Causal Inference

Joaquin Alvarez Galnares

March 2023

Abstract

This report corresponds to the second assignment of Causal Inference, taught by Prof. Mauricio Romero. The associated code for replication and data can be found **here**. We explore causal inference in the context of public health (COVID 19) and public education (standardized exams).

1 Education experiment

We want to consider the effect of increasing the salary of the teachers on education outcomes (e.g. by measuring the performance of students in standardized tests). There are various alternatives for conducting the experiment which we should discuss and consider.

1.1 Designing the experiment

a) Consider picking teachers randomly across all schools and giving them an increase in their salary. Then for a given school, perhaps some teachers would receive an increase in their salary and some won't. SUTVA assumption requires no externalities and homogenoeus dosage. The problem is that if we consider outcomes (Y_i) to be the average score of a standardized test for students per school we would not capture the overall effect of the treatment. For example suppose that in school A the Physics teacher gets an increase in her salary, but the Math and Geography teachers don't. And suppose that the standardized test consists of questions regarding Physics, Math and Geography. Then the average of the standardized test for students in school A won't allow us to capture the effect of the increase in the salary for the teacher in Geography, since the average is affected by other subjects. Externalities could also be a problem: for example, the Math and Geography teacher styles (hence, potentially student grade outcomes) may change if they find out that their salaries didn't increase while (unfairly) the Physics teacher salary were doubled. Moreover, we would be ignoring the correlation in outcomes: the grades of students form the same school would typically be correlated.

b) Consider picking schools randomly and giving a a salary increase (treatment) to *all* the teachers in the chosen schools. The idea of clustering (stratifying) is key here. The SUTVA assumption basically implies that there are

no externalities, meaning that if an individual is treated then that doesn't affect the outcome for other individuals. Also, it implies homogeneous dosage, meaning that all treated units are treated are given the exact same treatment. Now, if we pick schools randomly and give *all* the teachers in those schools the exact same increase in their salary, it is fair to say that the SUTVA assumption is satisfied: if all professors at school A get a salary increase, that would not affect the performance of students in school B. Moreover, the salary increase (size/amount of treatment dose) is something we control, since we decide how much money we increase the salary to the teachers. Hence we can make the salary increase (the treatment) homogeneous across all treated units.

c) Taking into account the arguments above, I would prefer implementing the experiment by assigning the treatment at the school level (for the randomly chosen schools, treating all the professors in the school) because a stratification in such a way would allow us to consider clustered errors and satisfying SUTVA. Clustered errors are important because student grades are not independent for each school!

1.2 Assigning schools randomly. Would it be a useful strategy to compare treated against non-treated schools after some years of implementing the policy?

I believe that if we compared treated against non-treated schools after some years of increasing the salary of teachers in the treated schools we could not indicate whether the policy was effective or not, because the long-term behaviour of the teachers could make them be just as they were before they received a salary increase. That is, the duration of the effect of increasing the salary of a teacher in his/her behaviour may be short. For example, it is reasonable to think that during the first months after the teacher received an increase in his salary he may be happier and more motivated to teach (possibly leading to better grades in standardized tests of his students). But after some time, the initial motivation of the teacher may fade. I think workers tend to be ambitious and unsatisfied with their salaries. So an increase in the salary only produce short-term satisfaction.

1.3 Regression model

There are various alternetives we can consider. Here we present 2 as a matter of example.

1.3.1 Correlated scores within students of the same school

We assume heteroskedasticity (heterogeneous variance) and correlated scores betweeen students from the same school. To achive somethign like that we propose a model of the form

$$Y_{i,j} = \beta_0 + \beta_1 X_{i,j} + \beta_2 Tj + \omega_j + \epsilon_{i,j} \quad ,$$

where $Y_{i,j}$ corresponds to the grade (outcome) of the i-th student in the j-th school.

$X_{i,j}$ can be some control variable such as the age or the gender or the current GPA of the i-th student in the j-th school.

$$T_j = \begin{cases} 1, & \text{if j-th school is treated} \\ 0, & \text{in other case} \end{cases} \quad , \quad j \in \{1, \dots, J\}$$

ω_j is an error that affects all the members of the j-th school and considers correlation between individuals of the same school, $\mathbb{V}(\omega_j) = \tau^2 \quad \forall j \in \{1, \dots, J\}$. This variable accounts for clustering the data, and introduces correlation in the scores obtained by students form the same school.

$\epsilon_{i,j}$ an individualized error for the i-th student in the j-th school, $\mathbb{V}(\epsilon_{i,j}) = \sigma^2$

1.3.2 Heteroskedasticity but non-correlated errors

An alternative model, for example, is

$$Y_{i,j} = \beta_0 + \beta_1 X_j + \beta_2 Tj + \epsilon_{i,j} \quad ,$$

where $Y_{i,j}$ corresponds to the grade (outcome) of the i-th student in the j-th school.

