PS2

12265092

04/05/2022

```
library(knitr)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(haven)

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   2.0.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stargazer)
```

```
## Warning: package 'stargazer' was built under R version 4.1.2

##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(broom)

library(kableExtra)
```

```
## Warning: package ’kableExtra’ was built under R version 4.1.3

##
## Attaching package: ’kableExtra’

## The following object is masked from ’package:dplyr’:
##
##     group_rows
```

Question 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?).

Let i be individual farmers where $i \in \{1, 2, ...N\}$

Treatment indicator D_i: D_i be the treatment indicator where $D_i \in \{0, 1\}$ When $D_i = 1$, It means providing the farmer i with rainfall_index insurance, i.e Treated unit When $D_i = 1$, It means not providing the farmer i with rainfall_index insurance i.e Untreeated unit

Outcome Y_i: Y_i be the outcome $Y_i(D_i = 1)$, Outcome(Profit) made by the farmer i with rainfall_index insurance, i.e outcome in a bad rainfall year incase of treatment $Y_i(D_i = 0)$, Outcome(Profit) made by the farmer i without rainall_index insurance, i.e outcome in a bad rainfall year incase of no treatment

Impact/Effect of treatement τ_i: τ_i is keeping everything else constant, the difference between outcomes(profits) made by a farmer i in a bad rainfall year when rainfall_index insurance is provided and in a bad rainfall year when the rainfall_index insurance is not provided, i.e difference in outcomes of a farmer i when treatment and when no treatment.

$\tau_i = Y_i(D_i = 1) - Y_i(D_i = 0)$

While we need both the outcomes at a given time to compute the impact of treatment, the problem is that at a given time, we cannot observe both the outcomes $ or $ we can only observe either $Y_i(D_i = 1) $ or $ Y_i(D_i = 0)$ at a given time.

In detail:

When a farmer i is treated: Observed outcome would be $Y_i(D_i = 1)$(i.e profit made by the farmer i during a bad rainfall year when rainfall_index insurance is provided) and the unobserved outcome would be $Y_i(D_i = 0)$ (i.e profit made by the farmer i during a bad rainfall year when rainall_index insurance is nor provided)

When a farmer i is not treated: Observed outcome would be $Y_i(D_i = 0)$ (i.e profit made by the farmer i during a bad rainfall year when rainall_index insurance is nor provided) and the unobserved outcome would be $Y_i(D_i = 1)$(i.e profit made by the farmer i during a bad rainfall year when rainfall_index insurance is provided)

Due to the un-observable outcome *or* not being able to observe both the outcomes at a given time, measuring τ_i is impossible.

Average Treatment effect τ^{ATE}:

ATE measures the average effect of treatment across a population of units i.e across a population of farmers $\tau^{ATE} = E[Y_i(D_i = 1)] - E[Y_i(D_i = 0)]$

Even when we consider ATE, we face the fundamental problem of not observing both the $Y_i(D_i = 1) $ and $ Y_i(D_i = 0)$ at the same time which makes calculation tau^{ATE} not possible.

We can conduct an RCT where we assign the treatment randomly to the farmers where the distribition of observables and unobservables is same among both the treated and untreated farmers. This helps us in assuming that there is no problem of selection by design.

Calculating the average outcomes(Profits) of both sets (the ones who were provided with rainfall_index insurance and the ones who were not provided with rainfall_index insurance) and subtracting them, we are determining a Naive Estimator τ_N

$\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$

where $\bar{Y}(D = 1)$ is the average outcome (Profitss) for farmers with treatment status 1 i.e providing with rainfallindex insurance and $\bar{Y}(D = 0)$ is the average (Profits) for farmers with treatement status 0 i.e not providing with rainfallindex insurance.

This brings us to the assumptions that the expectation of Y is same as (conditional expectation of Y that D_i is 1)and same as the (conditional expectation of Y given D_i is 0). We are assuming that the average of Y given $D_i = 1$ is a good counterfactual for when $D_i = 0$.

In other words, the expectation of the error term (unobservable), conditional on treatment, is zero.
i.e., D_i is exogenous

$E[Y_i(1)] = E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0]$
and
$E[Y_i(0)] = E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$

Results in

Estimated $\tau^{ATE} = Naive\ Estimator\ \tau_N = \bar{Y}(D_i = 1) - \bar{Y}(D_i = 0)$ ^Above its supposed to be Tau_hat_ATE instead of Tau_ATE. Thus through an RCT where the treatment is assigned randomly i.e the distribution of outcomes(observables and unobservables) are same for the farmers with treatment status 1 and for farmers with treatment status 0.With this, it can be assumed that there is no selection problem by design.

We can estimate the ATE by taking a difference of means of the treated group and the untreated group of farmers

Assumptions: Outcome is influenced by the treatment alone.
Full compliance, i.e for all i $R_i = D_i$

###Question 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.

HARRIS is asking to calculate the Average Treatment Effect on the Treated (ATT) τ^{ATT}.

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

$E[Y_i(1)|D_i = 1]$ is the average effect i.e profit of the treatment (providing rainfall_index during bad rainfall year) for farmers in the treatment group (actually provided with rainfall_index insurance) which is observable $E[Y_i(0)|D_i = 1]$ is the average effect of the treatment (providing rainfall_index insurance) for farmers in the control group (who actually didnt get the rainfall_index insurance), the counterfactual that is unobservable in the real world

Average Treatment Effect (ATE) is the average effect of the individual treatment of the population whereas Average Treatment Effect of the Treated (ATT) is the average of the individual treatment effects of those treated only and not the total population.

Since we have unobservable in the calculation of τ^{ATT}, it is not possible to calculate the ATT Average treatment effect of the treated (τ^{ATT}) in the real world.

###Question 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.

We are determining a Naive Estimator τ_N by comparing the two sets of outcome (average profit in a bad rainfall year), FIONA farms to non-FIONA farms

$\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$

where $\bar{Y}(D = 1)$ is the average outcome (Profits) for farmers with treatment status 1 and $\bar{Y}(D = 0)$ is the average (Profits) for farmers with treatement status 0. The Naive estimator τ_N(a sample average) is calculated based on observed outcomes where as the ATE(Average of population E[]) is calculated on potential outcomes.

Here we are observing $Y_i(D_i = 1)$ and $Y_j(D_j = 0)$, knowing that i is not equal to j.

