

Project Report

CS396-4 Causal Inference

March 18, 2022

Instructions

This assignment is due on Thursday, March 17 at 11:59pm CST. It will **only be accepted up to 24 hours late**, but with a 14.3% (1/7th) penalty if late. Grades are due the following Monday, and we need time to read everything each group has done. This assignment is worth 25% of your final grade. Please upload your (group's) update to Canvas as a single PDF. You can use this assignment TeX to fill in your answers.

By submitting this assignment, you affirm that you have neither given nor received any unauthorized aid on this assignment, and that your code and write-up are only the work of your group. Any violation of Northwestern's academic integrity policies will result in you receiving a 0 on this assignment and a report to your dean's office. If you're ever worried about whether you are at risk of violating these policies, please ask – we can help you follow the rules, but we can't retract a report of suspected cheating. If you are working in a group and are concerned that some members of your group are not contributing equally, please email me or let me know in an anonymous course survey.

1 Group members

Tyler Maule and Jipeng Sun

2 Code (10 points)

Github Link of our project:

https://github.com/JipengSun/Causal_Inference_Anticonflict

2.1 Running your code (4 points)

Our code is written in python and run as Python Jupyter Notebook files. The project is organized in the following way:

- `school_level_causal_inference_pipeline.ipynb`

Estimate causal effect between training assignment treatment and anti-conflict level outcome at the school level. Based on data aggregated from the student level, analyze the school level RCT.

- `student_level_causal_inference_pipeline.ipynb`

Estimate causal effect between training attendance treatment and anti-conflict level outcome at the student level. Includes network effects, noncompliance adjustment (with the "Attended Treatment" compliance indicator), and the Baseline Anti-conflict Score.

- `pre-processing/`

Data preprocessing. You can ignore the files in that folder since we have already provided pre-processed data files. But you can run these files to get pre-processed data from original dataset.

- preprocessing_pipeline_v0.ipynb
Identify missing values, whether missing completely or with error codes, and drop rows and columns with complete missingness or redaction. Implement a MICE algorithm to iteratively impute missing values with random forests, using the miceforest package (with documentation linked here). Calculate baseline and final "anti-conflict scores" from students' survey results. Filter dataframe based on selected relevant variables for the causal graph. Organized covariate/confounder, response, and treatment variables.
- network_analysis.ipynb
Build social network for students in same school and measure the treatment interference.
- causal_discovery.ipynb
Use causallearn package to construct causal graph automatically (exploratory work).
- data/
All the data files including pre-processed ones.
- output_imgs/
Output images of the project including causal graph
- Linear_Parameters.log
Parameters log for linear regression model used in student level analysis.
- requirement.txt
- README.md

2.2 Documentation (4 points)

General guidance to run the project can be found in README.md.

To start the project "from scratch" (without any of the saved files under the "data" folder), please visit the ICPSR website to find the data sharing portal. Use the "Download" button to select a file format for your dataset download. You'll need to register to download data, but it's not a major barrier (approval is instant in our experience). We downloaded the data as a RData file, loaded it in R, and immediately saved it as a csv file for use with Python. The ICPSR website also has plenty of documentation, from a codebook explaining variable names, to an in-depth description of the study, to the actual questionnaire that students were issued.