X_j is the control variable. For example, we can use X_j the marginalization score of school j.

$$T_j = \begin{cases} 1, & \text{if j-th school is treated} \\ 0, & \text{in other case} \end{cases} \quad , \quad j \in \{1, \dots, J\}$$

$\epsilon_{i,j}$ an individualized error for the i-th student in the j-th school, $\mathbb{V}(\epsilon_{i,j}) = \sigma_j^2$. We assume the same variance in the score for students in the same school (clustered errors!).

Notice that for a fixed given marginalization score x_j and students from the same school j, we have,

$$\mathbb{E}[Y_{i,j}|T_j = 1] - \mathbb{E}[Y_{i,j}|T_j = 0] = \beta_0 + \beta_1 x_j + \beta_2 + \underbrace{\mathbb{E}(\epsilon_{i,j})}_{=0} - \beta_0 - \beta_1 x_j - \underbrace{\mathbb{E}(\epsilon_{i,j})}_{=0} = \beta_2$$

Hence, $ATE = \mathbb{E}[Y_{i,j}|T_j = 1] - \mathbb{E}[Y_{i,j}|T_j = 0] = \beta_2$.

We verify this in the code by running some simulations!

1.4 Implementing a model with data from Mexico's PLANEA exam

First we consider using a linear regression model with the classical assumptions: non-correlated and homoscedastic errors. As we'll see, the performance of the model has limitations when the data has a clustered structure.

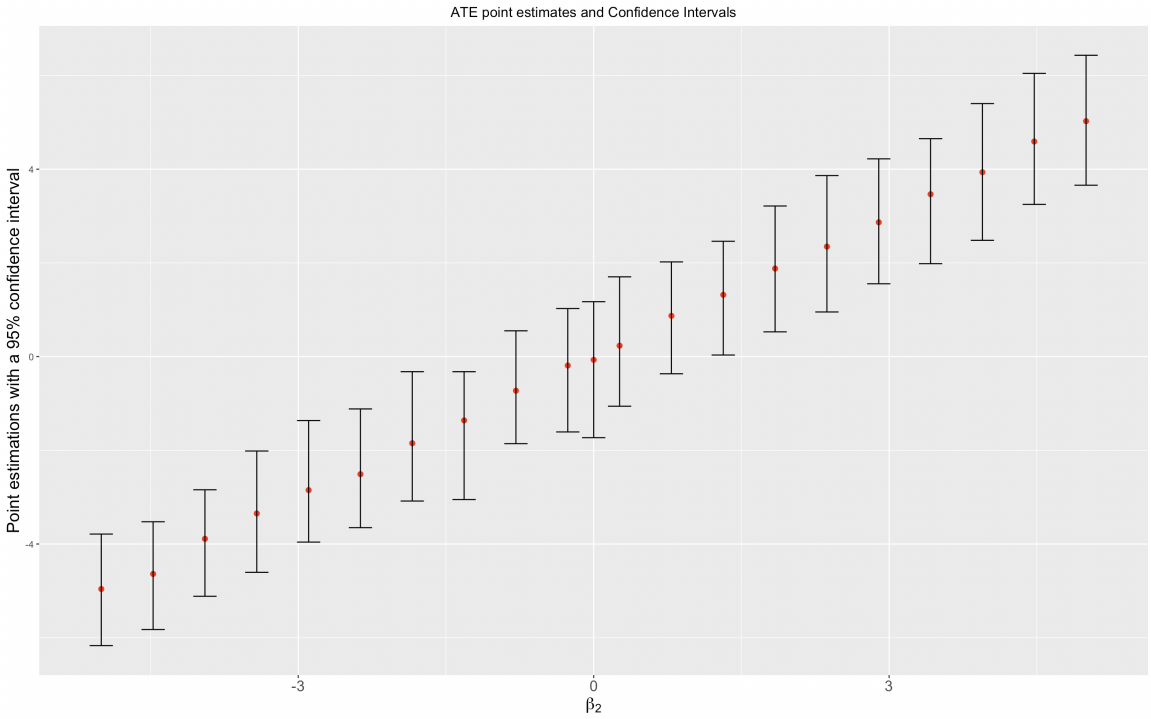

Figure 1: Unbiasedness of the OLS estimator despite the fact that the assumptions of the distribution of the errors aren't satisfied.

Something we should notice that when the ATE is zero the Confidence Interval at 95% contains the 0. So in terms of this perspective, which is a global one that contemplates the distribution of the estimator, we shouldn't reject the null hypothesis. However on practice we make decisions without the access to so many simulations (unless we use fancy techniques such as the bootstrap). And without knowing the variance of the estimators or without good estimators of the variance of the estimators, we could be making lots of type one errors (wrong rejections of the null) , as we see in the following figure.

