Introduction

Econometrics is measuring and analysing data for economic things. It's based on the development of statistical thinking to estimate economic relationships, theories and the evaluation and implementation of data.
There are 6 stages of the analysis of business or economic problems:

1. Understand the problems
2. Formulate and appropriate model to tackle the problem
3. Collect appropriate data
4. Look at the data
5. Estimate the model and make inference, predictions and policy prescriptions as appropriate
6. Evaluate, learn and improve the steps and repeat for each new improvement until the problem is solved.

Empirical analysis: uses data to test a theory or relationship.

A model is a function. We want to look at a theory and build a model from it.,

Sample: collection of observations.

Econometrics is used to predict or prescribe.
Prediction is where we use variables to guess what a target variable will do.
Prescription is finding the causes of a variable to control it with policy

The prediction of future values is also known as forecasting and the tools for it allows us to build forecasting models.

Prescription requires causation, not just correlation

Theories in economics and finance suggest how variables are related, which can be useful when building models. We can also use econometrics to test these theories.

OLS is a linear function with an intercept and slope that has been mapped to data.

We need to pick parameters using data. To do this we need to pay attention to the possibility of spurious regression, better estimations, exploiting the persistence of errors and feedback of data.

In business and economics it's unethical and expensive to use experimental data, so we have to use observational data. But this can lead to confounding factors.

The data structures of econometrics:

- Cross- sectional data: observations of variables in one point in time
- Time series data: observations of one or more variables at different points in time
- Panel: the same cross-sections at different points in time. This is done for each member and may also be called a longitude survey
- Pooled data: a cross-sectional survey at different times. This is useful for finding the effects of policy.

Time series data is ordered, while cross-sectional is not. Cross-sectional is assumed to be independent, while time series as temporal dependence. The data frequency also needs to be considered for time series.

Cross sectional regression: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i$
Time series regression: $y_t = \beta_0 + \beta_1 x_{t_1} + \beta_2 x_{t_2} + \cdots + \beta_k x_{tk} + u_t$
T denotes unit of time.
In both cases Y is the outcome/dependent variable, x is the independent variable, β is the slope, with β_0 the intercept, and u is the error term. We don't normally have this for the population so we need to estimate it, which means we put hats on everything, except the xs (because we know what they are) and get rid of u as there is no longer an error.

We need to think of all variables as random, meaning they have several outcomes, with a probability distribution. If there's no information a centre of its distribution is the best prediction.

The aim is to find one variable that has a casual effect on another.

It's important to get a feel of the data when we get it. To do this we do a preliminary data analysis. For this we figure out what type of data we have, and make sure the data looks right. If it's cross section we look at descriptive statistics and maybe a histogram then do the regression. If it's a time series it's more helpful to make it a line graph and look at things like seasonality, trends, structural changes and temporary shocks.
Seasonality happens inside a year (micro data), while cycles are longer than a year (macro data). They are both fluctuations.
When doing a line graph we need to look at which variable should be on which axis first. Then we ask ourselves, is there a relationship, is it strong, is it linear and is it positive or negative. These are all essential for the regression.

Different patterns means different types of data

Descriptive statistics tells us about our data. However we may need to adjust the data first if it's skewed or there's an outlier.
There are three main types:

- Measures of central tendency: this includes the mean, median and mode.
- The spread of the data: looks at the min, max, standard deviation and range
- The shape: looks at the skew and mode.

In all cases we have to put it in layman's terms and consider the meaning.
With skew we ignore the number and just look at if it's positive or not and look at the mode and the extreme values.