Pset 2 Program Eval

Andrew Caughey

1. The ideal experiment would be a Randomized Control Trial (RCT), in which individual Bangladeshi farmers would be randomly assigned rain-based insurance or not. In a perfect world, our treatment group would be mandated to comply and accept this insurance, while the control group would be barred from receiving this insurance. HARRIS should also consider if spillover effects are endogenous or exogenous to treatment. In other words, because insured farmers market transactions might effect the prices of other insured farmers and even uninsured farmers, in a perfect world, Harris may want to either 1. estimate these effects and incorporate them into the policy rollout or 2. design an RCT where non-adjacent countiess, not individuals, are assigned to treatment to minimized spillover effects. Still, totally elimiating spillover effects may not be possible, and indeed HARRIS may prefer to actually estimate these and incorporate them into their planning.

We would like a data-set with pre-treatment demographic and financial characteristics for all individuals (to make sure randomization was done correctly), a treatment indicator, and data on post-treatment outcomes for each individual, including the price they sold specific crops at.

Even in a perfect world, the causal impact of treatment for specific individuals cannot be observed because we don't see how they would respond if they had been assigned a different treatment group, but we *can* still estimate the average treatment effect (ATE) of offering this insurance to farmers as a group.

In math, if we let Y be the outcome for an indivdual farmer i, and D be an indicator where 1 means an indiivudal was assigned to recieve treatment and 0 otherwise, we'd like to estimate the causal impact of treatment with

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

However, if if indivdual i recieved treatment $Y_i(1)$, that means we don't see the universe in which he did not recieve treatment $Y_i(0)$. With randomization, though, we can ensure that, in expectation, the only different on the whole between our treatment and control groups is their assignment to treatment; in this case, if we randomize a large sample of farmers, the only difference between the farmers in the two groups, on average, would be if they were assigned to recieve HARRIS's insurance or not. In this case, even though we only see

$$E[Y - i(1)|D_i = 1] \text{ and } E[Y - i()|D_i = 0]$$

we can exploit that these two are equal to each other and estimate our average treatment effect with some really easy math, which just compares the average outcome in both groups

$$\tau^{ATE} = \bar{Y}(1) - \bar{Y}(0)$$

2. If not every farmer who is offered treatment will participate, and Harris wants to know the impact on people who actually took up insurance, using the strategy above probably won't work.

Instead, we can try and estimate the Local Average Treatment Effect, which is equal to the Average Treatment Effects on the Treated (ATT). Assuming there are no defiers or selection into treatment, who simply do the opposite of whatever experimenters tell them too, we can recover $\tau^T = \tau^{ATT} = \tau^{LATE}$,

$$\hat{\tau}^T = \frac{\bar{Y}(R=1) - \bar{Y}(R=0)}{P_{R_i=1}^{D_i=1} - P_{R_i=0}^{D_i=1}}$$

here, the denominator looks awful but is just the fraction of the treatment group units who actually received treatment minus those who were assiged assigned to control but were actually treated, and the numerator is the mean outcome of those assigned to treatment minus the mean outcome of those assigned to control.

If there was no selection into treatment, this is the same as the Average Treatment Effect we esimated above. However, this is pretty unlikely in the real world. In other words, because some farmers are more likeley to opt our of treatment than others based on their underlying characteristics, the ATT we've estimated above isn't the same as the ATE, because the ATT still includes some selection bias.

Here, we're still fundmanetally unable to observe the missing counterfacutal we introduced in the potential outcomes framework: what would the outcome be for farmers who didn't take up the treatment if they had, what would the outcome be for those who took up treatment if they hadn't, and what would be the outcome for the control if they had been assigned to treatment? We also are assuming, pretty reasonably, there are no defiers, but we don't observe the reasons for selection so we can't know this for sure and its unobserved.

Depending on the data we gather, we may also not be able to observe the share of the control group that gets this kind of insurance outside of the program. We are also assuming, for now, that one units treatment status doesn't impact the other units treatment status, but for now this is unobserved.

3. If we just compared the average outcome to farmers who recieved insurance and those who don't, we might naively think we're getting the Average Treatment Effect. Sadly, this almost certainly isn't the case: our estimate is likely heavily distorted by selection bias. This is because we don't really know that the two groups of farmers, treated and untreated, are similar and good counterfactuals for each other on observable or unobservable characteristics. In math, if we let τ be a treatment effect, this approach gives us

$$\tau + E[\epsilon_i|D_i = 1] - E[\epsilon_i|D_i = 0]$$

In other words, we get the treatment effect... plus a bunch of selection bias, and they are never to be cleft in twain.

In this example, this could concretely manifest in a few different ways. Say that farmers in certain locals have different levels of rainfall that make this insurance more attrative - then, wehen these farmers opt-in to treatment, they look meaningfully different from the control group. Perhaps farmers in rainer counties have higher crop yields than the contorl group anyway, and are also more likeley to enroll in treatment. Or, say that farmers who opt-in tend to be older and do so, perhaps, because they remember previous droughts. Suppose that these older farmers are also less productive because of their age. Further, while rainfall-index insurance tries to avoid adverse selection issues, where people with low-yields are more likeley to select in, we shouldn't just assume that this isn't a problem. In each of these three examples, there is potentially another difference between the control and treatment groups (besides assignment to treatment) that could have a causal impact on the outcome. If we just use the naive estimator, we won't be able to know what the impact of the treeatment was and what the impact of the underlying diffferences between the two groups was.

4. HARRIS can still get an estimate of the programs impact.

Since we can't pull off an RCT, Harris might consider a selection on observables design, but should be cautious since this makes some strong assumptions. In these designs, we assume that the observable differences between control and treatment groups capture everything that makes the two groups different from each other. In this case, we're assuming that farmers who received the insurance are only different from those who did not because of their assignment to the treatment or control groups *and* their age, the district they live in, the group they grow, and the profits they made in 2005. That's probably not a reasonable assumption in the real world, but if it is true, once we can mathematically account for these differences the two groups will be similar to each other and a good counterfactural for each other. This allows us to make causal inferences about the impact reciving insurance has on their profits.

Regression adjustment proposes that farmers who recieved insurance and those who did not are similar, but only *conditional* on the observable traits that make them different. Building on this, we can estimate the impact of treatmetn as:

$$\hat{\tau} = (\bar{Y_T} - \bar{Y_U}) - \hat{\gamma}(\bar{X_T} - \bar{X_U})$$

While there's some fancy letters in here, the equation obove boils down to this: our estimate of the effect of treatment comes from the difference in average outcomes among treatment and control groups, but only once we remove the observed, covaried differrences between these groups. In this way, the second term here is trying to control away the selection bias that scuttled our first-attempt at the naive estimator. In this way, once we account for the differences in farmers ages, crops, districts, and profits, the control and treatment groups can be thought of as good counterfacutals for each other, so the difference between the two is the impact that receiving this insurance has on their profits.

$$\hat{\tau}^{ATT} = \frac{1}{N_T} \sum_{i \in T} (Y_i(1) - \hat{Y}_i(0))$$

There are a couple different flavors of matching, and the approach is slightly different depending on the perameter of interest, but the equation above captures the basic idea. For each treatmented individual, we estimate an exact or nearest match in the control group based on both individuals observed characteristics. By pairing individuals who have exactly or nearly the same traits, we hope we've found a good counterfactual that lets us recover the causal impact of treatment with little to no selection bias. Here, we're hoping to find exact or near matches for farmers who recieved insurance and those who did not: we'd like to compare farmers who have the same pre-treatment profits, similar ages, grow the same crops, and live in the same district. The difference in their profits after one received insurance, then, should be the causal impact of receiving treatment.

```
#let's inspect data and look for outliers

#issues: 1973 and 1972 spelled out, need to be cleaned
harris_data$farmer_birth_year<-gsub("nineteen seventy-two", 1972, harris_data$farmer_birth_year)
harris_data$farmer_birth_year<-gsub("nineteen seventy-three", 1973, harris_data$farmer_birth_year)
harris_data$farmer_birth_year<-as.numeric(harris_data$farmer_birth_year)

#other formatting checks: seems normal + consistent format
unique(harris_data$fiona_farmer)
```

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

```
unique(harris_data$district)
```

```
## [1] KARUR       TENKASI     MADURAI     PUDUKKOTTAI THANJAVUR   DINDIGUL
## Levels: DINDIGUL KARUR MADURAI PUDUKKOTTAI TENKASI THANJAVUR
```

```
unique(harris_data$crop)
```

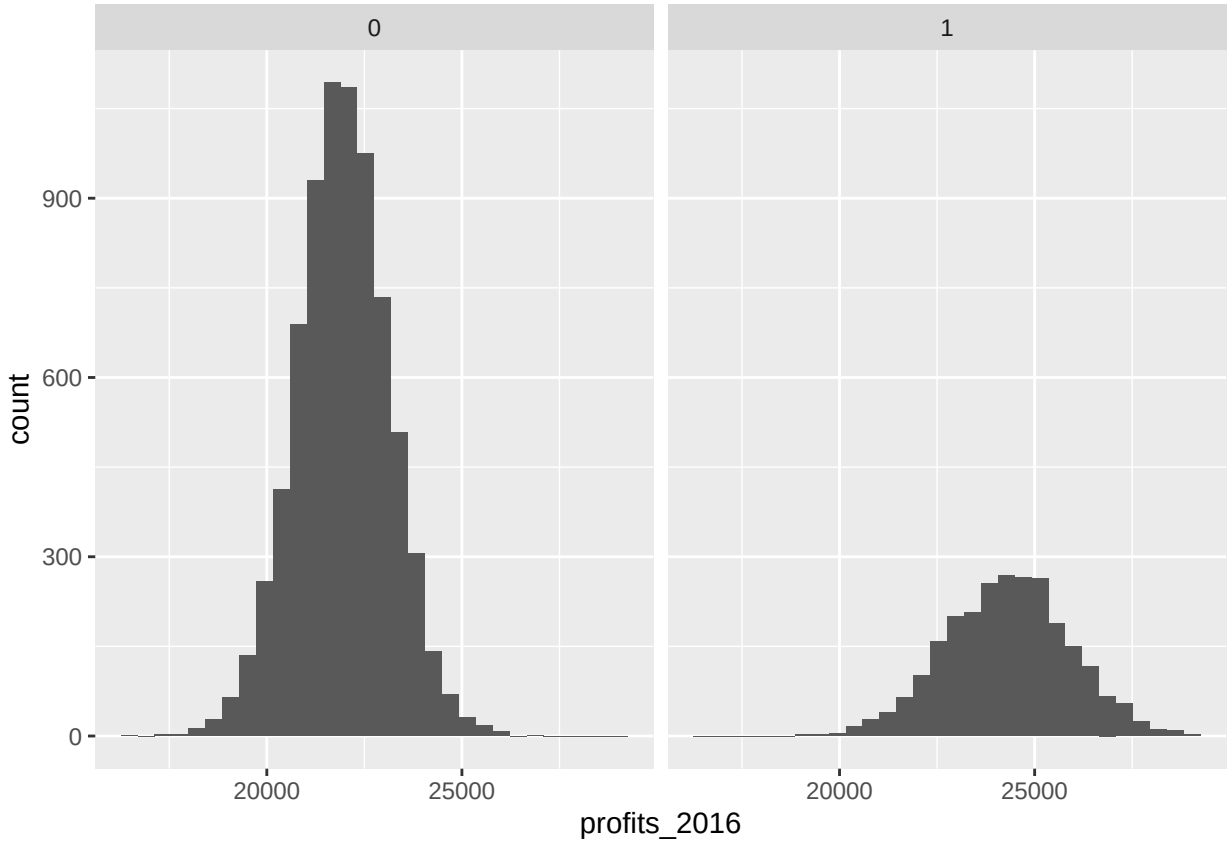
```
## [1] RICE    LENTILS WHEAT   COTTON
## Levels: COTTON LENTILS RICE WHEAT
```

```
unique(harris_data$fertilizer_use)
```

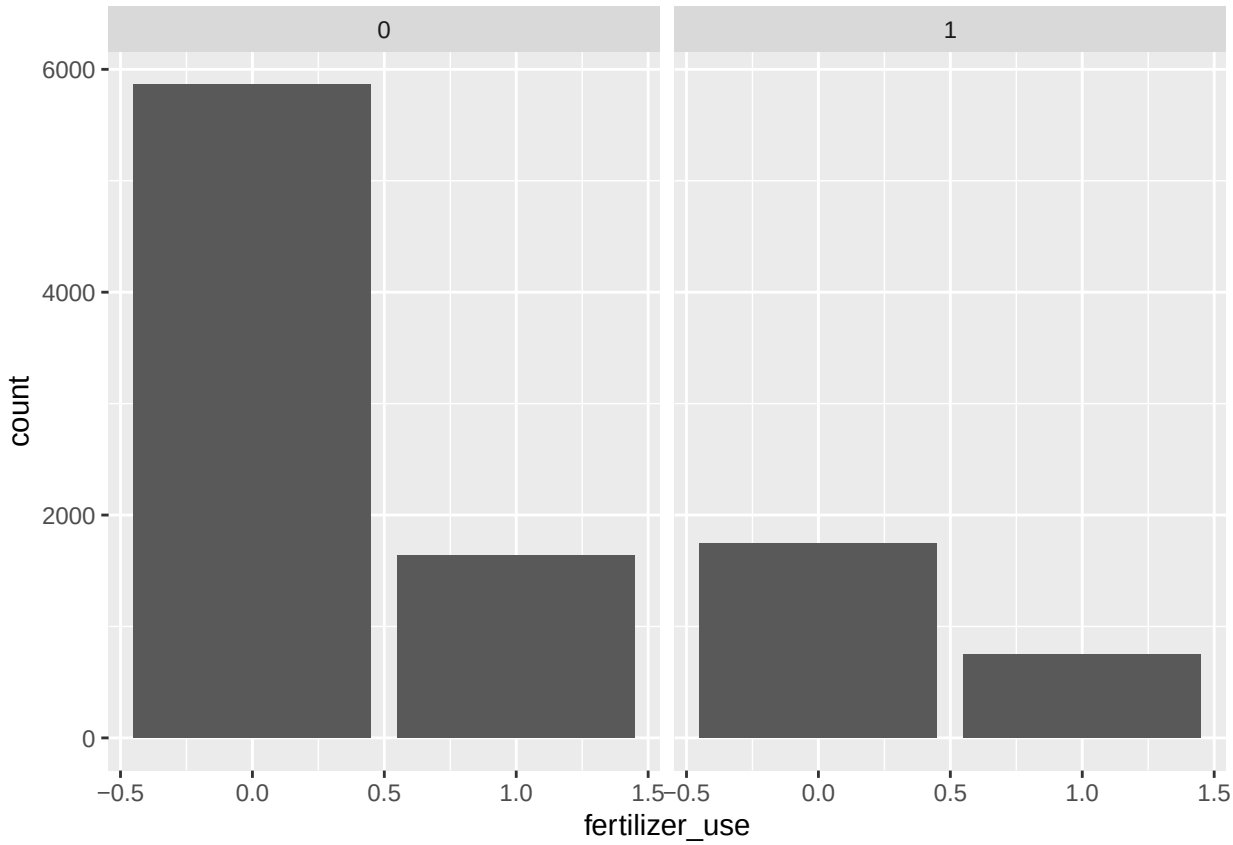
```
## [1] 1 0
```

```
#make sure we are working with factors
harris_data$crop<-as.factor(harris_data$crop)
harris_data$district<-as.factor(harris_data$district)

#quick check of data to get shape
ggplot(harris_data, aes(x=profits_2016)) +
  geom_histogram() +
  facet_grid(~fiona_farmer)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#fewer in treatment group overall, but average profits higher

#control group proportion dont use fertilizer much larger, treatment narrows gap
ggplot(harris_data, aes(x=fertilizer_use)) +
  geom_bar() +
  facet_grid(~fiona_farmer)
```



5. This is pretty troublesome, it looks like on some critical pre-treatment observed characteristics, the control and treatment groups look pretty different.

