Causal Effect of Online Learning during COVID Pandemic on Student Performance

Ayumu Yamagishi

ECON427 Introduction to Econometrics II: Prediction and Causal Inference

Introduction

COVID-19 pandemic had a significantly huge impact on human life. A trend graph, the image.1 in the visualization section below, indicated that the number of COVID-19 patients exponentially increased from the beginning of 2020 and gradually continued its power until the middle of 2022(CDC, COVID Data Tracker). Through this pandemic, we have experienced lots of changes in some fields. Educational fields were not the exception. Starting from the end of 2019 and beginning of 2020, a large number of schools/districts replaced their traditional in person class style with online learning or hybrid learning style. Even though the pandemic ended, the online learning style remains due to its accessibility, reducing instructors loadworks or efficiency.

However, the effectiveness of those online learning methods are still unclear. Some researchers suggested that students tended to be lazy not to do homework or cheat in the online assignments and exams. In my research, we will explore the effectiveness of online learning by comparing some subjects' scores before and after the COVID-19 on those who shifted to online learning and remained in person learning style. Considering the current situation, my research question is:

"Does online learning during the COVID-19 pandemic causally affect student academic performance compared to in-person learning?"

Firstly, we will define the treatment variable based on if the school/district switched to online learning from the traditional way. After setting the treatment variable, we will find potential outcomes under the assumptions such as average treatment effect(ATE). Once we have determined the potential outcomes, we would move on the difference in methodology for how online learning style changed students' performance in 2019 and 2022 via online learning during 2022 to2021. After the difference in difference, we would figure out whether the result varies depending on states or races of students or schools. Finally, for the visualization and evaluation purposes, we will go over the synthetic control approach. With these methodologies and approaches, we try to figure out how online learning gave a critical impact on student performance during COVID-19 pandemic.

Now, we start from determining the potential outcomes of student performance with online learning and traditional learning style, then set a dummy variable D for treatment. Learning share dataset has monthly data about if the school hired in person or online learning at a certain month. Also the data is identified by district ID and district names. The district implies the name of the school from elementary level to high school. However, since the SEDA dataset contains data only for two years 2019 and 2020. As the online learning is continuous leffect for the treatment, I decided to create a treatment variable D in the learning share dataset based on the in person or online learning share The threshold for the variable is 0.5, which means if the ratio is larger than 0.5(more than 50% of students in a district is taken the online learning), **the treatment variable D is set to 1, otherwise, it sets to 0.**

Potential Outcomes Framework

In causal inference, potential outcomes refer to the possible outcomes a unit (a district in this research) could have under each treatment condition. We assigned online learning (D=1) and in-person learning (D=0). However, we only observe one outcome per unit. The unobserved outcome is called the counterfactual. The goal is to estimate causal effects such as the Average Treatment Effect (ATE), which is the expected difference in outcomes between treated and untreated conditions. This framework is essential as it clarifies what we are estimating and emphasizes the assumptions needed to identify causal effects.

Difference-in-Differences (DiD)

To begin estimating causal effects, we implement the Difference-in-Differences (DiD) method. This approach compares the change in outcomes over time (from 2019 to 2022) between treated and control groups. We compute the ATE using:

$$ATE = E[Y_1] - E[Y_0] = (Y_{post}, T = 1 - Y_{pre}, T = 1) - (Y_{post}, T = 0 - Y_{pre}, T = 0)$$

In our case, we found that ATE computed using group means matched the DiD estimate, confirming robustness under the assumption that trends would have been parallel in the absence of treatment. The fact that both ATE and DiD produce nearly identical results strengthens the credibility of the treatment effect estimation. Before we estimate the DiD, we put some assumptions for the dataset.

Assumptions

1. **Treatment Variable(D):** 1 if share_virtual + share_hybrit >= 50 %, 0 for else.
2. **Online Months:** Aggregation(sum) of the treatment variable D grouped by (State, ID, Name), length of online learning periods.
3. **Treatment Group(T):** If online Months >= 1, 0 for else.
4. **Delta Score:** Difference of student score in 2022 and 2019.
5. **Evaluation:** OLS model representing how far student performance is from the average(in SE unit scale).
6. **Subcat:** Subcategory, referring to the subject either mth(MATH) or rla(Comprehension).
7. **Subgroup(race):** Another subcategory, but we ignored it.

The result of DiD showed that the school districts that switched to online learning, including hybrid, had test scores that were about 6% lower than average on a standard deviation basis, compared to those that continued with traditional face-to-face learning.

