Causal Inference Project Report

Dor Zehavi 211567706 , Ron Dagani 318170917

Introduction:

In this project we were given the chance to select any causal inference related question that we would like to explore.
We chose to attend the field of education, a subject that concerns us very much.
We decided to look into the elementary levels of education, and to ask: "How does the class size affect the student grades?" .
We assume that there should be a difference between students that are in larger classes and students that are in smaller classes.

The Data:

We are using the STAR dataset (Student Teacher Achievement Ratio- linked here), an RCT research on how class size and the usage of educational aides affect the performances of the students in different schools in Tennessee. When we first looked into this dataset, we saw that there are many different features and flags that were gathered, and there are also many N/A and missing values in the data. Furthermore there are several sub datasets within the STAR zip file. Hence, we believe that one of the biggest challenges will be to clean the data, choose relevant features and to combine the relevant datasets together.
Note: the data is an RCT, where the treatment is both the size of the class and the usage of educational aides in class. Here we refer to the usage of the aides as a feature rather than a treatment, and therefore it is no longer an RCT regarding the class size as the single treatment.

Challenges with the data:

- The data contains a lot of parameters, and it was difficult to decide which features are relevant to our question, because many things can affect the outcome on some levels and be considered as cofounders.
- In addition, we had to look carefully to make sure that our data will contain only relevant students. For example, as we mentioned before, the original data contains special education students that are most likely to learn in small classes for their own reasons, those students aren't relevant for our study and don't need to be on the dataset. Due to cases like this, we had to look very carefully at the parameters and make sure that we deal with each one correctly so the dataset would contain only relevant data.
- There are few unmeasured confounders: economic status of the student, education level of the student's parents, student's current mental stability. The dataset does contain parameters that gave us enough information to deal with those unmeasured confounders. Those will be discussed in detail (including the way we handled it) in the weakness section.

Causal graph:

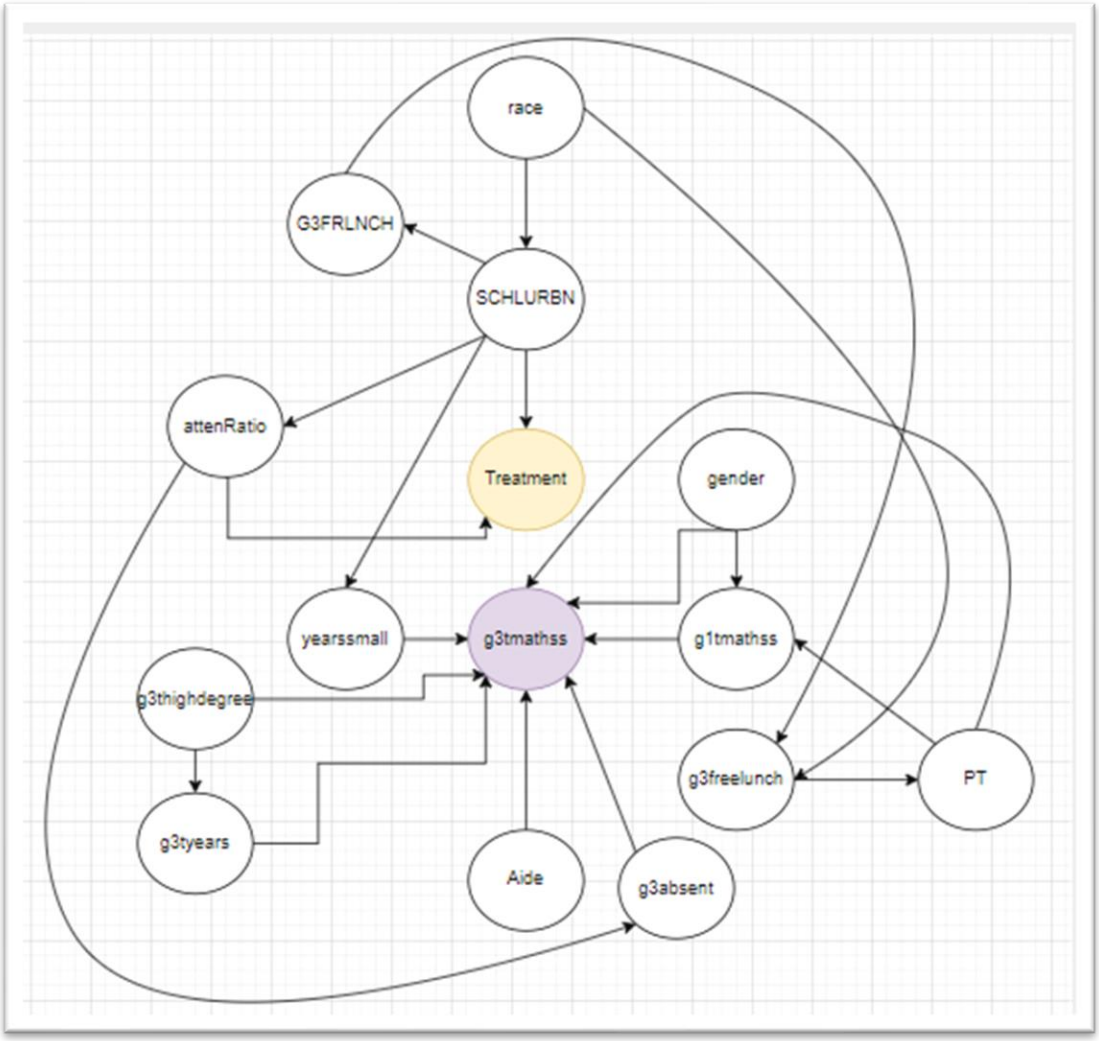
We used a causal graph and backdoor criterion for the identification step.
We modeled this causal graph, based on our domain knowledge and some help from friends and family that work in the education system, and we consider them as sort of domain experts. Needless to say that we can't assure that this is indeed the true causal graph, but we believe that it's close enough and that the differences are negligible.