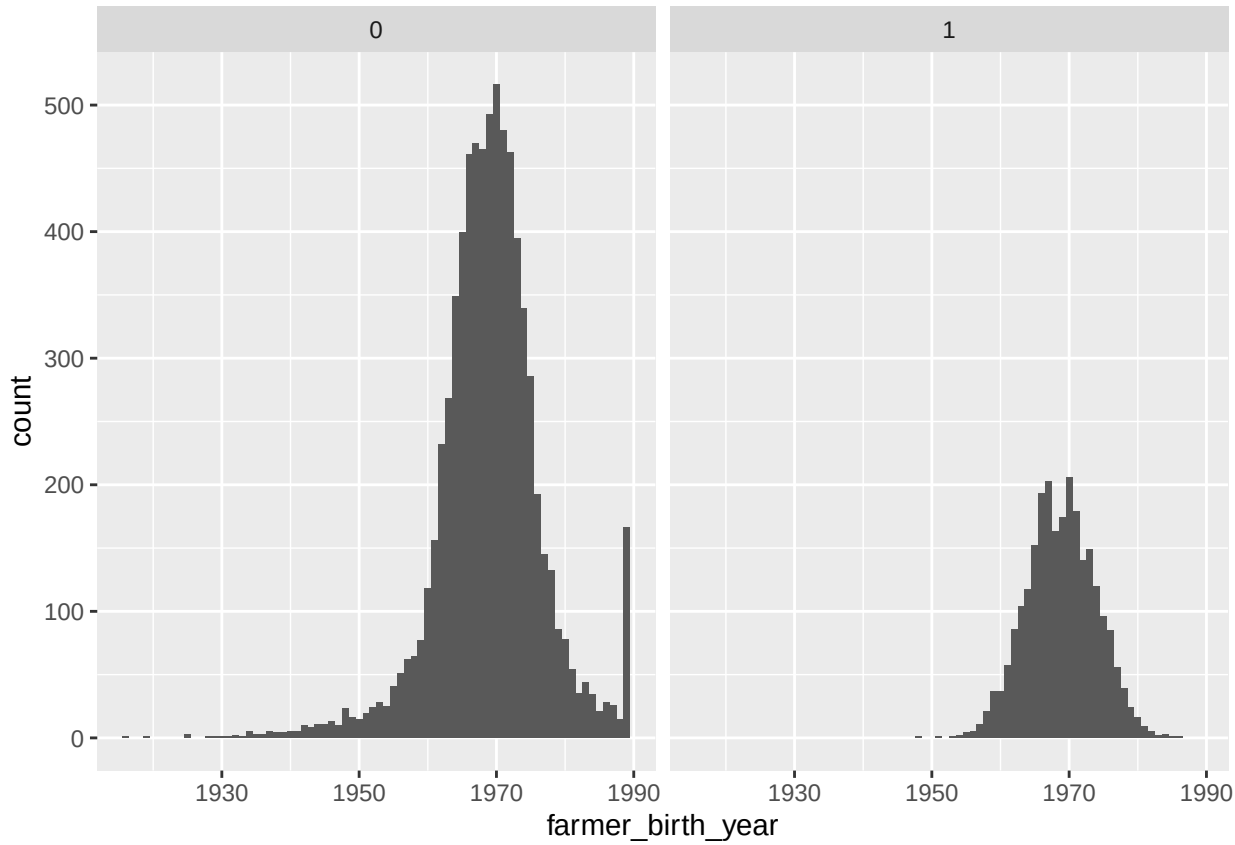
Let's start with the good news. Farmer's age and their profits in 2006 aren't statistically significant between the control and treatment groups, so it doesn't look like they are selecting into treatment based on how old they are or how much profit they had in previous years.

The bad news is there's heavy selection in our other pre-treatment variables. Farmer's choice of crop has a huge impact on wheter they select into treatment. Indeed, as the plot below makes clear, all the farmers who grow cotton selected into treatment, and no control farmers grow cotton at all.

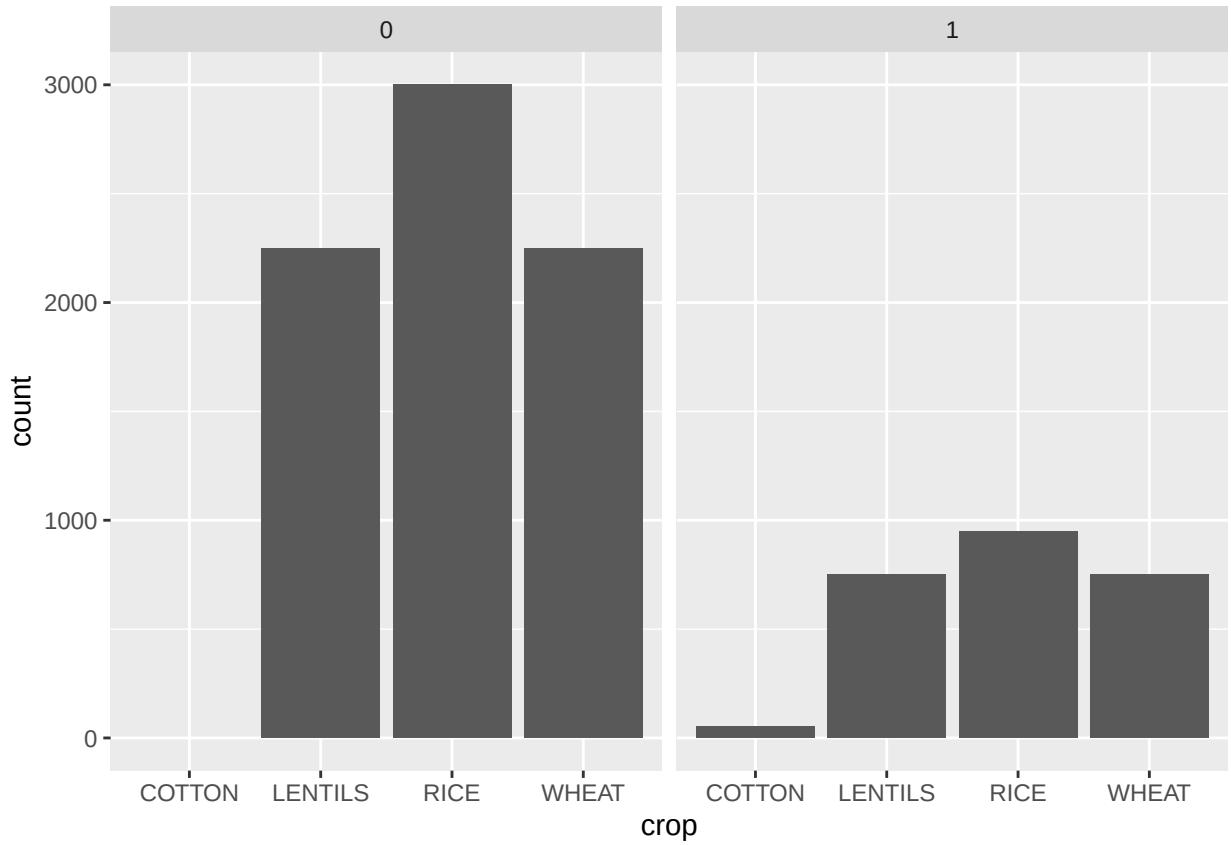
The really bad news, though, is that all of the farmers who opted into treatment live in the same district, Thanjavur, and there are no control farmers who live in that district. This means our regression actually fails to work at all - there is no covariation here, because the control and treatment districts don't overlap at all.

All in all, the bad news outweighs the good here, and our concerns in question three about selection bias are not alleviated at all - they are reinforced.