Synthetic Control Method

As a more refined approach, we applied the Synthetic Control Method. For each treated state, we constructed a synthetic control using a weighted average of other states with similar pre-treatment (2019) test scores. This method helps estimate what the treated state's outcome would have been in 2022 had it not experienced widespread online learning. For example, California's actual performance was compared to a synthetic CA created using AR, CO, and FL as donor states. The difference in 2022 scores represents the estimated causal effect. We repeated this process for all states and visualized treatment effects using bar plots. States with large negative values likely suffered more from online learning, while those with smaller or positive values might have had better infrastructure or policy responses. The synthetic control approach is particularly useful because it provides a credible counterfactual for each treated unit, especially in cases where traditional methods like regression may be biased due to unobserved heterogeneity.

For Math Interpretation, The result differed state by state. Some of the actual test scores in these states were lower than the scores of their synthetic counterparts. This result suggests that online learning likely had a negative causal effect, contributing to a decline in student performance in mathematics. On the other hand, other actual scores were higher than the synthetic scores. This implies that the negative impact of online learning was relatively small in these states, or that supportive measures and interventions may have been effective in mitigating learning loss.

Conclusion

This research explored the causal effect of online learning during the COVID-19 pandemic on student academic performance in the United States. By implementing both Difference-in-Differences (DiD) and Synthetic Control Method (SCM), we aimed to measure the magnitude of learning loss in mathematics and reading comprehension (RLA) attributable to the shift from in-person to online education formats.

Our DiD analysis showed that districts that switched to online or hybrid learning during the pandemic experienced a significant decrease in test scores—approximately 6% lower in standardized deviation terms—compared to those that remained in traditional face-to-face instruction. This finding provides a baseline estimate of the average treatment effect (ATE) of online learning.

To improve upon this estimate and account for state-level variation, we employed the Synthetic Control Method. This allowed us to construct credible counterfactuals for each treated state by creating synthetic versions from weighted combinations of untreated states with similar pre-treatment characteristics. The SCM results revealed substantial heterogeneity in treatment effects across states. In some states, such as Mississippi and New Mexico, students experienced much worse outcomes in 2022 than their synthetic counterparts, particularly in mathematics. This suggests a more severe negative impact of online learning in these areas. Conversely, states like California and Massachusetts showed smaller gaps or even slightly better performance than their synthetic equivalents in RLA, hinting at more effective adaptation or mitigation policies.

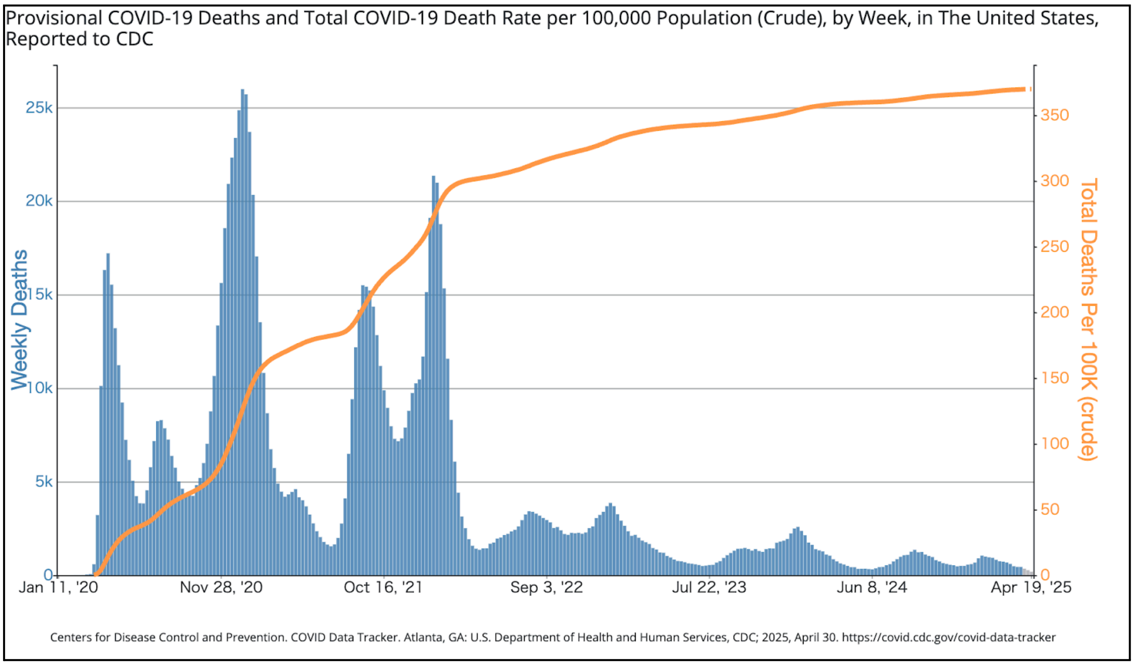
These results imply that the impact of online learning during the pandemic was not uniform and likely depended on a range of contextual factors, including infrastructure, socioeconomic conditions, and the speed and quality of local education policy responses. The contrasting effects between math and RLA further indicate that the nature of the subject may influence how well it adapts to remote learning environments.

