# Econometrics 1
# Lecture 1: Introduction

ENSAE 2014/2015

Michael Visser (CREST-ENSAE)

# Course outline:

1. Introduction
2. Simple linear regression (ch 2)
3. Multiple linear regression (ch 3)
4. Finite sample statistical inference: t-test, confidence intervals, F-test (ch 4)
5. Large sample theory and statistical inference (ch 5)
6. Functional form and specification (ch 6.2-6.4, 7.1-7.4, 9.1)
7. Heteroscedasticity (ch 8)
8. Repeated cross sections, panel data (ch13)
9. Panel data (ch 13/14)
10. Instrumental variables 1 (ch 15.1-15.6)
11. Treatment effects and instrumental variables 2 (ch 15)
12. Simultaneous equations (ch 16.1-16.4)

The chapters refer to chapters from the book “Introductory Econometrics: A Modern Approach” by Jeffrey Wooldridge (Thomson, South-Western, 2003). This is the main source used in the course.

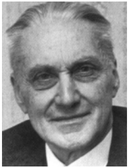
In today’s introduction (and occasionally in other lectures) I also use “Mostly Harmless Econometrics” by Joshua Angrist and Jörn-Steffen Pischke (Princeton University Press, 2009).

Throughout the course, the methods and techniques are illustrated with empirical results from the literature.

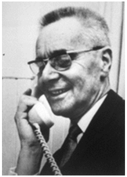
# Practicalities:

- 11 Weeks: 2 hours lecture + 2 hours exercise class
- Exercise class: standard exercises + computer-based exercises
- Final grade: 2/3 grade exam + 1/3 grade exercise class
- Grade in exercise class: 10 min quiz at beginning of each class
- Exam in mid-January

# First Nobel prize in economics was awarded to two econometricians



Ragnar Frisch



Jan Tinbergen

"To the layman, it may seem somewhat reckless to seek, without support from experiment, for laws of development within these extremely complicated processes of economic change, and to apply for this purpose the techniques of mathematical and statistical analysis." *The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 1969*

# What is econometrics?

Econometrics is a combination of two Greek words: “oikonomia” (administration or economics) and “metron” (measure).

It is the application of statistical and mathematical methods in the field of economics. Like “biometrics” is the application of these methods in the field of biology, like “psychometrics” is the application in the field of psychology, ...etc.

This mostly means:

- Testing economic theory
- Estimating relationship between economic variables
- Predicting future outcomes of economic variables

and it typically involves:

- economic models
- economic data
- mathematical statistics

Because econometrics is an interaction of economic theory, real data sets, and mathematical and statistical methods, it is often fascinating, but also challenging and sometimes difficult ...

# Why Econometrics?

- Econometric techniques and methods are needed to deal with the specific aspects of economic data ($\neq$ medical, psychological, or sociological data)
- The analysis of economic data plays a crucial role in understanding the world around us.
- Majority of published research in economics is empirical
  - For example, published research in labor economics: 80% was empirical in 1994-97 vs. 13% in 1965-69.
- Increasing availability of data in all fields in economics.

# Role of theory in econometric modeling

- Non-structural econometrics (also called reduced-form econometrics)
    - Framing of research question, guide for specification.
- Structural econometrics
    - Close link between functional form of econometric model and theory.
    - Aim:
        - Recovering economic primitives (preferences/technology)
        - Predicting outcomes under alternative economic settings
- Both share
    - Statistical techniques
    - Rely on theory to frame research question
- This course focuses on reduced-form econometrics.

# Example of a "structural model"

- Consider a first-price sealed bid auction (each bidder submits a sealed bid to the auctioneer, who collects all bids and assigns the object to the highest bidder; winner pays this highest bid).
- Bidder $i$ has a private (not observed by others) valuation $v_i$ for the object on sale.
- The $v_i$ are i.i.d. for $i = 1, ..., N$ (number of bidders), with a distribution function $F$ that has support $[0, \infty)$.
- Bidders are assumed to be risk-neutral
- It can be shown that the (unique) symmetric Bayesian Nash equilibrium is

$$b_i = s(v_i) = v_i - \frac{1}{(F(v_i))^{N-1}} \int_{p_0}^{v_i} F(u)^{N-1} du$$

  where $p_0$ is the reservation price of the seller.
- The equilibrium bid $b_i$ of the $i$th bidder is thus $b_i = s(v_i)$.

# Example of a "structural model" (cntd)

- Note that bidders bid below their valuations ($b_i < v_i$).
- Using data on bids from all bidders in a series of auctions, econometricians have developed methods to estimate $F$ using explicitly the above equilibrium relationship.
- Useful approach to predict the outcome in other auction mechanisms (English auction, second-price sealed bid auction, etc.).

# Example of a "reduced form model"

Suppose we are interested in the effect of job training on productivity. Economic theory tells us that:

- Education, experience and training all affect productivity.
- Wages are a function of productivity.

This would gives us the following "model":

$$productivity = f(educ, exp, train, other)$$

We could estimate the reduced form equation

$$wage = \gamma_0 + \gamma_1 educ + \gamma_2 exp + \gamma_3 train + v$$

where $v$ are other (unobserved) factors that determine wages.

# Econometric model: statistical inference

Once the econometric model is determined, we need:

- Estimation method to estimate the model (non-parametric, semi-parametric, parametric).
- The statistical properties of the estimators (small sample properties, asymptotic properties). Necessary to perform tests of statistical hypotheses and construct confidence intervals for parameters.
- Data. There are different types of data sets.

# Cross section data

- Data on individuals, households, firms, cities, countries, etc.
- For each unit (individual, household, ...) various variables are typically observed.
- Generally data are collected at one point in time (even if date of data collection varies somewhat across units it is ignored)
- Order of data does not matter (it does not matter whether unit $i$ appears before or after unit $j$ in the data set).
- Asymptotic properties of estimators are derived under the assumption that number of units tends to infinity and number of observations per unit is one: $N \to \infty$, $T = 1$.

# Time series data

- Repeated observation of the same variables over time (examples: GDP, interest rate, unemployment rate, prices)
- Order of data is important.
  - observations are typically not independent over time.
  - There is typically seasonality in the data.
  - Common trends over time in variables.
- These aspects need to be accounted for in the statistical analysis.
- $N$ fixed, $T \to \infty$.

# Panel data

- Panel data follows units over several time-periods.
- Advantages of panel data compared to cross-section data:
  - More observations and additional source of variation in the data.
  - Possibility to learn about dynamics
  - Multiple observations per individual allows the econometrician to obtain estimation that are robust to certain types of omitted variables bias.
- $N \to \infty,\ T > 1.$

# Repeated (pooled) cross sections

- Combination of several cross sections.
- Crucial difference with panel data is that we cannot follow units over time.
  - We cannot model and identify the dependence between observations of a given unit.
  - We cannot get rid of permanent unobserved heterogeneity of units.
- These types of datasets are more readily available than panel data (and cheaper to collect) and are in practice much larger than panel datasets.
- Useful to analyze the impact of some event (a policy chance for instance) where one is interested in behavior before and after the event.

# Causality and ceteris paribus

- The aim of econometric analysis is typically to estimate a causal relationship.
- Although purely descriptive studies are of interest, the most important research in economics is about questions of cause and effect.
- Causal relationships are useful for making predictions of what may happen following policy changes.
- The causal effect of a variable $x$ on a variable $y$ is the effect of the former on the latter keeping everything else constant (ceteris paribus).
- This is analogous to comparative statics in economic theory
    - compensated price elasticity: effect of price on demand keeping income or utility constant
- In many other sciences, causal effects are often measured using data from randomized experiments.

# Causality and ceteris paribus (cntd)

- In medicine, for instance, the effect of a new medical drug on the duration of a certain illness is typically analyzed by randomly assigning patients into a treatment group and a control group.
  - Patients in the treatment group receive the new drug.
  - Patients in the control group do not get the drug.
  - In both groups the illness duration is measured for each patient.
  - By contrasting the distribution of durations (or features of the distribution, such the mean, variance, deciles, etc.) in the two groups, one obtains the causal effect of the new drug on the illness duration.
- To illustrate the experimental approach consider another example.