This brings us to the assumptions that the expectation of Y is same as (conditional expectation of Y that D_i is 1)and same as the (conditional expectation of Y given D_i is 0). We are assuming that the average of Y given $D_i = 1$ is a good counterfactual for when $D_i = 0$. Below in mathematical expression form:

$E[Y_i(1)] = E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0]$
and
$E[Y_i(0)] = E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$

There can be a problem when i and j significantly differ from each other. Which is, the units that receive treatment differ a lot from the units that donot receive any treatment on observables and the unobservables. This leads to a bias called the Selection bias. This bias can be explained through the following example:

##Example 1 Consider selection problem with an unobservable characteristic of farmers Lets say that there is bad rainfall in the districts under consideration but the farmers in the districts where FIONA is not being implemented generally produce higher yield compared to the farmers in the districts where FIONA is being implemented. This will result in a selection problem while determining the Naive estimator where the unobservable characteristic will result in an underestimation of the average effect of treatment i.e insurance through FIONA.

τ_N given by $\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$

The similar case can be constructed where an overestimation of average effect of treatment through FIONA happens instead of underestimation.

##Example 2 Now lets consider a case where we take districts into account. Lets say that the districts where FIONA is implemented are the only ones where bad rainfall occured and the districts where FIONA is not implemented have rainfall in surplus. This will result in farmers from the bad rainfall districts receive insurance which may affect their outcome(profits), but the farmers outcomes(profits) from the districts where FIONA is not implemented are not affected as they are having good rainfall. Comparing these two through a naive estimator $\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$ would potentially underestimate the average effect of treatment i.e insurance through FIONA

##Example 3 Consider a case where there is non-compliance. For example, some farmers who are in the control group somehow got to know about the potential benefits of FIONA went ahead and registered for the same leads to Non compliance. Same way, farmers who are assigned treatment, but due to many reasons they did not avail the FIONA insurance. And the researcher doesnt have any knowledge of non-complianes in control group nor the treatment group. This invalidates the Naive estimator $\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$ and makes the estimate of ATE of FIONA inaccurate.

##Example 4 Parallel program case Consider that the bad rainfall year is common across all districts where FIONA is being implemented and the districts where FIONA is not being implemented. FIONA is being implemented in economically backward districts and not implemented in districts with good economy. Now consider a parallel program being run by the government providing subsidies on agricultural products for farmers in economically backward districts during this same period. Thus the farmers in economically backward districts during the bad rainfall year get insurance from FIONA and also their costs are reduced due to the government's program. And the farmers in the districts with good economy neither received insurance through FIONA, nor their costs reduced due to no presence of the government program. In this case the naiver estimator $\tau_N = \bar{Y}(D = 1) - \bar{Y}(D = 0)$ would overestimate the average effect of treatment i.e insurance through FIONA

###Question 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. 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.

```r
data <- read_csv('ps2_data.csv')
```

```
## Rows: 10000 Columns: 7

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): district, crop, farmer_birth_year
## dbl (4): fiona_farmer, fertilizer_use, profits_2005, profits_2016

##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
summary(data)
```

```
##   fiona_farmer    district             crop           farmer_birth_year
##  Min.   :0.00   Length:10000       Length:10000       Length:10000
##  1st Qu.:0.00   Class :character   Class :character   Class :character
##  Median :0.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.25
##  3rd Qu.:0.25
##  Max.   :1.00
##  fertilizer_use   profits_2005    profits_2016
##  Min.   :0.000   Min.   :15842   Min.   :16527
##  1st Qu.:0.000   1st Qu.:19721   1st Qu.:21409
##  Median :0.000   Median :20000   Median :22339
##  Mean   :0.239   Mean   :19993   Mean   :22535
##  3rd Qu.:0.000   3rd Qu.:20269   3rd Qu.:23461
##  Max.   :1.000   Max.   :24001   Max.   :29096
```

```
data$fiona_farmer <- as.factor(data$fiona_farmer)

data$crop <- as.factor(data$crop)


data %>%
  group_by(fiona_farmer) %>%
  dplyr::summarize(n=n())
```

```
## # A tibble: 2 x 2
##   fiona_farmer     n
##   <fct>        <int>
## 1 0             7500
## 2 1             2500
```

```
#7500 farmers with no FIONA i.e control group
#2500 farmers with FIONA i.e treatment group




data %>%
  group_by(fiona_farmer, crop) %>%
  dplyr::summarise(n=n())
```

```
## `summarise()` has grouped output by 'fiona_farmer'. You can override using the `.groups` argument.
```

```
## # A tibble: 7 x 3
## # Groups:   fiona_farmer [2]
##   fiona_farmer crop        n
##   <fct>        <fct>   <int>
## 1 0            LENTILS  2250
## 2 0            RICE     3000
## 3 0            WHEAT    2250
## 4 1            COTTON     53
## 5 1            LENTILS   750
## 6 1            RICE      947
## 7 1            WHEAT     750
```

```
#We can see in the groupby summary that only for the COTTON farmers, everyone is
#in the treatment group with no one in the control group.

data %>%
  group_by(fertilizer_use) %>%
  dplyr::summarise(n=n())
```

```
## # A tibble: 2 x 2
##   fertilizer_use     n
##            <dbl> <int>
## 1              0  7610
## 2              1  2390
```

```
data %>%
  group_by(fiona_farmer, crop, fertilizer_use) %>%
  filter(!is.na(crop)) %>%
  dplyr::summarise(n=n())
```

```
## ‘summarise()‘ has grouped output by ’fiona_farmer’, ’crop’. You can override using the ‘.groups‘ argu

## # A tibble: 14 x 4
## # Groups:   fiona_farmer, crop [7]
##    fiona_farmer crop    fertilizer_use     n
##    <fct>        <fct>            <dbl> <int>
##  1 0            LENTILS              0  1754
##  2 0            LENTILS              1   496
##  3 0            RICE                 0  2349
##  4 0            RICE                 1   651
##  5 0            WHEAT                0  1758
##  6 0            WHEAT                1   492
##  7 1            COTTON               0    32
##  8 1            COTTON               1    21
##  9 1            LENTILS              0   509
## 10 1            LENTILS              1   241
## 11 1            RICE                 0   672
## 12 1            RICE                 1   275
## 13 1            WHEAT                0   536
## 14 1            WHEAT                1   214
```

```
data %>%
  group_by(fiona_farmer, district) %>%
  filter(!is.na(district)) %>%
  dplyr::summarise(n=n())
```

```
## ‘summarise()‘ has grouped output by ’fiona_farmer’. You can override using the ‘.groups‘ argument.

## # A tibble: 6 x 3
## # Groups:   fiona_farmer [2]
##   fiona_farmer district        n
##   <fct>        <chr>       <int>
## 1 0            DINDIGUL     1500
## 2 0            KARUR        1500
## 3 0            MADURAI      1500
## 4 0            PUDUKKOTTAI  1500
## 5 0            TENKASI      1500
## 6 1            THANJAVUR    2500
```

7500 farmers with no FIONA i.e control group 2500 farmers with FIONA i.e treatment group We can see in the groupby summary that only for the COTTON farmers, everyone is in the treatment group with no one in the control group.