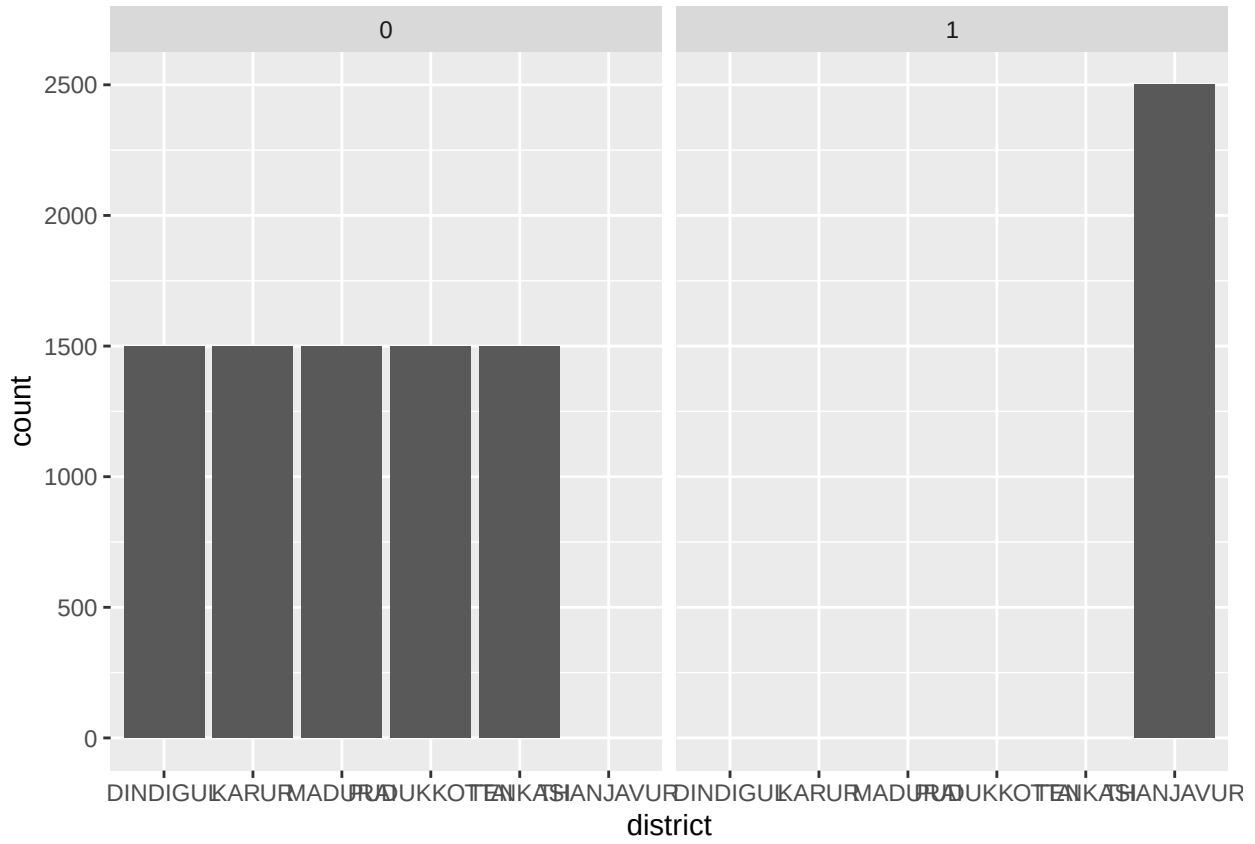
```
#before we dip into the regression lets just explore visually for a moment

#in the real world, i'd be concerned this data was manipulated because
#we have a very unnatural spike in 1 year only in the control group.
#This willthrow off the balance
ggplot(harris_data, aes(x=farmer_birth_year)) +
  geom_bar() +
  facet_grid(~fiona_farmer)
```




```
#roughly similar, but only treated farmers grow cotton. this will cause matching problems
ggplot(harris_data, aes(x=crop)) +
  geom_bar() +
  facet_grid(~fiona_farmer)
```

