Econometrics II

Chapter I
Introduction and Concepts

Marie Paul

University of Duisburg-Essen
Ruhr Graduate School in Economics

Summer Semester 2020

Outline of lecture

1. Introduction and Concepts
   - Course Outline
   - Structural and reduced form models
   - Identification Concepts
   - Data types

Self-study plan for Lecture I

- For our first virtual meeting, please work through all slides of Lecture I.
- Please read Part One (this includes Chapters 1, 2 and 3; in the book I use this corresponds to page numbers 3 to 62, but page numbering seem to vary between different prints) in Cameron and Trivedi (2005) as far as the text relates to the slides (in particular, you do not need to read those parts that relate to SEM, because we will not cover SEM in this course and these parts are impossible to understand if you have not studied SEM before).
- Look up cited papers, data sets etc. in case your are interested in them.
- On the slides, you sometimes find questions under the heading "your own research". These questions relate to your own research or your research ideas. Write down your answers and discuss them with at least one fellow PhD student.
- Collect remaining questions and topics to discuss them in our virtual meeting.

Outline

Syllabus

- Instructors

Marie Paul
marie.paul@uni-due.de

Fabian Dehos
fabian.dehos@uni-due.de

- Time and place
  - Lectures: self-study and virtual discussion until further notice.
  - Tutorials: self-study and virtual discussion until further notice.
- Exam and assignments.
- Slides, problem set etc. are sent by email.
- These slides build on earlier versions of this course. Many thanks go to my precursors in teaching this course for passing on their material.
- The slides mostly follow **Cameron and Trivedi** [2005] and **Angrist and Pischke** [2009], and to a lesser extent, and **Wooldridge** [2010] and **Wooldridge** [2019].
- Plus some topic-specific papers.

Why Microeconometrics?

Microeconometric analysis: analysis of individual level data on the economic behavior of individuals or firms using typically cross section or panel data or regional data,

- Econometric methods are widely applied in **research**, **public policy**, and **business** and thus relevant for many occupations.
  - Powerful computers and large data sets have dramatically increased the potential for **quantitative** research.
  - Misuse may generate **incorrect inferences** that could lead to **misleading** conclusions and policy advice.
  - Social science is not natural science: Inference based on **untestable assumptions** in an **uncontrolled environment**.
- This course gives an overview of the core applied econometrics **toolbox** with a **practical** focus.
  - **Do's and dont's** of commonly used methods and when to use which model.
  - **When** to apply which estimator and **how** to interpret the results.
- **Empirical research in dissertation**. Capacity to **read empirical papers** and follow empirical **presentations**.

Topics

- **Part 1 – Linear models**
  - Introduction and Concepts
  - OLS and the linear regression model
  - Panel Data I
  - Panel Data II, Clustering and Difference-in-differences Estimator
  - Instrumental Variables Estimation
  - Regression Discontinuity design
- **Part 2 – Non-linear models**
  - Discrete response models I
  - Discrete response models II
  - Non- and Semiparametrics
  - Matching

Outline

Model structures

- Applied econometricians use **statistical models** to analyze data.
  - A statistical model is essentially a set of assumptions on the **process** from which observed data points are **generated**.
  - Such assumptions on the **structure** of a model are typically **non-testable**.
- The difference between a **structural** and a **reduced form** model is the **interpretation** of the parameters of the statistical model.
  - **Structural** models are derived from **economic theory** and therefore parameters have economic meaning. **Theory** plays a major role and is not just used to generate a list of variables used in a more or less arbitrarily specified functional form.
  - In earlier times structural models in microeconometrics usually meant **simultaneous equations models (SEM)**, today this often refers to models involving **dynamic stochastic optimization**.
  - Estimation of structural models requires more **assumptions** but is also more **informative** than reduced form parameters.
  - Structural economists also call empirical analysis that is not firmly based on a theoretical model **reduced form analysis**.