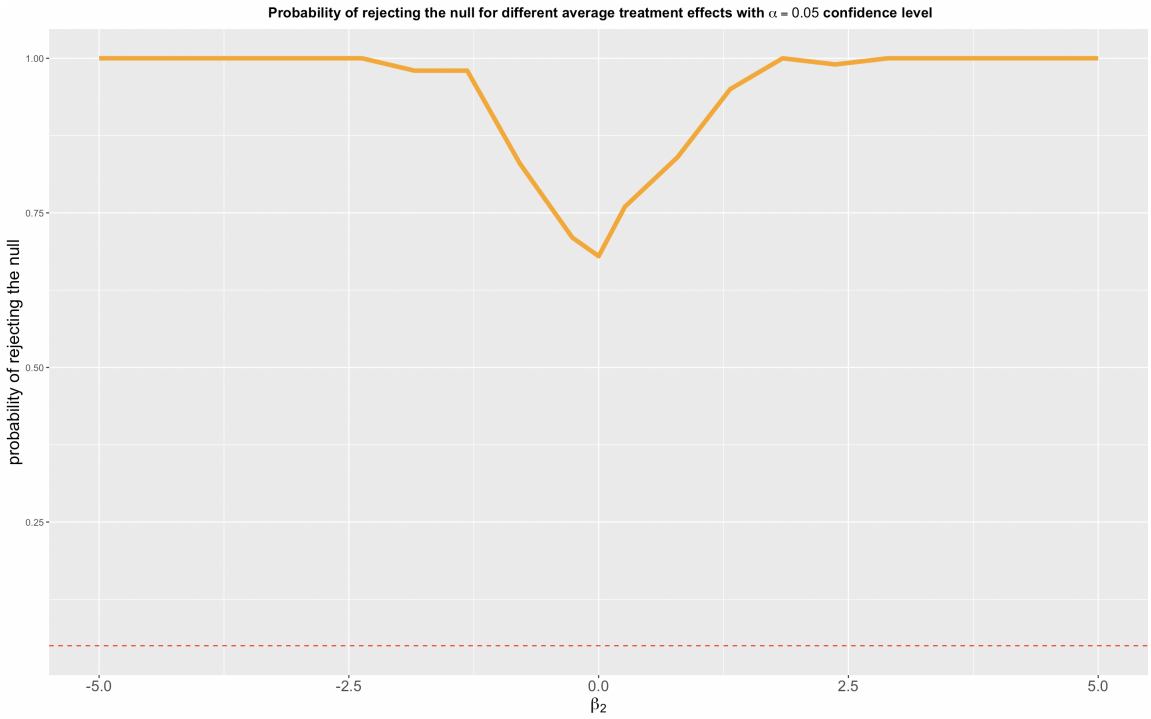

Figure 2: Dashed red line corresponds to the $\alpha = 0.05$ confidence level we pretended to achieve using a classic linear regression model to fit the data.

We see that using classic OLS leads to VERY non-conservative decisions. When the ATE is equal to zero we should be rejecting the null (concluding that there's a treatment effect) around 5 % of the times. However we see that we are rejecting the null 68% of the times. Why such a drastic difference? Well, a reasonable explanation is that we are making decisions with a model that is not appropriate for the data. The data generating process was based on correlated scores for students in the same school and heteroscedasticity. While the classical linear regression model assumes homoscedasticiy and no correlation between the errors.

This is something that was actually discussed in class. The **blog post** form Markus Konrad remarkably confirms what we are observing:

> In many scenarios, data are structured in groups or clusters, e.g. pupils within classes (within schools), survey respondents within countries or, for longitudinal surveys, survey answers per subject. Simply ignoring this structure will likely lead to spuriously low standard errors, i.e. a misleadingly precise estimate of our coefficients. This in turn leads to overly-narrow confidence intervals, overly-low p-values and possibly wrong conclusions.

This is particularly true when $\beta_2 = 0$. We are rejecting the null more often than we should when using a significance level of $\alpha = \frac{5}{100}$.

This reflects how important it is to understand the assumptions behind behind models and asking ourselves if those assumptions are satisfied in the data.

Next we present the results when implementing a linear regression model that accounts for the fact that the data has a clustered structure.