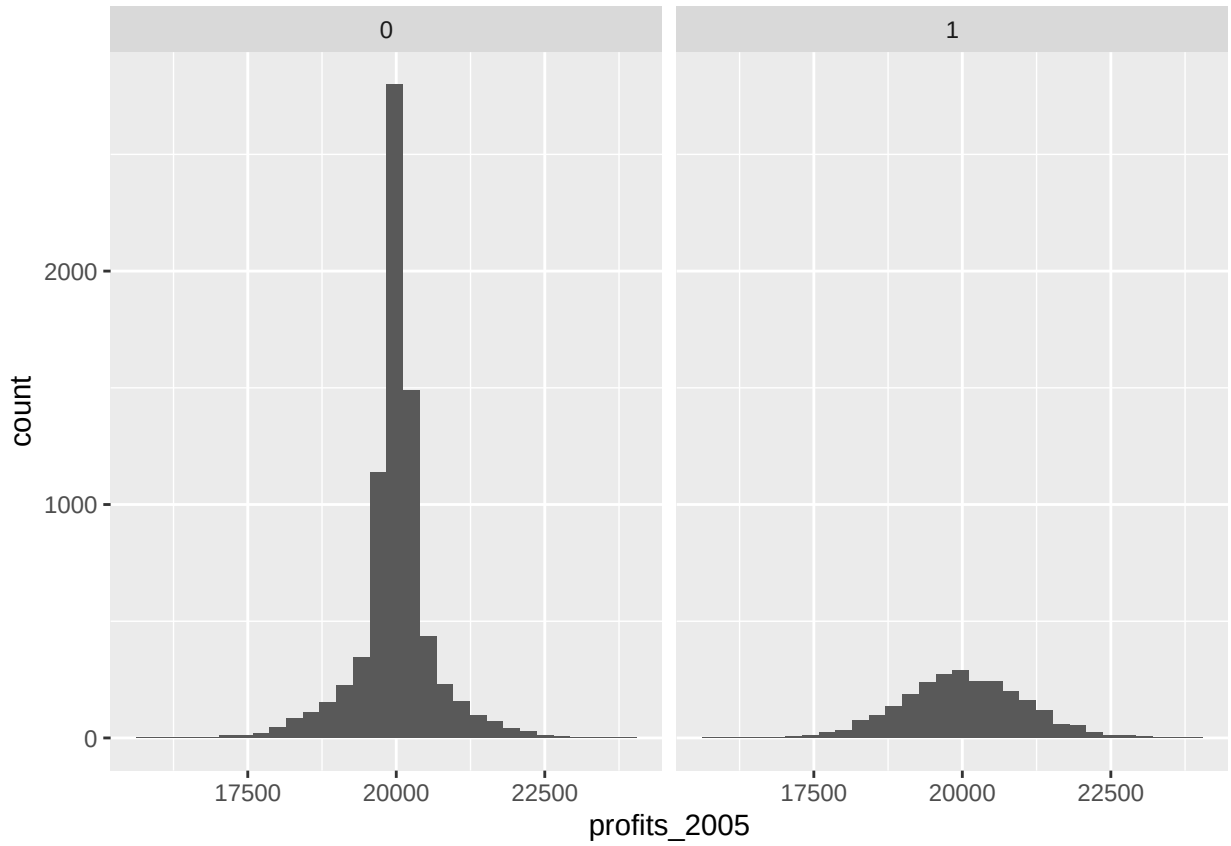



```
#lol well this is clearly a huge problem
#all treated farmers live only in one district, and control has none
ggplot(harris_data, aes(x=district)) +
  geom_bar() +
  facet_grid(~fiona_farmer)
```




```
#distributions have roughly the same center and tails before treat
#only variable without major concerns.
ggplot(harris_data, aes(x=profits_2005)) +
  geom_histogram() +
  facet_grid(~fiona_farmer)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
##balance table regressions
##dv is treatment varaible, IV is the baseline control

#age: ~0c oeff, pvalue of .49, not statistically significant
balance_check_farmer_birth_year<-lm(fiona_farmer ~ farmer_birth_year, data = harris_data) %>%
  tidy() %>%
  print()
```

```
## # A tibble: 2 x 5
##   term               estimate std.error statistic p.value
##   <chr>                 <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)        1.06      1.19         0.888   0.375
## 2 farmer_birth_year -0.000411  0.000606    -0.678   0.498
```

```
#crop: all of the crops are highly statistically significant
#cotton is the baseline - but, only treatment group grows it. no overlap in this obv
balance_check_crop<-lm(fiona_farmer ~ crop, data = harris_data) %>%
  tidy() %>%
  print()
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1.       0.0590      16.9 1.57e-63
## 2 cropLENTILS   -0.75     0.0595     -12.6 4.02e-36
## 3 cropRICE      -0.760    0.0594     -12.8 3.43e-37
## 4 cropWHEAT     -0.75     0.0595     -12.6 4.02e-36
```

```
#district: essentially perfect fit, no variation in the X for treatment so regression assumptions are v
#baseline district: Dindigul
balance_check_district<-lm(fiona_farmer ~ district, data = harris_data) %>%
  tidy() %>%
  print()
```

```
## Warning in summary.lm(x): essentially perfect fit: summary may be unreliable
```

```
## # A tibble: 6 x 5
##   term                 estimate std.error statistic p.value
##   <chr>                   <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)          1.48e-16  6.31e-32   2.34e15       0
## 2 districtKARUR       -1.19e-16  8.92e-32  -1.33e15       0
## 3 districtMADURAI     -1.53e-16  8.92e-32  -1.71e15       0
## 4 districtPUDUKKOTTAI -1.33e-16  8.92e-32  -1.49e15       0
## 5 districtTENKASI     -1.08e-16  8.92e-32  -1.21e15       0
## 6 districtTHANJAVUR    1.00e+ 0  7.98e-32   1.25e31       0
```

```
#profit: n~0 coeff, not statistically significant,
balance_check_profit<-lm(fiona_farmer ~ profits_2005, data = harris_data) %>%
  tidy() %>%
  print()
```

```
## # A tibble: 2 x 5
##   term           estimate  std.error statistic p.value
##   <chr>             <dbl>      <dbl>     <dbl>   <dbl>
## 1 (Intercept)  0.170      0.116          1.46    0.145
## 2 profits_2005 0.00000401 0.00000582     0.689   0.491
```

```
#some data restructuring for balance table
#create dummies for categorical varaibles

harris_data_dummies<-harris_data %>%
  mutate(cotton = ifelse(crop=="COTTON", 1, 0),
         lentil = ifelse(crop=="LENTILS", 1, 0),
         rice = ifelse(crop=="RICE", 1, 0),
         wheat = ifelse(crop=="WHEAT", 1, 0),
         karur = ifelse(district=="KARUR", 1, 0),
         tenkasi = ifelse(district=="TENKASI", 1, 0),
         madurai = ifelse(district=="MADURAI", 1, 0),
         pudukkottai = ifelse(district=="PUDUKKOTTAI", 1, 0),
         thanjavur = ifelse(district=="THANJAVUR", 1, 0),
         dindigul = ifelse(district=="DINDIGUL", 1, 0))


#balance tables, summary stats and p values
balance_table<-rbind(balance_check_crop, balance_check_district, balance_check_farmer_birth_year, balan
  as_data_frame() %>%
  arrange(term, p.value) %>%
  filter(term != "(Intercept)") %>%
  print()
```

```
## Warning: `as_data_frame()` is deprecated, use `as_tibble()` (but mind the new semantics).
## This warning is displayed once per session.
```

```
## # A tibble: 10 x 5
```

```
##    term                 estimate std.error statistic  p.value
##    <chr>                   <dbl>     <dbl>     <dbl>    <dbl>
##  1 cropLENTILS         -7.50e- 1  5.95e- 2 -1.26e+ 1 4.02e-36
##  2 cropRICE            -7.60e- 1  5.94e- 2 -1.28e+ 1 3.43e-37
##  3 cropWHEAT           -7.50e- 1  5.95e- 2 -1.26e+ 1 4.02e-36
##  4 districtKARUR       -1.19e-16  8.92e-32 -1.33e+15 0.
##  5 districtMADURAI     -1.53e-16  8.92e-32 -1.71e+15 0.
##  6 districtPUDUKKOTTAI -1.33e-16  8.92e-32 -1.49e+15 0.
##  7 districtTENKASI     -1.08e-16  8.92e-32 -1.21e+15 0.
##  8 districtTHANJAVUR    1.00e+ 0  7.98e-32  1.25e+31 0.
##  9 farmer_birth_year   -4.11e- 4  6.06e- 4 -6.78e- 1 4.98e- 1
## 10 profits_2005         4.01e- 6  5.82e- 6  6.89e- 1 4.91e- 1
```

```
balance_table_sum_stats<-harris_data_dummies %>%
  group_by(fiona_farmer) %>%
  summarise(total_participants = n(),
            profits_2005_mean = mean(profits_2005),
            profits_2005_sd = sd(profits_2005),
            farmer_birth_year_mean = mean(farmer_birth_year),
            farmer_birth_year_sd = sd(farmer_birth_year),
            lentil_percent = mean(lentil)*100,
            rice_percent = mean(rice)*100,
            wheat_percent = mean(wheat)*100,
            karur_percent = mean(karur)*100,
            tenkasi_percent = mean(tenkasi)*100,
            madurai_percent = mean(madurai)*100,
            pudukkottai_percent = mean(pudukkottai)*100,
            thanjavur_percent = mean(thanjavur)*100,
            dindigul_percent = mean(dindigul)*100
  ) %>%
  print
```

```
## # A tibble: 2 x 15
##   fiona_farmer total_participa~ profits_2005_me~ profits_2005_sd
##          <int>            <int>            <dbl>           <dbl>
## 1            0             7500           19990.            616.
## 2            1             2500           20002.           1036.
## # ... with 11 more variables: farmer_birth_year_mean <dbl>,
## #   farmer_birth_year_sd <dbl>, lentil_percent <dbl>, rice_percent <dbl>,
## #   wheat_percent <dbl>, karur_percent <dbl>, tenkasi_percent <dbl>,
## #   madurai_percent <dbl>, pudukkottai_percent <dbl>, thanjavur_percent <dbl>,
## #   dindigul_percent <dbl>
```

6. As stated above, Harris might consider a selection on observables design, but should be cautious since this makes some strong assumptions that probably aren't borne out in the real world. These designs assume that the differences we see between control and treatment groups capture everything that matters, and totally explains how the two groups different from each other. In this case, we're assuming that farmers who received the insurance are only different from those who did not because of their assignment to the treatment or control groups *and* the district they live in and the group they grow. If that's so, using regression adjustment we can mathematically account for these differences between the two groups and draw causal inferences. Recall that the basic expression we're working with is

$$\hat{\tau} = (\bar{Y_T} - \bar{Y_U}) - \hat{\gamma}(\bar{X_T} - \bar{X_U})$$

Where X is a vector of observed variables shared by the control and treament groups - here, age, crop, district,

and pre-treatment profits.

If we assume that we have properly identified *all* the variables that make the groups different, we also need to assume that we have properly described the functional relationship variables have on outcomes. This is because we're trying to control for the conditional expectation of treatment based on observed characteristics $E[D_i|X_i]$, which is subtly different from just controlling for characteristics X_i. In other words, we are assuming that age has some kind of linear impact profits, and that we have generally picked the right expontent to reflect this relationship. Otherwise, if $X_i \neq E[D_i|X_i]$, then we end up with $\gamma(X_i - E[D_i|X_i])$ in our error term, and the expected value of our error is no longer 0, rendering the excercise pointless. Fortunately, we can include multiple functional forms to show our results are robust, though this does increase the chances of false-positive as we run more and more regressions. Mathematically, for example, we might pick age^2 or age^3 or age^4 and test each of these forms.