Model structures

Definition 1.1 (Model)

A **model** is the specification of the **probability distribution** for a set of observations. Examples:

1. a **full** specification of the probability distribution of the observations.
2. a **partial** specification of distributional properties such as **moments**.

Definition 1.2 (Structure)

A **structure** is the specification of the **parameters** of that distribution consisting of

1. a set of variables $\mathbf{W}$ (**data**) partitioned into $[\mathbf{Y}, \mathbf{Z}]$.
2. a joint probability distribution of $\mathbf{W}$, $F(\mathbf{W})$.
3. a priori ordering and restrictions of the variables in $\mathbf{W}$.
4. a parametric, semiparametric or nonparametric specification of the **functional forms** and **parameter restrictions** of the model.

Structural and reduced form

Definition 1.3 (Structural model)

For an observable, **interdependent** vector $\mathbf{y}' = (y_1, ..., y_G)$ and a set of i observations, a **structural model** is defined by

$$\mathbf{g}(\mathbf{y}_i, \mathbf{z}_i, \mathbf{u}_i | \boldsymbol{\theta}) = 0 \tag{1.1}$$

where $\mathbf{g}(\cdot)$ is a known function, $\mathbf{z}$ are observable factors and $\mathbf{u}$ are unobservable factors and $\boldsymbol{\theta}$ are structural parameters.

Definition 1.4 (Reduced form)

The **reduced form** of the structural model is the **explicit** form of (1.1) with $\mathbf{y}_i$ as a function of $(\mathbf{z}_i, \mathbf{u}_i)$

$$\mathbf{y}_i = \mathbf{f}(\mathbf{z}_i, \mathbf{u}_i | \boldsymbol{\pi}). \tag{1.2}$$

where the vector $\boldsymbol{\pi}$ is a function of $\boldsymbol{\theta}$. If $\mathbf{f}(\cdot)$ is a known function additively separable in $\mathbf{z}_i$ and $\mathbf{u}_i$ we can predict $\mathbf{y}$ by the linear **regression**

$$\mathbf{y}_i = \mathbf{f}(\mathbf{z}_i, |\boldsymbol{\pi}) + \mathbf{u}_i = E[\mathbf{y}_i | \mathbf{z}_i] + \mathbf{u}_i \tag{1.3}$$

Structural and reduced form: Model example

Example 1.5 (Supply and demand)

Structural model of demand and supply as a function of price:

$$Q^s = a_s + b_s P \tag{1.4}$$

$$Q^d = a_d + b_d P, \tag{1.5}$$

where $Q^s = Q^d = Q$ in equilibrium. To get the reduced form one must solve for the **endogenous** variables in terms of the **exogenous** variables

$$Q = \pi_0, \qquad P = \pi_1, \tag{1.6}$$

where

$$\pi_0 = (a_d b_s - a_s b_d)/(b_s - b_d) \tag{1.7}$$

$$\pi_1 = (a_d - a_s)/(b_s - b_d). \tag{1.8}$$

The 2 reduced form parameters (π_0, π_1) are functions of the 4 structural model parameters (a_s, a_d, b_s, b_d). Without additional structure we cannot retrieve the latter – the system is **underidentified**.

Structural or reduced form

Example 1.6 (Your own research)

Think of one example for structural analysis and one example for reduced form analysis in the economic field you are mostly interested in.

Why has the non-structual approach improved in recent two decades?

e.g. Angrist and Pischke (JEP, 2010):

- **Better and more data**, e.g. administrative records, long survey panels.
- Fewer distractions: functional form, heteroskedasticity tests etc. have lost importance.
- **Better research designs** and more transparent exposition: identification assumptions play a major role, (quasi-)experiments.
- Focus on **threats to validity** instead of sometimes erratic sensitivity analysis.

Example 1.7 (Non-structural analysis in the past and today)

- **Labor Supply Effect of Wages**: In the past: estimate effect of wage on hours using Heckman selection approach identified by normality assumption. Recently: field experiments in Africa offering different wages to randomly selected groups.
- **Wage Effect of Training:** In the past: Fixed-effects regressions using small survey data. Recently: matching or hazard-models using large administrative data with precise definition of training programs. Dealing with dynamic selection into training, common support etc..

