Assignment 2

Selection on Observables

4/30/2022

Contents

Background

A well-meaning NGO, Harvest Aid and Rainfall Risk Insurance Services (HARRIS) is interested in improving agricultural yields for farmers in Bangladesh. They are particularly excited about the idea of providing farmers with rainfall-index insurance. While most insurance suffers from adverse selection problems (people who already have low yields are likely to select in) and moral hazard issues (when you have insurance, you may be less likely to work hard on your fields), rainfall-index insurance avoids these issues by tying insurance payouts to how much it rains, rather than individual crop yields. Their idea is that, if farmers have insurance, they can invest in riskier – but more profitable – up-front inputs like fertilizer, with less worry of losing money during a bad rainfall year (I'm not just making this up; this is a real thing), thereby raising overall profitability.

They are working on putting together proposal documents to fundraise to pilot this type of insurance product, and have discovered that the Indian government has already implemented a program to do exactly this, called the Farmer's Insurance On Non-rainy Annums (FIONA) scheme. This scheme ran from 2014 to 2016. HARRIS has asked you to help them evaluate the effectiveness of the FIONA scheme in order to understand whether this is a good idea for Bangladesh. (Assume for the purposes of this problem set that farmers do not change crops in response to FIONA.)

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?).

Answer

To estimate the effect of FIONA program on farmers' profits, ideally, we would have an experiment where each farmer is randomly assigned to treatment and control groups. The treatment group would be enrolled in the FIONA program, in other words, they would be insured against undesirable levels of rainfall. Whereas, the control group would be left out from the program. Assuming that there is no selection bias, i.e., control and treatment groups are do not differ from each other significantly besides the treatment, we would be estimate a reliable average treatment effect (ATE). In order to estimate it, we would need data on profits both before and after the administration of treatment as well as a variable, D, indicating the assignment of treatment (0 if control, 1 if treated). In this experiment, the unit of analysis would ideally be an individual farmer. With such a dataset and random assignment, we could estimate the ATE as:

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

Under the potential outcomes framework, we would want to estimate:

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$$

But we observe the farmer, i, only once and do not know its outcomes under the both treatment conditions simultaneously. And randomization allows us to estimate $\hat{\tau}^{ATE}$ ensuring that:

$$\mathbb{E}[Y_i(1)|D_i = 1] = \mathbb{E}[Y_i(1)]$$

$$\mathbb{E}[Y_i(0)|D_i = 0] = \mathbb{E}[Y_i(0)]$$

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.

Answer

If we cannot get every farmer to participate, this would violate random assignment of treatment. In this case we would ideally want to estimate the average treatment effect on the treated (ATT) as:

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

This would is different than the ATE in the sense that it would capture the effect of FIONA only on the farmers who took up insurance. But we still cannot estimate this directly since we do not and cannot observe $\mathbb{E}[Y_i(0)|D_i = 1]$. In words, we do not observe the outcome of farmers who took up insurance if they did not participate. And farmers who chose to participate could be different than the farmers who chose not to, potentially leading to heterogeneous treatment effects due to selection.

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.

Answer

If we compared FIONA farms to non-FIONA farms on average, we would recover the naive estimator $\overline{Y(1)} - \overline{Y(0)}$. This could be problematic if the FIONA program was not offered to districts randomly. For instance, if the program was offered to districts where insurance levels were low, we could expect that the naive estimator would underestimate the treatment effect since it would be likely that farmers and districts who were not insured before the program were less susceptible to fluctuations in rainfall. Conversely, this could overestimate the average treatment effect if the program was offered to districts where crops are especially susceptible to rainfall. In this case, the program would have a bigger impact than we would expect from the population average.

Finally, if only certain districts were randomly offered FIONA and within those districts not every farmer chose to participate in the program, what we might be estimating with the naive estimator could be the intent to treat (ITT) instead of ATE due to some farmers not complying (picking up the insurance).

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.

