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 (m2), 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. This is relevant since more children affect the time that can be spend on the field. 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 ($educ_i = 0$).

Amount harvested (kg): $harv_i \sim$ U(30, 800)
Cultivated area with maize (m2): $area_i \sim$ U(100, 5000)
Sun-hours during growing period (h): $sun_i \sim$ U(400, 600)
total rainfall during growing period (mm): $rain_i \sim$ U(600, 850)
Number of children under 10 years: $child_i \sim$ U(0, 5)
Years of experiences: $exp_i \sim$ U(1, 35)
Education: $educ_i \sim$ U{0,1}, where 1 is good and 0 bad reading and writing skills
Distance to the supply stations of a household (km): $dist_i \sim$ U(1, 100)

⇒ Idea: more children influence the time, which can be spend on the field to look after the plants; can change in both direction in 5.)

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 Rubin(1970?) 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

⇒ increases in 5.)

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 would look the following

$$harv_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 area_i + \boldsymbol{\beta}_2 sun_i + \boldsymbol{\beta}_3 rain_i - \boldsymbol{\beta}_4 child_i + \boldsymbol{\beta}_5 exp_i + \boldsymbol{\Delta}_i D_i + e_i. \tag{4}$$

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. Unobserved: Soil quality, topographic information OVB
For this reason, we follow an alternative instrument variable approach following

$$D_i^* = \boldsymbol{\alpha} + \boldsymbol{\beta} Z_i + V_i. \tag{5}$$

As instrument variable we use the 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

$$D_i^* = \boldsymbol{\alpha} + \boldsymbol{\beta} vouch_i + V_i. \tag{6}$$

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

$$\boldsymbol{E}\{U_i(1)\} = \boldsymbol{E}\{U_i(0)\} = \boldsymbol{E}\{V_i(1)\} = 0. \tag{7}$$

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)

$$D_i = \begin{cases} 1 & \text{if } D_i^* \geq 0 \\ 0 & \text{if } D_i^* < 0 \end{cases}$$

$$D_i^* = \gamma_0 - \gamma_1 dist_i + \gamma_2 educ_i - \gamma_3 child_i + \gamma_4 exp_i + u_i. \tag{8}$$

$$D_i = \begin{cases} 1 & \text{if } D_i^* \geq 0 \\ 0 & \text{if } D_i^* < 0 \end{cases}$$

However, there are several reason, why the assignment into treatment could be not random. For example, ... This leads to unbalanced characteristics of treated and untreated individuals. Education (can read/write) $\sim$ U(1, 0) $\Rightarrow$ is supposed to have an influence on farmers evaluation regarding the effectiveness / advantage of improved seeds + fertilizer package.

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	Treatment	No Treatment
harv	651.95	727.03	626.61	890.82	577.89
area	2549.94	2544.39	2551.82	2507.06	2572.54
sun	500.04	499.41	500.25	500.68	499.70
rain	725.10	725.33	725.02	725.39	724.95
exp	17.90	18.10	17.84	17.97	17.87
dist	50.59	51.51	50.28	48.02	51.94
child	2.50	2.52	2.50	1.90	2.82
educ	0.50	0.51	0.49	0.55	0.47
Observations	10000	2524	7476	3451	6549

c) Model estimation

At this point we assume that we know the data generating process for the two assignment cases (random and self-selection). In the case of random assignment we obtain the following regression:

In the case of the self-selection we obtain the following regression:

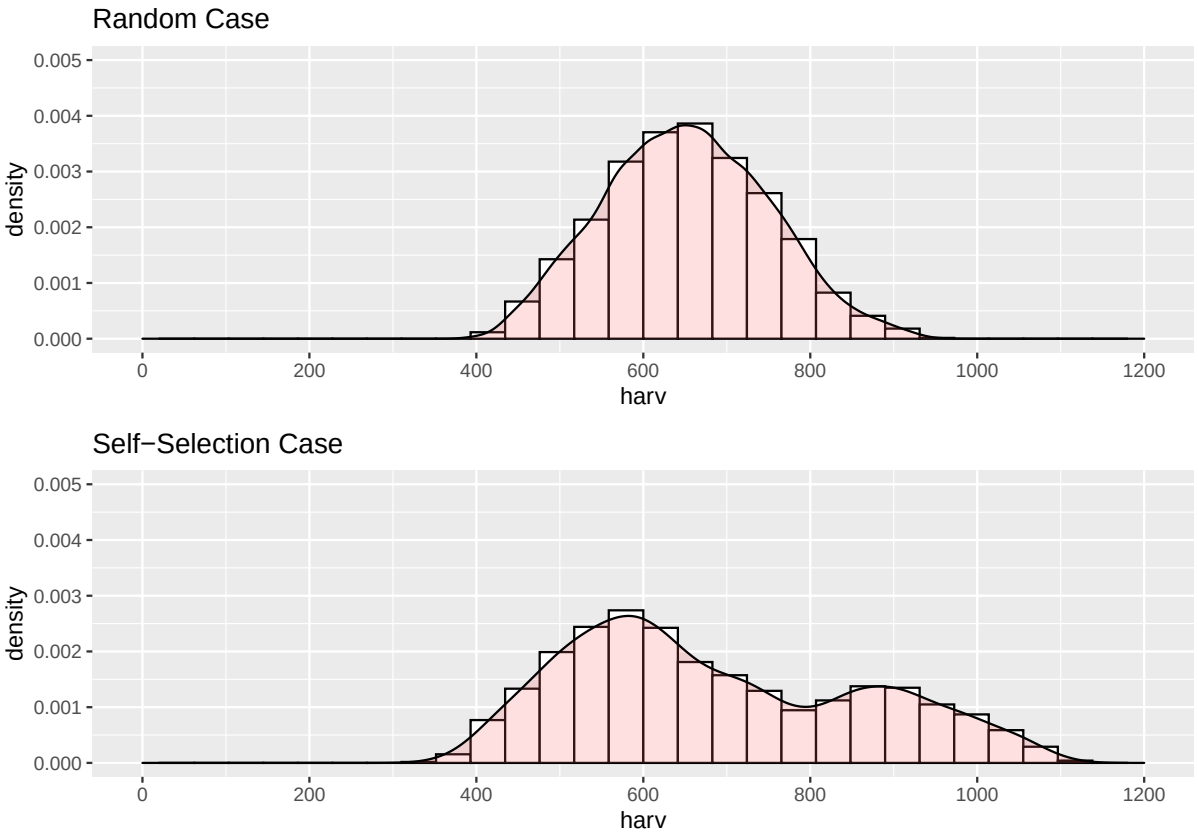

Figure 1: Density function

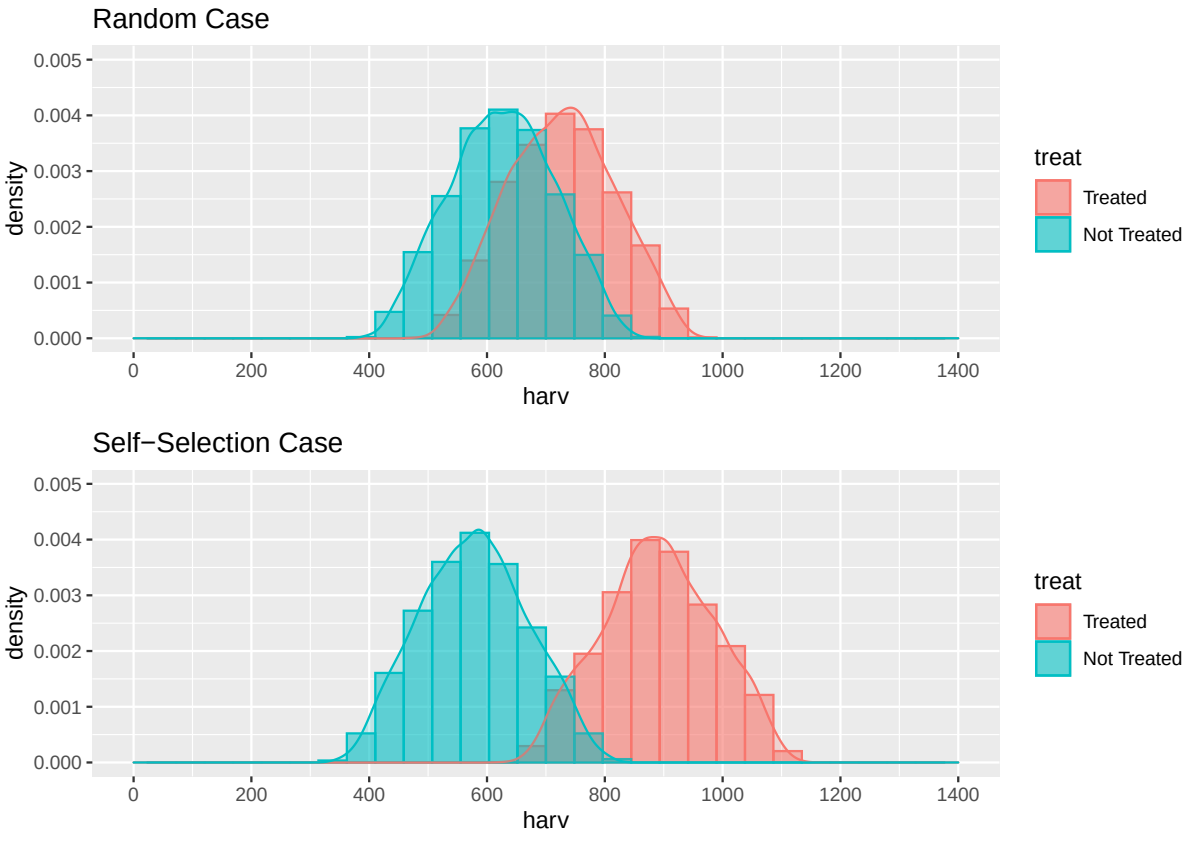

Figure 2: Density function grouped by treatment

4 Selection Bias

text...

5 Difference in Difference

Test citation:
bla bla Foroni et al. (2011) bla bla.

References

Foroni, C., Marcellino, M., and Schumacher, C. (2011). U-MIDAS: MIDAS regressions with unrestricted lag polynomials. Discussion Paper Series 1: Economic Studies 2011,35, Deutsche Bundesbank.