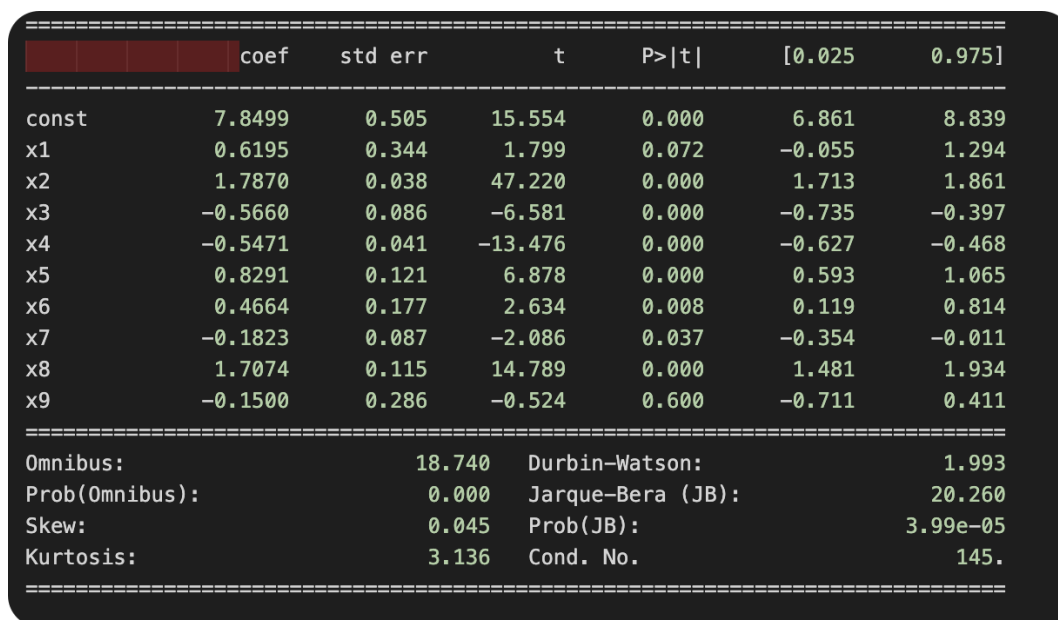
For the causal inference functions, please run the school level and student level data separately. The preprocessed dataset is already located under data folder.

For preprocessing part, please run preprocessing_pipeline_v0.ipynb first to get preprocessed data from original dataset. Then, run network_analysis.ipynb to get social network interference effect.

Specific code comments can be found in the jupyter notebook files.

2.3 Estimating function (2 points)

The way to get linear regression model parameters from DoWhy package is not obvious. There is no related documentation. However, by reading the source code of the DoWhy, we find that even though there is no direct way to get model info in DoWhy/causal_estimators/linear_regression_estimator.py, in its parent class DoWhy/regression_estimator.py, it provides DEBUG level log info to get the model parameters. (I used to modify the source code in linear_regression_estimator to let it report parameters but sometimes it will break the project due to None assignment to the estimators in initial method). Notice that writing the log stream directly to stdout of console will break the program due to the extensive output. Thus, I write the log stream to file instead. (Linear_Parameters.log) Figure1 is the final parameters of linear regression model for backdoor estimator.



	coef	std err	t	P> t	[0.025	0.975]
const	7.8499	0.505	15.554	0.000	6.861	8.839
x1	0.6195	0.344	1.799	0.072	-0.055	1.294
x2	1.7870	0.038	47.220	0.000	1.713	1.861
x3	-0.5660	0.086	-6.581	0.000	-0.735	-0.397
x4	-0.5471	0.041	-13.476	0.000	-0.627	-0.468
x5	0.8291	0.121	6.878	0.000	0.593	1.065
x6	0.4664	0.177	2.634	0.008	0.119	0.814
x7	-0.1823	0.087	-2.086	0.037	-0.354	-0.011
x8	1.7074	0.115	14.789	0.000	1.481	1.934
x9	-0.1500	0.286	-0.524	0.600	-0.711	0.411
Omnibus:		18.740	Durbin-Watson:			1.993
Prob(Omnibus):		0.000	Jarque-Bera (JB):			20.260
Skew:		0.045	Prob(JB):			3.99e-05
Kurtosis:		3.136	Cond. No.			145.

Figure 1: Parameters of Linear Regression Model for Backdoor Estimator

Based on the parameters info, we can get our linear regression model for our backdoor estimator:

$$\begin{aligned} \text{composite_anticonflict_score} = & 0.62 * \text{training_attendance} + 1.79 * \text{baseline_anticonflict_score} \\ & - 0.57 * \text{phone_internet} - 0.54 * \text{age} + 0.83 * \text{go_this_school} \\ & + 0.47 * \text{network_effect} - 0.18 * \text{treatment_school} \\ & + 1.71 * \text{yes_college} - 0.15 * \text{treatment_student} \end{aligned}$$

3 Updates since your update (5 points)

3.1 Add social network interference score

See comment: Q3.1 Subsection: Add social network interference score in student_level_causal_inference_pipeline

To analyze the interference of the training effect flowing by student's social network in individual level. We add an independent python notebook file for social network analysis as /pre-processing/network_analysis.ipynb.

