Problem Set 4: Bias

The purpose of this assignment is to help you understand what variables you should control for when fitting a multiple regression model.

If you've taken a course on regression before, you may have been taught to control for any variable that affects your dependent variable. Or maybe you learned a procedure known as step-wise regression were variables are iteratively added and removed from a model based on p-values. These are generally considered bad practice now, and this assignment will help you understand why.

If regression is new to you, this assignment will provide a theoretical framework to help you think about modeling observational data.

Part 1: Reading

Read the following two short papers and answer the questions below in 2-5 sentences:

Building a better model: abandon kitchen sink regression

Causal Inference Is Not Just a Statistics Problem

1. What is the difference between a confounder and a collider? How should you address each in your models?
2. How can conditioning on a collider create bias?
3. Why can't statistical summaries or correlations alone tell us whether to control for a variable?
4. What is meant by a "kitchen sink" regression, and what is wrong with this approach to modeling?
5. What is a "backdoor path" and how does multiple regression help block these paths?

Part 2: Simulation

Think of some social causal relationship that you are interested in studying. Simulate a dataset for this phenomenon that contains the following:

- A treatment and outcome variable
- A confounder
- A mediator
- A collider
- An independent variable that has an exogenous effect on the outcome variable (i.e. it affects Y but does not affect any other variable in your DAG)
- An independent variable that has an exogenous effect on the treatment variable (this is also known as an instrument).

Refer to figure 2 in Causal Inference Is Not Just a Statistics Problem to help with generating the data, or draw your own DAG to illustrate the causal relationships.

Start by generating random data for variables that are not causally affected by any others in your DAG (e.g. the confounder and exogenous variables). Then, generate the remaining variables as linear functions of the variables that causally affect them. Each linear function should have beta coefficients that represent the true effect size, and a random error term.

1. Fit a model that recovers the direct effect of the treatment on the outcome variable. Which variables are necessary to recover the direct effect?
2. Fit a model that recovers the total effect of the treatment on the outcome variable. How does your model change to estimate the total effect?
3. How do your results change when you control for the collider, the exogenous independent variable, or the instrument (individually, not all simultaneously)?
4. Given the reading and simulation results, how should you choose which variables to include in a model?