Further, the average difference in observed characteristics between the control and treatment groups can't be too large (there needs to be some overlap). Remeber that, in

$$\hat{\tau} = (\bar{Y_T} - \bar{Y_U}) - \hat{\gamma}(\bar{X}_T - \bar{X}_U)$$

we're trying to discern treatment effects from everything else. In this approach, if the second term (the everything else) is very large, that's going to be a problem. In this example, our district varaibles are so heavily selected that they have no overlap between treatment and control groups, so this could be a serious issue.

At this point, I don't think we're going to get very credible answers to this question. The lack of any overlap in district is a particularly big hurdle to overcome: we won't be able to include it in our regression adjustment, and we won't be able to find any exact matches. This might be a huge source of the variation we see in outcomes, but because our control and sample groups don't share any overlap, there's not much we can do to discern what the effect of being in that particular district is from the effects of treatment. If we knew more about the districts themselves, we could make use bandwidth matching to find similar-ish districts to compare, and that could improve our estimate - but right now we don't have that.

As an aside, it would also be nice to know which farmers used fertilizer in 2005. This might give us a better understanding of the ATT/LATE. Fromt the chart above, it seems like more farmers who recieved insurance used fertilizer and this likely correlated with increased profits. But if we had data from 2005 on fertilizer use, we could learn more about this impact on the treated and explore this possible mechanism in more detail.

7. The model below suggests that, once we account for farmer's differences in age, crops, and 2005 profits, receiving the insurance is casually associated with an increase in profits of, on average $\$2,358.35$, give or take about \$24.34. Moreover, this effect is highly statistically significant, so we can be very confident that we aren't just observing this relationship due to random chance. In this model, I did not include the district variable because, unfortunatley, there is no overlap and therefore no variation between the control and treatment group's districts. This violates our assumptions, and means that running a regression won't really tell us anything, so including it in our running variable of controls isn't a viable option. This is unfortuante, because it seems likeley that these a lot of selection bias based into the insurance program on where farmers live, so we really need to be cautious about interpreting these results. I also did not include fertilizer use, because this is a post-treatment outcome, and we don't want to try and control for those.

Comparing this regression to the naive estimator really reinfornces the strengths and weaknesses of this design. On the one hand, we do get a different estimate that looks slightly more refined, and we can see the standard deviation for that estimate, and interpret its statistical signficance. Those are all improvements from just lopping off the means between the two groups.

Still, even though we have good reason to worry about selection bias, the final estimates are really similar - indeed, the naive estimator is just \$11 from of the regression adjustment estimator, well within one standard deviation. This might be because our regression really isn't adding much new information that wasn't captured already by the naive estimator.

In our balance chart, we saw that the control and treatment groups were actually pretty similar in terms of age and 2005 profits. Selection bias for these variables wasn't a big concern, so including them in our r egression shouldn't have a big impact. We are able to make some improvement by controlling for crop-type, which could explain why this estimte seems a bit more refined, but remember that no one in the control group grew cotton at all so overlap remains a concern. Critically, most of the balance differences came from districts, which we unfortunately had to leave out of the model because there is no overlap at all. This regression can't account for this differences, so we know they are still lurking out in the error term, just as they are for the naive estimator.

We've also assumed, here, that age has a linear relationship with profit, but that might not be the case. We should test multiple function forms - perhaps aging up from 18 to 35 does have a positive impact, but aging from 55 to 85 has a negative impact. We should check this to be robust.

In other words, though this regression adjustment is certaintly an improvement on the naive estimator, there's good reason to think it's not doing all-too-much better in this example. This is particularly true becuase there isn't enough overlap in the district variable to make inferences and deal with the selection bias from this critical characteristic.

```
#regression adjustment to estimate
#we want to estimate profit 2016 as the DV
#include age, crop, 2005 profit and treatment indicator as IV
#dont include district because there's no overlap at all

#treatment indicator highly significant, estimates $2,358.35 with $24.34 sd
lm(profits_2016 ~ fiona_farmer + farmer_birth_year + crop +profits_2005,data=harris_data) %>%
  tidy() %>%
  print()
```

```
## # A tibble: 7 x 5
##   term              estimate std.error statistic p.value
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)       1239.    2900.         0.427 0.669
## 2 fiona_farmer      2358.      24.3       96.9   0
## 3 farmer_birth_year    0.193    1.46       0.132 0.895
## 4 cropLENTILS        359.     146.         2.46  0.0139
## 5 cropRICE            31.2    146.         0.214 0.831
## 6 cropWHEAT          428.     146.         2.93  0.00338
## 7 profits_2005         1.00     0.0141    71.4   0
```

```
#naive estimator
harris_data_dummies %>%
  group_by(fiona_farmer) %>%
  summarise(avg_profit = mean(profits_2016),
            sd_profit = sd(profits_2016)) %>%
  print
```

```
## # A tibble: 2 x 3
##   fiona_farmer avg_profit sd_profit
##          <int>      <dbl>     <dbl>
## 1            0     21942.     1188.
## 2            1     24312.     1582.
```

```
 naive_estimate <- (24311.90 - 21942.34) %>%
   print()
```

```
## [1] 2369.56
```

```
difference_regression_and_naive<- (naive_estimate - 2358.35) %>%
  print()
```

```
## [1] 11.21
```

8.

Unlike regression adjustment, matching doesn't require the difference between $(\bar{X}_T - \bar{X}_U)$ to be small, and it doesn't make functional form assumptions. This will allow us to incorporate more data and make fewer assumptions about how variables impact the profits of farmers.

Still, exact matching imposes some restrictions. We can't really use continuous variables like profit because we just won't have enough exact matches. Even when I round to the nearest decimal place (getting rid of cents), we still have only about 100 matches - and we've voilated the "exact matching" principle. Moreover, we still can't include district because there are no pairs at all between treatment and control groups. If we have more information on traits shared by villages, perhaps we could put them into buckets, but we're back to bandwidth, non-exact matching again. I also exclude fertilizier since that's a post-treatment variable.

I estimate an ATE of 2384.02 and an ATT of 2151.45. These are relatively similar, and since our estimates are still imperfect because we couldn't include continuous variables and those with no overlap, they may be even more similar statistically than we know in the real world. When we think about the magnitude of a difference, it's important to think about who this variable matters to. If I was a farmer I would really want that extra 230 bucks or so - that's about 9% of the total ATE, so it is a pretty significant difference to them.

As discussed in question 1, we wouldn't really expect these to be the same. The ATE and the ATT are only equivelant when there is no selection into treatment, but our balance table showed this is not the case in this example. Even though our pairing generally did a good job on observed differences, the balance table of paired observations showed that there were still a few important differences: even when paired, treated farmers were more likely to grow lentils than their counterfactuals. All this would lead us to think the ATE and ATT should be at least slightly different, not identical.

This approach seems to be a general improvement over regression adjustment. We don't have to assume a functional form, and we dont' have to worry about the average underlying characteristics between control and treatment not being close enough. However, this method introduecs new problems. We need a lot of observations for this to work, and that is only true when we add more variables. Further, exact matching struggles to deal with continuous variables. We can fix some of that by using bandwidth paired matching, but even though this gets close it can never be perfect because it is, afterall, not exact.

Our estimates here are actually pretty similar, once again, to the naive estimator. The ATE is only $14 off, and the ATT is only $218.06 off.