In this part, we did following steps:

1. Group the students by their school.
2. Recalculate out the true student ID number and the valid relationship edges for every student to eliminate the data entry and NaN errors.
3. Use python’s igraph package to construct a social network graph of that school based on the preprocessed vertices and edges information.
4. Add ID and Treatment Status as vertices attributes in the graph.
5. For every vertex in the graph, calculate all the shortest distance between this source point to the treatment target points set.
6. Count the number of 1-jump and 2-jump treated vertices.
7. Based on our social network effect formula, $\sum_i bf_i * \frac{1}{2^i}$, calculate the final social network interference score.
8. Merge the new column back to the original dataset based on the UID key.

3.2 Evaluate baseline anti-conflict score and training attendance in student level

See comments (Q 3.2 – Baseline Anti-Conflic Score) in the preprocessing_pipeline_v0.ipynb file and (Q 3.2 – changing compliance framework) in the student_level_causal_inference_pipeline.ipynb file.

At the time of our project update, we incorporated the results of the study’s second survey as our response variable. However, we left out the results of the survey that students took at the beginning of the academic year. In doing so, we omitted a potentially important confounder—perhaps students who experienced the highest level of conflict were most likely to attend anti-conflict training sessions, and thus the treated individuals were not representative of the student population’s attitude about school conflict. The final version of our project takes this first survey into account, adding it to the causal graph as the Baseline Conflict Score. From the DoWhy backdoor estimation results, we see that the Baseline Anti-Conflict Score does have a statistically and practically significant relationship with the Final Conflict Score, with a p-value of < 0.001 .

We also changed the way we incorporated treatment compliance into our causal graph, removing the discrete variable "Roots" and replacing it with the binary categorical variable "Training Attendance." Since the nine anti-conflict training sessions (called meetings of the "Roots" program) occurred over the course of the academic year, there was a longitudinal component which we could not take into account without adding significant complexity to our model. Instead, we decided to take a simpler route by using "Training Attendance," where 1 signifies attendance of at least one Roots meeting, and 0 signifies that the student never attended a Roots meeting (regardless of whether they were assigned treatment, aka invited to the meetings). So we now measure treatment compliance as True/False, rather than a 0-9 meeting attendance count.

3.3 Define a new causal graph and run its estimation

See comment (Q 3.3 – causal graph revision) in the student_level_causal_inference_pipeline.ipynb file.

We propose a more reasonable causal graph considering the baseline anti-conflict level of the student's environment, the student attendance, and the social network interference to the treatment and outcome. (Figure 2). Based on that, we re-estimate our causal effect using linear regression model and double machine learning model. We also run random common cause, placebo, and subset refuters to test each estimation.

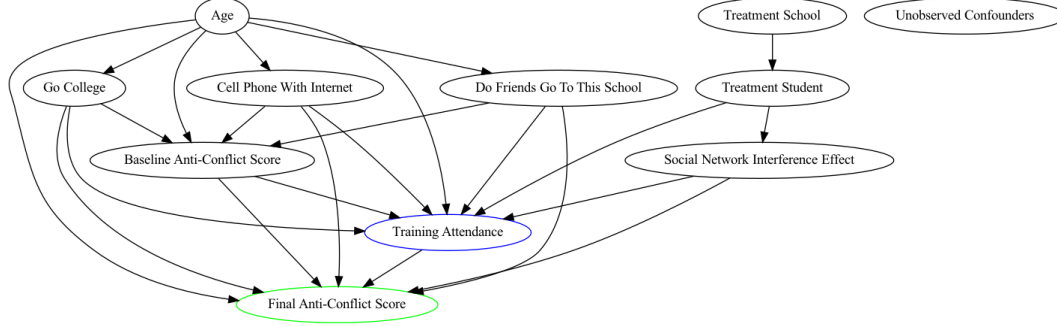


Figure 2: Causal Graph Between Causal Effect of Training Attendance and Anti-Conflict Score

3.4 Connect the causal estimation pipeline to the imputed data

See comment (Q 3.4 – missing data imputation) in the preprocessing_pipeline.v0.ipynb file.