We observe that the there is only one treatment district THANJAVUR and remaining districts are in the control group. We can assume that the bad rainfall year is common across all these districts

The inference 2 can cause a problem of selection bias as none of the COTTON farmers are in the control group. So we may need to remove the COTTON data. We observe from the summary of columns that the columns crop has null values. So we need to clean that as well.

```
cleaned_data <-
  data %>%
    filter(crop != "COTTON") %>%
    filter(!is.na(crop))

summary(cleaned_data)
```

```
##  fiona_farmer   district              crop      farmer_birth_year
##  0:7500       Length:9947        COTTON :   0   Length:9947
##  1:2447       Class :character   LENTILS:3000   Class :character
##               Mode  :character   RICE   :3947   Mode  :character
##                                  WHEAT  :3000
##
##
##  fertilizer_use    profits_2005    profits_2016
##  Min.   :0.0000   Min.   :15842   Min.   :16527
##  1st Qu.:0.0000   1st Qu.:19724   1st Qu.:21407
##  Median :0.0000   Median :20000   Median :22333
##  Mean   :0.2382   Mean   :19994   Mean   :22528
##  3rd Qu.:0.0000   3rd Qu.:20268   3rd Qu.:23450
##  Max.   :1.0000   Max.   :24001   Max.   :28903
```

Farmer_birth_year is in strings. We need to convert them to numbers

```
cleaned_data$farmer_birth_year[cleaned_data$farmer_birth_year == "nineteen seventy-three"] <- 1973
cleaned_data$farmer_birth_year[cleaned_data$farmer_birth_year == "nineteen seventy-two"] <- 1972

cleaned_data$farmer_birth_year <- as.numeric(cleaned_data$farmer_birth_year)


summary(cleaned_data)
```

```
##  fiona_farmer   district              crop      farmer_birth_year
##  0:7500       Length:9947        COTTON :   0   Min.   :1916
##  1:2447       Class :character   LENTILS:3000   1st Qu.:1965
##               Mode  :character   RICE   :3947   Median :1969
##                                  WHEAT  :3000   Mean   :1969
##                                                 3rd Qu.:1973
##                                                 Max.   :1989
##  fertilizer_use    profits_2005    profits_2016
##  Min.   :0.0000   Min.   :15842   Min.   :16527
##  1st Qu.:0.0000   1st Qu.:19724   1st Qu.:21407
##  Median :0.0000   Median :20000   Median :22333
##  Mean   :0.2382   Mean   :19994   Mean   :22528
##  3rd Qu.:0.0000   3rd Qu.:20268   3rd Qu.:23450
##  Max.   :1.0000   Max.   :24001   Max.   :28903
```

With the cleaned data, we can implement SOO (Selection on observables) design using X_i assuming the outcomes i.e the profits of the farmer are independent of assignment D_i of FIONA , conditional on this covariate X_i.

The following SOO approaches can be possible here:

Regression adjustment:

We need to estimate: $Y_i = \alpha \ + \ \tau D_i + \gamma X_i + v_i$ where $E[\epsilon_i] = E[\gamma X_i + v_i] = 0$ i.e., $Y_i \perp D_i | X_i$ We get, $\hat{\tau} = \tau^{ATE}$

Replacing Y_i D_i with terms in our context and then we estimate $profit_i = \alpha \ + \ \tau(fiona_{farmer}) + \gamma X_i + v_i$ to get $\hat{\tau}$ which is closely equal to τ^{ATE} ie., ATE (Average Treatment Effect) of FIONA
Another SOO approach possible:

Matching We compare untreated units to treated units having identical X_i's. As we are comparing units having identical X_i's, the functional form is not relevant anymore. Thus the difference in outcomes will be the $\hat{\tau}$ which is closely equal to τ^{ATE} ie., ATE (Average Treatment Effect) of FIONA. Following is how we implement Matching. First divide data into unique cells categorized by covariates such that for each cell, we calculate $\bar{Y}_T$ and $\bar{Y}_U$. Then to estimate the ATE i.e $\hat{\tau}^{ATE}$, we take the difference between $\bar{Y}_T$ and $\bar{Y}_U$ for each cell as a weighted average.

###Question 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
#Determing possible observable characterisitcs in order of column names
Columns_names <- c("fertilizer_use", "profits_2005", "profits_2016",
                   "iswheat", "isrice", "islentils", "isyoung",
                   "thanjavur", "dindigul", "karur", "madurai", "pudukkottai", "tenkasi")


cleaned_data <-
  cleaned_data %>%
  mutate(iswheat = ifelse(crop=="WHEAT", 1, 0)) %>%
  mutate(isrice = ifelse(crop=="RICE", 1, 0)) %>%
  mutate(islentils = ifelse(crop=="LENTILS", 1, 0)) %>%
  mutate(isyoung = ifelse(farmer_birth_year >= 1969,1,0)) %>%
  mutate(thanjavur = ifelse(district=="THANJAVUR", 1, 0)) %>%
  mutate(dindigul = ifelse(district=="DINDIGUL", 1, 0)) %>%
  mutate(karur = ifelse(district=="KARUR", 1, 0)) %>%
  mutate(madurai = ifelse(district=="MADURAI", 1, 0)) %>%
  mutate(pudukkottai = ifelse(district=="PUDUKKOTTAI", 1, 0)) %>%
  mutate(tenkasi = ifelse(district=="TENKASI", 1, 0))