Also the **structural approach** has made huge improvement during this period of time!

Outline

Causal analysis

Example 1.8 (Examples for Causal Claims)

- Smoking causes lung cancer.
- Seatbelt saves lives.
- Education increases earnings.

Example 1.9 (Your own research)

Can you formulate a research question you are interested in (or from your Master thesis) as a causal claim?

Three steps in causal analysis

- Define the set of hypotheticals or **counterfactuals** (e.g. economic theory).
- **Identify causal parameters** from hypothetical population data (e.g. identification results, natural experiment).
- Identify parameters from **real data** (e.g. coefficient estimation, inference).

Identification: Concepts

The goal of empirical research is to consistently estimate parameters of interest and conduct statistical inference.

- **Identification** is concerned with **determination** of a parameter **given** sufficient number of observations.
  - Identification is an **asymptotic** concept, logically occurring **prior to** and **separate from** statistical estimation.
  - Statistical inference is hence **not** concerned with identification.
  - Identification depends on untestable **identifying assumptions**.

- **Point identification** has traditionally been the focus in most of the econometric literature.
  - Point identification refers to a **single value** of a parameter.
  - This is in contrast to **set** identification where the parameter is estimated around an **interval**.

- Identification **can concern all parameters** in a model, a subset of parameters or only a particular function of the parameters.

Identification: Definitions

Definition 1.10 (Observational Equivalence)

Two structures of a model $\Pr[\mathbf{x}|\theta]$, $\mathbf{x} \in \mathbf{W}$, $\theta \in \Theta$ are **observationally equivalent** if $\Pr[\mathbf{x}|\theta^1] = \Pr[\mathbf{x}|\theta^2]$, $\mathbf{x} \in \mathbf{W}$.

In other words, if two structural models, given the same data, imply **identical joint probability distributions** of the variables, then the two structures are observationally equivalent.

Definition 1.11 (Identification)

A structure θ^0 is **identified** if there exist no other observationally equivalent structure in Θ given the same set of observations.

Hence, the definition concerns **uniqueness** of the structure given the data. Example of nonidentificattion: perfect collinearity between regressors in linear regressions.

Identification issues in SEM

Identification requires being able to obtain unique estimates of structural parameters. **A priori restrictions**, or **identifying assumptions**, have to be imposed to rule out the existence of other observationally equivalent structures.

- Identification in the SEM framework focusses on necessary and sufficient **conditions for identification** of parameters in a linear system of equations. Many of the identification results from SEM are helpful in other model context like 2SLS models or discrete choice models.
- Possible restrictions e.g. in the SEM context are **normalization** (e.g. setting diagonal elements to zero), **covariance restrictions**, **exclusions restrictions** (model contains variables that has zero impact on some endogenous variables), and **distributional assumptions**. Often also a rank condition is needed.
- In a **full-information structural approach** the entire distribution is identified. Issues like simultaneity and selection are addressed **explicitly**. Obtain a lot of information, but risk of misspecification may be large.

Identification issues using the CEF

An alternative to identifying a full model is to identify a **conditional expectation function (CEF)** of interested (also called structural conditional expectation) $E(y|\mathbf{w}, \mathbf{c})$.

- If we are willing to make an assumption about its **functional form** and if there are **no threats** like measurement error or unobserved variables such a CEF is under standard assumptions identified and the parameters can be estimated.
- If such threats occur: under additional identification assumptions we can sometimes still recover the CEF.
- Or can at least estimate parameters of it to obtain e.g. **marginal effects** (example: fixed effects model).

Identification strategies

Frequent practical threats to identification:

- Functional form misspecification
- Omitted variables bias
- Measurement errors and misclassification
- Reverse causality
- Sample selection

Popular identification strategies:

- Exogeneization: use data generated by natural experiments and quasi-experiments
- Elimination of Nuisance Parameters
- Controlling for Confounders
- Creating Synthetic Samples (e.g. comparison individuals as proxies for controls)
- Instrumental Variables
- Reweighting Samples

Identification strategies

Example 1.12 (Your own research)

Which identification strategies are typically used in empirical work in your research area?

