ECON-122
Introduction to Econometrics

Agnieszka Postepska

July 11, 2011

What is econometrics

- No generally accepted answer (there is a entire journal article in *Econometrica* devoted to this question: Tintner, 1953)
- Some definitions:
  - "Econometrics is what econometricians do"
  - "Econometrics is the study of the application of statistical methods to the analysis of economic phenomena"
- In a nutshell: *Econometrics is a study about the use of statistical methods for investigating economic relationship, testing economic theories, and using estimated models to evaluate policy intervention in both the public and private sector*
- It's a tool that one can use in any field of economics

Who are econometricians

- economists
- mathematicians
- accountants
- applied statisticians
- theoretical statisticians

⇒ Bottom line: econometricians focus on ways to deal with the dirty data

Empirical Economic Research in Theory

- MOTIVATION: about what and why should we care?
- ECONOMIC MODEL: explanation of the process by which the relationships you are talking about occur
- ECONOMETRIC MODEL & DATA: turning theory into observables and finding appropriate method for the data
  → writing down the estimable equation
- ECONOMETRIC ANALYSIS: obtaining meaningful results

Empirical Economic Research in Practice

- MOTIVATION:
  - WHAT: anecdotal evidence that children of immigrants born in Germany perform worse at school than children of immigrants that arrived to Germany with their parents
  - WHY: this is in contradiction with economic theory: standard theory predicts that immigrants are disadvantaged due to cultural and language problems but these differences decline with economic assimilation → children born in Germany should perform better
- ECONOMIC MODEL:
  - immigrant newcomers are over-optimistic about their children's opportunities in Germany - they believe returns from education are very high
  - this optimism declines with time - parents that reside in Germany longer believe that returns are lower than parents that reside in Germany shorter
  - children form beliefs about returns to education based on information from their parents
- ECONOMETRIC MODEL & DATA: $S_i = X_i'\beta + \beta_0 A_i + \epsilon_i$

Theory vs. Intuition

- Econometrics can be used to test economic theory or just to investigate economic relationship - theory vs. intuition as starting points
- Example: decision to attend college
  - formally: utility maximizing problem conditioned on family budget, GPA, location etc.
  - start from empirical model - build an equation based on many theories and existing empirical studies

CROSS-SECTIONAL DATA

- sample of individuals, households, firms, countries (any kind of unit) taken at a given point in time
- examples: data on wages, occupations, labor force participation, happiness, pollution etc
- even if data is collected at different points of time we ignore it
- important feature: *random sampling* from the underlying population
  For example (reminder): we want to estimate the influence of children on labor force participation of women aged 20-30 in DC → the underlying population is all women aged 20-30 years old in DC and we draw a sample of say 1000 ■
- widely used in applied micro, but also in other social science

TIME SERIES DATA

- observations on a variable or several variables over time
- examples: stock prices, money supply, CPI, GDP, annual homicide rates, product sales etc
- observations cannot be assumed independent over time - need to take this correlation into account
- frequency of collecting the data matters - seasonal patterns in the data
- widely used in macro and finance

POOLED CROSS SECTION

- repeated cross-sections from the same population (not repeated observations on the same individuals)
- useful to examine effect of a policy: compare data from before policy was introduced and after
- often allows to increase sample size
- analysis much like cross section but account for changes over time ■

PANEL or LONGITUDINAL DATA

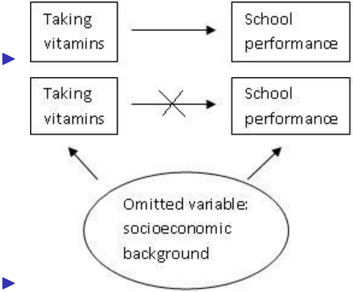
- repeated observations on the same individuals
- usually very hard to obtain and most of the time small number of repetitions
- very useful as allow to control for unobservable characteristics (if constant over time)

Causality and notion of *ceteris paribus*

- Causality
  - identifying *causal effect* is among main goals of empiricists
  - often a very difficult task: correlation vs. causality
  - correlation does not imply causality!!!
- Ceteris paribus
  - meaning: "other (relevant) factors being equal" or "holding everything else equal"
  - in practice - hardly impossible to satisfy this - relevant question: have enough other factors been held fixed to make a case for causality
  - simple graphical illustration

Simple example of why causality can be confused with correlation

Researcher asks a question whether taking vitamins by school age children lead to better results at school



Why is it worth looking at SLRM

- explaining one variable in terms of another (eg. wage as a function of education) "explain y in terms of x"
- not widely used in applied economics (why?)
- very useful in understanding regression model in general
- three issues to keep in mind when writing a model that will explain "y in terms of x"
    - the relationship is never exact - how do we allow for other factors to affect y?
    - what is the functional relationship between x and y?
    - how to capture ceteris paribus between x and y?