Variables:

- **gender** - student's gender.
- **race** - student's race.
- **yearssmall** - student's number of years in a small class.

- **g1tmathss** - student's grade in math in first grade.
- **g3tyears** - Teacher's teaching experience (in years) of student's teacher when he was in third grade.
- **g3thighdegree** - Teacher's highest degree of student's teacher when he was in third grade.
- **g3freelunch** - Free/reduced lunch status in third grade.
- **g3absent** - Number of days the student was absent from school in third grade.
- **g3tmathss** - Student's math score in third grade.
- **Aide** - Student's class had aide (or had no aide)
- **SCHLURBN** - School urbanicity.
- **attenRatio** - School average daily attendance on third grade divided by School enrollment on third grade.
- **G3FRLNCH** - School percentage of students receiving free/reduced price lunch in third grade.
- **treatment**: the treatment is the class size.
    - group 0 - smaller than median class size.
    - group 1 - larger than median class size.
- **outcome**: student's math score.

Note: we added a node- PT: The student is using a private tutor.
We added this node because we wanted to prove (via backdoor) that this confounder has no impact on the results (given the data we do have).



BackDoor Criterion:

To isolate the direct effect of T on Y and eliminate any other spurious correlations in the graph, we employed the back-door criteria in our causal graph by discarding unnecessary confounders. This reduction in the number of confounders will aid us in better generalizing our propensity score predictions at a later stage.