Answer

```
farm <- read_csv("ps2_data.csv")
skimr::skim(farm)
```

Table 1: Data summary

Name	farm
Number of rows	10000
Number of columns	7
Column type frequency:	
character	3
numeric	4
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
district	0	1	5	11	0	6	0
crop	0	1	4	7	0	4	0
farmer_birth_year	0	1	4	22	0	67	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
fiona_farmer	0	1	0.25	0.43	0.00	0.00	0.00	0.25	1.00	
fertilizer_use	0	1	0.24	0.43	0.00	0.00	0.00	0.00	1.00	
profits_2005	0	1	19992.84	743.83	15842.34	19721.39	19999.87	20268.83	24000.59	
profits_2016	0	1	22534.73	1654.58	16526.75	21408.55	22338.90	23460.87	29096.45	

`farmer_birth_year` seems to be a character field and contains string values instead of digits for some farmers, we will need to fix that.

```
farm <- farm %>%
    mutate(farmer_birth_year = ifelse(farmer_birth_year == "nineteen seventy-three",
        1973, ifelse(farmer_birth_year == "nineteen seventy-two",
            1972, farmer_birth_year))) %>%
    mutate(farmer_birth_year = as.integer(farmer_birth_year))
```

```r
unique(farm$crop)
unique(farm$district)
```

```
[1] "RICE"    "LENTILS" "WHEAT"   "COTTON"
[1] "KARUR"       "TENKASI"     "MADURAI"     "PUDUKKOTTAI" "THANJAVUR"
[6] "DINDIGUL"
```

```r
farm %>%
    group_by(district, fiona_farmer) %>%
    summarise(n()) %>%
    knitr::kable()
```

district	fiona_farmer	n()
DINDIGUL	0	1500
KARUR	0	1500
MADURAI	0	1500
PUDUKKOTTAI	0	1500
TENKASI	0	1500
THANJAVUR	1	2500

We can make use of the other observable characteristics of farmers to undertake selection-on-observables designs by making the assumption that the participation in the FIONA program is as good as random once we control for observable characteristics, namely the crop type, birth year, and profits prior to program.

One approach could be regression adjustment. We include the farmer birth year, pre-treatment profits, and crop type into the regression model to mitigate the selection bias. If our assumptions hold, this would allow us to isolate the effect of treatment on profits by accounting for fluctuations in observables. The regression in this case would take the form

$$Y_i = \alpha + \tau D_i + \beta_1 birthyear + \beta_2 profit2015 + \sum_j \beta_j crop_{j,i} + v_i$$

where crops are represented as dummy variables that take the value of 1 if the farmer i grows the crop and 0 otherwise. The unit of analysis in this regression is an individual farmer.

Another approach to doing this would be matching. Through exact matching, we can compare the profit of FIONA farmers to profit of non-FIONA farmers with identical X_is. In our case, for instance, we could compare the FIONA farmers who grow rice, and were born in 1970, to a same age non-FIONA farmer who also grow rice. In other words, we would calculate $\overline{Y}_T - \overline{Y}_U$ for each $X = x$. Then we would estimate τ^{ATE} by taking the weighted average of each slice. Though incorporating profit into this exact matching process would be problematic due to it being continuous.

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)?

Answer

```
balance_table_1 <- farm %>%
    group_by(fiona_farmer) %>%
    mutate(cotton = crop == "COTTON", rice = crop == "RICE",
        wheat = crop == "WHEAT", lentils = crop == "LENTILS") %>%
    select(-district, -crop, -profits_2016, -fertilizer_use) %>%
    summarise_all(mean) %>%
    select(-fiona_farmer) %>%
    t() %>%
    round(4) %>%
    as.data.frame() %>%
    rownames_to_column("variable")

names(balance_table_1) <- c("variable", "non-FIONA", "FIONA")

balance_table_2 <- farm %>%
    mutate(cotton = crop == "COTTON", rice = crop == "RICE",
        wheat = crop == "WHEAT", lentils = crop == "LENTILS") %>%
    select(-district, -crop, -profits_2016, -fiona_farmer, -fertilizer_use) %>%
    lapply(., function(i) tidy(lm(i ~ farm$fiona_farmer))) %>%
    do.call(rbind, .) %>%
    rownames_to_column("variable") %>%
    filter(term == "farm$fiona_farmer") %>%
    select(-term) %>%
    mutate(variable = str_remove(variable, "\\.2")) %>%
    mutate_if(is.numeric, round, digits = 4)

knitr::kable(balance_table_1, caption = "Balance Table, part 1",
    )
knitr::kable(balance_table_2, caption = "Balance Table, part 2",
    )
```