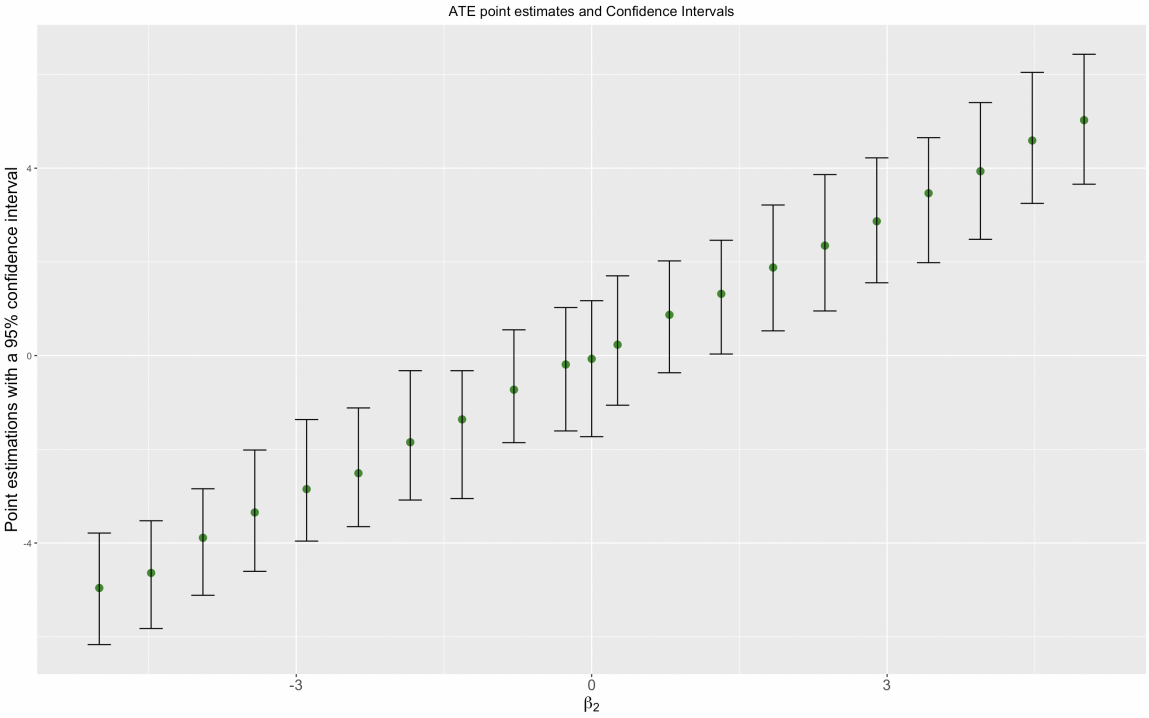

Figure 3: Point estimates with a confidence interval using a linear regression model with clustered data.

We see that the estimations are still unbiased. For values of β_2 that are close to zero out confidence intervals include the 0. If we were to make decisions based on these intervals, we see that we would not reject the null for values of β_2 that are very close to zero. Something interesting is that the length of the intervals is almost constant across the different values of β_2.

