1 Introduction

The first step in answering a causal question is to be precise about the question you want to answer. In this first step, you will get to brainstorm a causal questions you want to investigate. Picking a topic that you are interested in will make the project a lot more fun! It could even turn into a senior thesis eventually.

We will want to analyze data to answer the causal question, so we will need to make sure some data on the topic is available. You have two options. You can either take a look at the Add Health Data (described below) and think of a causal question involving the variables from that data set; or, you can come up with a question and find the appropriate data. A few potential data sources are listed out below to get you started. Also, make sure you take a look at the project overview so you can plan ahead as you pick your causal question.

Fill in the answers to the "Action Items" section below using the provided .Rmd file. Turn in Task 1 on canvas by **Oct 3** at 5pm.

2 Data

As described above, students may either choose to answer a causal question using the Add Health Data or find a data set of their own.

2.1 Add Health

The National Longitudinal Study of Adolescent to Adult Health (often referred to as "Add Health") is a survey of roughly 20000 individuals which was first administered to high-school aged participants in 1994-1995. The survey followed up with the same participants at 4 subsequent time points, as recently as 2018 when the original participants were in their mid 20s or early 30s. The questions asked in the survey cover a range of topics including: Crime/Delinquency and Victimization, Demographic Characteristics, Education, Family, Medication and Substance Use and Abuse, Psychological Well-being and Cognition, Reproductive Health, SES, Labor Market and Occupation. You can find a more comprehesive list of topics and specific survey questions here. To get an idea of the types of questions people have used the data to answer, take a look at journal articles citing Add Health here.

2.2 Other potential data sources

We want you to be creative in the causal questions you are asking! If you have a causal question of interest but aren't sure if you can get data, feel free to ask us for pointers. We've provided a few ideas to get you started below.

- Detailed data is available for many major sports. See this page for a good list of potential resources. As examples, we have listed two applications of causal inference to sports questions below
  - Yam and Lopez (2019) consider data from the NFL and ask what is the causal effect of trying convert on fourth down more often?
  - Cummiskey et al (2024) consider data from major league baseball and ask what is the effect of bunting on the probability of scoring at least 1 run?
- State and federal governments have open data portals which provide a variety of data sets which may be of interest
  - The Environmental Protection Agency (EPA)
  - NY State

  - The Federal Govnernment makes has a data clearinghouse data.gov
  - The Federal Reserve makes many macroeconomic data sets available through their FRED website.
- Opportunity Insights is a team of economists who study inequality and social mobility. They have their data publicly available here. They also have interesting data on COVID 19.
- The Inter-university Consortium for Political and Social Research (ICPSR) holds a number of political and social science data sets and organizes them into themes here. This includes topics like: Health and Medical Care Archive, National Addiction & HIV Data Archive Program, National Archive of Criminal Justice Data, and the Child and Family Data Archive.

3 Action Items

These are the questions you should answer and turn in for Task 1.

- (7 pts) Describe your causal question in a way that someone who has not taken this class would understand. Why are you interested in this question? How could answering this question allow for better decision making?
- (3 pts) What is the treatment? What is the outcome? Write out the potential outcomes using the notation we have used in class.
  - If your treatment is a variable that can take many different values, you could consider making it binary by simplifying the treatment in some way. For instance, if the treatment is the number of hours spent studying each day, you could dichotomize the treatment in the following way

$$A_i = \begin{cases} 0 & \text{if Hours } \leq 2 \\ 1 & \text{if Hours } > 2 \end{cases}$$

- (5 pts) How does the fundamental problem of causal inference apply to your question?
- (10 pts) We want to make sure there's some data available which gives you a chance of answer your question of interest. If you are using ADD health data list out the variable name for the treatment and outcome below
  - Treatment variable name:
  - Outcome variable name:

  If you are using an original data set, provide a link to the data which includes observations on both the treatment and the outcome: