Topic 1: Introduction

1.1 Introduction I

- Econometrics is a discipline which uses mathematical and statistical techniques to analyze data of various types.
- Historically, econometrics developed as a sub-discipline of economics and econometric models were initially used to analyze economic data.
- However, modern econometric techniques are used to model data from a wide variety of disciplines, including
  - finance
  - marketing
  - management
  - actuarial studies
  - politics

1.2 The linear regression model - an overview I

1.2.1 The architecture of the linear regression model

- By far the most commonly used model in econometrics is the **linear regression model**.
- In this unit we will study in detail the properties of the linear regression model.
- For the purpose of illustration, assume that we are interested in analyzing the relationship between three random variables, y, x_1 and x_2 (random variables are discussed in Topic 2), and also assume that we have available a sample consisting of n observations on each of these variables.

1.2 The linear regression model - an overview II

1.2.1 The architecture of the linear regression model

- In the linear regression mode, we assume that the relationship between y, which is called the **dependent variable**, and x_1 and x_2, which are called the **regressors or explanatory variables**, can be represented by the following equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i, i = 1, 2, ...n. \quad (1)$$

- In (1) y_i denotes the ith observation on the variable y, x_{i1} denotes the ith observation on the variable x_1 and x_{i2} denotes the ith observation on the variable x_2.
- For example, the first three equations of the model given by (1) are

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + u_1, \\ y_2 &= \beta_0 + \beta_2 x_{21} + \beta_2 x_{22} + u_2, \\ y_3 &= \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + u_3. \end{aligned}$$

- In the equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i, i = 1, 2, ...n. \quad (1)$$

- $$\begin{aligned} y_i, \ i &= 1, 2, ...n, \\ x_{i1}, \ i &= 1, 2, ...n, \\ x_{i2}, \ i &= 1, 2, ...n, \end{aligned}$$

  are all **observed.**
- The n observations on each of y, x_1 and x_2 constitute the **data** upon which our empirical analysis is based.
- The quantities β_0, β_1 and β_2 are **unobserved**, fixed numbers (not random variables) which we call the **regression coefficients**. The regression coefficients are also called **parameters**.

- $$u_i.i = 1, 2, ...n,$$

  is a set of **unobserved** random variables.
- A primary goal of regression analysis is to use the data to estimate the unobserved regression coefficients β_0, β_1 and β_2.
- In the model given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i, i = 1, 2, ...n, \quad (1)$$

the relationship between y, x_1 and x_2 is assumed to have a systematic part, given by

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}, .i = 1, 2, ...n$$

and a purely random part given by

$$u_i.i = 1, 2, ...n.$$

- Note that even if we knew the values of $\beta_0, \beta_1, \beta_2, x_{i1}$ and $x_{i2,}$ we would not know the exact value of y_i because we would not know the value of the random error term u_i.
- The model given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i, i = 1, 2, ...n, \tag{1}$$

  is called a **linear** regression model because it is linear in the parameters β_0, β_1 and β_2.
- That is, β_0, β_1 and β_2 all appear in (1) raised to the power of 1.
- By contrast, the model given by

$$y_i = \beta_0 + \beta_1^2 x_{i1} + \sqrt{\beta_2} x_{i2} + u_i . i = 1, 2, ...n,$$

  is not linear in the parameters because β_1 and β_2 do not appear raised to the power of 1.

1.2 The linear regression model - an overview VI

1.2.1 The architecture of the linear regression model

- A reasonable question to ask is "how do we know that the relationship between y, x_1 and x_2 is linear in the parameters β_0, β_1 and β_2"?
- The answer is we don't know, we simply assume that it is linear. (As you will discover, we love to "make up stuff" in econometrics!).
- However, in many applications it is often reasonable to assume that the relationship between the dependent variable and the regressors is linear in the parameters, or at least that the assumption of linearity is a good approximation to the true relationship.
- In the linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i, i = 1, 2, ...n : \tag{1}$$

  - β_1 measures the change in y in response to a small change in x_1, holding the value of the variable x_2 fixed.

- β_2 measures the change in y in response to a small change in x_2 holding the value of x_1 fixed.

- For those of you who are familiar with multivariable calculus,

$$\frac{\partial y_i}{\partial x_{i1}} = \beta_1, \frac{\partial y_i}{\partial x_{i2}} = \beta_2.$$

- For example, if y_i is hours worked by individual i, x_{i1} is the wage rate faced by individual i and x_2 is individual $i's$ years of experience, then:
  - β_1 measures the change in hours worked by individual i in response to a small change in the wage rate of individual i, holding constant individual $i's$ years of experience.
  - β_2 measures the change in hours worked by individual i in response to a small change years of experience of individual i, holding constant individual $i's$ wage rate.

- Because they measure the change in the dependent variable in response to small change in x_1 and x_2 respectively, β_1 and β_2 are called **marginal** or **partial** effects.

- As previously mentioned, a major goal of regression analysis is to measure the marginal or partial effects of the explanatory variables on the dependent variable.
- While there are several estimation methods available to accomplish this task, the simplest and most commonly used method is called the method of **Ordinary Least Squares** (OLS).
- The associated estimator is called the **Ordinary Least Squares (OLS) estimator.**
- In this unit we will restrict our attention to the study of the OLS estimator.