While we'd already imputed missing values in our dataset, at the time of the project update we hadn't yet rerun the causal estimation functions on this new, less naively imputed data. The project update worked with median-imputed data, which might've altered the estimated effect of confounders. The use of MICE always carries some risk of introducing nonexistent effects into the model, but we'd argue that use of MICE better aligns with the true values of missing data (on the whole) than median-imputed data. In the future, we'd love to expand on our work in MICE (and the miceforest package) to fine-tune the imputation process.

4 Interpreting your results (5 points)

4.1 Before

In the project update report. We analyze the causal effect between anti-conflict training treatment and the anti-conflict level score outcome both in school level and student level.

4.1.1 Training Assignment Causal Effect in School Level

The final average risk difference result we get from backdoor estimator is 0.165. The result means the treatment for anti-conflict training session will increase the anti-conflict level of the school by 0.165. Since the treatment sessions are only assigned randomly to 15% of the students and the average score of the control group is about 4.95, this training session does have positive effect on anti-conflict level of the school. You can interpret this result as around 16 students from 100 students will report they met 1 less conflict event in the school in a past period of time comparing the counterfactual world where they have not assigned to the training session.

4.1.2 Training Session Attendance Causal Effect in Student Level

The final average mean difference result we get from backdoor estimator is -0.293. The result means the treatment for anti-conflict training session will increase the conflict perceived by students who attended the training sessions by 0.293. We interpret this result by concluding that regular discussions of peer-to-peer conflict and students' responsibilities to mitigate it heighten session participants' awareness of conflict and teach them to view their community in a critical light. However, earlier results showed that on the school level, the presence of anti-conflict training sessions at a given school increases the mean anti-conflict composite scores at the school. Therefore, we argue that while the sessions promote community-building and reduce perceptions of conflict at large (due to a truly lower level of conflict), the sessions do so by increasing perceptions of conflict by session participants (heightened awareness at a given level of conflict).

4.2 After

We fully focus on student level causal analysis at this final stage. Our purpose is trying to answer the conflict we found in project update report. Even though the treatment rate in school level will bring positive effect on anti-conflict score, however, the training session attendance in individual level shows negative causal effect. This incoherence will lead us to two different conclusions when it comes to the question 'Should we promote this training session to larger group of people in that school?' since the treatment rate for now is only 15%.

Inspired by Prof. Zach's feedback on project update report, we made three improvements to better reason this problem in student level.

- Measure Interference Effect Through Individual Social Network

In our previous measurement, the attendance of one student only affect him/herself, with the assumption that the training will only interfere the attendees and they won't affect other students, which is unlikely to hold in this situation.

To measure how would the training student affect others, we use well the answers in the questionnaire that ask student to indicate his/her top 2 best friends. Based on this information, we build the social network graph for students in one school. We assume the treatment will affect student by social network by a 0.5 decreasing term. Thus, we add one more variable to measure the training interference effect for students.

- Calibrate Results with Baseline Conflict Score

As indicated in (3.2) above, we incorporated results from the study's first survey to decrease confounding of the treatment effect. With the Baseline Conflict Score in our model, we expect to see decreased bias and a smaller variance associated with our estimated treatment effects.

- Remeasure the Student Attendance Metrics

By changing the measure of compliance (as explained in 3.2), we better answer the question at hand with a more interpretable framing of treatment. Using a binary treatment effect allows for better generalization to other training sessions and schools, as a nine-session study design is not common to all researchers or possible for all school districts.

After the adjustment, we finally remeasure our causal effect using linear regression and double machine learning method for backdoor estimator in student level.

