ECO220Y

Economic Questions and Data

Readings: Section 7.5 and Handout

Fall 2011

Lecture 5
Part 1 of 1

Economic Questions

- Remember that correlation treats X and Y symmetrically: corr(x,y)=corr(y,x).
- This can be clearly seen from $r = \frac{\sum z_x z_y}{n-1}$
- But what we really care is **"causal"** effect.
  - Causal means that change in one variable **causes** change in another variable
  - Correlation is not causation
  - Correlation is association
- Why do we care about causal effect?
- Policy questions: should the government invest to reduce the class size in schools? What is the impact of extended maternity leave on child health?

Step 1: Dependent and Independent Variables

Years of Education	$\Rightarrow$	Wage
Smoking	$\Rightarrow$	Lung cancer
Class size	$\Rightarrow$	Test scores
Peers behaviour	$\Rightarrow$	One's own behaviour

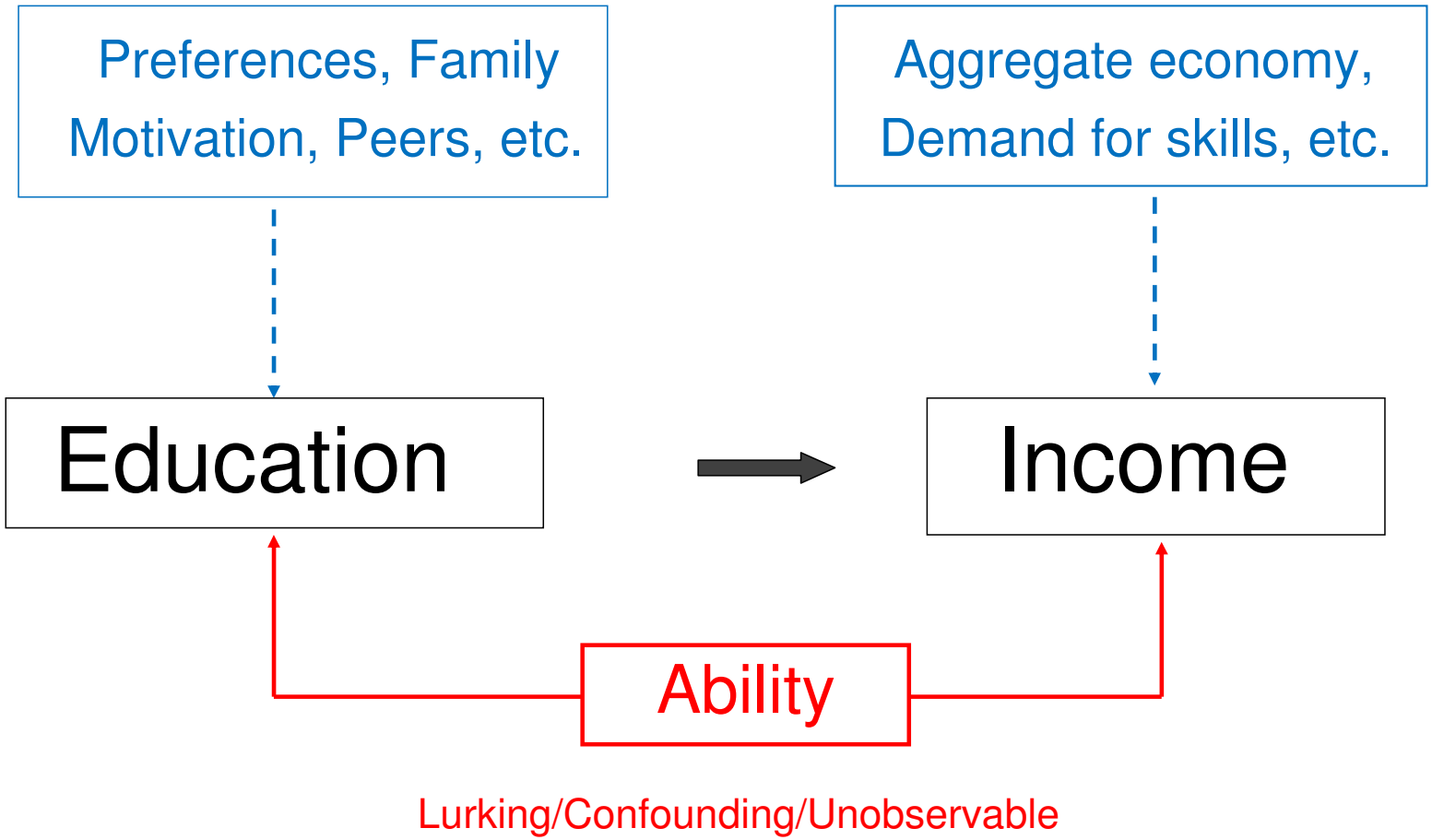
X	Y
predictor	response
independent	dependent
regressor	regressand
	outcome
Starts the process	Created by the process
What do we change?	What do we observe?