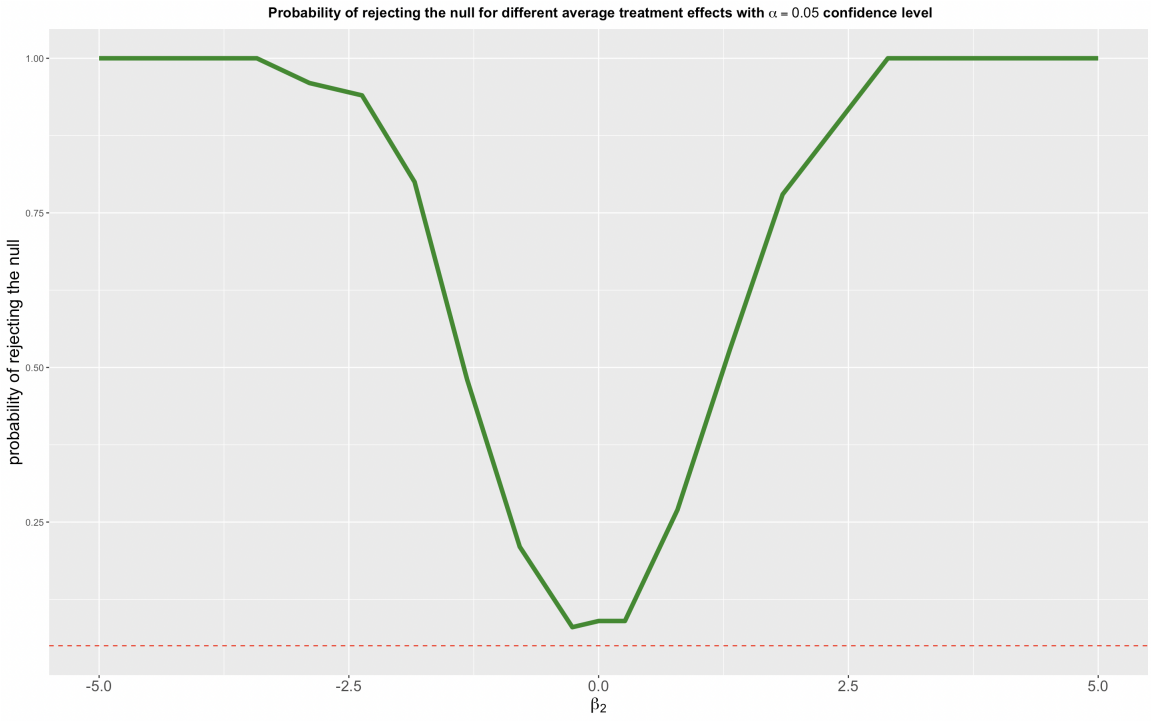

Figure 4: These are the probabilities of rejecting the null hypothesis (which is that the treatment effect is zero) using a linear regression model that contemplates clustered observations. The dashed red line corresponds to the $\alpha = 0.05$ confidence level we pretended to achieve and is the one we used to simulate rejections of the null in our simulations.

We see that the coverage when $\beta_2 = 0$ got way better: when using clustered variance, we obtained a probability of rejecting the null hypothesis that is close to the significance level that we chose (this is true coverage: if the probability of making a false positive is 0.05 we expect to reject the null 5% of the times when $\beta_2 = 0$). Hence, this model has makes a smaller Type-I error (probability of rejecting the null when it is true) compared to the simple/classic linear regression model.

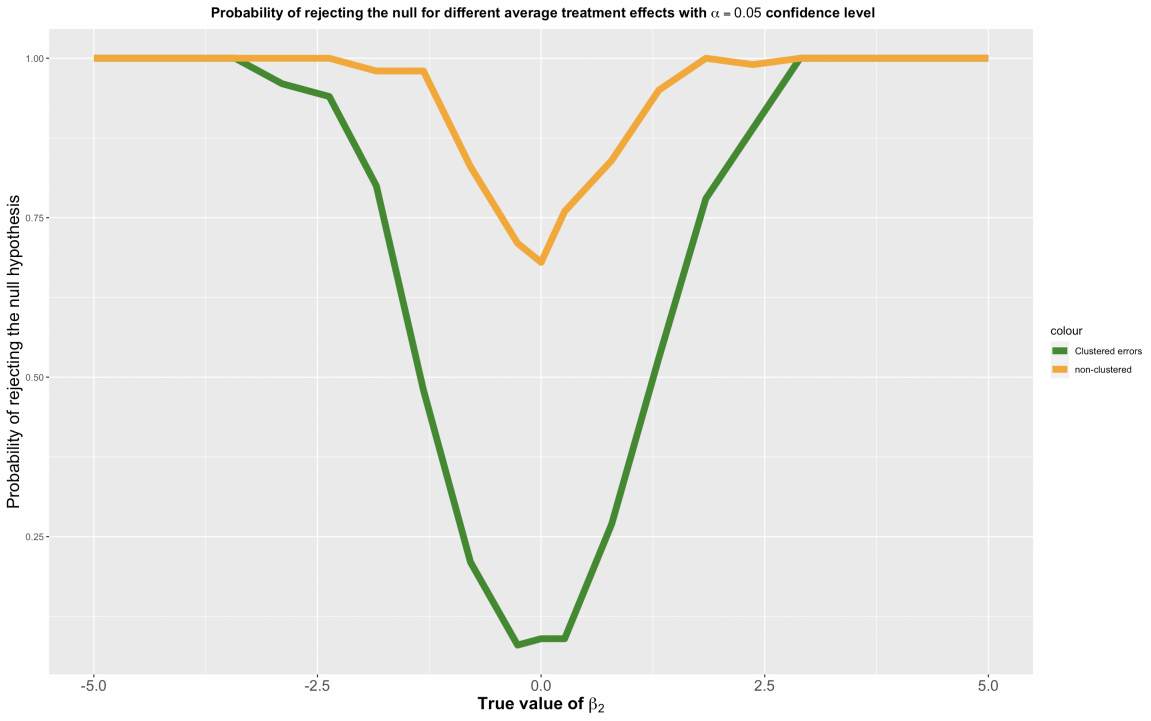

Figure 5: Comparison of the power using two different models.

It is true that with clustered errors the Type I error gets reduced. However that comes at the expense of bigger type II errors for values of β_2 such that are close to zero and $\beta_2 \neq 0$, e.g. in a region $\beta_2 \in (-2.5, 0) \cup (0, 2.5)$ we observe that the probability of rejecting the null is smaller when using a linear regression model with clustered errors. This sort of trade-off between Type-I error and Type-II errors occurs very often in Machine Learning. In many applications we see this phenomena: making less type-I errors usually comes at the expense of making more Type II errors.

