1. Introduction

1.1 Phases of an econometric analysis

Empirical economic or econometric analyses:

Use of estimation and testing methods to test economic hypotheses or quantify economic relationships with data

Single phases of econometric analyses:

- Formulation of a research problem
- Formulation of an economic model or economic plausibilities and intuition
- Transition to an econometric model
- Formulation of hypotheses about the unknown parameters
- Estimation of the econometric model and testing of hypotheses with corresponding data

---

Example: Relationship between training and wages

The effect of in-company training on individual labor productivity and thus on wages is examined. It is assumed that especially education (educ), work experience (exper) and in-company training (training) can have an influence on wages.

Economic model:

$$\text{wage} = f(\text{educ}, \text{exper}, \text{training})$$

The following applies, for example:

wage: hourly wage, educ: number of years of schooling, exper: work experience in years, training: number of weeks in in-company training

Econometric model:

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{training} + \varepsilon$$

The betas represent the parameters and ε the error term (which contains other unobserved factors). Using individual data, these parameters are then estimated with appropriate estimation procedures and hypotheses about the parameters are statistically tested with appropriate test procedures.

1.2 Structure of economic data

Cross-sectional data:

- Data which are collected from units of the underlying population at a given time period (which may vary occasionally)
- The arrangement of the units in the data set does not matter for the analysis
- Starting point is mostly the implicit assumption that the data have been collected by random sampling
- Examples: Individual or household data (e.g. income, expenditures), firm data (e.g. sales, production, employment), city or country data (e.g. unemployment, gross value added)

Example (Earnings Structure Survey, Verdienststrukturerhebung 2018 – Research Data Center of the Statistical Offices):

Observation number	Monthly net income (in Euro)	Tenure (in years)	Weekly working hours	Age (in years)
1	848	18	20	60
2	1,570	13	40	44
:	:	:	:	:
1,456	2,444	32	20	56

Time series data:

- Data which are collected for one or more variables during several successive time periods
- Time is an important dimension (i.e. observations are often correlated over time) so that the arrangement of the observations in the data set contains potentially important information
- The frequency of data collection over time may vary strongly, e.g. daily, weekly, monthly, quarterly, and annual data with possible seasonal effects for intra-year data
- Examples: Macroeconomic data (e.g. income, consumption, investment, money supply, price index), financial market data (e.g. stock prices)

Example:

Observation number	Year	Fertility rate West Germany	Unemployment rate West Germany (in %)
1	1955	2.11	5.6
2	1956	2.20	4.4
:	:	:	:
68	2022	1.51	5.4
69	2023	1.41	5.8

Aggregated (pooled) cross-sectional data:

- Data which have both cross-sectional as well as time series characteristics since several cross-sectional data sets are collected independently over different time periods and linked to increase sample size
- Although the arrangement of observations in the data set is not essential, the corresponding period is recognized as an important variable
- The data are mostly analyzed like conventional cross-sectional data
- Examples: Individual or household data (e.g. income, expenditures) during several years

---

Example: (Earnings Structure Survey, 2014, 2018):

Observation number	Year	Monthly gross Income (in Euro)	Weekly working hours
1	2014	400	10
2	2014	4637	35
:	:	:	:
1262	2014	5003	40
1263	2018	1510	20
:	:	:	:
2524	2018	3535	20

Panel data:

- Data which have both a time series and a cross-sectional dimension where (in contrast to pooled cross-sectional data) the same units (e.g. individuals, firms, countries) are observed over several time periods
- The number of units is often much higher than the time dimension
- The data are often first sorted by units and then by periods
- The data provide the opportunity to control for unobserved characteristics of the units and to examine lagged variables
- Examples: Individual or household panel data (e.g. SOEP), firm panel data (e.g. MIP), country panel data

---

Example (SOEP 2018, 2019):

Observation number	Individual	Year	Monthly gross income (in Euro)	Gender
1	1	2018	1,567.21	Female
2	1	2019	1,649.59	Female
3	2	2018	2,362.25	Male
4	2	2019	3,970.55	Male
:	:	:	:	:
2751	1376	2018	2,195.89	Male
2752	1376	2019	3,116.80	Male

1.3 Causality and *ceteris paribus* term

→ In many empirical applications, the goal of an analysis is to find, confirm, and quantify causal effects

Causality:

Describes the sequence of interrelated events and states, whereby the effect (for the dependent variables) always follows the cause (for the explanatory variable)

Ceteris paribus term:

All other relevant factors remain constant. If all other factors are not assumed to be constant, no causal effect of a single variable can be measured

Measurement:

Ideally, an effect would be measured in a randomized, controlled experiment. This, however, is not possible for many economic questions.

→ Under certain conditions, econometric methods can simulate a *ceteris paribus* experiment