Table 5: Balance Table, part 1

variable	non-FIONA	FIONA
farmer_birth_year	1968.932	1968.8200
profits_2005	19989.878	20001.7126
cotton	0.000	0.0212
rice	0.400	0.3788
wheat	0.300	0.3000
lentils	0.300	0.3000

Table 6: Balance Table, part 2

variable	estimate	std.error	statistic	p.value
farmer_birth_year	-0.1119	0.1649	-0.6783	0.4976
profits_2005	11.8349	17.1786	0.6889	0.4909
cotton	0.0212	0.0017	12.7441	0.0000
rice	-0.0212	0.0113	-1.8782	0.0604

variable	estimate	std.error	statistic	p.value
wheat	0.0000	0.0106	0.0000	1.0000
lentils	0.0000	0.0106	0.0000	1.0000

Based on the balance tables, we can notice that the baseline profits and age seems to be balanced across FIONA and non-FIONA farmers. The same is true for the crops wheat and lentils too–farmer's treatment status is not associated with the farming of these two crops. However, there is imbalance in rice and cotton crops. All farmers who produce cotton seem to be included in the FIONA program. This might be due to the program being implemented only in the Thanjavur region which is the exclusive producer of cotton. These imbalances are problematic and confirms some of our concerns in (3). Furthermore, these can be indicative that there might be potential imbalances in unobservable characteristics as well.

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.

Answer

We would have to make two major assumptions which need to be true for these designs to be valid.

Firstly, we would need to assume that

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

where X_i are the observables. So in words, this would mean that we need to assume once we account for the observables in the data, potential outcomes are independent of the assignment of treatment (i.e. conditional independence).

Secondly, we would need to assume that:

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

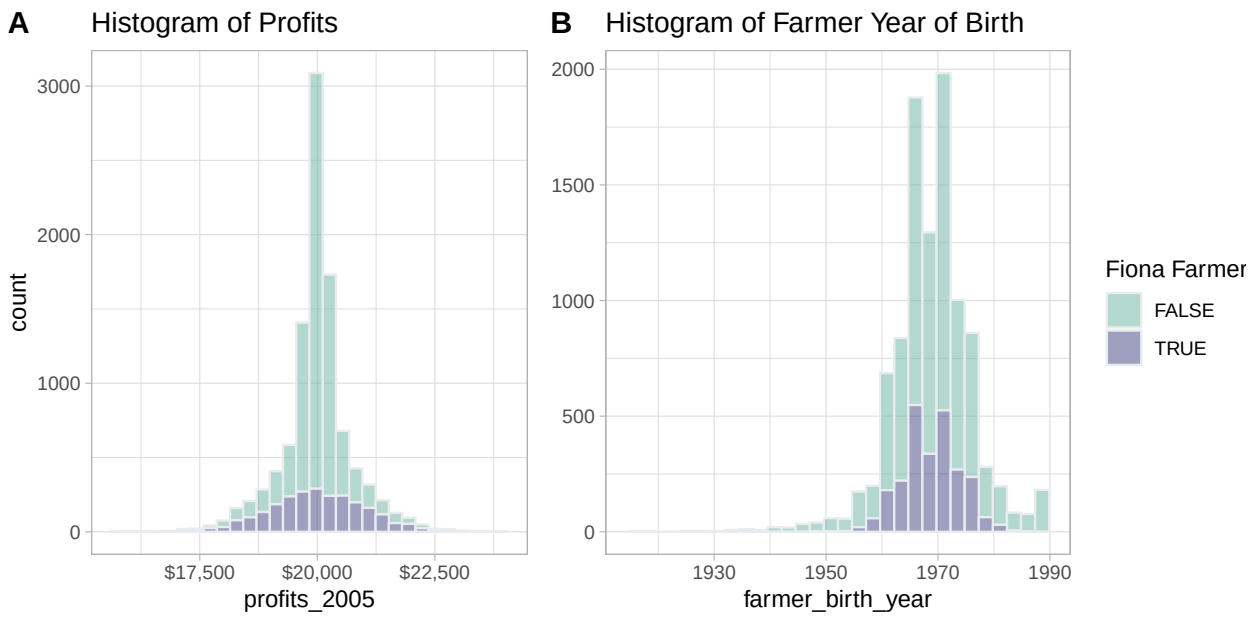
In other words, at each level of X, there are both treated and untreated units. We can test for this assumpton to see if it holds.

```
farm %>%
    group_by(crop, fiona_farmer) %>%
    summarise(n = n()) %>%
    knitr::kable()
```

crop	fiona_farmer	n
COTTON	1	53
LENTILS	0	2250
LENTILS	1	750
RICE	0	3000
RICE	1	947
WHEAT	0	2250
WHEAT	1	750

As we can see from the table above, all farmers who produce cotton crop are included in the FIONA program, therefore we would not be able to match those farmers with untreated counterparts.

```
library(cowplot)

hist_profit <- farm %>%
    ggplot(aes(x = profits_2005, fill = as.logical(fiona_farmer))) +
    geom_histogram(color = "#e9ecef", alpha = 0.5) + scale_fill_manual(values = c("#69b3a2",
    "#404080")) + theme_light() + labs(fill = "Fiona Farmer",
    title = "Histogram of Profits") + scale_x_continuous(labels = scales::dollar_format())

hist_yob <- farm %>%
    ggplot(aes(x = farmer_birth_year, fill = as.logical(fiona_farmer))) +
    geom_histogram(color = "#e9ecef", alpha = 0.5) + scale_fill_manual(values = c("#69b3a2",
    "#404080")) + theme_light() + labs(fill = "Fiona Farmer",
    title = "Histogram of Farmer Year of Birth") + ylab(NULL)

legend <- get_legend(hist_yob + theme(legend.box.margin = margin(0,
    0, 0, 12)))

plots <- plot_grid(hist_profit + theme(legend.position = "none"),
    hist_yob + theme(legend.position = "none"), labels = "AUTO")

plot_grid(plots, legend, rel_widths = c(1, 0.15))
```



Since profit is continuous, it is difficult to find both treatment statuses at each level, but if we relax the exact match criterion, we can find pairs of members of both groups at reasonable proximity. When it comes to year of birth, on the other hand, farmers who has been a part of the FIONA program seem to be more average aged. Both groups have a mean around 1970, but non-treated group tends to be more spread out across the years. This would be a problem for this assumption, as we would not be able to match possibly very experienced (or novice) non-treated members with treated counterparts.

We should also note that here I investigated these variables individually, it is more likely that we will run into

the curse of dimensionality when we look at these variables cumulatively (i.e. it will be even more difficult to find counterparts when we include the number of unobservables we use in splitting the data).

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?

Answer

```
farm_lm <- lm(profits_2016 ~ fiona_farmer, data = farm)
farm_lm_covs <- lm(profits_2016 ~ fiona_farmer + farmer_birth_year +
    profits_2005 + crop, data = farm)

stargazer(farm_lm, farm_lm_covs, type = "latex", column.labels = c("Naive Estimator",
    "Regression Adjustment"), header = FALSE)
```

I included all baseline variables besides `COTTON` dummy as it is the only one that does not appear within both groups, therefore it would not be possible to add it as a covariate to our model. I included imbalanced variables such as `WHEAT`, and `LENTILS` to account for the variation in these observables when estimating τ^{SOO}. I also added balanced variables like `farmer_birth_year`, `profits_2005`, and `RICE` to shrink the standard errors of our estimation. I did not include `fertilizer_use` as it is measured after treatment and can lead to bias in our estimation of $\hat{\tau}$ when included as a covariate.