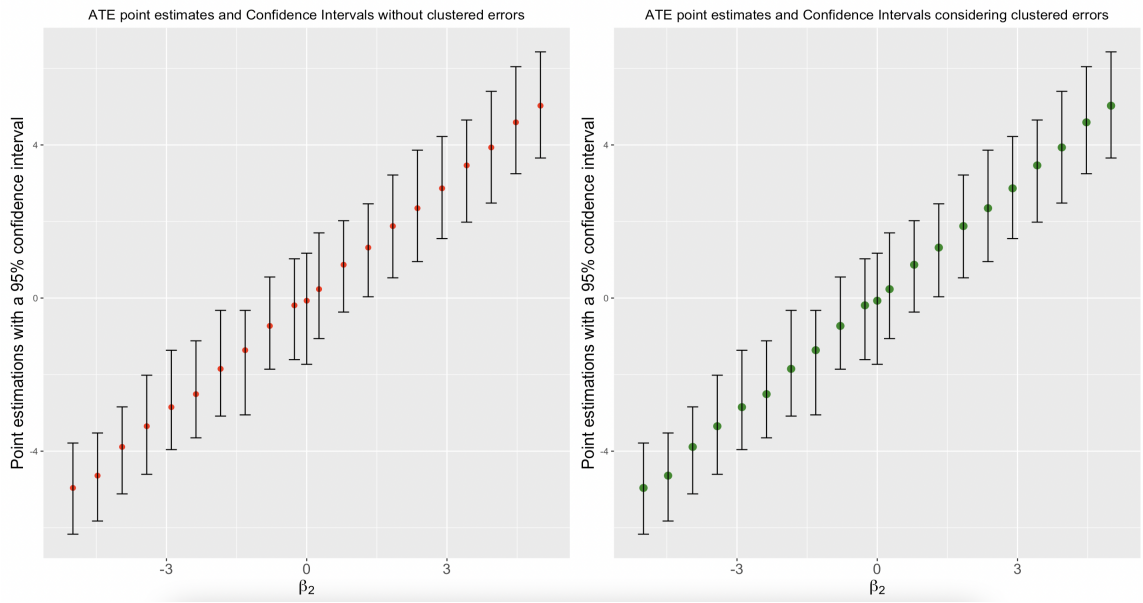

Figure 6: Unbiasedness is independent of the assumption of the distribution of the errors.

Under both models we observe pretty confidence intervals in terms of their length. Both plots look almost identical.

1.5 Inference via randomization experiments

We consider a model of the form $Y_{i,j} = Y_{0,i,j} + \delta_{i,j} T j$, where $\delta_{i,j}$ is the treatment effect. $Y_{0,i,j}$ is the score in the math test obtained by the $i-th$ student in the $j-th$ school in the absence of treatment. Our null hypothesis will be that the treatment effect is zero. As a test statistic we will use the difference in means in the math test score between the treated and the controls. We implemented some simulations and obtained the following results