# Effect of crop fertilizer on yield

- Question: what is the causal effect of a fertilizer on the yield of a crop.
- Select plots of land, vary quantity of fertilizer and compare yield across plots and relate this to fertilizer use.
- Yield depends on other things than fertilizer alone: land fertility, amount of rain, hours of sunshine, etc...
- The fertilization decision should not depend on these variables.
- Instead one should randomly decide how much fertilizer each plot receives. Comparison of the yields accross plots then allows to identify causal effects.

# Measuring causal effects in economics

- In economics, randomized experiments are rare because of
  - High costs
  - Moral issues (can we exclude someone from a job training program?)
  - Practical problems (non-participation, attrition)
- To measure causal effects, econometricians therefore almost always use observational data (survey data, administrative data).
- Identification of causal effects is then difficult.
- Suppose, for example, that we are interested in the effect of class size on the math results of primary school children.
- There are data sets (at the Ministry of education for instance) that record, for a sample of children, their math results and the size of the class to which they belonged.

# Measuring causal effects in economics (cntd)

- The difficulty with these data is that generally children have not been randomly assigned to classes of different sizes.
- Instead, school heads are typically the ones who decide which child goes to which class (if, as often, there are multiple classes of the same level).
- They usually decide to put the better and more mature children in large classes, and children who perform less well in smaller ones.
- Comparing school results in the small classes with those in large classes does not in this case identify the causal effect of class size. The naive comparison reflects in part a selection effect, i.e., the fact that school heads assigned the intrinsically better pupils to large classes.
- To illustrate the problem of identifying causal effect using observational data, let us consider two additional examples.

# Example 1: Effect of ENSAE on wages

- Question: what is the effect of ENSAE on wages of young people.
- Ideally we would like to have experimental data like in the fertilizer example (here people are plots, and ENSAE is like a fertilizer ...):
    - Choose a sample of young people, and randomly make half of them do the ENSAE.
    - Measure wages and identify the causal effect by comparing wages of those who did ENSAE and those who didn't.
- There are of course many reasons why such a randomized experiment is inconceivable in practice.
- We can instead collect observational data (e.g., survey data on wages earned by young people and their attained schooling levels).
- These data allow us to construct two groups of people: ENSAE graduates and those who didn't.

# Example 1 (cntd)

- The people in these two groups differ not only in their education level.
- For instance, ENSAE graduates are probably more motivated and ambitious than people who only did high school.
- On the other hand young people with just a high school diploma have more work experience than ENSAE graduates.
- Unlike in an experiment we cannot balance these other factors between those who do the ENSAE in reality and those who do not.
- The difference in earnings between the two groups also reflects the difference in the factors between these two groups, and not just the difference in education level.
- Hence a naive comparison of the two groups does not all allow to identify the causal effect of doing ENSAE.

# Example 2: Do hospitals make people healthier?

- The National Health Interview Survey records the answer to the following question: "During the past 12 months, was the respondent a patient in a hospital overnight?".
- Repondents are also asked to indicate what their current health status is (from 1=excellent to 5=poor).
- These data are summarized in the following table.

| Group | Sample Size | Mean health status | Std. Error |
|---|---|---|---|
| Hospital | 7,774 | 2.79 | 0.014 |
| No Hospital | 90,049 | 2.07 | 0.003 |

Source: Angrist & Pischke (2009)

# Example 2 (cntd)

- These results suggest that hospitals make you sick.
- The effect that we find (2.79-2.07=0.72) is probably not the causal effect of hospitalization on health, but captures in part that those who spent time in hospital in the past 12 months are people who are anyway less healthy (even if they hadn't gone to hospital).

# More formal definition of causality

- The concept of causality can be defined more precisely using the counter-factual framework pioneered by Rubin (1974).
- Let $D$ be the treatment indicator, equal to 1 if an individual received a given treatment and 0 otherwise. In the ENSAE example, $D = 1$ if the individual has a ENSAE diploma, and 0 otherwise. In the hospital example, $D = 1$ if the individual was hospitalized in the past 12 months, 0 otherwise.
- Let $Y_1$ be the outcome with treatment (salary of a person with ENSAE diploma, health status of a person who went to hospital) and $Y_0$ the outcome without treatment (salary without ENSAE, health without hospitalization).
- The causal effect of the treatment is $Y_1 - Y_0$, i.e., the difference of the outcome with and without treatment.
- Since a person cannot be in both states, we cannot observe both $Y_0$ and $Y_1$. That is why these two variables are called potential outcomes.