While our analysis provides meaningful insights, it is important to acknowledge limitations. Our treatment definition simplifies online exposure and does not capture nuances such as teaching quality or student engagement. Moreover, unobserved confounding variables and only two years of outcome data restrict our ability to identify long-term trends.

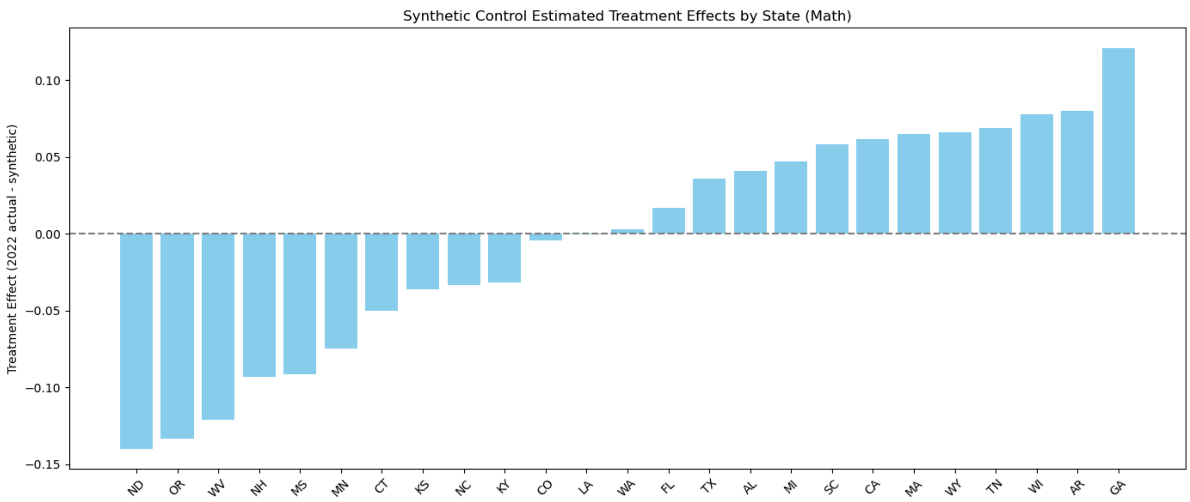
Despite these limitations, our study contributes to the growing literature on pandemic-era education and highlights the value of combining multiple causal inference methods. Future work may focus on extending the time window, incorporating student-level data, or evaluating specific intervention strategies. Ultimately, understanding what worked and what failed during the pandemic is crucial for building a more resilient and equitable education system for the future.

Visualizations

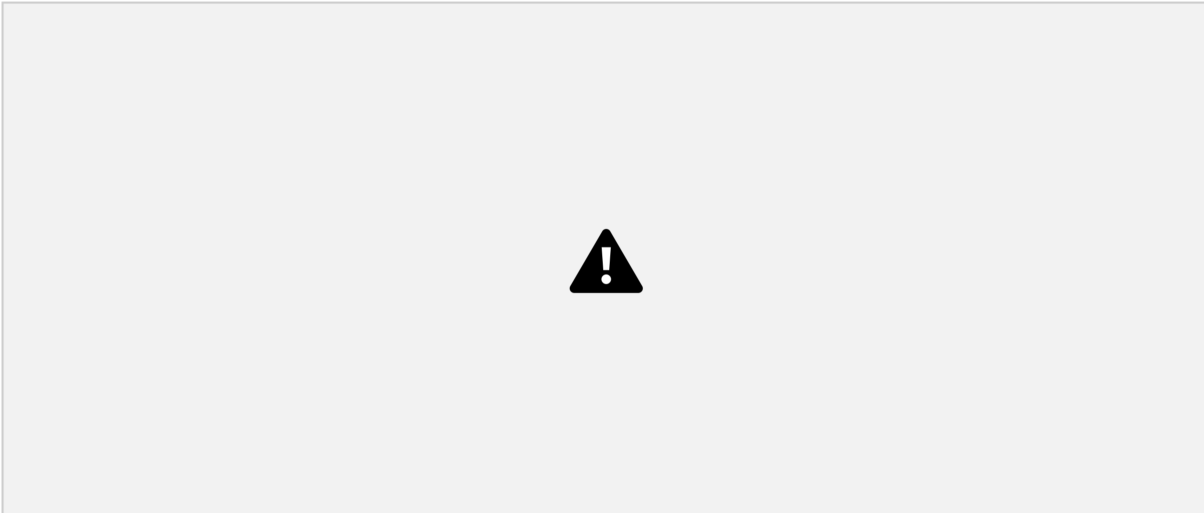
1. Trend of COVID-19 in the U.S. from 2019 to 2025



