The Causal Effect of Online Instruction on Academic Performance

1. Causal Research Question

Does taking courses online affect students' academic performance compared to in-person instruction? Specifically, I aim to estimate the causal effect of online learning on students' final grades or test scores.

2. Data Source and Description

I will use data from the Stanford Education Data Archive (SEDA) and/or publicly available **COVID school datahub** datasets from the COVID-19 period (2019-2021), when many schools and universities moved to online learning. Additional sources may include surveys or administrative education records from state or national databases.

Coverage: Students from multiple schools, including both those who switched to online learning and those who remained in-person.
Time Frame: Data from before and during the COVID-19 pandemic, e.g., academic years right before the COVID-19 2018-2021 or after the COVID-19.
Units of Observation: Individual students or schools (depending on data availability).
Repeated Measures: In many datasets, student performance is tracked over multiple years, allowing for panel data analysis.

3. Study Design

The essential comparison will rely on a difference-in-differences (DiD) approach. I will compare changes in academic performance before and after the shift to online learning between:
Treatment group: Students who shifted to online learning due to COVID-19.
Control group: Students who continued in-person instruction or returned earlier.

This method helps control for time-invariant differences between groups and broader trends affecting all students. If possible, I will check robustness using matching methods or fixed effects.

Data Source & Reference

- The Stanford Education Data Archive: https://edopportunity.org/
- Covidshooldatahub: District-Monthly Percentage In-Person, Hybrid, or Virtual

I would use the "**SEDA2022_admindist_poolsub_YS**" dataset from the Stanford Education Data Archive. The main reasons I chose this dataset are that the dataset compares the student test performance between 2019 and 2022. The matrix "YS" is good for comparison for time difference associated results.
It uses the average score for each year to compare students of the same grade and same subject

Approaches

1. Find potential outcomes: D=1 when school/district switched to online/ hybrid format. D=0 for the else
2. Estimate ATT, ATE, and ATU based on the variable share_virtual(online learning format) and treatment variable D.
3. **Difference in difference**
    a. Aggregate the overall scores, math scores, and English scores
    b. Evaluate the scores before/after the COVID-19 period
4. **Subclassification**
    a. Group by state
    b. Group by race
    c. Group by the length of the year they continue online learning
    d. Observe the evaluation result for each sub group
5. **Synthetic Control**
    a. Take an example of states(e.g. California)
    b. Aggregate the overall scores, math scores and English scores
    c. Evaluate the
6. Analysis Interpretation
7. Conclusion
8. Further research chance & Limitation
9. Resources