Econometrics
Week 2: main notions

- *Econometrics model*
- *Correlation vs. Causation*
- *Ceteris paribus*
- *Random sampling*
- *Data structures*

Seminar 2: Introduction

Problem 1. Task: Exploring Simpson's Paradox with Berkeley Admissions Data Background: In 1973, UC Berkeley's graduate admissions data revealed a surprising statistical pattern. Your job is to investigate whether there was gender discrimination in the admissions process. The dataset *berkeley.csv* contains admission decisions (Admitted/Rejected) for male and female applicants to six graduate departments (A-F) at UC Berkeley in Fall 1973.

1. Overall admission rates: Calculate admission rates for males and females across all departments combined. Does there appear to be gender discrimination? Against whom?
2. Department-level rates: Calculate admission rates by gender for each department separately. In how many departments do males have higher rates? Females? Does this match part (1)? How do you explain any contradiction?
3. Application patterns: Examine (a) the number of male and female applicants per department, and (b) each department's overall admission rate. Which departments are most/least selective? Which departments did each gender apply to most frequently? How do these patterns relate to selectivity?
4. Explain the paradox: Using your findings from (1)-(3), explain in your own words what caused any reversal between overall and department-level patterns.

Create bar plots to illustrate your findings at each step.

Problem 2. Will Rogers paradox. The dataset in *will_rogers_paradox.csv* contains test scores for 18 students who were initially classified into two groups (Group A - Advanced, Group B - Standard) based on a preliminary assessment. After a more thorough evaluation, some students were reclassified to better reflect their actual performance level. Columns:

- ***student_id***: Unique identifier for each student (1-18)
- ***initial_score***: The student's test score (range: 52-95)
- ***initial_group***: Initial classification (A = Advanced, B = Standard)
- ***final_group***: Final classification after reclassification (A or B)

1. Calculate initial statistics: Mean score for students where initial_group = 'A'; Mean score for students where initial_group = 'B'; Overall mean score for all students.
2. Calculate final statistics: Mean score for students where final_group = 'A'; Mean score for students where final_group = 'B'; Overall mean score for all students
3. Create a comparison table showing: Group A average (before and after); Group B average (before and after); Overall average (before and after)
4. Answer these questions:
   - Which students changed groups?
   - Did Group A's average increase or decrease? By how much?
   - Did Group B's average increase or decrease? By how much?
   - Did the overall average change?
   - Did any individual student's score change?
   - How is it possible for both group averages to increase when no scores changed?
5. Create visualizations (box plots) showing the distribution of scores in each group before and after reclassification.

Problem 3. Suppose you are interested in estimating the effect of hours of private tutoring *hours* on final exam scores in calculus *exam_score*. The population is all first-year university students taking introductory calculus during the current academic year.

(i) Suppose you receive a research grant to conduct a controlled experiment. Explain how you would structure the experiment in order to estimate the causal effect of *hours* on *exam_score*.

(ii) Consider the more realistic case where you can only randomly sample *exam_score* and *hours* from the population. Write the population model as

$$exam\_score = \beta_0 + \beta_1 hours + u$$

where, as usual in a model with an intercept, we can assume $E(u) = 0$. List at least two factors contained in u.

Problem 4. Let *reading_score* denote the average reading comprehension score of students at an elementary school. Suppose we wish to estimate the effect of a free after-school tutoring program on student achievement. If anything, we expect the tutoring program to have a positive ceteris paribus effect on achievement: if a student whose parents work long hours and cannot help with homework gains access to after-school tutoring, his or her reading scores should improve. Let *tutor_prg* denote the percentage of students at the school participating in the free tutoring program.

(i) Suppose you are given funding to run a controlled experiment. Explain how you would structure the experiment in order to estimate the causal effect of *tutor_prg* on *reading_score*.

(ii) Consider the more realistic case where you can only randomly sample *reading_score* and *tutor_prg* from the population. Write the population model as

$$reading\_score = \beta_0 + \beta_1 tutor\_prg + u$$

where, as usual in a model with an intercept, we can assume $E(u) = 0$. List at least two factors contained in u.

Problem 5. Suppose you are hired by a fitness center chain to analyze the feasibility of adding evening yoga and pilates classes (7-9 PM) to their schedule. The company conducted a survey asking current members whether they would attend these new evening classes. You have data on 8,000 individuals collected over four months, including: age, sex, employment status, current membership type (basic or premium), frequency of gym visits per week, preferred workout time (morning, afternoon, evening), and whether they would be interested in attending evening yoga/pilates classes (yes/no). When building your model, you observe that 75% of respondents visit the gym primarily in morning hours (6-10 AM). Your colleague argues that your model will be biased and won't represent the preferences of the overall membership base. Is your colleague's criticism valid? Explain your reasoning.

Problem 6. Suppose you are hired by a tech company to study the factors that determine whether candidates who receive job offers actually accept them. You have a large random sample of candidates who were extended offers last year. Your dataset includes: whether each candidate accepted the offer, previous work experience, salary offered, benefits package details, location of the position, candidate's current city, and education level. A colleague comments, "Any analysis of this data will produce biased results because it's not a random sample of all job seekers in the tech industry, but only those who applied and were selected by this specific company." What is your assessment of this criticism?

Problem 7. A researcher wants to know if studying abroad improves language skills. They survey 200 students who studied abroad last year and ask them to self-rate their language improvement on a scale of 1-10.

1. Identify all the problems with this design.
2. Redesign the study to address these issues.
3. Specify data structure, variables, and sampling approach.
4. Explain what causation vs correlation issues remain.