balance_table <- cleaned_data %>%
  select(all_of(Columns_names)) %>%
  lapply(., function(i) tidy(lm(i ~ cleaned_data$fiona_farmer))) %>%
  do.call(rbind, .) %>%
  rownames_to_column("variable") %>%
  filter(term == "cleaned_data$fiona_farmer1") %>%
  select(-term)
balance_table$variable <- str_remove(balance_table$variable, ".2")
knitr::kable(balance_table, digits=3, caption = "Balance Table FIONA", "latex")
```

Column by column, lets analyse the p-value and determine if the variables are balanced across the control and treatment groups

1) fertilizer use using fiona_farmer as treatement variable From the p-value for 7.74e-16, we can reject the null hypothesis that the Differences in means = 0.However we cannot reject the alternate hypothesis

Table 1: Balance Table FIONA

variable	estimate	std.error	statistic	p.value
fertilizer_use	0.080	0.010	8.072000e+00	0.000
profits005.2	17.596	17.178	1.024000e+00	0.306
profits016.2	2380.718	30.144	7.897800e+01	0.000
iswheat	0.006	0.011	6.080000e-01	0.543
isrice	-0.013	0.011	-1.141000e+00	0.254
islentils	0.006	0.011	6.080000e-01	0.543
isyoung	-0.016	0.012	-1.344000e+00	0.179
thanjavur	1.000	0.000	9.721356e+14	0.000
dindigul	-0.200	0.008	-2.473100e+01	0.000
karur	-0.200	0.008	-2.473100e+01	0.000
madurai	-0.200	0.008	-2.473100e+01	0.000
pudukkottai	-0.200	0.008	-2.473100e+01	0.000
tenkasi	-0.200	0.008	-2.473100e+01	0.000

that the Differences in means != 0. Thus we can say that farmers that are insured through FIONA use fertilizer more 0.08 at 99% signifincae level. The difference is statistically significant, hence we can say that the treatment group and control group are not balanced w.r.t variable fertilizer_use

2)profits_2005 using fiona_farmer as treatment variables From the above p-value for 0.306, we cannot reject the null hypothesis (Differences in means = 0). The differences are not statistically significant. Thus we can say that the treatment group and control group are balanced, which means that the assignment is random across farmers.

3)crop variety using fiona_farmer as treatment variable From the above p-value for 0.543, 0.254, 0.543 respectively for iswheat,isrice, islentils, we cannot reject the null hypothesis that the Differences in means = 0 . This means that differences are not statistically significant. Thus we can say that the treatment group and control group are balanced, which means that the assignment is random across farmers across the crop variety. We analyse further by determining average profits of treated farmers in 2005 and 2016 by crop variety.

```
#Crop with profits in 2005

cleaned_data %>%
  filter(fiona_farmer == 1) %>%
  group_by(crop) %>%
  summarise_at(vars(profits_2005), list(name = mean))
```

```
## # A tibble: 3 x 2
##   crop      name
##   <fct>    <dbl>
## 1 LENTILS 19985.
## 2 RICE    19985.
## 3 WHEAT   20058.
```

```
#Crops with profits in 2016
cleaned_data %>%
  filter(fiona_farmer == 1) %>%
  group_by(crop) %>%
  summarise_at(vars(profits_2016), list(name = mean))
```

```
## # A tibble: 3 x 2
##   crop      name
##   <fct>    <dbl>
## 1 LENTILS 24729.
## 2 RICE    23443.
## 3 WHEAT   25028.
```

Now lets observe how pre and post profits vary with crop across treted and untreated farmers

```
treated_data <- cleaned_data %>% filter(fiona_farmer == 1)

reg_2005_lentils <- lm(profits_2005 ~ islentils, data = treated_data)
reg_2016_lentils <- lm(profits_2016 ~ islentils, data = treated_data)

reg_2005_wheat <- lm(profits_2005 ~ iswheat, data = treated_data)
reg_2016_wheat <- lm(profits_2016 ~ iswheat, data = treated_data)

reg_2005_rice <- lm(profits_2005 ~ isrice, data = treated_data)
reg_2016_rice <- lm(profits_2016 ~ isrice, data = treated_data)


untreated_data <- cleaned_data %>% filter(fiona_farmer == 0)

reg_2005_lentils_untreated <- lm(profits_2005 ~ islentils, data = untreated_data)
reg_2016_lentils_untreated <- lm(profits_2016 ~ islentils, data = untreated_data)

reg_2005_wheat_untreated <- lm(profits_2005 ~ iswheat, data = untreated_data)
reg_2016_wheat_untreated <- lm(profits_2016 ~ iswheat, data = untreated_data)

reg_2005_rice_untreated <- lm(profits_2005 ~ isrice, data = untreated_data)
reg_2016_rice_untreated <- lm(profits_2016 ~ isrice, data = untreated_data)


stargazer(reg_2005_lentils_untreated, reg_2005_rice_untreated, reg_2005_wheat_untreated, header = FALSE
          type = "latex", title = "2005 profits cropwise - untreated farmers ",
          dep.var.labels = c("Lentils", "Rice", "Wheat"))
```

```
stargazer(reg_2016_lentils_untreated, reg_2016_rice_untreated, reg_2016_wheat_untreated, header = FALSE
          type = "latex", title = "2016 profits cropwise - untreated farmers ",
          covariate.labels = "Crop",
          dep.var.labels = c("Lentils", "Rice", "Wheat"))
```

```
stargazer(reg_2005_lentils, reg_2005_rice, reg_2005_wheat, header = FALSE,
          type = "latex", title = "Pre treatment profits cropwise - treated farmers - 2005",
          dep.var.labels = c("Lentils", "Rice", "Wheat"))
```

```
stargazer(reg_2016_lentils, reg_2016_rice, reg_2016_wheat, header = FALSE,
          type = "latex", title = "Post treatment profits cropwise - treated farmers - 2016",
          covariate.labels = "Crop",
          dep.var.labels = c("Lentils", "Rice", "Wheat"))
```

Table 2: 2005 profits cropwise - untreated farmers

	Dependent variable:		
	Lentils		
	(1)	(2)	(3)
islentils	−1.670 (15.528)		
isrice		6.158 (14.525)	
iswheat			−5.367 (15.528)
Constant	19,990.380*** (8.505)	19,987.410*** (9.186)	19,991.490*** (8.505)
Observations	7,500	7,500	7,500
R^2	0.00000	0.00002	0.00002
Adjusted R^2	−0.0001	−0.0001	−0.0001
Residual Std. Error (df = 7498)	616.254	616.247	616.250
F Statistic (df = 1; 7498)	0.012	0.180	0.119
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 3: 2016 profits cropwise - untreated farmers

	Dependent variable:		
	Lentils		
	(1)	(2)	(3)
Crop	2.641 (29.944)		
isrice		−22.330 (28.009)	
iswheat			22.880 (29.943)
Constant	21,941.550*** (16.401)	21,951.280*** (17.715)	21,935.480*** (16.401)
Observations	7,500	7,500	7,500
R^2	0.00000	0.0001	0.0001
Adjusted R^2	−0.0001	−0.00005	−0.0001
Residual Std. Error (df = 7498)	1,188.382	1,188.332	1,188.336
F Statistic (df = 1; 7498)	0.008	0.636	0.584
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 4: Pre treatment profits cropwise - treated farmers - 2005

	Dependent variable:		
	Lentils		
	(1)	(2)	(3)
islentils	-32.359 (44.920)		
isrice		-36.216 (42.518)	
iswheat			72.779 (44.901)
Constant	20,017.390*** (24.869)	20,021.490*** (26.450)	19,985.170*** (24.858)
Observations	2,447	2,447	2,447
R^2	0.0002	0.0003	0.001
Adjusted R^2	-0.0002	-0.0001	0.001
Residual Std. Error (df = 2445)	1,024.466	1,024.423	1,024.025
F Statistic (df = 1; 2445)	0.519	0.726	2.627
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 5: Post treatment profits cropwise - treated farmers - 2016

	Dependent variable:		
	Lentils		
	(1)	(2)	(3)
Crop	585.655*** (68.146)		
isrice		$-1,434.998^{***}$ (58.688)	
iswheat			1,015.929*** (66.045)
Constant	24,143.560*** (37.727)	24,878.410*** (36.510)	24,011.680*** (36.564)
Observations	2,447	2,447	2,447
R^2	0.029	0.196	0.088
Adjusted R^2	0.029	0.196	0.088
Residual Std. Error (df = 2445)	1,554.148	1,414.011	1,506.247
F Statistic (df = 1; 2445)	73.860***	597.867***	236.616***
Note:			*p<0.1; **p<0.05; ***p<0.01

First we saw the relation between treated farmers and fertilizer use, a post treatment variable (thus endogenous). We observed that the farmers who were treated or insured under FIONA use more fertilizer frequently. THis factor mightve affected the profits for these treated farmers.

Also from the above tables, we see that for all the crops, difference in means of profits for treated farmers is close to 0 in 2005. However this is not the case for profits in 2016 post treatment. In 2016 i.e post treatment, the difference in means of profits for treated farmers across all crops is not 0 in 2016. Thus we can doubt that selection of crop by the farmer has some effect on the post treatment profits in 2016.

This table makes it worse about the earlier concerns in (3).

###Question 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. 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.

As discussed above, two assumptions

1) Common Support: For all the possible covariate X's, we should be able to observe both the treated and untreated as we have a significantly large sample. As we can observe both the treated and untreated, treatment effects can be estimated.

$0 < Pr(D_i|X = x^0) < 1, \forall x^0$

To check for validity, lets analyse the C.S assumption across covariates

