The Effect of Improved Maize Seeds and Fertilizers on Crop Yields

Assignment: Identification and Causal Inference

Jonas Schmitt, Sina Streicher, Owen Cortner, Philipp Kronenberg

February 18, 2021

1 Introduction

We examine the effect of a package of improved maize seeds and fertilizers on the harvested amount of maize of small-scale agricultural households. Maize is an important source of carbohydrates in the imaginary country, which we examine. The result of the study should evaluate if the package is manageable and effective with a limited amount of extension service and if it can help to improve the food self-supply situation of those households.

The package of improved maize seeds and fertilizers can be picked up for free at several supply station throughout the research area. Agricultural households have been informed in person by a group of young students about the process and where they can pick up the package. The information was spread by considering a federal list of all agricultural households in the study area. The total number of agricultural households in the list was 10.000. 5.000 households were randomly picked and received a voucher to pick up the seed + fertilizer package.

However, we observe that only a subgroup of agricultural households takes the treatment by picking up the seed + fertilizer package. We guess that e.g. the distance to the supply stations of a household, the education (as read/write dummy) or number of very young children ($<$ 10 years) influences the choice to pick up the package and that the selection of treatment is therefore not random.

Unfortunately, we have no information about the soil quality, which has an influence of the amount harvested even without any treatment. Therefore, we also face the issue of omitted variable bias.

a) Experiment

   In the experiment, we would randomly distribute the package of improved seeds + fertilizers to a treatment group of agricultural households e.g. by an address-list and measure the causal effect by comparing the harvested amount between treated and untreated households. The treated households additionally receive a voucher for farm equipment after the harvest, if they apply the package to their best of knowledge and without sharing it with their neighbours or any family members. Thereby, we want to avoid any spill-over effects between treatment and control group.

b) Optimal Data Set

   The optimal dataset would reveal all the relevant variables from an agricultural perspective (soil quality, topographic information, weather data on each farm etc.) and socioeconomic variables like income situation, education or family composition.

(Here: if we need to adjust, we can also assume that the soil quality of the study area has no differences between the farms. However, I thought

c) Threats to Identification

Omitted variable bias ⇒ control for all confounders, factors that determine assignment to treatment correlated with the outcome. The second challenge is the validity of the instrument (internal and external validity) ⇒ different weather effects during the experiment on the farms?

2 Data Generation and Model Specification

In this experiment we consider a sample consisting of 10.000 observations. Each observation represents one individual agricultural household, indicated by the subscript i. The outcome variable $harv_i$ represents the amount harvested (kg) by individual i. Our treatment $D_i \in \{0, 1\}$ is a binary variable that indicates whether individual i used the package of improved maize seeds and fertilizers or not. Therefore, $D_i = 1$ means that the individual received the treatment and $D_i = 0$ means no treatment. To control for other relevant variables, we collected the following covariates: $area_i$ is the cultivated area with maize by farmer (m^2), which is uniformely distributed between 100 and 5000 square meters. sun_i is the sun-hours during growing period (h), which is uniformly distributed between 400 and 600 hours. $rain_i$ is the total rainfall during growing period (mm), which is uniformly distributed between 600 and 850 millimetres. $child_i$ is the number of children under 10 years, which is uniformly distributed between 0 and 5. exp_i is the years of experience and is uniformly distributed between 1 and 35 years. $educ_i$ is a binary variable indicating whether individual i is able to write and read ($educ_i = 1$) or not ($educ_i = 0$). In this study, we assume that the variable $educ_i$ is unobserved.

I did not include soil quality yet!

