Motivation for regression analysis

2025/2026, Semester 1

Ralf A. Wilke
Copenhagen Business School

Before we start with more formal analysis….

- What is econometrics?
- Who is using it and for which purpose?
- Data structures
- The problem of causality

Econometrics is …

- estimating partial economic relationships
- testing economic theories
- evaluating and implementing government and business policies.

Examples:

- Evaluate the effectiveness of a publicly funded job training program
- Test different investment strategies of a bank to decide whether they comply with implied economic theory

- Formal economic modelling , e.g. a utility maximization framework, is often the starting point for empirical analysis.
- The economic model or our intuition provide us a mathematical relationship (equation).

<u>Example:</u>
Effect of training on the productivity of workers
(= higher wage)?

$$wage = f(educ, exper, training)$$

with *educ*=education, *exper*=experience
and a functional *f(.)*

- Turn the economic model into an econometric model:
  - Specify the functional *f(.)*.
  - How to deal with unobserved variables?
  - Introduce parameters β of the econometric model.

Example (cont.): a complete econometric model might be

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + u \ ,$$

where

- u contains all unobserved factors that can influence the person's wage. Examples?
- *f(.)* is a linear functional
- there are four parameters $\beta_0, \beta_1, \beta_2, \beta_3$.

Formulate a hypothesis for the unknown parameters: $\beta_3 > 0$

Typical Data Structures

- A big variety of data structures and how data is generated (experiments, interviews, administrative purposes, business activity).

- In economics, data is typically nonexperimental, i.e. not collected in laboratory environments.

- In the following some common data structures are presented:
  - Cross section
  - Time series
  - Pooled cross section
  - Panel or longitudinal data

- Cross sectional data:
  - Consists of a sample of individuals, firms, states or a variety of other units.
  - At a given point in time.
  - Requirement: Random sample of the underlying population.

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
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

- Time series data:
  - Consists of observations on a variable or several variables over time.
  - Examples include stock prices, consumer price index and automobile sales figures.
  - Chronologial ordering of observations conveys potentially important information.

obsno	*year*	*avgmin*	*avgcov*	*unemp*	*gnp*
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

avgmin: average minimum wage, *avgcov*: average coverage rate

- Pooled cross section data
  - Has both cross sectional and time series features.
  - For example, several cross section with the same variables at different point of time are pooled to one data set.

- Panel or longitudinal data
  - Consists of a time series for each cross-sectional member in the data set.
  - This implies that one can follow each cross-sectional unit over time.

A Two Year Panel Data Set

obsno	*city*	*year*	*murders*	*population*	*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	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

Causality…

- In most cases one wants to identify whether one variable has a causal effect on another variable, such as
  - Education on worker productivity
  - Price on quantity demanded
  - Job training for unemployed on employability or wages

- Ceteris paribus means “all other relevant factors being equal”.
- Most economic problems are ceteris paribus by nature.
- If other factors are not held fixed, then we cannot identify the causal effect.
- In empirical work the key question is: have enough other factors been held fixed to make the case for causality?

- Example: **Effect of Fertilizer on Crop Yield.**

  - Keep in mind that many factors determine crop yield. Which?

  The following experiment is conducted:

  - Choose several one-acre plots of land.
  - Apply different amounts of fertilizer to each plot and measure the yields.
    - This produces what kind of data?
  - Measure the association between yields and fertilizer amounts by statistical methods.

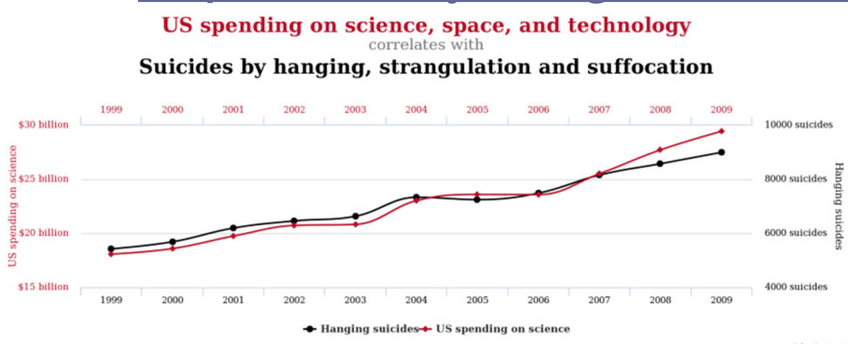
  - Why this may not be a good experiment?
    - Unclear how plots of land are chosen.
    - Other important factors are not observed.

  - When is this experiment useful to measure the causal effect?
    - If the levels of fertilizer are assigned to plots independently of other plots characteristics

Causality vs. correlation

- What can multivariate regression analysis add over competing approaches?

- When comparing only two variables, unrealistic bivariate correlations may be observed:
  - For example time series data: spurious correlations
  - Good resource: http://www.tylervigen.com/spurious-correlations



  - Solution: panel data analysis (if data available).

- Classification analysis is very common in business analytic.
  - May be based on high dimensional data structures.
  - Data fitting (maximise correlation between outcome and a number of predictors).
  - Does not reveal causation.
  - Classification may lead to discrimination and undesired inequalities based on spurious data artefacts. (Reference: Math Panic, Significance, 2016, Bursting Big Data Bubbles, 2017)
  - Solution: Define economic hypotheses and an economic model. Focus on consistent estimation of partial effects.
- By determining partial relationships, it is possible to dive deeper into the puzzle.

Econometrics and AI

- Econometrics is machine learning and therefore AI.

- Modern machine learning algorithm are increasingly used within classical econometric models to improve the accuracy of the fit of a model or to select model features such as variables.

- LLM can be used for code writing, interpretation of results (regression outputs, plots).
  - For example, the R-package **tidyllm** provides R an interface for interacting with the most common LLMs.

- As usual with AI tools, there is probably something right and something that is incorrect.
  - Code may not call the most appropriate methods, and it can be inefficient.
  - Result interpretation is superficial and does not inform about risks of the analysis (violation of model assumptions).

- To be able to assess the quality of these outputs, it is important to know the methods, whether assumptions hold and how to interpret the results.
  - This is what we do in this course.

Summary

The purpose and scope of econometric analysis:

- Used in all applied economic fields to test economic theories.

- Different data structures (cross section, time series).

- The notions of ceteris paribus and causal inference.

- While most hypothesis in the social sciences are ceteris paribus in nature, the nonexperimental nature of most data collected makes the estimation of causal relationships very challenging.