For linear regression model, the final expected average treatment effect is 0.76. The model is specified in equation in Section 2.3. Which means we would expect a 0.76 increase of the anti-conflict score of that student if this student attend the training session. The coefficient estimates of the model are consistent with our expectations, since training attendance, baseline anticonflict score, whether friends will go this school, network effect whether they will go to college are beneficial to improve the anti-conflict level of the student’s environment and phone internet, and age will increase the conflict level they perceived.

However, the p-value associated with training attendance is 0.072, so at the $\alpha = 0.05$ we fail to find a statistically significant average treatment effect of the anti-conflict score. We are 95% confident that the true value of the ATE lies between -0.055 and 1.294; the estimate’s standard error is 0.344. The coefficient related to treatment assignment also fails to differ from 0 to a statistically significant degree (with a p-value of 0.6). The remainder of the coefficients are statistically significant, with the next-highest p-value at 0.037 (associated with treatment at the school level). The network effect has an estimated coefficient of 0.47, which with a p-value of 0.008 is significant at the $\alpha = 0.05$ level.

To interpret these results, we reflect on our findings in the project update. From our perspective, it makes sense that a school on the whole and a complier’s best friends would benefit from, and see clearly, a student community with decreased conflict and increased cohesion as a result of the Roots training sessions. However, students in the training sessions receive two conflicting effects: on the one hand, they are more equipped to deal with and overcome conflict. On the other, they spend more time critically analyzing and taking responsibility for the school’s conflict environment. So there’s a complex set of outcomes which makes it understandable that the ATE is positive, but statistically insignificant at $\alpha = 0.05$.

For double machine learning backdoor estimator, we get a higher average treatment effect of 1.36. Which is more reasonable with the data distribution since the relationships between the treatment, covariates and outcome are nonlinear. Since the second estimator uses lasso regularization, the results in this case are less interpretable.

4.3 Comparing the two

Yes. The estimates do change and become more coherent with the school level analysis. The change is mainly due to our better representation for attendance and more precise measurement on the interference effect.

After taking interference effect into consideration, we would expect this will enlarge the scope of affect of the treatment since more indirect treatment effect will be counted in. Thus, the positive effect the training will bring will be counted through the interfered student, which we would count them as no treatment before. It will bring more higher anti-conflict cases into account and thus make the causal effect higher. At the same time, increased attribution of treatment effects to students’ social networks leaves the direct treatment compliance coefficient statistically insignificant. So with this analysis we gain a more nuanced understanding of the treatment’s effect on treated students, their close friends, and the school overall.

Besides the enlargement of the higher anti-conflict group for treatment, measures of the treatment effect in social network are also more coherent with our outcome measurement. Our outcome is to measure the anti-conflict level in the world of that student perceived. The world of that student perceived is largely formed by that student’s best friends. Thus, taking this interference into

account could help us improve our measurement on how treatment will affect that perceived world.

5 Reflections (5 points)

5.1 What you learned

5.1.1 Feedback from Jipeng Sun

For me, the most interesting part of this project is its limitless meaning of using quantitative causal methods to understand a traditionally qualitative subject such as social science. Before this project, my answer to the question 'How modern data science will change social science study?' remains in applying big data associative analysis and machine learning prediction. However, this causal inference project totally changes my mind. I learned how to directly measure and model the causal relationship from data rather than using association to guide the causal decision. Besides, social science has its intrinsic interference nature. Learning how to face this challenge directly rather than ignoring it is also a really interesting topic for me.

The most challenging part of this project for me is to learn how to solve the interference problem by taking social network effect into consideration. Using graph to model the social network and using graph algorithm to discover graph properties down to the code level is a challenge I want to take for long time. Combining this type of information into our causal inference process to get a better estimation is also an important challenge for me.

5.1.2 Feedback from Tyler Maule

The most interesting part of this project for me was its theoretical dimensions—that is, exploring the theory we'd learned in class and understanding how to implement it clearly. Many of the topics we explored in this project I'd learned about beforehand, so it was especially enjoyable to see the results of methods and experiment with many different "additional challenges." Coming from a statistics background, it's very compelling to me to learn about more computer science-driven approaches. I was especially impressed by the rigor of some of these methods, like the refutation methods and identifiability detection capabilities of DoWhy. It's great to see that we can challenge the assumptions we're making during model construction with real, empirical techniques.