```
cleaned_data %>%
  group_by(fiona_farmer, crop) %>%
  summarise(n=n())
```

```
## ‘summarise()‘ has grouped output by ’fiona_farmer’. You can override using the ‘.groups‘ argument.
```

```
## # A tibble: 6 x 3
## # Groups:   fiona_farmer [2]
##   fiona_farmer crop        n
##   <fct>        <fct>   <int>
## 1 0            LENTILS  2250
## 2 0            RICE     3000
## 3 0            WHEAT    2250
## 4 1            LENTILS   750
## 5 1            RICE      947
## 6 1            WHEAT     750
```

The assumption holds for when crop as covariate

```
cleaned_data %>%
  group_by(fiona_farmer, district) %>%
  summarise(n=n())
```

```
## ‘summarise()‘ has grouped output by ’fiona_farmer’. You can override using the ‘.groups‘ argument.
```

```
## # A tibble: 6 x 3
## # Groups:   fiona_farmer [2]
##   fiona_farmer district        n
##   <fct>        <chr>       <int>
## 1 0            DINDIGUL     1500
## 2 0            KARUR        1500
## 3 0            MADURAI      1500
## 4 0            PUDUKKOTTAI  1500
## 5 0            TENKASI      1500
## 6 1            THANJAVUR    2447
```

Similarly, the assumptions holds for district

```
cleaned_data %>%
  group_by(fiona_farmer, isyoung) %>%
  summarise(n=n())
```

```
## `summarise()` has grouped output by 'fiona_farmer'. You can override using the `.groups` argument.
```

```
## # A tibble: 4 x 3
## # Groups:   fiona_farmer [2]
##   fiona_farmer isyoung     n
##   <fct>          <dbl> <int>
## 1 0                  0  3472
## 2 0                  1  4028
## 3 1                  0  1171
## 4 1                  1  1276
```

Similarly, the assumption holds for isyoung

2) Conditional Independence: When conditioned on the X_i's, the potential outcomes of a unit are orthogonal to the treatment. In this context, when a given X_i i.e crop or district or isyoung , potential profits of a farmer are orthogonal (or) are independed of treatement with FIONA. This assumptions gives a safe base for comparision of estimates with other units. As we work with potential outcomes and not observed, we cannot check for validity of this assumption.

$(Y_{1,i}, Y_{0,i}) \perp D_i | X$

ATE (Average Treatment Effect) τ^{SOO}:

$\tau^{SOO} = E[Y_i(1)|X_i = x] \ - \ E[Y_i(0)|X_i = x]$

As said above, we cannot check for validity of the conditional independence assumption.

The following SOO approaches can be possible here:

Regression adjustment:

We are estimating: $Y_i = \alpha \ + \ \tau D_i + \gamma X_i + v_i$

where $E[\epsilon_i] = E[\gamma X_i + v_i] = 0$ i.e., $Y_i \perp D_i | X_i$

We get, $\hat{\tau} = \tau^{ATE}$

Thus we need to estimate $profit_i = \alpha \ + \ \tau(fiona_{farmer}) + \gamma X_i + v_i$

to get τ^{ATE} i.e ATE (Average Treatment Effect) of FIONA from $\hat{\tau}$ which is closely equal to τ^{ATE}

As we observe a good overlap between X_i for control($\bar{X}_u$) and treatment($\bar{X}_t$) for crop and isyoung, we can say that the assumption holds good for these variables. Also as we observed $D_i = E[D_i|X_i]$, this assumption also holds from the balance tables produced above.

From the above analysis, we can say we can use Regression adjustment approach as credible estimate for τ^{ATE} average treatment effect of FIONA on farmer's profit.

Matching

We compare untreated units to treated units having identical X_i's. As we are comparing units having identical X_i's, the functional form is not relevant anymore. Thus the difference in outcomes will be the $\hat{\tau}$ which is closely equal to τ^{ATE} ie., ATE (Average Treatment Effect) of FIONA. We perform the same as below:

First divide the data into cells uniquely as defined by the covariates and for each cell, determine $\bar{Y}_t$ and $\bar{Y}_u$ i.e for treated and untreated respectively. Now to determine $\hat{\tau}^{ATE}$, take the weighted average difference for each cell i.e $\bar{Y}_T$ - $\bar{Y}_U$

```
cleaned_data %>%
  group_by(fiona_farmer, crop) %>%
  filter(!is.na(crop)) %>%
  dplyr::summarise(n=n())
```

```
## ‘summarise()‘ has grouped output by ’fiona_farmer’. You can override using the ‘.groups‘ argument.
```

```
## # A tibble: 6 x 3
## # Groups:   fiona_farmer [2]
##   fiona_farmer crop        n
##   <fct>        <fct>   <int>
## 1 0            LENTILS  2250
## 2 0            RICE     3000
## 3 0            WHEAT    2250
## 4 1            LENTILS   750
## 5 1            RICE      947
## 6 1            WHEAT     750
```

```
cleaned_data %>%
  group_by(fiona_farmer, district) %>%
  filter(!is.na(crop)) %>%
  dplyr::summarise(n=n())
```