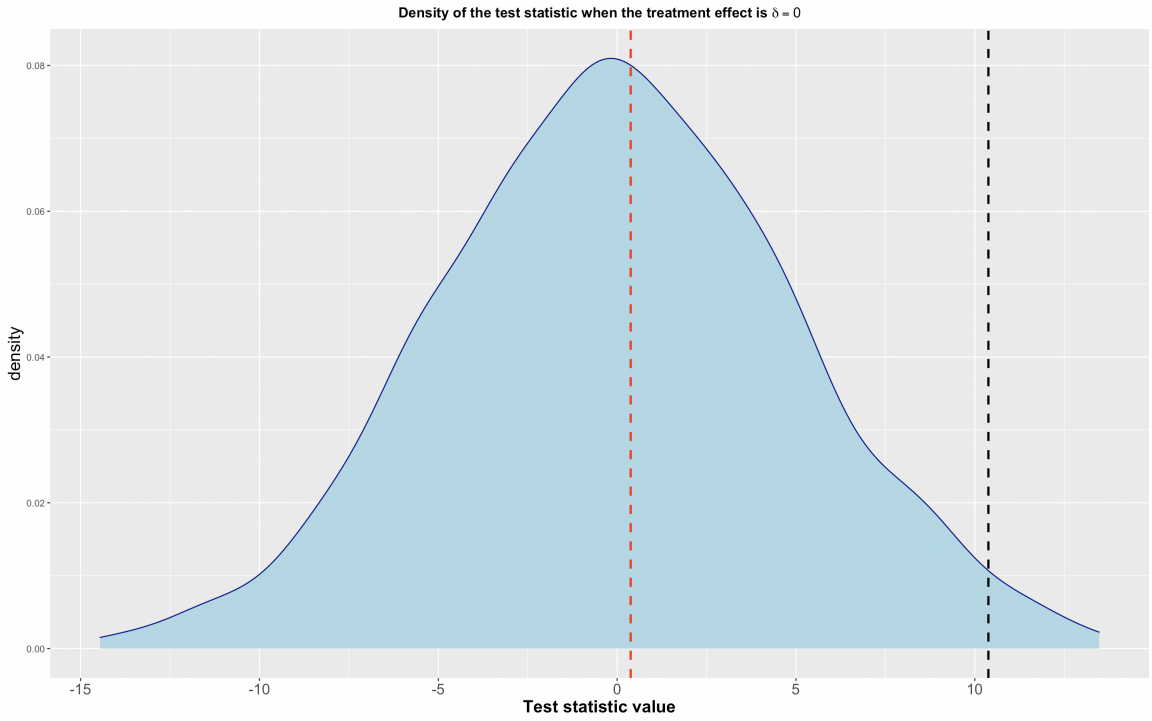

Figure 7: Dashed lines correspond to simulated observations of our test statistic with different treatment effects. The red line was an observation we simulated of the test statistic when the treatment effect is zero. The black line is a simulated observation of the test statistic when the treatment effect is $\delta = 10$.

The p-value we obtained for the simulation under the null was 0.459; hence we wouldn't reject the null hypothesis using a significance level of $\alpha = 0.05$. The p-value we obtained by simulating an experiment with an average treatment effect of $\delta = 10$ was 0.02. Under this setting we would reject the null using a significance level of $\alpha = 0.05$. The way to interpret each of these p-values visually is that they are the area under the the blue density curve accumulated at the right of each dashed line respectively (the right tail of the distribution).

How to interpret these results in the context of the education problem? Well, if an increase in the salary of the teachers leads to an average increase in 10 points in the math score, then we will likely be able to discover a positive effect of the policy and find that the salary increase has positive effect in students' math performance. This is also the type of reasoning that we use to interpret the following figure.

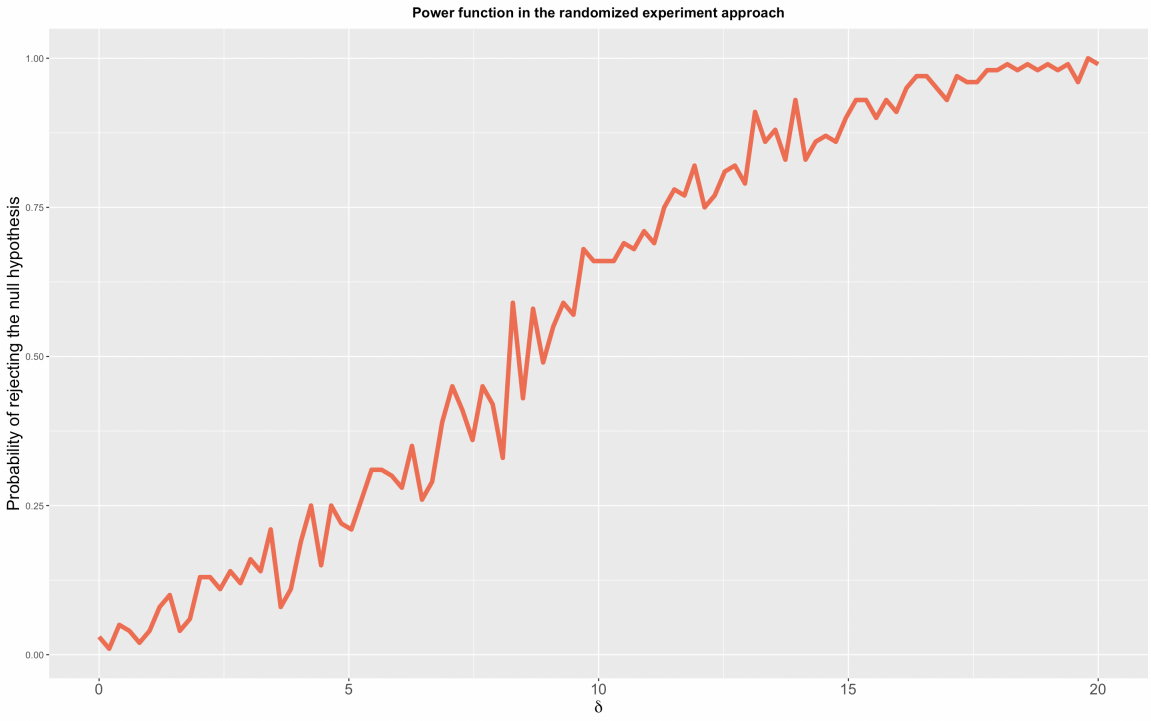

Figure 8: Point estimates for the probability of rejecting the null hypothesis for different values of the treatment effects, using a significance level $\alpha = \frac{5}{100}$.