For me, data pre-processing was a significant challenge for this project. Since many of the data points were missing, we decided to use MICE methods with random forests. It took a some time to get the data into a format that worked with the imputation packages. In addition, the creation of new variables such as the Baseline/Final Anti-Conflict Scores required plenty of data cleaning. The complexity of our data and graph was also a challenge, since it took the packages a long period of time to run the causal graphs we initially planned to use. There are a few variables I would've liked to add, but doWhy's computational capabilities and our personal computers limited the time we had for causal discovery. Finally, working with poorly documented libraries like doWhy made for some frustrating guess-and-check procedures.

5.2 Unaddressed challenges

Think back to your critiques of the academic and media sources in Homework 3. How would you similarly critique your project? That is, what are two reasons why the causal effect you estimated for this project might not accurately reflect the true causal relationship in the world? These could

be extensions to the additional challenges you considered, or new challenges. Of these two reasons, which do you think has a bigger effect on your results? Why?

We list following problems as our unaddressed challenges:

1. Better interference measurement for social network

In the survey, the questionnaire also asked students to indicate who are their top 3 students to spend most time with and with whom they have conflicts with in the past half of a year. In the current stage, we only measure the interference of the student by their best friend network. However, this might not be a perfect representation of the world they perceived since people listed in the above two groups may have larger interference to students considering the survey is for anti-conflict level measurement. Thus, the bias in interference may bring us a biased measurement on causal effect. In the future, we would try to build a better representation for the social network of the student.

2. Exploring unknown confounders in causal graph

The questionnaire actually consists of seven hundred columns. Even after our first stage filtering, there is still around two hundred variables that may be related with our causal outcome. For the simplicity of measurement, we only use or aggregate around fifty variables into our causal graph. It is possible that our measured effect is biased due to the so many unknown confounders in this problem(such as GPA, video game, demography, etc)

5.3 What's left to do?

We would focus on increasing the complexity of our variables and causal graph, using sensitivity analyses to determine the effect of changes to the graph, changes in the calculation of the response variable, and changes to our model interference impact our results. On the whole, I think that our variables and graph as defined do a fairly appropriate job of estimating the underlying effect, but I think we'd need to better evaluate our approach vs. other approaches before we could consider the estimate a truly actionable estimated effect representative of the true effect. We'd also love to continue exploring the causal discovery packages—this was something we'd mentioned in our project update, but other more pressing "challenges" took priority.

5.4 A follow-up study

If we could study this problem further and collect new data, we'd choose to (1) collect survey data more routinely and (2) administer surveys to teachers as well as students. Since the two surveys in our current study were at the beginning and end of a whole academic year, it's hard to clearly prove that a year-long intervention changed the level of conflict at the schools. More longitudinal data points, however, would allow us to more precisely measure the impact of sustained anti-conflict training sessions on the level of conflict. We'd also be better able to adjust for the response patterns of students (i.e. we may be able to detect which student's results are representative of others' missing or unreliable results). In addition, getting input from teachers would allow us to have more confidence in the assumption that conflict data truly represents the school environment, rather than students' more self-centered responses.

Our underlying study already randomly assigned treatment, but there were noncompliance issues that we incorporated into our causal graph ("Training Attendance"). If we could assign treatment randomly, leading to mandatory session attendance for treated students, we would

do so in order to better distinguish and estimate the ATE from the average effect on treated compliers. While we do adjust on multiple covariates (including initial survey results) to adjust for the likelihood of participation in anti-conflict trainings, having more confidence in our treatment variable and not having to adjust for noncompliance would improve our estimates.

This new dataset would allow us to be more certain about the assumptions we're making, it would decrease the variance of our causal estimates by eliminating noncompliance (to the extent possible), and it would allow us to use more comprehensive, longitudinal data in our analysis.