```
## ‘summarise()‘ has grouped output by ’fiona_farmer’. You can override using the ‘.groups‘ argument.
```

```
## # A tibble: 6 x 3
## # Groups:   fiona_farmer [2]
##   fiona_farmer district        n
##   <fct>        <chr>       <int>
## 1 0            DINDIGUL     1500
## 2 0            KARUR        1500
## 3 0            MADURAI      1500
## 4 0            PUDUKKOTTAI  1500
## 5 0            TENKASI      1500
## 6 1            THANJAVUR    2447
```

```
cleaned_data %>%
  group_by(fiona_farmer, isyoung) %>%
  filter(!is.na(crop)) %>%
  dplyr::summarise(n=n())
```

```
## `summarise()` has grouped output by 'fiona_farmer'. You can override using the `.groups` argument.
```

```
## # A tibble: 4 x 3
## # Groups:   fiona_farmer [2]
##   fiona_farmer isyoung     n
##   <fct>          <dbl> <int>
## 1 0                  0  3472
## 2 0                  1  4028
## 3 1                  0  1171
## 4 1                  1  1276
```

From the above analysis, we can say we can use Exact matching approach as credible estimate for τ^{ATE} average treatment effect of FIONA on farmer's profit.

###Question 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?

1) district - We observed earlier that the only treatment district is THANJAVUR and remaining(DINDIGUL, KARUR, MADURAI, PUDUKOTTAI, TENKASI) are in the control group. Thus the variable district cannot be used as a covariate, as the treatment at district level causes imbalance. This is true assuming that these above districts are similar in all other conditions and have similar bad rainfall year.
2) isyoung or farmer_birth_year: We have derived isyoung variable from farmer_birth_year, a continuous variable and we have observed that the year 1957 posed an imbalance in treatment and control group, but whereas when we mutated the data to filter with 1969 birth year, it has resulted in a balance across treatment and control group. We can safely reject any role of isyounf, as a result farmer_birth_year in the regression by assuming that age may not impact the outcomes of profits i.e age cannot significantly impact the τ^{ATE}. Thus isyoung cannot be used as a covariate.
3) profits_2005: We observed earlier that treatment group and the control group are balanced with the pre-treatment variable profits_2005. Hence it is not used as a covariate in the regression
4) fertilizer_use: We have observed that the fertilizer_use may have effect on the profits of farmers. THis is because we saw farmers who are insured under FIONA used fertilizer more frequently than the farmers who are not insured under FIONA. The use of fertilizer can have caused increase in profits for the farmers by affecting the yield. Also the variable fertilizer_use is endogenous.
5) crop: As we have observed in the balance tables, variety of crop(lentils, rice, wheat) also has effect on the farmer profits in 2016. Thus this variable can be used as a covariate in the regression.

Lets run a few regressions to determine the statistical significance of the variables in determining the profits

Regression Test 1 with crop, profits_2005 and isyoung variables

```
reg_test1 <- lm(profits_2016 ~ fiona_farmer + iswheat + isrice + profits_2005 + isyoung , data = cleane
summary(reg_test1)
```

```
##
## Call:
## lm(formula = profits_2016 ~ fiona_farmer + iswheat + isrice +
##     profits_2005 + isyoung, data = cleaned_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3792.8  -703.1     5.0   725.7  3844.1
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1974.56775  285.08331   6.926 4.59e-12 ***
## fiona_farmer1 2358.21140   24.35784  96.815  < 2e-16 ***
## iswheat         68.76216   27.01118   2.546   0.0109 *
## isrice        -328.24291   25.33855 -12.954  < 2e-16 ***
## profits_2005     1.00463    0.01422  70.664  < 2e-16 ***
## isyoung         -7.44653   21.02622  -0.354   0.7232
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1046 on 9941 degrees of freedom
## Multiple R-squared:  0.5991, Adjusted R-squared:  0.5989
## F-statistic:  2971 on 5 and 9941 DF,  p-value: < 2.2e-16
```

Summary:

profits_2005: the p-value of 2e-16 at 100% confidence level indicates the statistical significance of the variable in determining profits. We also observe the difference in means of 1.00463 that is not significant. Thus we can say that both the groups are balanced around this variable and thus the variable does not have significant impact on τ^{ATE}

crop variety iswheat: the p-value of 0.0109 at 99% confidence level for iswheat, shows that the groups are balanced around this variable and thus it can be a valid covariate isrice: the p-value of 2e-16 at 100% level for isrice shows that the groups are balanced around this variable and thus it can be a valid covariate

islentils: As we include iswheat and isrice in the regression, we would not require islentils

isyoung: the p-value 0.7232 says that the groups are balanced with this variable and the difference in means shows that its not statistically significant to impact the τ^{ATE}

We are estimating: $Y_i = \alpha \ + \ \tau D_i + \gamma X_i + v_i$

where $E[\epsilon_i] = E[\gamma X_i + v_i] = 0$ i.e., $Y_i \perp D_i | X_i$

We get, $\hat{\tau} = \tau^{ATE}$

Replacing the variable and outcomes relevant to our context, we estimate $profit_i = \alpha \ + \ \tau(fiona_{farmer}) + \gamma X_i + v_i$

to get τ^{ATE} i.e ATE (Average Treatment Effect) of FIONA from $\hat{\tau}$ which is closely equal to τ^{ATE}

Using only iswheat and isrice as the variables in regression

```r
SOO_reg_crop <- lm(profits_2016 ~ fiona_farmer + iswheat + isrice , data = cleaned_data)
summary(SOO_reg_crop)
```

```
##
## Call:
## lm(formula = profits_2016 ~ fiona_farmer + iswheat + isrice,
##     data = cleaned_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5224.7  -836.8     3.2   841.5  4953.2
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    22046.46      24.57 897.404   <2e-16 ***
## fiona_farmer1  2375.95      29.85  79.601   <2e-16 ***
## iswheat           85.22      33.10   2.575     0.01 *
## isrice          -324.21      31.05 -10.440   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1282 on 9943 degrees of freedom
## Multiple R-squared:  0.3977, Adjusted R-squared:  0.3975
## F-statistic:  2188 on 3 and 9943 DF,  p-value: < 2.2e-16
```

The p-value of 2e-16 at 100% confidence level shows tha statistical significance. Thus the stat that on an average, farmers with FIONA make a profit of 2375.95 INR more than the farmers without FIONA can be considered significant. Thus the ATE ($\hat{\tau}_{ATE}$ of FIONA on the profits is 2375.95 INR. As far as for iswheat, the p-value of 0.01 at 99% confidence level and the difference in means of 85.22 says that it can be statistically significant but maynot be economically significant when compared to average farmer profit in 2016. Similar to iswheat is the case with isrice where the p-value of 2e-16 at 100% confidence level and difference in means of -324.21 is statistically significant but may not be economically significant as compared to average farmer profit in 2016. Hence we can conclude that the crop variety is not effecting the profits to vary significantly in 2016.

Strenghts and Weaknessess of Regression method:

Constant Treatment Effects: As the outcomes are linear in X_i, the $\hat{\tau}_i$ gives us an unbiased yet consistent estimate of ATE The $\hat{\tau}_i$ will give a linear approximation to the average causal response $E[Y|D=1,X_i] - E[Y|D=1,X_i]$. Approximation in this case can be inaccurate and results in a biased $\hat{\tau}_i$ for the ATE. Heterogenous Treatment Effects: In case the outcomes are linear in X_i and $\hat{\tau}_i$ is different for different value of X, $\hat{\tau}_i$ results in an unbiased and consistent estimator for the conditional variance weighted average of the causal effects which is not same as ATE.

Estimating using Naive Estimator: Estimated $\hat{\tau}^ATE = Naive\ Estimator\ \tau_N = \bar{Y}(D_i = 1) - \bar{Y}(D_i = 0)$

