APPLIED ECONOMICS

Introduction: Economic Data

Outline

What is an empirical research about?

In general

- An empirical analysis uses data to test a theory or to estimate a relationship.

In this course in particular:

- The effect of class size on students' achievements
- The return to education
- Effect of minimum wage on employment
- Discrimination in the labor market
- Test economic theories (Growth determinants)

Stages of Empirical Research

- Make your research question clear. What is the causal relationship you want to analyze?
- Formulation of a model (or intuition)
- Formulation of an econometric specification
- Data collection
- Estimation
- Interpretation of the results

Typical Structure of a Research Output - Empirical Paper

- Introduction
- Literature Review
- Methodology (empirical/identification strategy)
- Data Description
- Results
- Good advice: Wooldridge, chapter 19

Economic Variables

What is an economic variable?

- An economic variable is a measure of some aspect of the economic reality

Aggregated Economic Variables

- National unemployment rate
- Regional Gross Domestic Product

Agent-Specific Variables

- Years of Schooling and Wages
- Firm's investment levels

The Economic Model

What is an Economic Model?

- the relation derived from basic economic principles—or assumptions—between a set of variables and another set of variables
- mathematically, an economic model is just a system of equations which relate a set of variables to another set of variables
- a model is a simplification/abstraction of reality which attempts to capture an important feature of reality

The Econometric Model

After we specify an economic model, we need to turn it into an econometric model.

- The form of the equations in the model must be specified.
- Decide which observable variables are appropriate.
- Decide how to deal with variables we cannot observe.

Endogenous vs Exogenous Variables

Endogenous variables

- the variables which are explained in the economic model
- in a model for labour force participation, the variable which measures the decision to participate is endogenous

Exogenous variables

- the variables which are taken as given in the economic model
- sometimes, the distinction is open to debate
- is a variable measuring the husband's labour force participation exogenous in a model of female labour force participation?

A Simple Human Capital Example

Human Capital Model

- wage equation: $wages = f_1(education, age, ability)$
- education equation:
  $education = f_2(wages, father's\ ed, mother's\ ed, ability)$

what is endogenous?

- wages are conditioned by productivity, which in turns also depends on years of schooling
- years of schooling are conditioned by the expected returns to education in wages

Observational Data

- first agents make economic decisions with measurable effects
- then economists observe (some of) these decisions: each observation is processed as a set of economic variables
- because
  1. researchers do not observe all relevant economic variables
  2. economic models are just models (incomplete descriptions of reality).
- applied economists work with probability models: a joint probability description of a set of random variables
- hence, the data is a realization of a set of random variables

An example of Observational Data

Returns to Education

- We are interested in the return to education: if a person is chosen from the population and given another year of education, by how much will her wage increase?
- Can we think of an experiment that allows us to answer this question (ceteris paribus)?
- Choose a group of people, randomly give each person an amount of education and then measure their wages
- BUT, individuals **choose** their educational levels and whether to work or not at market wages
- In practice, economists observe educational levels, wages, experience, parent's educational levels,...
- BUT **do not observe**, among other things, their ability

Types of Observational Data

cross sections: different individuals, firms, etc, at one period

- the ordering of the data in the dataset is usually not important.

time series: one market, region, etc. along many periods

- the frequency (yearly, monthly, weekly,...) and order are important features of the data.

panel data: a time-series for each individual in a cross-section

- usually, the number of individuals is much larger than the number of periods
- do not confuse with a repeated cross-section

Examples of Cross-Sections in Spain (1/2)

Encuesta de Población Activa

- quarterly survey at household level
- approx. 60,000 interviewed households (around 180,000 individuals)
- main objective: to obtain data on the labour force (which is subcategorized by employed and unemployed), and on the people outside the labour market

Examples of Cross-Sections in Spain (2/2)

Encuesta Anual de Estructura Salarial

- yearly survey
- information obtained from Social Security and income tax files at firm level
- over 15,000 firms and 150,000 workers
- main objective: to obtain data on labour earnings by different characteristics of the worker, such as sex, occupation, etc.