In the following analysis we want to estimate the causal effect of the treatment on the outcome variable as described in the following equation:

$$Y_i = \boldsymbol{\mu}(0) + \boldsymbol{\Delta}_i D_i + U_i(0), \tag{1}$$

where Y_i is the outcome by individual, $\boldsymbol{\mu}(0) = Y_i(0)$ is the outcome by individual without treatment and $\boldsymbol{\Delta}_i$ is the causal effect of the treatment for individual i.

$$\begin{aligned} \boldsymbol{\Delta}_i &= Y_i(1) - Y_i(0) \\ &= [\boldsymbol{\mu}(1) - \boldsymbol{\mu}(0)] + [U_i(1) - U_i(0)] \\ &= \boldsymbol{E}\{\boldsymbol{\Delta}_i\} + [U_i(1) - U_i(0)], \end{aligned} \tag{2}$$

represents the causal effect of treatment for individual i, which is composed of the common gain and the idiosyncratic gain. As indicated by Imbens and Rubin (2015) these potential outcomes by individual i according to treatment can only be interpreted as causal effect, if the assignment between treatment and no treatment is random and there arises no selection bias such that we can compare outcomes of treated and untreated individuals with same characteristics as unobserved counterfactual. If we use this model in a regression, we receive the following causal equation

$$\begin{aligned} Y_i &= \boldsymbol{\mu}(0) + \boldsymbol{E}\{\boldsymbol{\Delta}_i\} D_i + U_i(0) + \boldsymbol{\Delta}_i [U_i(1) - U_i(0)] \\ &= \boldsymbol{\mu}(0) + \boldsymbol{\beta} D_i + \boldsymbol{\epsilon}_i, \end{aligned} \tag{3}$$

which only holds if $\boldsymbol{E}\{D_i \boldsymbol{\epsilon}_i\} \neq 0$, therefore we do not have exogeneity. Assuming we could observe all relevant covariates, our regression of the true model would look the following

$$harv_i = 400 + 0.05 area_i + 0.04 sun_i + 0.03 rain_i - 5 child_i + 5 exp_i + 10 educ_i + 100 D_i + e_i, \tag{4}$$

where $e_i \sim \mathcal{N}(0, 1)$. The coefficients are chosen such that the data generated is reasonable in a stylized real life scenario.

3 Assignment into Treatment

a) Two different assignments into treatment

Keeping the idiosyncratic treatment effect $\boldsymbol{\Delta}_i$ fixed, the assignment of individuals into treatment can play a major role in causal inference. Therefore, we need to discuss different cases. In the optimal case we would like to observe a random assignment into treatment. In that case we would observe that characteristics of treated and untreated individuals are balanced. The equation would look like following:

$$D_i^* \sim U(-1, 1), \qquad \text{and} \qquad D_i = \begin{cases} 1 & \text{if } D_i^* \geq 0 \\ 0 & \text{if } D_i^* < 0 \end{cases}. \tag{5}$$

However, there are several reason, why the assignment into treatment could be not random. For example, we expect that the distance to the supply stations of a household decreases the probability to take the treatment. Education (can read/write) is supposed to have an influence on farmers evaluation regarding the effectiveness / advantage of improved seeds + fertilizer package. Thus, a better education (as read/write dummy) would increase the probability to take the treatment. In this study, we assume that $educ_i$ is unobserved. Furthermore, the number of very young children (< 10 years) decreases the possibility to pick up the package and therefore decreases the probability to take the treatment. Therefore, the selection of treatment is not random. This leads to unbalanced characteristics of treated and untreated individuals. The data generation in the self-selection case can be described as followed:

$$D_i^* = 1 - 0.02dist_i + 2educ_i - 0.4child_i + u_i \qquad \text{and} \qquad D_i = \begin{cases} 1 & \text{if } D_i^* \geq 0 \\ 0 & \text{if } D_i^* < 0 \end{cases}. \tag{6}$$

b) Descriptive statistics of raw data

In the following table the mean values of the outcome and the covariates of the simulated raw data are illustrated by treatment status (treated and untreated) and assignment mechanism (random and self-selection).