Because we were only working with a limited number categorical variables, the Curse of Deimenstionality did not seriously impact our approach, but it did rear its head. Matching was particularly difficult on crops because we had no matches in the treatment and control group. The Curse of Dimenstionality really rearered its head, though, when I rounded 2005 profits to the nearest decimle and tried to match on that. We actually did get about 100-200 observations, but we lost thousands of observations in the process. This made using the rounded profit data unworkable, so I had to back off of the idea. This shows that, as we introduce variables to matching evaluations, we are going to get noisier and noisier estimates because we're losing observations. We need to be really mindful of this and weigh the importance of a variable against the loss of observations associated with introducing it.

```
#use exact matching to estimate effect of FIONA on farmer profits
#what should we include in matching procedure? Only categorical variables with matches
#we can't match on continuous

#simple summary with mean and sd by treatment groups
harris_data_dummies %>%
```

```r
  group_by(fiona_farmer) %>%
  summarise_all(funs(mean(., na.rm = T), sd(., na.rm = T)))
```

```
## Warning: funs() is soft deprecated as of 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))
## This warning is displayed once per session.

## Warning in mean.default(district, na.rm = T): argument is not numeric or
## logical: returning NA

## Warning in mean.default(district, na.rm = T): argument is not numeric or
## logical: returning NA

## Warning in mean.default(crop, na.rm = T): argument is not numeric or logical:
## returning NA

## Warning in mean.default(crop, na.rm = T): argument is not numeric or logical:
## returning NA

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling var(x)
##   Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling var(x)
##   Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling var(x)
##   Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling var(x)
##   Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.

## # A tibble: 2 x 33
##   fiona_farmer district_mean crop_mean farmer_birth_ye~ fertilizer_use_~
##          <int>         <dbl>     <dbl>            <dbl>            <dbl>
## 1            0            NA        NA            1969.            0.219
## 2            1            NA        NA            1969.            0.300
## # ... with 28 more variables: profits_2005_mean <dbl>, profits_2016_mean <dbl>,
## #   cotton_mean <dbl>, lentil_mean <dbl>, rice_mean <dbl>, wheat_mean <dbl>,
## #   karur_mean <dbl>, tenkasi_mean <dbl>, madurai_mean <dbl>,
## #   pudukkottai_mean <dbl>, thanjavur_mean <dbl>, dindigul_mean <dbl>,
## #   district_sd <dbl>, crop_sd <dbl>, farmer_birth_year_sd <dbl>,
## #   fertilizer_use_sd <dbl>, profits_2005_sd <dbl>, profits_2016_sd <dbl>,
## #   cotton_sd <dbl>, lentil_sd <dbl>, rice_sd <dbl>, wheat_sd <dbl>,
## #   karur_sd <dbl>, tenkasi_sd <dbl>, madurai_sd <dbl>, pudukkottai_sd <dbl>,
## #   thanjavur_sd <dbl>, dindigul_sd <dbl>
```

```r
#matchit: we lose 501 control and 24 treated this way
match_model<-matchit(fiona_farmer ~ farmer_birth_year + crop, method = "exact", data = harris_data_dummi

#matches dataset with weights
exact_matches_data<-match.data(match_model) %>%
  mutate(avg_outcome_weighted_ATE = profits_2016 * weights)


treated_outcomes_matchit <- exact_matches_data %>%
  filter(fiona_farmer == 1) %>%
  select(avg_outcome_weighted_ATE)

control_outcomes_matchit <- exact_matches_data %>%
  filter(fiona_farmer == 0) %>%
  select(avg_outcome_weighted_ATE)

#####ATE ESTIMATE #####: + 2384.02
mean(treated_outcomes_matchit$avg_outcome_weighted) - mean(control_outcomes_matchit$avg_outcome_weighte
```

```
## [1] 2384.022
```

```r
#if we want to match by age exactly

(nrow(filter(harris_data_dummies, fiona_farmer == 1))) #2500 treated in total
```

```
## [1] 2500
```

```r
#counts and weights for each cell
summary_by_age_crop<-harris_data_dummies %>%
  group_by(fiona_farmer, crop, farmer_birth_year) %>%
  summarise(avg_profit_2016_outcome = mean(profits_2016),
            sd_profit_2016_outcome = sd(profits_2016),
            total_farmers = n(),
            ATE_weight = total_farmers/nrow(harris_data_dummies),
            ATE_weighted_avg = avg_profit_2016_outcome * ATE_weight) %>%
  mutate(ATT_weight = ifelse(fiona_farmer ==1, (total_farmers/
                                                  250), NA),
            ATT_weighted_avg = ATT_weight * avg_profit_2016_outcome) %>%
  print()
```

```
## # A tibble: 268 x 10
## # Groups:   fiona_farmer, crop [7]
##    fiona_farmer crop  farmer_birth_ye~ avg_profit_2016~ sd_profit_2016_~
##           <int> <fct>            <dbl>            <dbl>            <dbl>
##  1            0 LENT~             1916           20915.             NA
##  2            0 LENT~             1928           22746.             NA
##  3            0 LENT~             1930           22472.             NA
##  4            0 LENT~             1933           22253.             NA
##  5            0 LENT~             1934           21589.           1362.
##  6            0 LENT~             1935           21860.             NA
##  7            0 LENT~             1936           22404.             NA
##  8            0 LENT~             1937           21919.           1257.
##  9            0 LENT~             1938           21451.             11.2
## 10            0 LENT~             1939           21944.           1057.
## # ... with 258 more rows, and 5 more variables: total_farmers <int>,
```

```
## #   ATE_weight <dbl>, ATE_weighted_avg <dbl>, ATT_weight <dbl>,
## #   ATT_weighted_avg <dbl>
```

```
#getting our terms of interest ready and vectorized, only looking at treated group
treatment_match_outcomes <- summary_by_age_crop %>%
  filter(fiona_farmer == 1) %>%
  select(avg_profit_2016_outcome, ATE_weight, ATE_weighted_avg, ATT_weight, ATT_weighted_avg, total_far
```

```
## Adding missing grouping variables: `fiona_farmer`, `crop`
```

```
######## ATT estimate ########: + 2151.496
mean(treatment_match_outcomes$ATT_weighted_avg)
```

```
## [1] 2151.496
```

```
#####balance testing####
exact_matches_data %>%
  group_by(fiona_farmer) %>%
  summarise_all(funs(mean(., na.rm = T), sd(., na.rm = T))) %>%
  print()
```

```
## Warning in mean.default(district, na.rm = T): argument is not numeric or
## logical: returning NA

## Warning in mean.default(district, na.rm = T): argument is not numeric or
## logical: returning NA

## Warning in mean.default(crop, na.rm = T): argument is not numeric or logical:
## returning NA

## Warning in mean.default(crop, na.rm = T): argument is not numeric or logical:
## returning NA

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling var(x)
##   Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling var(x)
##   Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling var(x)
##   Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling var(x)
##   Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.

## # A tibble: 2 x 39
##   fiona_farmer district_mean crop_mean farmer_birth_ye~ fertilizer_use_~
##          <int>         <dbl>     <dbl>            <dbl>            <dbl>
## 1            0            NA        NA            1969.            0.219
## 2            1            NA        NA            1969.            0.298
## # ... with 34 more variables: profits_2005_mean <dbl>, profits_2016_mean <dbl>,
## #   cotton_mean <dbl>, lentil_mean <dbl>, rice_mean <dbl>, wheat_mean <dbl>,
## #   karur_mean <dbl>, tenkasi_mean <dbl>, madurai_mean <dbl>,
## #   pudukkottai_mean <dbl>, thanjavur_mean <dbl>, dindigul_mean <dbl>,
## #   weights_mean <dbl>, subclass_mean <dbl>,
## #   avg_outcome_weighted_ATE_mean <dbl>, district_sd <dbl>, crop_sd <dbl>,
## #   farmer_birth_year_sd <dbl>, fertilizer_use_sd <dbl>, profits_2005_sd <dbl>,
## #   profits_2016_sd <dbl>, cotton_sd <dbl>, lentil_sd <dbl>, rice_sd <dbl>,
```

```
## #   wheat_sd <dbl>, karur_sd <dbl>, tenkasi_sd <dbl>, madurai_sd <dbl>,
## #   pudukkottai_sd <dbl>, thanjavur_sd <dbl>, dindigul_sd <dbl>,
## #   weights_sd <dbl>, subclass_sd <dbl>, avg_outcome_weighted_ATE_sd <dbl>
```

```
#our pre-treatment observables means and sds look pretty similar now,
#though our outcomes don't (that's expected), and where there is no overlap we have 0s

#a battery of regressions to see if we have bad balance still
summary(lm(farmer_birth_year ~ fiona_farmer, data = exact_matches_data)) #not significant
```

```