As expected, the probability of detecting a treatment effect increases when the treatment effect gets bigger. Intuitively this is because for large values of the average treatment effect, it becomes more obvious in the data that there is a positive treatment effect, since students whose teachers received a salary increase start getting way better grades than students whose teachers didn't get the salalary increase.

2 Ivermectina & COVID

2.1 Possible problems when the person who evaluates a program is the same as the person who created and implemented it

I would say the main problem with that is a conflict of interest: the person who evaluates and judges the program has a benefit from the good evaluations of it.

2.2 The data

The main issue with the data is ambiguity. No specification for the population that was considered. The data set was formed from different sources of data. But how were those original data sets constructed. How was the population sampled and under which methodology? Is it valid to just fusion the data from different sources without rigorous considerations? The **original paper** talks about it, but not with sufficient detail. The methodology section of the report doesn't mention whether people were vaccinated or not. Nor if they received a vaccine before or after the kit treatment. For some variables, we can use common sense to deduce what they measure/ what they mean. But other variables don't have a trivial interpretation. So what are we supposed to do with them? There's no dictionary to clarify the variables in the data set. There's no specification about the use of the kit (the 'treatment') and the methodology which treated people followed to use it. Hence, we can not conclude "homogeneous dosage", which is a key assumption when studying causality in these type of settings. Even if the kits are all the same, perhaps there are differences in how each person uses it. Confoudedness is something we should also question.

2.3 Checking selection bias

We want to determine whether there's selection bias or not. For example, if we find that 90% of the treated units are men, whereas 50% in the global data set are men, then this would represent a bias in the treated units, since the treated units would not respect the proportions of the original set, giving rise to the question: are men less likely to get hospitalized for COVID while the treatment is just a highly correlated variable with gender? We explore such type of things in the R script in the Git repo.

2.4 Exploring some regression models to gain an insight about the data

Comments and implementation of some models are available **here**

2.5 By definition of the kit, it included things apart of Ivermectina

This represents a huge problem if the purpose is to conclude that ivermectin reduces chances of hospitalization. Mainly because we cant attribute the whole effect of the treatment to one of the components of th kit. Maybe there's another object on the kit that's the main responsible for the effect on lower hospitalization rates on treated individuals.

2.6 Any other issues?

According to the FDA, which is an important authority in the topic, ivermectin has other uses that are not directly related to COVID and there's no scientific evidence that supports the use of the tablets for healing/preventing COVID. So it's not only the considerations that we've pointed out about the data set and the underlying methodology presented in **the paper**. It's also a representative authority in the business of health pointing out a huge flaw in 'scientific findings' about ivermectin.

3 Summary of articles

Ivermectin is being used to treat COVID in various places. However there's no scientific evidence that it is useful for those purposes. The article cites a study that claims to give scientific evidence that ivermectin is effective Such study says in its conclusions: "Early addition of Ivermectin to standard care is very effective drug for treatment of COVID-19 patients with significant reduction in mortality". However, the Medium article claims that such affirmations are misleading because there were wrong uses of statistical procedures and a quite ambiguous methodology, just as the report we previously discussed for the case of Mexico. The data has considerable date mismatches!

The paper titled *Ivermectin for preventing and treating COVID-19* published by the Cochrane Library implemented a more formal/detailed methodology in order to determine if ivermectin is an effective treatment against COVID. It makes an analysis based on a compilation of studies that investigated the impact of ivermectin in COVID patients. After conducting causal experiments and analyzing research publications on the topic, the authors concluded that they don't have enough evidence to conclude that ivermectin is effective. Hence, they emphasize that most serious studies on the topic do not support the use of ivermectin as a treatment to prevent or treat COVID. The authors mention that they don't know if ivermectin leads to fewer deaths compared to a control group. This study also emphasizes the importance of taking differences between studies into consideration when comparing them: size of doses is not the same in every study, types of participants differ, interventions and methodologies also have significant differences. The authors evaluated other studies taking into consideration the quality of the evidence, meaning that if the studies followed a rigorous methodology, with unambiguous data and specified procedures then such studies were considered of lower risk. Despite the fact that there were a couple of studies trying to determine the effectiveness of ivermectin, the authors mention that the experiments had a low number of participants, which represents a problem if the objective is to generalize results. Overall, after a meticulous analysis, the authors claim that there's not sufficient evidence to conclude that ivermectin is effective for preventing or curing COVID-19.