Potential Outcome Model (POM)

- **Statistical framework for the estimation of causal parameters** departing from treatment evaluation and building on e.g. medical science: **Rubin** causal model.
- Causal analysis refers to comparing a **factual** with a **counterfactual**, but we only observe one (fundamental problem of causal inference).
- **Potential outcomes**: for each individual we imagine at least two outcomes. A set of counterfactuals defined for outcomes.
- Compare outcome under treatment with outcome under non-treatment. Only one is observed: use control persons. Derive **average treatment effect** or other parameter of interest.
- To compare outcomes of treated and untreated need random assignment or other assumptions on assignment (e.g. **conditional independence assumption)**.
- POM extended to non-binary treatments, dynamic assignment etc.

Potential Outcome Model

Example 1.13 (Your own research)

Can you formulate a research question you are interested in the POM framework? I.e. define the outcome, the factual situation, counterfactual situations, and the treatment effect to be estimated.

Identification issues in the POM framework

- In the POM framework identification is mostly concerned with threats to identification arising from e.g. **self-selection** or non-random sampling. Often estimation is semi-parametric.
- But also in the POM framework more involved models are used and this requires fundamental discussion of identification (e.g. dynamic treatment effect evaluation).
- Researcher try to **avoid functional form restrictions** (e.g. normality) as a requirement for **identification** (non-parametric identification) although for estimation these are often imposed.
- Role of **natural experiments** to search for credible sources of identifying information.

French and Taber (2011, *Handbook of Labor Economics, p.538-539*): *"If a feature of the model is not nonparametrically identified, then one knows it cannot be identified directly from the data. Some additional type of functional form assumption must be made. As a result, readers of empirical papers are often skeptical of the results in cases in which the model is not nonparametrically identified. [...] Some aspects of the data that might be formally identified could never be estimated with any reasonable level of precision. Instead, estimators are usually only nonparametric in the sense that one allows the flexibility of the model to grow with the sample size."*

Identification issues in the POM framework

Example 1.14 (Dynamic framework for program evaluation)

- Fitzenberger, Osikominu, Paul (2019) propose a dynamic evaluation approach in discrete time to estimate the effects of both training incidence and planned training duration on employment transitions.
- **Identification developed in the POM framework** based on results of other papers in the dynamic treatment effects literature.
- Identification relies on a no-anticipation condition, sequential randomization, time-varying covariates, and exclusion restrictions. As well as mild functional form restrictions (e.g. independence structure for error terms).
- Estimation is very flexible in some aspects: equations specified with (241 model parameters plus individual-specific effect-based parameters) but uses functional form assumptions in other aspects (normal distribution of the error terms $\rightarrow$ probit specification).

Outline

Experimental and observational data

Broadly speaking, data can be categorized into either **observational** or **experimental** data.

- Observational data refers to data generated in an **uncontrolled** environment; for example a **household survey**.
- Experimental data refers to data collected in a **controlled** environment; a **laboratory experiment** or a **field experiment**.

The **program evaluation approach** has cross-fertilized the two data types by exploiting quasi- or natural experiments as exogenous variation.

- The main problem with an experimental approach is that it is inevitably based on an **artificial** setting with often little scope for extrapolation of results.
- In contrast, a natural experiment is based on a **real-world setting**, both increasing external validity **and** retains a credible research design.

Experimental data: Overview

Experiments involve comparisons of outcomes between a **randomly selected treatment** group with those of a **control** group.

The main benefit is that the researcher can control the variation determining **treatment assignment** with a **minimum** of assumptions.

In addition, experimental data can be tailor-made to analyze and **predict** impacts of **hypothetical** policies not yet implemented.