```
naive_reg <- lm(profits_2016 ~ fiona_farmer, data = cleaned_data)

summary(naive_reg)
```

```
##
## Call:
## lm(formula = profits_2016 ~ fiona_farmer, data = cleaned_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5449.5  -842.8     4.7   838.7  5142.6
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    21942.34      14.95 1467.60   <2e-16 ***
## fiona_farmer1  2380.72      30.14   78.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1295 on 9945 degrees of freedom
## Multiple R-squared:  0.3854, Adjusted R-squared:  0.3854
## F-statistic:  6237 on 1 and 9945 DF,  p-value: < 2.2e-16
```

We observe from the naive estimator results that FIONA farmers make an average profit of 2380.72 INR more than farmers without FIONA. From the p-value 2e-16, we can say that the result is statistically significant. Thus the estimated ATE $\hat{\tau}_{ATE} = Naive\ Estimator\ \tau_N = \bar{Y}(D_i = 1) - \bar{Y}(D_i = 0)$ of FIONA on farmer profits in 2016 is 2380.72 INR. Earlier we saw that the Difference in means in farmer profits in 2016 between the crop varieties wheat and rice are statistically significant. However they are not economically significant in effecting the farmer profits in 2016. This being one of the reasons, we see that the $\hat{\tau}_{ATE}$ estimated in SOO approach is nearly equal to the estimated $\hat{\tau}_{ATE}$ using the Naive Estimator.

###Question 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.

```r
library(MatchIt)
```

```
## Warning: package 'MatchIt' was built under R version 4.1.3
```

```r
set.seed(9999)

covariate_data <- c('profits_2016', 'iswheat', 'isrice', 'islentils')

# First, we check the covariate and outcome means in the two groups
cleaned_data %>%
  group_by(fiona_farmer) %>%
  select(one_of(covariate_data)) %>%
  summarise_all(funs(mean(., na.rm = T)))
```

```
## Adding missing grouping variables: ‘fiona_farmer‘
```

```
## Warning: ‘funs()‘ was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with ‘tibble::lst()‘:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
```

```
## # A tibble: 2 x 5
##   fiona_farmer profits_2016 iswheat isrice islentils
##   <fct>               <dbl>   <dbl>  <dbl>     <dbl>
## 1 0                  21942.   0.3    0.4       0.3
## 2 1                  24323.   0.306  0.387     0.306
```

```r
cleaned_data_exact_match <- cleaned_data %>%
  select(profits_2016, fiona_farmer, one_of(covariate_data)) %>% na.omit()
```

```
mod_match_1 <- matchit(fiona_farmer ~  iswheat + isrice + islentils,
method = "exact",
estimand = "ATE",
data = cleaned_data_exact_match)

cleaned_data_exact_match_2 <- match.data(mod_match_1)

cleaned_data_exact_match_2 %>%
  group_by(fiona_farmer) %>%
  select(one_of(covariate_data)) %>%
  summarise_all(funs(mean(., na.rm = F)))
```

```
## Adding missing grouping variables: ‘fiona_farmer‘
```

```
## # A tibble: 2 x 5
##   fiona_farmer profits_2016 iswheat isrice islentils
##   <fct>               <dbl>   <dbl>  <dbl>     <dbl>
## 1 0                  21942.   0.3     0.4       0.3
## 2 1                  24323.   0.306   0.387     0.306
```

```
cleaned_data_exact_match_3 <- cleaned_data %>%
  select(profits_2016, fiona_farmer, one_of(covariate_data)) %>% na.omit()

mod_match_2 <- matchit(fiona_farmer ~  iswheat + isrice + islentils,
method = "exact",
estimand = "ATT",
data = cleaned_data_exact_match_3)

cleaned_data_exact_match_4 <- match.data(mod_match_2)

cleaned_data_exact_match_4 %>%
  group_by(fiona_farmer) %>%
  select(one_of(covariate_data)) %>%
  summarise_all(funs(mean(., na.rm = F)))
```

```
## Adding missing grouping variables: ‘fiona_farmer‘
```

```
## # A tibble: 2 x 5
##   fiona_farmer profits_2016 iswheat isrice islentils
##   <fct>               <dbl>   <dbl>  <dbl>     <dbl>
## 1 0                  21942.   0.3     0.4       0.3
## 2 1                  24323.   0.306   0.387     0.306
```

```
match_reg_iswheat = lm(iswheat~fiona_farmer, data = cleaned_data_exact_match_2)


summary(match_reg_iswheat)
```

```
##
## Call:
```

```
## lm(formula = iswheat ~ fiona_farmer, data = cleaned_data_exact_match_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3065 -0.3000 -0.3000  0.6935  0.7000
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.300000   0.005300  56.604   <2e-16 ***
## fiona_farmer1 0.006498   0.010686   0.608    0.543
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.459 on 9945 degrees of freedom
## Multiple R-squared:  3.718e-05,  Adjusted R-squared:  -6.337e-05
## F-statistic: 0.3698 on 1 and 9945 DF,  p-value: 0.5431
```

```
match_reg_isrice = lm(isrice~fiona_farmer, data = cleaned_data_exact_match_2)


summary(match_reg_isrice)
```

```
##
## Call:
## lm(formula = isrice ~ fiona_farmer, data = cleaned_data_exact_match_2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.400 -0.400 -0.387  0.600  0.613
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.400000   0.005649  70.804   <2e-16 ***
## fiona_farmer1 -0.012996   0.011390  -1.141    0.254
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4893 on 9945 degrees of freedom
## Multiple R-squared:  0.0001309,  Adjusted R-squared:  3.034e-05
## F-statistic: 1.302 on 1 and 9945 DF,  p-value: 0.2539
```

