566 proposal

He Ren, Kirara Kura, Matthieu Liger, Fumin Li

May 2024

1 Introduction

Starting from a dataset that record passive and automatic sensing data of Dartmouth students across five years (2017 - 2022), we want to figure out how much impact COVID has had on students' depression levels during this special period. To this end, we leverage DID method with appropriate covariate balancing to estimate the causal effect of the outbreak of COVID on the level of mental health.

2 Dataset

The dataset we used is from College Experience Study Nepal et al. [2024], which contains automatic sensing data and survey data collected from a total of 200 Dartmouth College students across five years (2017-2022) using their smartphones. These sensing data include sleeping hours, physical activity hours, and audio/phone usage. A survey on mental health (e.g. PHQ-4) was taken multiple times by each student throughout the study.

3 Goals and Hypothesis

The causal relations we plan on exploring include:

Goal 1: Causal relation between mental health and COVID pandemic stage. In this case, COVID would be the treatment.

Goal 2: Causal relations between mental health and other covariates e.g. physical activity or phone activity, conditioned by COVID. In this case, the treatment would be a physical activity indicator. The same approach could be used for some of the other covariates.

For our Goal 1, we need several assumptions:

Assumption1: We assume Dartmouth is stable except for the outbreak of COVID(i.e. For those students entering 2017 and 2018, they studied in a similar college environment before the outbreak of COVID).

Assumption2: There are no unmeasured confounding besides the covariates we list.

4 Data preparation

We will downselect and aggregate some of the data. Each observation will be a (individual $\times$ dated survey) pair. The pruned dataset will have a dimension in the order of ≈ 20 features, which will include

(1) Health survey metrics (e.g. could be the 4 question of PHQ-4)
(2) COVID (discrete/binary field indicating the stage of the COVID pandemic)
(3) Demographics
(4) Physical activity metrics (e.g. time active, running, walking)
(5) Sleep
(6) Social activity metrics (e.g. conversations, phone usage, geographic history)

Data on each feature would be an aggregate that is specific in some way to the (individual $\times$ dated survey). For example, a physical activity feature could be "average daily hours of exercise in the $< x >$ days preceding the survey". This approach gives us more samples than the original dataset which is only 200 individuals. One drawback would be a shorter time horizon. That is, we would not track individuals through the entire time of the pandemic.

References

Subigya Nepal, Wenjun Liu, Arvind Pillai, Weichen Wang, Vlado Vojdanovski, Jeremy F. Huckins, Courtney Rogers, Meghan L. Meyer, and Andrew T. Campbell. Capturing the college experience: A four-year mobile sensing study of mental health, resilience and behavior of college students during the pandemic. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1), mar 2024. doi: 10.1145/3643501. URL `https://doi.org/10.1145/3643501`.