SLRM

$$y_i = \beta_0 + \beta_1 x_i + u_i \qquad \text{where } i = 1, ..., n$$

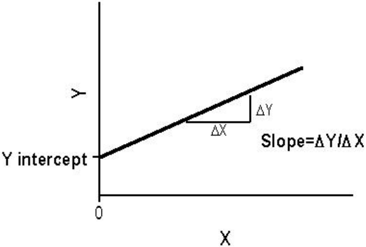
- n is the sample size
- i denotes individual observation
- y_i - the dependent (explained, regressand) variable (*observable*)
- x_i - the independent (explanatory, regressor) variable (*observable*)
- u_i denotes the error term: random component that changes for every observation and allow to account for other factors that affect y besides x (*unobservable*)
- β_0 is the **constant term** - to be estimated
- β_1 is the **coefficient** on x - to be estimated

SLRM cont.

The goal of this exercise is to infer how much our dependent variable (y) changes when we vary the independent variable (x).
In other words, if I'll give you the value of x you will be able to give me the corresponding average value of y (even if the x is not supported in the data) $\rightarrow E[y|x]$.
To be able to do that we need to know the value of β_0 and β_1 first - these are the two **parameters** of the model that we will estimate.
What does the following equation remind you of?

$$y_i = \beta_0 + \beta_1 x_i$$

Graphical interpretation



- β_0 corresponds to the intercept of the regression line: it tells us what the average value of y would be if x was zero
- β_1 corresponds to the slope of the regression line: it tells us by how much y changes when we vary x by one unit

Error term or disturbance

- disturbance represents all other factors that can affect y and we do not control for that - you can think of it as *unobservables*
- ceteris paribus in this context mean that when we look at the change of y as we vary x we hold u constant, so that $\Delta u = 0 \longrightarrow$ in this case x has a linear effect on y:

$$\begin{aligned} \Delta y &= \beta_1 \Delta x \quad \text{if } \Delta u = 0 \\ \beta_1 &= \frac{\Delta y}{\Delta x} \end{aligned}$$

This is exactly the slope of the function. Note that β_0 does not enter the equation above. Why?

Error term and Constant term

- Constant term is rarely core to the analysis but rather useful
- if we exclude the constant term from the regression equation, what is the expected value of y if x is zero?
- we need some assumptions about the expected value of u

Assumption 1:

The average value of u in the population is 0

$$E(u_i) = 0$$

- Not very restrictive - consider the following model:

$$\begin{aligned} E(u_i) &= a \\ y_i &= \beta_0 + \beta_1 x_i + u_i \\ \text{rewrite it as: } y_i &= \beta_0 + a + \beta_1 x_i + u_i - a \\ y_i &= \gamma_0 + \beta_1 x_i + v_i \\ \text{where } \gamma_0 &= \beta_0 + a \qquad v_i = u_i - a \qquad E(v_i) = 0 \end{aligned}$$

Error term and Constant term cont.

- it's important in the example above that a is constant - i.e. independent of x_i, the same for all observations
- it's not a statement about the relationship between x and u but about the distribution of the unobservables
- the next assumption, crucial
  in LRM, is a statement about the former and is MUCH more restrictive

  Assumption 2:

  The average value of u in the population doesn't depend on x:

  $$E(u|x) = E(u) = 0$$

- Note that this is an assumption-we cannot test it in our data. Consider the following example:

  $$\text{wage}_i = \beta_0 + \beta_1 \text{education}_i + \text{ability}_i$$

  What does Assumption 2 imply in this example?

Why do we do all that?

- we want to be able to make statements about the underlying population based on the sample
- "give me any value of x from the *population* and I will give you the corresponding **average** value of y
- in the example above: pick an individual from the population, tell me how many years of education this person has and I'll give you the average wage of individuals with the same level of education
- How does this work?

$$\begin{aligned} E(y|x) &= E(\beta_0|x) + E(\beta_1|x)E(x|x) + E(u|x) \\ E(y|x) &= \beta_0 + \beta_1 x \end{aligned}$$

- we would not be able to say anything about the average value of y without Assumption 2

Table: Cross-Sectional Data Set on Women Labor Force Participation

obsno	age	edu(years)	married	family inc	hours of work	children
1	22	11	1	20000	0	1
2	23	12	0	35000	20	2
3	21	8	0	70000	40	0
4	27	18	1	120000	15	2
5	29	14	1	0	60	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
999	20	11	1	80000	0	1
1000	30	13	0	75000	40	1

return

Table: Data on Wage, Unemployment, and Related Data for Puerto Rico

obsno	year	avgmin	avgcov	unemp	GNP
1	1950	0.2	20.1	15.4	878.7
2	1951	0.21	20.7	16	925
3	1952	0.23	22.6	14.8	1015.9
⋮	⋮	⋮	⋮	⋮	⋮
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

return

Table: Pooled Cross Sections: Two Years of Housing Prices

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85500	42	1600	3	2
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮
250	1993	243600	41	2600	4	3
251	1995	65000	16	1250	2	1
252	1995	182400	20	2200	4	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
520	1995	57200	16	1100	2	1.5

return

Table: A Two-Year Panel Data Set on Crime Statistics

obsno	city	year	murders	pop	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
⋮	⋮	⋮	⋮	⋮	⋮	⋮
299	150	1986	10	260700	9.6	286
300	150	1990	6	245000	9.8	334

return