##
## Call:
## lm(formula = farmer_birth_year ~ fiona_farmer, data = exact_matches_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.0111  -3.8406  -0.0111   3.9889  17.1594
##
## Coefficients:
##                Estimate Std. Error   t value Pr(>|t|)
## (Intercept)  1969.01114    0.06509 30250.685   <2e-16 ***
## fiona_farmer   -0.17059    0.12790    -1.334    0.182
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.445 on 9443 degrees of freedom
## Multiple R-squared:  0.0001883,  Adjusted R-squared:  8.246e-05
## F-statistic: 1.779 on 1 and 9443 DF,  p-value: 0.1823
```

```
summary(lm(cotton ~ fiona_farmer, data = exact_matches_data)) #no variation, regression is busted
```

```
##
## Call:
## lm(formula = cotton ~ fiona_farmer, data = exact_matches_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##      0      0      0      0      0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0          0      NA       NA
## fiona_farmer        0          0      NA       NA
##
## Residual standard error: 0 on 9443 degrees of freedom
## Multiple R-squared:    NaN,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 1 and 9443 DF,  p-value: NA
```

```
summary(lm(lentil ~ fiona_farmer, data = exact_matches_data)) #significant at .0489
```

```
##
## Call:
## lm(formula = lentil ~ fiona_farmer, data = exact_matches_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -0.3066 -0.2856 -0.2856  0.6934  0.7144
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.285612   0.005429  52.607   <2e-16 ***
## fiona_farmer 0.021011   0.010669   1.969   0.0489 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4542 on 9443 degrees of freedom
## Multiple R-squared:  0.0004106,  Adjusted R-squared:  0.0003047
## F-statistic: 3.879 on 1 and 9443 DF,  p-value: 0.04893
```

```
summary(lm(wheat ~ fiona_farmer, data = exact_matches_data)) #not significant
```

```
##
## Call:
## lm(formula = wheat ~ fiona_farmer, data = exact_matches_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3208 -0.3208 -0.3208  0.6792  0.6938
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.320760   0.005562  57.671   <2e-16 ***
## fiona_farmer -0.014546   0.010929  -1.331    0.183
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4653 on 9443 degrees of freedom
## Multiple R-squared:  0.0001875,  Adjusted R-squared:  8.166e-05
## F-statistic: 1.771 on 1 and 9443 DF,  p-value: 0.1833
```

```
summary(lm(rice ~ fiona_farmer, data = exact_matches_data)) #not signficant
```

```
##
## Call:
## lm(formula = rice ~ fiona_farmer, data = exact_matches_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3936 -0.3936 -0.3872  0.6064  0.6128
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.393628   0.005836  67.450   <2e-16 ***
## fiona_farmer -0.006465   0.011468  -0.564    0.573
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4882 on 9443 degrees of freedom
## Multiple R-squared:  3.365e-05,  Adjusted R-squared:  -7.224e-05
## F-statistic: 0.3178 on 1 and 9443 DF,  p-value: 0.5729
```

```r
summary(lm(karur ~ fiona_farmer, data = exact_matches_data)) #village regressions all busted
```

```
##
## Call:
## lm(formula = karur ~ fiona_farmer, data = exact_matches_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2136 -0.2136 -0.2136  0.0000  0.7864
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.213602   0.004218   50.65   <2e-16 ***
## fiona_farmer -0.213602   0.008288  -25.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3528 on 9443 degrees of freedom
## Multiple R-squared:  0.06572,    Adjusted R-squared:  0.06562
## F-statistic: 664.2 on 1 and 9443 DF,  p-value: < 2.2e-16
```

```r
summary(lm(tenkasi ~ fiona_farmer, data = exact_matches_data))
```

```
##
## Call:
## lm(formula = tenkasi ~ fiona_farmer, data = exact_matches_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1459 -0.1459 -0.1459  0.0000  0.8541
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.145878   0.003632   40.16   <2e-16 ***
## fiona_farmer -0.145878   0.007138  -20.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3039 on 9443 degrees of freedom
## Multiple R-squared:  0.04236,    Adjusted R-squared:  0.04226
## F-statistic: 417.7 on 1 and 9443 DF,  p-value: < 2.2e-16
```

```r
summary(lm(madurai ~ fiona_farmer, data = exact_matches_data))
```

```
##
## Call:
## lm(formula = madurai ~ fiona_farmer, data = exact_matches_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.213 -0.213 -0.213  0.000  0.787
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.213030   0.004214   50.56   <2e-16 ***
```

```
## fiona_farmer -0.213030   0.008280  -25.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3525 on 9443 degrees of freedom
## Multiple R-squared:  0.06551,    Adjusted R-squared:  0.06541
## F-statistic:   662 on 1 and 9443 DF,  p-value: < 2.2e-16
```

```
summary(lm(pudukkottai ~ fiona_farmer, data = exact_matches_data))
```

```
##
## Call:
## lm(formula = pudukkottai ~ fiona_farmer, data = exact_matches_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2137 -0.2137 -0.2137  0.0000  0.7863
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.213745   0.004219   50.67   <2e-16 ***
## fiona_farmer -0.213745   0.008290  -25.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3529 on 9443 degrees of freedom
## Multiple R-squared:  0.06577,    Adjusted R-squared:  0.06567
## F-statistic: 664.8 on 1 and 9443 DF,  p-value: < 2.2e-16
```

```
summary(lm(thanjavur ~ fiona_farmer, data = exact_matches_data))
```

```
##
## Call:
## lm(formula = thanjavur ~ fiona_farmer, data = exact_matches_data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.667e-14 -4.000e-16 -4.000e-16  0.000e+00  2.817e-12
##
## Coefficients:
##               Estimate Std. Error   t value Pr(>|t|)
## (Intercept)  1.725e-14  3.465e-16 4.980e+01   <2e-16 ***
## fiona_farmer 1.000e+00  6.809e-16 1.469e+15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.899e-14 on 9443 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.157e+30 on 1 and 9443 DF,  p-value: < 2.2e-16
```

```
summary(lm(dindigul ~ fiona_farmer, data = exact_matches_data))
```

```
##
## Call:
## lm(formula = dindigul ~ fiona_farmer, data = exact_matches_data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2137 -0.2137 -0.2137  0.0000  0.7863
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.213745   0.004219   50.67   <2e-16 ***
## fiona_farmer -0.213745   0.008290  -25.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3529 on 9443 degrees of freedom
## Multiple R-squared:  0.06577,    Adjusted R-squared:  0.06567
## F-statistic: 664.8 on 1 and 9443 DF,  p-value: < 2.2e-16
```

```
######check for curse of dimensionality (too few obs after adding vars)
harris_data_dummies_rounded <- harris_data_dummies %>%
  mutate(profits_2005_rounded = round(harris_data_dummies$profits_2005, 0),
         profits_2016_rounded = round(harris_data_dummies$profits_2016, 0))


#if we try to use profits, even when we round, we lose 7377 control and 2390 treated
#we only match 123 control and 110 treated
#this is a small demonstration of the curse.
#If we used real bandwidth matching it'd be a little better, but the concept would hold
match_model_profit_rounded<-matchit(fiona_farmer ~ farmer_birth_year + crop + profits_2005_rounded, meth
```

9. Even though our point estimates from both regression adjustment and matching suggest that these programs are effective, as a relatively conservative analyst I'd have to recommend they do not implement the program.

I just don't have much faith in these estimates. Principally, this is because the selection issues here are severe, and the methods we've employed to deal with them don't completeley address the problem. It seems like district of residence might have a large impact on farmer's profits, but because there is no overlap between control and treatment groups, neither of these tools can really grapple with this important difference. If we had truly even minimal overlap, we could be more comfortable with our estimates, but because there is none whatsoever, it's best to remain skeptical.

Further, I don't think we can really be confident that we've captured all the variables here that might have an impact on farmer's profits. We would like to know more about the differeneces in weather, soil quality, labor, irrigation, education - all of those likeley have an impact. It would be pretty naive of us to assume that this analaysis had really captured all the observables that make a difference. Since we've likeley failed to do so, our estimate is probably still heavily influenced by selection bias, which may explain why it is still pretty similar to the naive estimator.