```
match_reg_islentils = lm(islentils~fiona_farmer, data = cleaned_data_exact_match_2)


summary(match_reg_islentils)
```

```
##
## Call:
## lm(formula = islentils ~ fiona_farmer, data = cleaned_data_exact_match_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3065 -0.3000 -0.3000  0.6935  0.7000
```

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.300000   0.005300  56.604   <2e-16 ***
## fiona_farmer1 0.006498   0.010686   0.608    0.543
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.459 on 9945 degrees of freedom
## Multiple R-squared:  3.718e-05,  Adjusted R-squared:  -6.337e-05
## F-statistic: 0.3698 on 1 and 9945 DF,  p-value: 0.5431
```

```
match_reg_iswheat_2 = lm(iswheat~fiona_farmer, data = cleaned_data_exact_match_4)


summary(match_reg_iswheat_2)
```

```
##
## Call:
## lm(formula = iswheat ~ fiona_farmer, data = cleaned_data_exact_match_4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3065 -0.3000 -0.3000  0.6935  0.7000
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.300000   0.005300  56.604   <2e-16 ***
## fiona_farmer1 0.006498   0.010686   0.608    0.543
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.459 on 9945 degrees of freedom
## Multiple R-squared:  3.718e-05,  Adjusted R-squared:  -6.337e-05
## F-statistic: 0.3698 on 1 and 9945 DF,  p-value: 0.5431
```

```
match_reg_isrice_2 = lm(isrice~fiona_farmer, data = cleaned_data_exact_match_4)


summary(match_reg_isrice_2)
```

```
##
## Call:
## lm(formula = isrice ~ fiona_farmer, data = cleaned_data_exact_match_4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.400 -0.400 -0.387  0.600  0.613
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.400000   0.005649  70.804   <2e-16 ***
## fiona_farmer1 -0.012996   0.011390  -1.141    0.254
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4893 on 9945 degrees of freedom
## Multiple R-squared:  0.0001309,  Adjusted R-squared:  3.034e-05
## F-statistic: 1.302 on 1 and 9945 DF,  p-value: 0.2539
```

```
match_reg_islentils_2 = lm(islentils~fiona_farmer, data = cleaned_data_exact_match_4)


summary(match_reg_islentils_2)
```

```
##
## Call:
## lm(formula = islentils ~ fiona_farmer, data = cleaned_data_exact_match_4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3065 -0.3000 -0.3000  0.6935  0.7000
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.300000   0.005300  56.604   <2e-16 ***
## fiona_farmer1 0.006498   0.010686   0.608    0.543
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.459 on 9945 degrees of freedom
## Multiple R-squared:  3.718e-05,  Adjusted R-squared:  -6.337e-05
## F-statistic: 0.3698 on 1 and 9945 DF,  p-value: 0.5431
```

p-value for islentils = 0.5431 p-value for isrice = 0.2539 p-value for iswheat = 0.5431

The p-values shows that the mathching worked well for all the covariates as expected

```
match_reg_profits_2016 <- lm(profits_2016 ~ fiona_farmer + iswheat +
                                  isrice + islentils, data = cleaned_data_exact_match_2,
                                weights = weights)
summary(match_reg_profits_2016)
```

```
##
## Call:
## lm(formula = profits_2016 ~ fiona_farmer + iswheat + isrice +
##     islentils, data = cleaned_data_exact_match_2, weights = weights)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -5278.0  -834.3     4.9   843.9  4966.9
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   22047.13      24.53  898.71   <2e-16 ***
## fiona_farmer1  2366.59      29.85   79.29   <2e-16 ***
## iswheat           84.09      33.10    2.54   0.0111 *
## isrice          -327.81      31.05  -10.56   <2e-16 ***
## islentils            NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1282 on 9943 degrees of freedom
## Multiple R-squared:  0.395,  Adjusted R-squared:  0.3948
## F-statistic:  2164 on 3 and 9943 DF,  p-value: < 2.2e-16
```

```
#The above summary would give us ATE
```

```
match_reg_profits_2016_2 <- lm(profits_2016 ~ fiona_farmer + iswheat +
                                  isrice + islentils, data = cleaned_data_exact_match_4,
                                weights = weights)
summary(match_reg_profits_2016_2)
```

```
##
## Call:
## lm(formula = profits_2016 ~ fiona_farmer + iswheat + isrice +
##     islentils, data = cleaned_data_exact_match_4, weights = weights)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -5222.8  -833.7     6.0   842.9  5010.5
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   22043.73      24.34 905.539   <2e-16 ***
## fiona_farmer1  2380.43      29.84  79.784   <2e-16 ***
## iswheat          84.09      32.82   2.562   0.0104 *
## isrice         -327.81      31.07 -10.550   <2e-16 ***
## islentils           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1282 on 9943 degrees of freedom
## Multiple R-squared:  0.3978, Adjusted R-squared:  0.3976
## F-statistic:  2190 on 3 and 9943 DF,  p-value: < 2.2e-16
```

```
#The above summary would give us ATT
```

The estimated ATE $\hat{\tau}_{ATE}$ of farmer profits by FIONA in 2016 using the exact matching method is 2366.59 INR Similarly, the estimated ATT $\hat{\tau}_{ATT}$ using the exact matching method is 2380.43 INR. We observed fairly similar values using the Regression method and the Naive estimator, 2375.95 INR and 2380.72 INR respectively. We did not fall into the problem of Curse of Dimensianality. This is due to the reason that for each combination of crop variety i.e iswheat, islentils, isrice and the fiona_farmer, we had enough samples to estimate ATE and ATT for each cell. If there are more covariates, that would add up to the dimensionality which would make the available data sparse by increasing the volume of space, thus making it very difficult to find an exact match. In our case, the p-value of iswheat 0.01 and p-value of isrice 2e-16 says that the Difference in means with profits due to other crops is significant meaning that the crop selection has some effect on the profits in 2016. While its statistically significant, it may not be economically significant. Not economically significant becayse the profits difference is 84.09 for iswheat and -327.81 for isrice, but the mean profits in 2016 actually are approximately 22528 INR.

###Question 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.

In India, we observed in the data that farmers insured with FIONA tend to use fertilizer frequently that improved profits. This means that FIONA program in India has enabled insured farmers engage in direct profitable inputs. Also, we saw that the variety of crop selected by the farmer, whether the farmer is insured under FIONA or not, while it proved to statistically significant is not economically significant to impact the profits in 2016. Assuming that Bangladesh is similar to India in the characteristics(observables and unobservables), we can say that farmers in Bangladesh would also benefit from FIONA-like program. Thus HARRIS should implement FIONA-like program in Bangladesh.