All possible paths from T to Y are:

1. $T \leftarrow attenRatio \rightarrow g3absent \rightarrow Y$
2. $T \leftarrow attenRatio \leftarrow SCHLURBN \rightarrow G3FREELNCH \rightarrow g3freelnch \rightarrow PT \rightarrow Y$

3. $T \leftarrow attenRatio \leftarrow SCHLURBN \rightarrow G3FREELNCH \rightarrow g3freelnch \rightarrow PT \rightarrow g1tmathss \rightarrow Y$
4. $T \leftarrow attenRatio \leftarrow SCHLURBN \rightarrow G3FREELNCH \rightarrow g3freelnch \rightarrow PT \rightarrow g1tmathss \leftarrow gender \rightarrow Y$
5. $T \leftarrow attenRatio \leftarrow SCHLURBN \rightarrow race \rightarrow g3freelnch \rightarrow PT \rightarrow Y$
6. $T \leftarrow attenRatio \leftarrow SCHLURBN \rightarrow race \rightarrow g3freelnch \rightarrow PT \rightarrow g1tmathss \rightarrow Y$
7. $T \leftarrow attenRatio \leftarrow SCHLURBN \rightarrow race \rightarrow g3freelnch \rightarrow PT \rightarrow g1tmathss \leftarrow gender \rightarrow Y$
8. $T \leftarrow SCHLURBN \rightarrow G3FREELNCH \rightarrow g3freelnch \rightarrow PT \rightarrow Y$
9. $T \leftarrow SCHLURBN \rightarrow G3FREELNCH \rightarrow g3freelnch \rightarrow PT \rightarrow g1tmathss \rightarrow Y$
10. $T \leftarrow SCHLURBN \rightarrow G3FREELNCH \rightarrow g3freelnch \rightarrow PT \rightarrow g1tmathss \leftarrow gender \rightarrow Y$
11. $T \leftarrow SCHLURBN \rightarrow race \rightarrow g3freelnch \rightarrow PT \rightarrow Y$
12. $T \leftarrow SCHLURBN \rightarrow race \rightarrow g3freelnch \rightarrow PT \rightarrow g1tmathss \rightarrow Y$
13. $T \leftarrow SCHLURBN \rightarrow race \rightarrow g3freelnch \rightarrow PT \rightarrow g1tmathss \leftarrow gender \rightarrow Y$
14. $T \leftarrow SCHLURBN \rightarrow yearssmall \rightarrow Y$
15. $T \leftarrow SCHLURBN \rightarrow attenRatio \rightarrow g3absent \rightarrow Y$

Now we will find the minimal set of nodes that satisfies the Back Door criterion:
Note: there is no node that is a descendent of T, therefore the first condition of the criterion is already satisfied.

$$X = \{SCHLURBN, attenRatio\}$$

The above set blocks (in the d-separation manner) every path from T to Y, and it's the minimal set that satisfies the Back Door criterion:

- $SCHLURBN$ blocks the paths 8 to 15 because of the fork: $T \leftarrow SCHLURBN \rightarrow x$, where x is one from:$\{G3FREELNCH, race, yearssmall, attenRatio\}$.
- $attenRatio$ blocks the paths 1 to 7.
  1 is blocked by the fork: $T \leftarrow attenRatio \rightarrow g3absent$, and the rest are blocked by a chain.
- It's a minimal set because if we take one of them out, the set will no longer block all the paths.

Data Preprocessing:

- Creating a new column named “attenRatio” on the schools dataset, that contains the division of the average daily attendance by the enrollment count, for third grade for each school.
- Use the g1speced to find out if the student is on special education, if he is, we’ll remove it’s record’s from our data because it’s an unusual cause of learning in a small class (in that case, the median size class is irrelevant).
- Define new column “Aide” by mapping this information from the g3classtype, that contains information about the class, including if it had aide or not.
- Define new column treatment as described above (0 if it’s smaller than median class size, else 1).
- Normalizing the students grades ('g1tmathss',' g3tmathss') using minmax normalization.
- Creating dummy variables for categorial values and mapping binary value into zero and one.
- Drop rows that has missing information and columns that were used for creating new columns and filtering so there are no longer needed in the data.
- Joining the schools and the students datasets on the g3schid(student’s school id) and SCHID(school id) columns to get all relevant information about the student and the school it went to.

Assumptions:

- ignorability

  Our assumption is that the majority of confounders are present in the data, and those that are not are most likely to have a negligible impact.

- Consistency

  For each treatment group we observe the corresponding outcome: for each class size group we observe the corresponding math scores - the student learns in one specific class (in that size that matches the treatment group they're in) all year and then gets a math score.