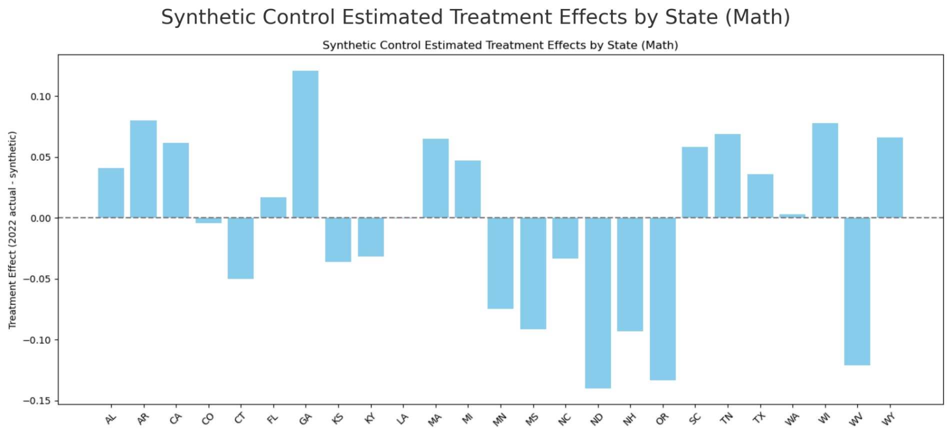
2. Synthetic Control Model for Math by Treatment Effects

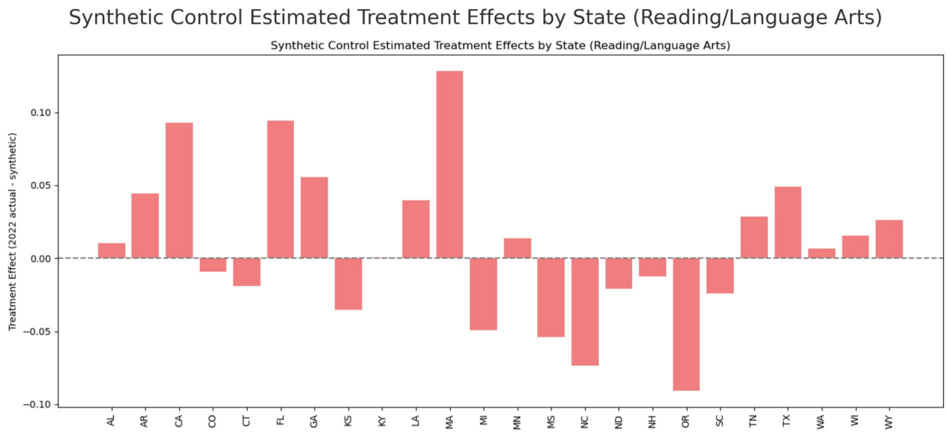


3. Synthetic Control Model for RLA by Treatment Effects



4. Synthetic Control Models for RLA and Math by State





References

- **Centers for Disease Control and Prevention(CDC), Trends in United States COVID-19 Deaths, Emergency Department (ED) Visits, and Test Positivity by Geographic Area** **https://covid.cdc.gov/COVID-DATA-TRACKER/#trends_weeklydeaths_totaldeathratecrude_00**
- The Stanford Education Data Archive: https://edopportunity.org/
- Covidshooldatahub: District-Monthly Percentage In-Person, Hybrid, or Virtual
-