Week 1

Econometrics

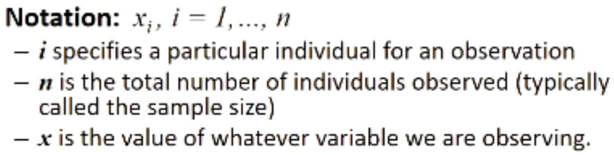
- Use of statistical methods to answer economic questions
- Describing the economic “landscape”
    - Annual growth rate of GDP
    - Unemployment growth rate
    - Do people with higher levels of education earn more

Data

- Set of measurements taken on a set of individual units, stored and presented in a *dataset*
    - *Variable*: any characteristics that is recorded for each case
    - Generally each case makes up a row in a dataset and each variable makes up a column

Types of Data

- A dataset records the value of one or more variables on several units of observation
    - A *variable*
        - A characteristic that we are interested in (e.g. GDP, unemployment status)
    - A *unit of observation* or case
        - Unit on which we measure each variable (e.g. country, person, company)
- Three classification criteria
    - Type of variable (numerical or categorical)
        - **Numerical:** naturally recorded in numbers (continuous or discrete)
        - **Categorical:** data recorded in groups (gender, religion, birth place)
    - Type of unit observation (cross section, timeseries, panel data)
        - **Cross section:** data collected on different entities at a common point in time (e.g. single year census data, unemployment rates by state for a particular year)

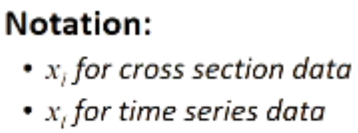
          

          Notation: x_i, $i = 1, \dots, n$

          – ***i*** specifies a particular individual for an observation

          – ***n*** is the total number of individuals observed (typically called the sample size)

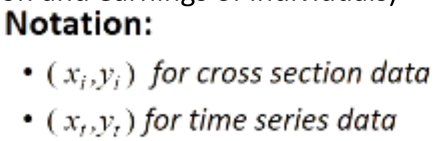
          – ***x*** is the value of whatever variable we are observing.
        -
        - **Time series data:** data on the same quantity at different points in time (order of observations is meaningful as it is based on time)
            - Examples: GDP of a country overtime, daily averages of S&P500, monthly unemployment rate
        - **Panel data:** data on *different entities* with each entity observed at *multiple points intime* (hybrid of cross section + time series)
    - Number of variables (univariate, bivariate)
        - **Univariate data:** single data series containing observations of only one variable (e.g. earnings of graduates in 2012, inflation rate from 1960-2000)



Notation:

- x_i *for cross section data*
- x_i *for time series data*

- **Bivariate data**: data composed of two potentially related data series (e.g. education and earnings of individuals)



Notation:

- (x_i, y_i) *for cross section data*
- (x_t, y_t) *for time series data*

- **Multivariate data:** data composed of three or more potentially related data series

Data Summary

- Typically use a combination of visual representations of the data and statistics
  - A variety of tables, graphs, and charts (scatterplots, histograms)
- Use statistics to measure characteristics of:
  - A single variable (mean, median, variance)
  - Relationships between multiple variables (covariance, linear regression)

Statistical Inference

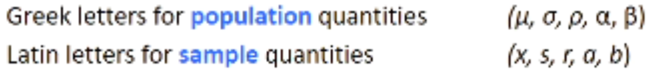
- Basic idea of **statistical inference** is to draw conclusions about a relationship we cannot observe
  - No definitive conclusion as sample is only observed, not population
- **Statistical inference**
  - Using what we know about the sample & probabilities of reaching certain conclusions - make statement about probably characteristics of variables at population level

Interpretation

- Back to the economic model - what does this mean *economically*

Statistics Assumptions

- Assume that our dataset is a ***sample*** taken from the population
- From this dataset, we calculate an ***estimator*** for the true but unknown parameter
- Standard practice



Greek letters for **population** quantities	$(\mu, \sigma, \rho, \alpha, \beta)$
Latin letters for **sample** quantities	(x, s, r, a, b)

Univariate Data Summary

- **Univariate data** are a single series of data that are observations on one variable
  - E.g. numerical data, categorical data

Types of Summary statistics

- Central tendency
  - Where is the center of the distribution of the data (mean, median, mode)
- Dispersion
  - How spread out is the data
- Skewness (asymmetry)
  - How symmetric is the distribution
- Kurtosis (Peakedness)

    - How fat are the tails, how tall is the peak

Sample Mean

- Most common way to measure central tendency (sample average)
    - $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample Median

- Value that divides sample into two halves
    - 50% above 50% below
- Orders data from lowest to highest value
- Less sensitive to outlier than sample average

Sample Mode

- Most frequently occurring value in sample
- Does not make sense for continuous data: all observations are going to be different
- Useful with discrete data where particular values are meaningful

Measures Dispersion

- Characterise the dispersion, spread or width of the distribution
- Sample variance:
    - We use squared deviations so that we only get positive differences
    - $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$