The nature of econometrics and economic data

Empirical Methods, Lecture 1

Tunga Kantarcı, FEB, Groningen University, Spring 2025

What is econometrics?

Econometrics is the use of statistical methods to analyze economic data.

Typical goals of econometric analysis

Estimating relationships between economic variables.

Testing economic theories and hypotheses.

Evaluating and implementing government and business policy.

Steps in econometric analysis

1. Economic model – This step is often skipped

2. Econometric model

Econometrics, what is it?

Economic models

Maybe micro- or macro models

Often use optimizing behaviour, equilibrium modeling

Establish relationships between economic variables

Examples: demand equations, pricing equations

Economic model of crime (Becker, 1968)

Derives equation for criminal activity based on utility maximization:

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$$

where

- y is hours spend in criminal activities
- x_1 is wage for criminal activities
- x_2 is wage for legal employment
- x_3 is other income
- x_4 is other income
- x_5 is probability of getting caught
- x_6 is probability of conviction if caught
- x_7 is expected sentence

Here this equation is not specified.

Econometric model of crime

An equation can postulated without economic modeling:

$$crime = \beta_0 + \beta_1 wage + \beta_2 frepri + \beta_3 avesen + \beta_4 age + u$$

where

- *crime* is a measure of criminal activity
- *wage* is wage for legal employment
- *frepri* is frequency of prior arrests
- *avesen* is average sentence length after conviction
- *age* is age
- *u* is unobserved determinants of criminal activity such as moral character, and family background.

Most of econometrics deals with the specification of the error u. Econometric models are used also for hypothesis testing. E.g., the parameter β_5 represents the effect of average sentence length on crime. How large is this effect? Is it different from zero?

Econometric analysis requires data

There are several different kinds of economic data sets:

- Cross-sectional data
- Time series data
- Pooled cross sections
- Panel/Longitudinal data

Econometric methods depend on the nature of the data used.

Use of inappropriate methods may lead to misleading results.

Data, cross-sectional

Cross-sectional data

These may include samples of individuals, households, firms, cities, states, countries, or other units of interest at a given point of time or in a given period.

Cross-sectional observations are more or less independent.

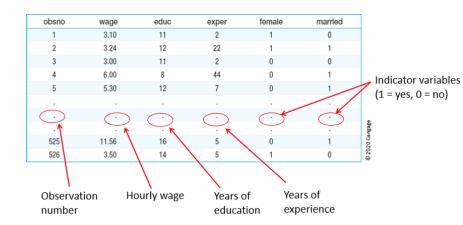An example is pure random sampling from a population.

Sometimes pure random sampling is violated, for example, people refuse to respond in surveys, or sampling may be characterized by clustering.

Cross-sectional data is typically encountered in applied microeconomics.

Cross-sectional data

Cross-sectional data set on wages and other characteristics

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

Indicator variables
(1 = yes, 0 = no)

© 2020 Cengage

Observation number

Hourly wage

Years of education

Years of experience

Time series data

This includes observations of a variable, or variables, over time.

Examples include stock prices, money supply, consumer price index, gross domestic product, annual automobile sales, and etc.

Time series observations are typically serially correlated.

Ordering of observations conveys important information.

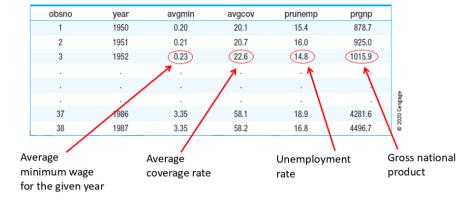Data frequency may include daily, weekly, monthly, quarterly, annually, etc.

Typical features of time series include trends and seasonality.

Typical applications include applied macroeconomics and finance.

Time series data

Time series data on minimum wage, unemployment, and related data for Puerto Rico

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
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

© 2020 Cengage

Average minimum wage for the given year

Average coverage rate

Unemployment rate

Gross national product

Data, pooled cross sections

Pooled cross sectional data

Two or more cross sections are combined in one data set.

Cross sections are drawn independently of each other.

Pooled cross sections are often used to evaluate policy changes.

Example:
- Evaluating effect of change in property taxes on house prices.
- Random sample of house prices for the year 1993.
- A new random sample of house prices for the year 1995.
- Compare before/after (1993: before reform, 1995: after reform).

Data, pooled cross sections

Pooled cross sections on two years of housing prices

Property tax

Size of house in square feet

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85,500	42	1600	3	2.0
2	1993	67,300	36	1440	3	2.5
3	1993	134,000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243,600	41	2600	4	3.0
251	1995	65,000	16	1250	2	1.0
252	1995	182,400	20	2200	4	2.0
253	1995	97,500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57,200	16	1100	2	1.5

Number of bedrooms

Number of bathrooms

Before reform

After reform

Panel or longitudinal data

The same cross-sectional units are followed over time.

Panel data have a cross-sectional and a time series dimension.

Panel data can be used to account for time-invariant unobservables.

Panel data can be used to model lagged responses.

Example:

- City crime statistics; each city is observed in two years.
- Time-invariant unobserved city characteristics may be modeled.
- Effect of police on crime rates may exhibit time lag.

Data, panel data

Two-year panel data set on city crime statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350,000	8.7	440
2	1	1990	8	359,200	7.2	471
3	2	1986	2	64,300	5.4	75
4	2	1990	1	65,100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260,700	9.6	286
298	149	1990	6	245,000	9.8	334
299	150	1986	25	543,000	4.3	520
300	150	1990	32	546,200	5.2	493

Each city has two time series observations

Number of police in 1986

Number of police in 1990

Data, observational vs experimental data

Econometricians often do not have experimental data so they have to rely on observational data.

Observational data is collected by observing and recording events or behaviors as they naturally occur, without any manipulation or intervention by the researcher. Think of administrative data on wages and hours worked which appear on salary slips. Think of survey data coming from interviews with people.

Experimental data is collected through controlled experiments where the researcher manipulates one or more independent variables to observe their effect on a dependent variable.

Effect of fertilizer on crop yield

By how much will the production of soybeans increase if one increases the amount of fertilizer applied to the ground?

Implicit assumption: All other factors that influence crop yield such as quality of land, rainfall, presence of parasites, and so on are held fixed.

Experiment: Feasible

Choose several one square meter plots of land, randomly assign different amounts of fertilizer to the different plots, compare yields.

The experiment works because amount of fertilizer applied is unrelated to other factors influencing crop yields.

Data, observational data

Effect of education on wages

If a person is chosen from the population and given another year of education, by how much will his or her wage increase?

Implicit assumption: All other factors that influence wages such as experience, family background, intelligence, and so on are held fixed.

Experiment: Infeasible

Choose a group of people, randomly assign different amounts of education to them (infeasable!), compare wage outcomes.

Problem without random assignment, amount of education is related to other factors that influence wages (e.g. intelligence).