Table 1. Summary Statistics

		Random Assignment			Self-Selection		
	Total	Treatment	No Treatment	P-Value	Treatment	No Treatment	P-Value
harv	677.25	752.78	651.75		978.75	624.12	
area	2549.94	2544.39	2551.82	0.82	2481.00	2566.55	0.02
sun	500.04	499.41	500.25	0.53	501.51	499.68	0.22
rain	725.10	725.33	725.02	0.85	724.28	725.30	0.58
exp	17.90	18.10	17.84	0.24	17.99	17.88	0.67
dist	50.59	51.51	50.28	0.06	39.25	53.32	0.00
child	2.50	2.52	2.50	0.54	1.92	2.64	0.00
educ	0.50	0.51	0.49	0.08	0.94	0.39	0.00
Obs	10000	2524	7476		1941	8059	

If treatment was randomly assigned, there is no significant difference in the mean of the covariates for the treated and untreated sample. Thus, a random assignment would allow us to identify the causal treatment effect by calculating the differences in the mean for the outcome variable between treated and untreated sample. Taking a look at the p-values of the unpaired two-samples t-test, we can see that the means of the covariates are mostly not significantly different from each other. However, in the self-selection case, we see that the variables $child_i$, $educ_i$, and $dist_i$ are not balanced between treated and untreated sample and the p-values show that the means for those variables are highly significantly different from each other. Since these variables have an effect on the outcome variable $harv_i$, the distribution of the outcome variable looks different as we can observe in the density plots.

why some of them still are?

why also some other variable have low p-values?

Figure 1: Density function

c) Model estimation

At this point we assume that we do not know the data generating process for the two assignment cases (random and self-selection) and we want to estimate the model for the two cases. The regression contains all the relevant covariates except for $educ_i$, which we assume to be unobserved. For both cases, we estimate the following model:

$$harv_i = \beta_0 + \beta_1 area_i + \beta_2 sun_i + \beta_3 rain_i - \beta_4 child_i + \beta_5 exp_i + \mathbf{\Delta}_i D_i + e_i. \tag{7}$$

The regression perfectly fits the data generating process and estimates the treatment effect correctly, even though the unobsreved variable $educ_i$ is left out, since it is balanced across the two treatment groups. In the case of the self-selection we obtain the following results:

Here, we can see a large bias. The treatment effect is estimated much larger than it is in real. This is due to the omitted variables leading to self-selection. For the variable $educ_i$