Note: We assign X and Y variables based on our subjective knowledge about the data generating process

Step 2: Endogenous and Exogenous Variables

- Choice variable vs pre-determined variable
  - Years of Education, Smoking - choice?
- Endogenous variable - choice variable
- Exogenous variable - chosen for us
  - Compulsory schooling - choice or pre-determined?
- Y-variable is always endogenous
- X-variable may be exogenous or endogenous
- If X is endogenous, we cannot be certain about the magnitude and sign of the effect of X on Y, i.e. cannot infer causality.

If X is endogenous, we cannot infer causality from the data. The reason is that there are other factors which affect our choice and at the same time affect the outcome, Y. We call these factors lurking variables, confounding variables or unobserved/unobservable variables.



Experiments

- Economists resort to experiments, or randomized controlled experiments (RCE) to manipulate the independent variable.
- Randomized controlled experiments are borrowed from drug trials when patients do not choose the drug dose, but rather the dose is chosen for them by the researcher.
- This allows estimating the effect of the drug independent of the patient's choice of the drug dose.
- Consider: a patient have trouble sleeping $\longrightarrow$ takes a pill $\longrightarrow$ sleeps better $\longrightarrow$ can we attribute this effect entirely to the drug?

Randomized Controlled Trials

- Divide sample randomly into two groups: treatment and control
- Control group serves as a comparison group for the treatment
- "Treat" the treatment group - manipulate, or randomly assign X variable
  - Example of treatment - drug dose
  - All individuals in the control group get placebo
  - All individuals in the treatment group get the drug
- Compare outcomes for control and treatment group$\longrightarrow$ causal effect

Types of Data (Again!)

- Observational Data
  - The most common type of data in economics, business and finance
  - All variables might be affected by choices and behaviour of agents
  - Unobserved/lurking variables$\longrightarrow$ Endogeneity bias
    - ★ Endogeneity bias - we incorrectly attribute the effect of some unobserved/lurking variables to X
- Experimental Data
  - Experimental data offer the most clean estimates
  - Frequently infeasible due to cost or moral objections
    - ★ Can we randomly assign smoking to individuals to assess health risks?
    - ★ Can we randomly assign marital status to parents to measure impact on children?
- Natural experiments - observational data with a flavour of experiment
  - A set of conditions that mimic the experiment

Sampling and Non-Sampling Errors

- Sampling Error - random difference between sample and population
  - Reason: sample is only a subset of the population.
  - Remedy: increase in the sample size.
- Non-sampling Error - systematic difference between sample and population
  - Reason: systematic errors in data collection.
  - Remedy: specific to the type of the non-sampling error.

Types of Non-Sampling Errors

1. Non-response bias
   - low response rate
   - non-respondents are not random
2. Selection bias
   - sampled population $\neq$ targeted population
3. Measurement error
   - systematic lying
   - poor survey design
   - recording errors