Examples of Time Series in Spain

Indice General de Precios al Consumo

- monthly survey of average consumer prices

IBEX 35

- daily stock market index of the Bolsa de Madrid, comprising the 35 most liquid Spanish stocks traded

Examples of Panel Data in Spain

Central de Balances del Banco de España

- annual voluntary reports from around 9,000 firms

aggregated unemployment rates from EPA

- Quarterly Unemployment Rates at regional level

Causality and Observational Data

When economic data is observational, correlation between two economic variables does not necessarily imply a causal effect

Returns to Education

- we want to know the effect of education as if it was assigned randomly (that is, not chosen by individuals)
- however, in reality individuals with higher ability will tend to have higher levels of education
- are their wages a result of their higher ability, their higher educational level, or both?

The Selection Problem 1/2

Do Hospitals Make People Healthier?

- consider the population that uses emergency rooms for primary care
- exposure to other sick patients by vulnerable people might have a negative impact on their health
- Using data about hospitalizations and health status of individuals from the National Health Survey in the US, we get the following table: (Health Status ranges between 1 -Bad and 5-Excellent):

		Health Status	
	Sample Size	Mean	Std. Dev.
Hospital	7,774	3.21	0.014
No hospital	90,049	3.93	0.003

Difference in means $= -0.72$

t-statistic $t = 58.9$

Source: Angrist and Pishcke (2008)

The Selection Problem 2/2

- in reality, people who go to hospitals are, on average, less healthy than people who do not go to hospitals
- the comparison is unfair: people go to hospitals because they already have poor health conditions!
- this problem is known as "selection bias"

The Selection Problem 3/3

- it is useful to ask about the ideal experiment (think about the hospital example)
- clearly this ideal experiment will be most likely impossible to carry out.
- then it will be necessary to face a selection problem.
- knowing the selection mechanism is key to find a solution.

Random Assignment Solves Selection Bias (1/3)

Do Subsidized Training Programs Help People Find Jobs?

- in reality, people who need training programs later have higher unemployment rates.
- the comparison is unfair: people who need training programs already have less ability to find jobs than people who don't.
- ideally, we would like to compare them with people with the same problem.

Random Assignment Solves Selection Bias (2/3)

- we can make the comparison fairer if we compare individuals with the same observed characteristics: years of education, labor experience,etc.
- however, some important characteristics, like innate ability or social connections, are usually not available.
- this problem is sometimes called "unobservable heterogeneity".

Random Assignment Solves Selection Bias (3/3)

a randomized field trial in a training program

- define the population of interest as "eligible".
- suppose that you can randomly assigned the training program among those eligible.
- those randomly excluded are, on average, a fair control group for comparison.
- randomization is common in medicine, and OLS techniques are enough to obtain good estimates of the policy effect.
- there is actually no need to control for any variable (such as the father's education) in addition to the policy variable.

Limitations of Randomization

- very costly to run.
- sometimes "randomization" is not "clean": there may be many alternative settings of interest, extrapolation may be important.
- actual assignment must ex–post not be related to output and eligibility (bureaucrats must cooperate!).
- sometimes attrition from treated and controls is not random.
- controls may be receiving another type of treatment (e.g. stigma effect).
- partial equilibrium in nature.

Examples of Natural Exogenous Interventions (1/2)

example 1: twins data and the returns to education

- mono-zygotic twins are genetically identical and share the family background.
- their differences in schooling are not due to innate ability or a different social background.
- causal interpretation in correlation between wage differences across twins and differences levels of education.

Examples of Natural Exogenous Interventions (2/2)

example 2: minimum wage effects on employment

- New Jersey increased minimum wages while neighboring Pennsylvania did not.
- NJ and Penn are not identical in many respects, so direct differences in employment not fair.
- but changes in employment in Penn a reasonable counterfactual of the changes in NJ had it not changed the legislation.