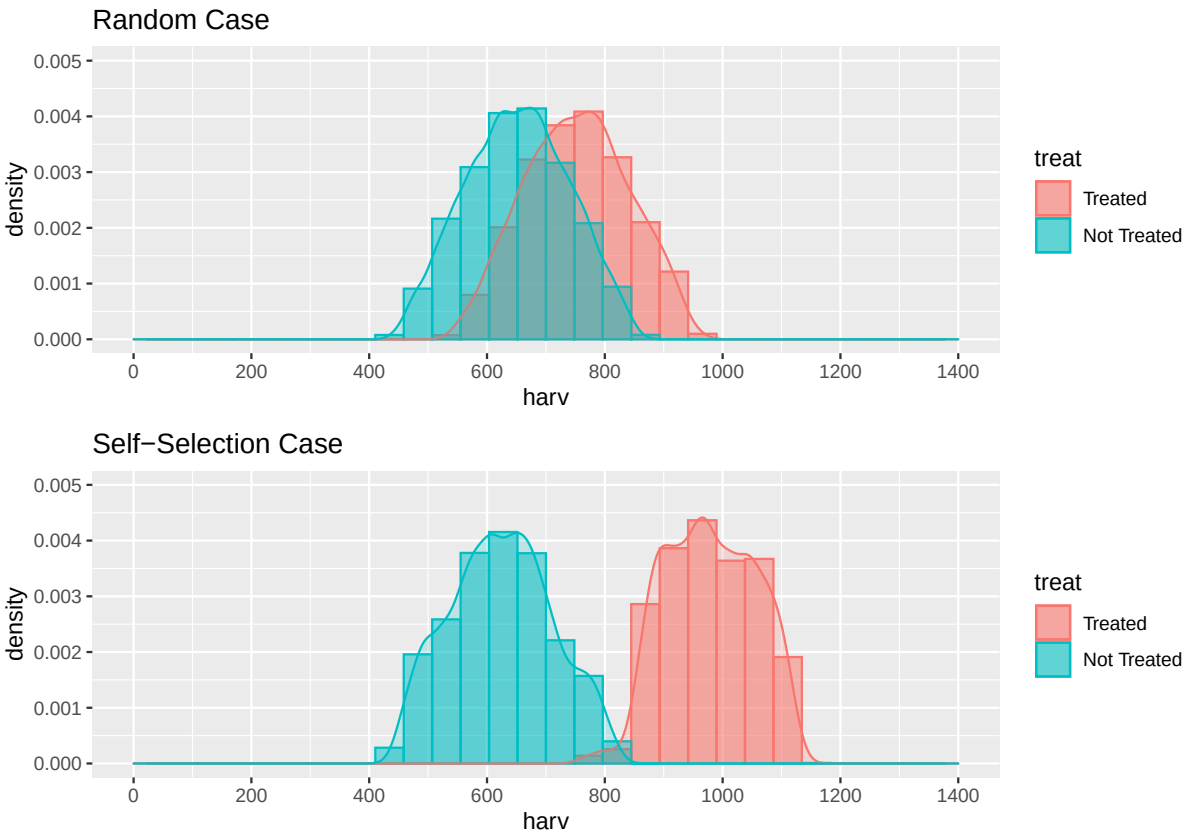

Figure 2: Density function grouped by treatment

Table 2. OLS Regression Random Case

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	405.7869	0.6809	595.99	0.0000
area	0.0500	0.0000	1418.31	0.0000
sun	0.0412	0.0009	48.01	0.0000
rain	0.0291	0.0007	42.07	0.0000
exp	4.9963	0.0051	982.34	0.0000
child	-4.9508	0.0334	-148.06	0.0000
treat	100.2027	0.1151	870.66	0.0000

Table 3. OLS Regression Self-Selection Case

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	433.6133	6.0982	71.11	0.0000
area	0.0499	0.0003	158.18	0.0000
sun	0.0472	0.0077	6.14	0.0000
rain	0.0235	0.0062	3.79	0.0002
exp	1.4605	0.0455	32.09	0.0000
child	-1.5903	0.3048	-5.22	0.0000
treat	357.5306	1.1528	310.15	0.0000

we expect the omitted variable bias to be positive, since it is positively correlated with the treatment and the output variable.

4 Selection Bias

However, we expect that e.g. the distance to the supply stations of a household, the education (as read/write dummy) or number of very young children (< 10 years) influences the choice to pick up the package and that the selection of treatment is therefore not random. Thus, self-selection into treatment depends on the unobserved variable education. As a result, the model suffers from omitted variable bias, since the treatment is not independent of the error term, $Cov(D_i, e_i) = 0$. As a result, the conditional independence assumption (CIA) is not fulfilled and a causal interpretation of the coefficients is not possible.

For this reason, we follow an alternative instrument variable approach following

$$D_i^* = \boldsymbol{\alpha} + \boldsymbol{\beta} Z_i + V_i. \tag{8}$$

As instrument variable we use a random assignment to receive a voucher to pick up the seed + fertilizer package. $vouch_i$ is a binary variable indicating whether individual i received a voucher ($vouch_i = 1$) or not ($vouch_i = 0$). This results in the equation

$$D_i^* = \boldsymbol{\alpha} + \boldsymbol{\beta} vouch_i + v_i. \tag{9}$$

where $v_i \sim \mathcal{N}(0, 1)$. This random assignment of the voucher is correlated with treatment, since we expect a positive share of individuals receiving the voucher will pick up the package and only individuals receiving the voucher are allowed to pick up the package. Furthermore, the instrument is correlated with the outcome variable, since receiving the treatment should have an effect on output. Finally, the instrument is uncorrelated with the error term $\boldsymbol{\epsilon}_i$ since the assignment of receiving the voucher is randomly picked. As a result, the following holds

$$\boldsymbol{E}\{\, U_i(1)\} = \boldsymbol{E}\{\, U_i(0)\} = \boldsymbol{E}\{\, V_i(1)\} = 0. \tag{10}$$

5 Difference in Difference

References

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.