- SUTVA

  the potential outcomes for any unit do not vary with the treatments assigned to other units. We assume that the fact that one student studies in a certain class size does not affect the other students' class size (more classes can be created).

  In addition, for each unit, there are no different forms or versions of each treatment level that could lead to different potential outcomes. We assume that the specific changes in the number of class members is negligible, therefore the categories can be separated to two size scales: large and small and not to a specific group by the exact number of class members.

- Common support:

  We assume that because the original data came from an RCT, there is a probability (greater than 0) that every student can be in every treatment group, $P(T = t|X = x), \forall t \in DOM(T), \forall x \in DOM(X)$. This is a reasonable assumption because the data came from an RCT, and there is no "feature based" separation of the research subjects to the different treatment groups, e.g: There is no case such that only male students were assigned to smaller classes, while female students were assigned to larger classes.

Estimation Methods

First, we checked which algorithm was able to learn the data the best way, by comparing the models score (F1, Accuracy). Then we chose the best model for our data and continued to the following calculations:

- IPW:

  We calculated the ATE using the propensity score we estimated, followed by this formula:

$$\widehat{ATE}_{IPW} = \frac{1}{n}\sum_{1}^{N}\frac{t_i * y_i}{\widehat{e(x)}} - \frac{1}{n}\sum_{1}^{N}\frac{(1-t_i) * y_i}{1-\widehat{e(x)}}$$

- Matching:

  This technique uses the Nearest Neighbor matching based on the covariates of the samples. For every observation that received a particular treatment, we identified and matched an observation from the other treatment group based on the distance of the covariates.

$$\widehat{ATE}_{Maching} = \frac{1}{n}\sum_{1}^{N}\widehat{ITE(\imath)}$$

- S-Learner:

  This method uses a single model to estimate the treatment effect, this model is fitted on the entire sample with the treatment being a regular feature. The estimation is calculated by averaging on the subtraction of the model prediction on every sample while setting the treatment to be 1 and then 0:

$$\hat{Y} \approx f(x,t)$$

$$\widehat{ATE}_{S-Learn} = \frac{1}{n}\cdot\sum_{\mathrm{i=1}}^{\mathrm{n}} f(x_i,1) - f(x_i,0)$$

- T-Learner:

  Like the S-Learner, except here we fit two models on the entire sample, but for one model we set the treatment for every sample to be 1, and for the other we set it to be 0. Then we calculate the average of the subtraction result on the models predictions:

$$\hat{Y}_1 \approx f_1(x), \hat{Y}_0 \approx f_0(x)$$

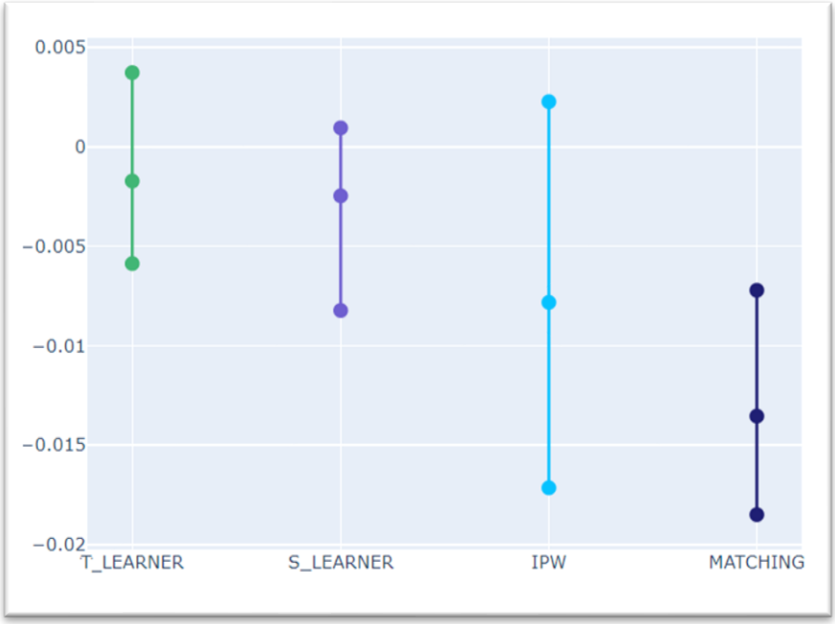
$$\widehat{ATE}_{T-learn} = \frac{1}{n} \cdot \sum_{i=1}^{n} f_1(x_i) - f_0(x_i)$$