# More formal definition of causality (cntd)

- Using data, we can possibly only estimate entities like $E[Y_1 - Y_0]$ (average causal effect) or $E[Y_1 - Y_0|D = 1]$ (average causal effect on the treated).
- Define the observed outcome $Y$ ($Y = Y_0$ if $D = 0$, and $Y = Y_1$ if $D = 1$):

$$Y = Y_0 + D(Y_1 - Y_0).$$

- By comparing the mean outcome of the treated and the mean outcome of non-treated (as we did in the ENSAE and hospital examples) we are in fact estimating the following entity

$$\begin{aligned} E[Y|D=1] - E[Y|D=0] &= E[Y_1|D=1] - E[Y_0|D=0] \\ &= \underbrace{E[Y_1|D=1] - E[Y_0|D=1]}_{\text{Treatment effect for the treated}} \\ &\quad + \underbrace{E[Y_0|D=1] - E[Y_0|D=0]}_{\text{Selection bias}} \end{aligned}$$

# More formal definition of causality (cntd)

- In other words, the difference in sample means only identifies the causal effect (on the treated) if there is no selection bias.
- As mentioned already, there is reason to believe that the ENSAE and hospital data do not identify causal effects. That is because there is probably a selection bias in both cases:

$$E[wage_{\text{no ENSAE}}|ENSAE = 1] \neq E[wage_{\text{no ENSAE}}|ENSAE = 0]$$

$$E[health_{\text{no hospital}}|hospital = 1] \neq E[healthe_{\text{no hospital}}|hospital = 0].$$

# What randomization does

It balances potential outcomes across treated and controls:

- $E[Y_0|D = 1] = E[Y_0|D = 0]$
- $E[Y_1|D = 1] = E[Y_1|D = 0]$

which means that we

1. have no selection bias, and
2. will estimate an average causal effect.

# What econometrics does ...

- As mentioned above, randomized experiments are rare in economics.
- Large majority of empirical studies are thus based on observational data.
- The objective of most empirical studies is to overcome the selection bias (such as described in the the hospital example) and other types of biases encountered in the analysis of observational data.
- Econometric research has produced a large set of techniques for this purpose:
  - Regression methods
  - Difference-in-differences methods
  - Instrumental variables methods
  - And many other methods ...
- But nothing is perfect ...

# What econometrics does ... (cntd)

- In an influential paper, Lalonde (1986, AER) compared experimental with non-experimental estimates.
- He analyzed the National Supported Work (NSW) demonstration program. This was a temporary employment program to help disadvantaged workers lacking basic job skills move into the labor market.
- The NSW program assigned qualified applicants to training positions randomly.
    - mid 1970s
    - AFDC women, ex-drug addicts, ex-criminal offenders, high school drop-outs of both sexes
    - guaranteed a job for 9 to 18 months
- Baseline earnings and demographics (1975)
- Post treatment earnings and demographics (1979)

# What econometrics does ... (cntd)

- Research question: what is effect of temporary NSW jobs on earnings in 1979.
- Experimental estimates are based on comparing earnings in the experimental treatment and control groups.
- To obtain non-experimental estimates, Lalonde had a control group from other data bases (PSID, CPS-SSA).
- Non-experimental estimates are based on comparing the non-experimental control group and experimental treatment group (using econometric techniques).
- Conclusion from paper: experimental and non-experimental estimates differ considerably. The latter are in addition very sensitive to the particular econometric method used.
- Recently the Lalonde data have been re-evaluated with more sophisticated econometric methods (matching techniques):
    - some (Dehejia and Whaba, 1999; Dehejia 2005) claim one can recover experimental estimates.
    - others (Smith and Todd, 2005) have emphasized sensitivity of this approach.