# First Look at Causal Inference

# Web Page Designers Again

- You want to set up a company in some city to help businesses design their websites. You will need to hire a lot of web page designers, and you want to estimate how much you will have to pay them to get the additional manpower needed. You decide to hire a consultant to estimate labor supply elasticity of web page designers in your city

- Your consultant collects data on the average wage of web designers in different cities in China and the number of workers doing this job in those cities. She has learned some statistics and knows how to run a regression. She estimates an equation of the form:

$$\log(\text{number of designers}) = \beta \log(\text{wage}) + \text{controls}$$

- The elasticity of supply is $\partial \log(\text{employment})/\partial \log(\text{wage})$. So the estimated value of $\beta$ is her answer in the consultancy report

- What can go wrong?

# What Can Go Wrong

- Why are wages different in the first place?
  - cross sectional variations
  - time series variations
  - a combination of both (panel/logitudinal data)
- Do we have adequate controls?
- What does the correlation mean?

# Correlation Does not Imply Causation

- Omitted variables bias:

$$\begin{aligned}\text{incidence of heat strokes} &= \beta(\text{ice cream consumption}) + \text{controls}\\ &\quad + (\text{omitted controls})\end{aligned}$$

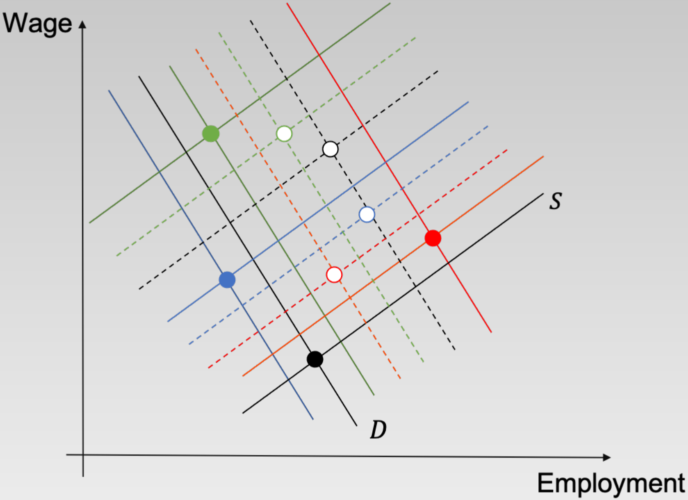
# Sample Selection

- Sample selection bias: The sample you use may not be representative of the population of interest:
  - conducting opinion survey using land-line telephone
  - in-hospital death rate of non-Covid patients in the U.S. increased from 2.1% to 2.6% between March 2020 and July 2020

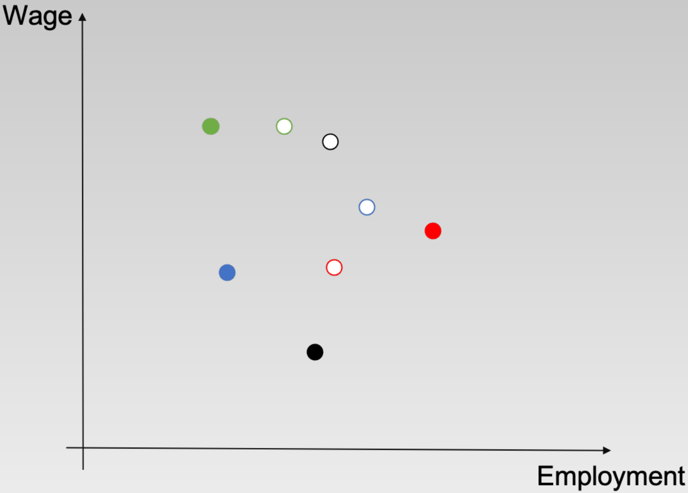
# Self Selection

- Self-selection bias: Is vaping bad for health?
    - compare health outcomes of smokers, vapers, and non-smokers after standardizing other observable characteristics
    - why does treatment status vary in the first place?
    - treatment status is the result of choice—the fact that seemingly identical subjects make systematically different choices suggests that their other (omitted) characteristics are different

# What Does the Correlation Mean



# But You Don’t Really See the Demand and Supply Curves



# Identification Problem

- Recall the regression equation:

$$\log(\text{number of designers}) = \beta \log(\text{wage}) + \text{controls}$$

  - If the consultant finds a positive estimate of $\beta$, she declares victory and pockets your consultancy fee
  - If she finds a negative value, she claims it is a demand elasticity and claims 80% of your consultancy fee
  - What actually is the estimated value of $\beta$ anyway?

- This is sometimes known as the identification problem/reverse causation problem/simultaneous equation system

- If you have two variables $X$ and $Y$ and you don't know which causes which, regressing $Y$ on $X$ doesn't make $X$ cause $Y$

# Large Natural Experiments

- Both labor demand and labor supply in Alaska shifts across years
- But our knowledge about the Alaskan economy tells us that the pipeline project is large and will induce a large rightward shift in demand
- It is unlikely that shifts in supply that also occur during this time period are as large
- We therefore expect the data to mostly reflect movement along a (roughly constant) supply curve induced by the large shift in demand → the observed changes (mostly) measure labor supply responses
- The black death example is the opposite: a large leftward shift in labor supply induces a movement along a (roughly constant) demand curve → the observed changes (mostly) measure demand responses