n = nrow(df_match)

# Calculate weights by cell (subclass)
cell_n <- df_match %>%
  group_by(subclass) %>%count()

# Convert to dataframe
cell_size_df <- cbind(data.frame(cell_n))
colnames(cell_size_df) <- c('cell_t', 'cell_n')

# Merge into a single dataframe
matching <-cbind(means, cell_size_df)

# Drop duplicate identifier columns used for merging dataframes
matching <- matching[ -c(5) ]

# Calculate weights by dividing cell_size/n
matching $weights <- matching $cell_n / n

# Calculate weights * difference in means
matching $w_x_diffm <- matching $diff_means * matching $weights

# Get top observations of the dataframe
head(matching)
```

```
##   cell_t treated_mean untreated_mean diff_means cell_n   weights w_x_diffm
## 1      1     23443.41       21928.95   1514.468   3947 0.3968031  600.9456
## 2      2     24729.22       21944.19   2785.022   3000 0.3015985  839.9583
## 3      3     25027.61       21958.36   3069.251   3000 0.3015985  925.6813
```

```
# Sum to get ATE estimate
ate <- sum(matching $w_x_diffm)
ate
```

```
## [1] 2366.585
```

As a result of this process, we get that our estimated $\tau^{ATE} = 2366.59$ rupees.

- Next, I do the weights calculation for (2), which is τ^{ATT}:

```
# STEP 4 (2) WEIGHTED ATT

# Store the treatment sample:
n_treat = nrow(treated)
```

```
# Calculate treated counts by cell (subclass)
n_treated<- matched_treated %>%
  group_by(subclass) %>%count()

# Convert to dataframe
cell_treated_df <- cbind(data.frame(n_treated))
colnames(cell_treated_df ) <- c('cell_t', 'cell_treated_n')

# Merge into a single dataframe
matching <-cbind(matching, cell_treated_df)

# Drop duplicate identifier columns used for merging dataframes
matching <- matching[ -c(1, 8) ]

# Calculate treated weights by dividing cell_treated_n/sample n_treat
matching $weights_treated <- matching $cell_treated_n/n_treat

# Calculate treated weights * difference in means
matching $wtreat_x_diffm <- matching $diff_means * matching$weights_treated

# Get top observations of the dataframe
head(matching)
```

```
##   treated_mean untreated_mean diff_means cell_n   weights w_x_diffm
## 1     23443.41       21928.95   1514.468   3947 0.3968031  600.9456
## 2     24729.22       21944.19   2785.022   3000 0.3015985  839.9583
## 3     25027.61       21958.36   3069.251   3000 0.3015985  925.6813
##   cell_treated_n weights_treated wtreat_x_diffm
## 1            947       0.3870045       586.1060
## 2            750       0.3064978       853.6029
## 3            750       0.3064978       940.7185
```

```
# Sum to get ATT estimate
att <- sum(matching $wtreat_x_diffm)
att
```

```
## [1] 2380.427
```

For our (2) estimate, we obtain that $\tau^{ATT} = 2380.43$ rupees.

COMPARISON BETWEEN (1) ATE AND (2) ATT WITH THE MATCHING ESTIMATOR

- My τ^{ATE} and τ^{ATT} are very close: ATE = 2366.59 rupees vs ATT = 2380.43 rupees.
- This tells us that the estimated effect of FIONA on endline 2016 profits for the average farmer, *given the matched covariates and assuming the SOO assumptions hold*, is approximately 2366.59 rupees, compared to 2380.43 rupees for the average farmer among *farmers that took up the insurance.* This represents only a 0.58% percentage difference between the ATE and ATT estimates, so they are **not meaningfuly different**.
- I would not have expected these results to be the so close because when there is selection bias $\tau^{ATE} \neq \tau^{ATT}$. Since there was no randomization in this case, it is very likely that farmers that opt for insurance are systematically different that those who do not, **ie. heterogenous effects.** As such, I was expecting that the average treatment effect for the those who took up FIONA insurance would be even higher than the ATE. This result does not mean that there is no selection bias but instead, that we are unable to capture it with the exact matching estimator.

```r
# Percentage change
((att - ate)/ate) *100
```

```
## [1] 0.5848982
```

ASSESSMENT OF THE EXACT MATCHING APPROACH:

1. **Strengths**

   - Based on the selected covariates, most of the observations had identical matches (ie. 9947). This creates observably identical comparison between the treated and untreated groups.
   - The balance table post matching showed that the exact matching procedure was successful in achieving balance in the three included covariates.
   - This approach does not make any functional form assumptions and is therefore more flexible than the adjusted regression.

2. **Weaknesses**

   - This approach doesn't work for continuous variables like baseline 2005 profits. This makes it impossible to find exact matches. When attempting to include it for testing purposes, I ran into **the Curse of Dimensionality** because no matches could be found. In this scenario, an alternative approach might be needed like matching by k-nearest neighbors. On one hand, we want to include more covariates to sustain the conditional independence assumption but, we the more covariates, the lower the likelihood of finding exact matches.
   - Since cotton is completely unbalanced, it's not possible to match on cotton, which limits our analysis. Same with the districts, these are not useful variables because they don't satisfy the common support assumption and as such, it is not sensible to include them in the matching procedure.
   - It still doesn't account for selection on unobservables, which an approach like an RCT is able to.

EXACT MATCHING VS. NAIVE ESTIMATOR

When I calculated the naive estimator in (7), I obtained an estimated ATE of 2369.56 rupees, compared to an ATE of 2366.59 rupees with the exact matching estimator, **the difference is minimal** (2.97). The fact that the matching estimated ATE is so close to the naive estimator points to the unreliability of SOO designs, given the high selection bias.

(9) Based on your results in (8), explain to HARRIS whether or not they should implement a FIONA-like program in Bangladesh. Be sure to tell them the reasoning behind your recommendation.

Based on my results in 8, I do not feel confident enough to tell HARRIS whether they should implement a FIONA-like program in Bangladesh.

- Due to the lack of an RCT in the FIONA data, I employed an exact matching approach as an alternative. However, only 3 covariates satisfied the common support assumption, limiting the scope of the technique.
- Second, the ATE estimate obtained with exact matching is essentially the same to the one obtained with the naive estimator. If we analyze the context, it is **very, very likely** that there is **selection bias** in the treatment variable (opting into the FIONA insurance). As such, farmers that opt for the scheme are statistically different from those who do not. The fact that the estimated ATE is so close to the naive estimator, shows how unreliable these results are.
- Third, the required SOO assumptions do not hold entirely.
  - The **conditional independence assumption** is unlikely to hold because the covariates we have available in the dataset are not the only ones that are correlated with the treatment variable, so we cannot safely assume we're in a *"as good as random"* situation.

  - I only included those covariates for which the **common support assumption** holds, by excluding cotton, the district dummies and 2005 profits. However, this means we are excluding precisely those problematic variables for which we would like to have counterfactuals. With the way FIONA was designed, offering insurance only in certain districts, this is simply not possible.

- The ATE and ATT matching estimates are very, very close to the naive estimator, showing this method fails to perform any better.
- An additional issue is that only 2005 profits are available for the baseline, when the study actually started in 2014. Profits in 2005 are unlikely to influence self-selection into treatment 9 years later. It would have been useful to have baseline profits in 2011, for instance and then conduct matching with KNN techniques or bandwidths. With what I estimated in (8), we are unable to compare farmers that had the exact same income before the intervention.

Given these reasons, I cannot determine causality nor be determinant in my conclusions. Instead, I would recommend to HARRIS that we run an RCT, even at a smaller scale through a pilot, to test if the rainfall-index insurance is effective, before doing a full-scale implementation. Alternatively, we could search for other similar projects employing insurance schemes that have successful evaluations to assess whether this should be implemented or not, taking into account the external validity of these studies.