The naive estimator and the regression adjustment show interesting consistency in results. Naive estimator indicates that there is $2,369.556 increase in farming profits after the implementation of the FIONA program. Whereas, the regression adjustment results with covariates indicate that the difference is $2,358.354. So there is roughly $11 difference between the two results–the naive estimator slightly overestimates.

Using this approach, we are able to account for the variation in our observables when estimating a treatment effect. But, due to selection on unobservables, it is possible that there may still be heterogeneous treatment effects that we cannot identify with the existing data. Without random assignment, we would need to make some critical assumptions to argue that this number is the population ATE.

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.

Table 8:

	Dependent variable:	
	profits_2016	
	Naive Estimator	Regression Adjustment
	(1)	(2)
fiona_farmer	2,369.556***	2,358.354***
	(29.977)	(24.346)
farmer_birth_year		0.193
		(1.464)
profits_2005		1.004***
		(0.014)
cropLENTILS		359.395**
		(146.087)
cropRICE		31.204
		(145.820)
cropWHEAT		428.280***
		(146.093)
Constant	21,942.350***	1,239.037
	(14.988)	(2,899.621)
Observations	10,000	10,000
R^2	0.385	0.601
Adjusted R^2	0.385	0.601
Residual Std. Error	1,298.041 (df = 9998)	1,045.628 (df = 9993)
F Statistic	6,248.246*** (df = 1; 9998)	2,507.267*** (df = 6; 9993)
Note:		*p<0.1; **p<0.05; ***p<0.01

Answer

To ensure that we do not lose most of our data while doing exact matching, I only include `crop` and `farmer_birth_year` variables in the matching process, if we include profits into the exact matching process, we struggle to find matches for most of the farmers in the dataset.

```
set.seed(1)

farm %>%
    matchit(fiona_farmer ~ crop + farmer_birth_year, method = "exact",
        data = .) %>%
    match.data() %>%
    group_by(subclass, fiona_farmer) %>%
    summarise(n = n(), mean_profit = mean(profits_2016)) %>%
    arrange(subclass, fiona_farmer) %>%
    ungroup() %>%
    pivot_wider(names_from = fiona_farmer, values_from = c(mean_profit,
        n)) %>%
    summarise(ATT = sum((mean_profit_1 - mean_profit_0) * n_1/sum(n_1)),
        ATN = sum((mean_profit_1 - mean_profit_0) * n_0/sum(n_0)),
        ATE = sum((mean_profit_1 - mean_profit_0) * (n_0 + n_1)/sum(n_0 +
            n_1))) %>%
    knitr::kable(caption = "Exact Matching Results")
```

Table 9: Exact Matching Results

ATT	ATN	ATE
2384.022	2365.609	2370.377

The difference between ATE (1) and ATT (2) is around $13.6. Since we know that treatment is not as good as random, even when we match records using the observables, it is expected that these two values are not the same. The results are different from what naive estimator and regression adjustment estimate (7) as well. Strength of this approach is the fact that it allows us to compare peers by matching on their observable characteristics, mitigating some (although not all) concerns of selection bias. Weaknesses of this approach is the fact that we cannot always exactly match treated units with untreated units when there are too many observables (curse of dimensionality) or when there are continuous observables. Also, even when we match on observables perfectly, it is possible that there are unobservables that result in selection bias which we cannot verify or mitigate using the data at hand. We can say that when we include more dimensionality (i.e. when we include `crop`, and `farmer_birth_year` variables into the exact matching process) the number of observations in the matched dataset decreases to 9,445 from 10,000. Although this was not a drastic decrease, it is possible that this decrease would become much larger with each additional observable.

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.

Answer

If there is evidence to validate that farmers in Bangladesh are not fundamentally different than the Indian (especially Thanjavur district) farmers included in this study (in terms of crops and other unobservables), I would recommend Bangladesh to implement a FIONA-like program to increase farmer profits. We see statistically significant positive values in each of the naive estimator, regression adjustment, and matching estimations. Although these methods have their downsides , as explained in the above sections, that could raise legitimate concerns about the external and internal validity of these estimations, unless there is highly critical unobservables left out from the dataset, it is likely that the effects of such a program on farmer profits will be net positive.