Prominent examples of experimental studies in economics:

- **Giné et al (2010, AEJ:AE):** Subjects randomly assigned to a commitment device to quit smoking to analyze **rational addiction**.
- **Bertrand and Mullatain (2004, AER):** Send out fake job applications randomly assigning race and measure call-backs. Study **discrimination**
- **Fehr and Goette (2007, AER):** Subjects randomly assigned to different wage rates to study **intertemporal labor supply** elasticity.

Experimental data: Advantages and limitations

Main advantages

- Controlled **randomization** removes any unobserved heterogeneity between participants and non-participants.
- **Exogenizing** of policy variable leads to a simplified (often non-parametric) analysis.
- Can, by virtue of controlling data generation, create **hypothetical** situations to predict not yet implemented policies.

Main limitations

- Often very **expensive** to run which may limit sample size and statistical precision and flexibility. Sometimes **impossible** due to ethical reasons or reasons inherent to the question (e.g. dynamic selection).
- Random assignment may be **compromised** due to non-compliance, sample selection, Hawthorne, randomization and substitution biases.
- Does not take into account general equilibrium (feedback and interaction) effects limiting the **external validity** of results.

Observational data: Overview

Observational data are **real-world** data from surveys, administrative registers or other processes or censuses collected in an **uncontrolled** environment.

Observational data are mainly categorized into **cross-sections**, **repeated cross-sections** and **panel or longitudinal** data.

Cheap to use and relatively easy to obtain but inherently **difficult to analyze**; *raison d'etre* for the **existence** of the econometrics field.

Some popular observational survey data sets:

1. Panel Study of Income Dynamics (PSID)
2. German Socio-Economic Panel (GSOEP)
3. British Household Panel Study (BHPS)
4. Survey of Health, Ageing and Retirement in Europe (SHARE)

Popular observational process-generated data:

1. Social security data
2. Health insurance data
3. Data drawn from internet platforms (e.g. ebay, twitter)

Observational data: Advantages and limitations

Main advantages

- Generated by real-world **economic behavior** implies that observed patterns have high external validity.
- Often **free to use** and applicable for many different contexts and research questions.
- Often contains many **observations** making statistical precision less of an issue.

Main limitations

- Data inherently suffers from potential problems of **selection bias** and **mismeasurement** due to its uncontrolled nature.
- Data collection may be subject to issues relating to **sample frame**, **sample design** and **sample scope**.
- Inferences hinge on **untestable assumptions** which often causes controversies regarding arbitrarinesses in, e.g., model choice.
- Sometimes difficult to get access and **tedious to prepare**.

Process generated data

Example 1.15 (Sample of Integrated Labour Market Biographies, SIAB)

Source: webpage of RDC at IAB

- The Sample of Integrated Labour Market Biographies (SIAB) is a 2 percent random sample drawn from the Integrated Employment Biographies (IEB).
- Daily information.
- The IEB includes all persons in Germany who are characterized by at least one of the following employment status: employment subject to social security (in the data since 1975), marginal part-time employment (in the data since 1999), benefit receipt (since 1975), officially registered as job-seeking (in the data since 2000).
- These data, which come from different sources, are merged in the IEB.

Natural Experiments

Sometimes experimental conditions can be approximately replicated in observational data where a **natural experiment** have occurred.

The main advantage of such **quasi-experiments** is that assignment to the policy variable of interest can be treated as **exogenous**.

However, critics of this approach argues that many of the **problems** of experiments, such as external validity, also carries over.

Some prominent examples are

- **Ashenfelter and Krueger (1994, AER):** Using twins to analyze **returns to schooling** while keeping ability constant.
- **Card and Krueger (1994, AER):** Uses a policy change in New Jersey to analyze the effect of **minimum wages on employment**.
- **Card (1990, ILR):** Using an unexpected inflow of Cuban immigrants in Florida to analyze the **effect of immigration** on wages.

Observational Data or Experiment

Example 1.16 (Your own research)

In your field of interest: Are lab experiments used? Observational data? Natural experiments? Field experiments? Would it make sense to use them?