Results

For each method we calculated the CI using bootstrap with confidence level of 97.5%. We used 100 random sub datasets that each one of them contained 80% samples of the original data.

method	lower bound CI	upper bound CI	ATE
T_LEARNER	-0.005876	0.003731	-0.001723
S_LEARNER	-0.008234	0.000952	-0.002466
IPW	-0.017148	0.002272	-0.007826
MATCHING	-0.018495	-0.007218	-0.013544

confidence intervals



As we can see, 3 out of 4 of the CI includes zero; that means that there is no significant effect of the treatment on the outcome. We find this result quite surprising, because when we first investigated the subject, we found a lot of research that claims that the class size has a significant effect on the students' grades, and that students that were in smaller classes had better scores than those who were in larger classes.

Weakness

- Unmeasured Confounders and how we handled them:

  While reviewing the data we encountered with few unmeasured confounders, such as:

  - Economical status of the student:

  We have no data regarding the financial status of the student's family. This may cause an issue because wealthier families can afford a private tutor while less wealthier families cannot.
  We handle this situation by using a feature that states whether the student had a free lunch at school or not.
  This may indicate the student's economical status, and cancel the effect of this unmeasured confounder.

  - The education level of the student's parents:

  There can be a correlation between the parents' education and the child's grades due to genetic reasons and due to the fact that they can get better help at home (regardless of getting private lessons from a tutor).

  We handled this situation by using the student's grades from the 1st grade, and the student's grade has a stronger indication about the student's abilities than what his parent's education has.

  - student's mental stability during 3rd grade: we assume that the percentage of those who suffer from mental health issues is very small, and that the effect it has is very minimal and negligible.

- It can be claimed that the difference between the exact number of students in two classes (of the same treatment group) is not negligible and therefore SUTVA assumptions may not hold.

- The data contains a lot of null values, this could lead to biased results.

- There is a possibility that the estimation on the propensity was not close enough, so estimation could be affected by that.

- There could always be things that we're not aware of that took place when this data was collected, or that we did not think about that could affect our outcome and causal graph. We did take all the measured data that we had and worked with what was relevant, so if there is something that we did not take into account, it has probably not been measured and could not be added to the data without adding significant bias.

Conclusions and Discussion:

In this project we tried to answer the question: "how does the size of the class affect the students' grades?". We did so by modeling a causal graph, limiting our unmeasured confounders and by estimating the ATE in few different methods. When looking at the results we got, it's easy to notice that the effect of the treatment (class size) on the outcome (students' grades) is negligible or even negative, most of the CI we got are close to zero and are tilted to the negative side. This may be due to the fact that the original data from the research contained a lot of null values which created some sort of bias, and therefore the results we got may not reflect the real effect of class size on the students' performances.

In conclusion, the answer we got for our research question is that smaller class does not lead to better grades.

In future research, we suggest working with a more extensive dataset, that contains a greater number of observations, which would offer better representation. Additionally, we believe that obtaining additional datapoints could potentially aid us in overcoming any potential missing confounders. Moreover, we suggest implementing advanced ATE estimation techniques that were not utilized in this project.

Bibliography

Experienced elementary school teachers.

Course lectures and tutorials slides.

L.Wang, L.Calvano, Class size, student behaviors and educational outcomes, 2022.

N.Ligembe , B. Peter ,Impact of Class Size and Students' Academic Performance in Public Secondary Schools in Kwimba District Council, Mwanza -Tanzania, 2022.