- More advanced estimation methods are studied in third-year units such as ETC3410 and ETC3400.
- Of course, we want an estimator which produces estimates β_0, β_1 and β_2 which are in some sense "accurate".
- In this unit, we will study the statistical properties of the OLS estimator and show that, under certain assumptions, the OLS estimator has good statistical properties and produces estimates of the regression coefficients (and other parameters) which are in some sense accurate.
- Note that an estimator of the unknown parameters β_0, β_1 and β_2 is a rule or formula which tells us how to use the data to produce **estimates** of β_0, β_1 and β_2.

- It is important to distinguish between an **estimator** of an unknown parameter, and an **estimate** of the parameter.
  - An estimator a parameter of interest is a formula or rule which specifies how to use the data to produce a numerical value which represents our "best guess" about the value of the parameter.
  - An estimate of a parameter of interest is the numerical value we get when we plug the data into the formula given by the estimator.
- For example, we will see later in the unit that the OLS estimator of the parameter β_1 in the linear regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i . i = 1, 2, ...n, \tag{2}$$

is given by the formula

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{3}$$

- Equation (3) is a rule which tells us how to use the data that we collect in order to product an estimate of β_1 in (2).
- When we collect our data and plug the values of x and y into (3) we obtain an estimate of β_1.
- We know the formula for the estimator before we even collect the data.
- However, we can only compute the estimate after we have collected the data.

- In some applications, particularly when working with time series data (see below), we may be able to use our estimates of β_0, β_1 and β_2 to produce forecasts of future values of y.
- For example, in the simple time series model given by

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t, t = 1, 2, ...n,$$

we could use the formula

$$y_{n+1}^f = \widehat{\beta}_0 + \widehat{\beta}_1 y_n$$

to forecast the value y in time period n+1, where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ denote the OLS estimates of β_0 and β_1 respectively.

1.2 The linear regression model - an overview II

1.2.3 Forecasting and hypothesis testing

- In addition to estimating the regression parameters in

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i, i = 1, 2, ...n, \quad (1)$$

  we often wish to test hypotheses about the behavior of the variables in the model which are of interest.
- For example, in the model given by (1), if y_i is hours worked by individual i, x_{i1} is the wage rate faced by individual i and x_2 is individual $i's$ years of experience, we may wish to test the null hypothesis hours worked is unresponsive to small changes in the wage rate, when we hold years of experience constant.
- Testing this null hypothesis is equivalent to testing the null hypothesis that

$$\beta_1 = 0$$

  equation (1).

1.2 The linear regression model - an overview III

1.2.3 Forecasting and hypothesis testing

- Notice that in order to test the null hypothesis that x_1 has no effect on y once we have controlled for x_2, we must first convert this null hypothesis expressed in words into a restriction on one of the parameters in the model.
- In this unit we will study how to test a variety of hypotheses about the parameters of the linear regression model.

- Another reasonable question to ask about the general linear regression model with k explanatory variables given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_k x_i.u_i.i = 1, 2, ...n, \tag{4}$$

  is "how do we decide which variables to include as regressors"?
- In an econometric analysis we often wish to give a causal interpretation to the regression coefficients.
- For example, we may wish to interpret β_1 as the change in y "caused" by a small change in x_1, holding the values of the other regressors constant.
- When conducting such causal analysis, we need to think very carefully about what other variables we need to control for in order to be able to plausibly interpret β_1 as the change in y **caused** by a small change in x_1.

1.2 The linear regression model - an overview II

1.2.4 Choosing regressors

- In some cases economic theory suggests what variables we should use as regressors.
- For example, if y_i is aggregate demand for a particular demand in time period i, economic theory suggests that we should include as regressors in our linear regression model variables such as the price of the good in time period i, the prices in time period i of goods which are close substitutes for y and aggregate income in time period i.
- In other cases we have to rely on common sense to guide our choice of explanatory variables.
- If , for example, y in

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_k x_i . u_i . i = 1, 2, ...n, \tag{4}$$

is the probability of an individual contracting lung cancer, then common sense would suggest including as regressors variable such as an individual's age and whether or not they are a smoker.

- There are four data structures which are commonly encountered in econometrics:
  1. **Cross-sectional data**: a cross-sectional data set consists of observations on one or more variables taken at the same point in time. In the case of cross-sectional data, all the observations are collected at the same point in time.
     - For example, observations on infant mortality and GDP per capita for all OECD countries in 2021.
  2. **Time series data**: a time series data set consists of observations on one or more variables taken at different points in time. In the case of time series data, the observations on the variables are collected over multiple time periods.
     - For example, monthly observations on the rate of inflation and interest rates in Australia from 1990 to 2021.

3. **Pooled cross-sectional data**: a pooled cross-sectional data set consists of observations on cross-sectional units, such as individuals, firms, countries, collected at different points in time.
    - For example, monthly observations on earnings of Victorian workers in 2020 and 2021.
4. **Panel data**: a panel data set consists of observations on the same cross-sectional units (such as individuals, firms, countries) collected at different points in time.
    - For example, monthly observations on earnings of the **same** 100 Victorian workers collected in 2020 and again in 2021.

- Notice that what distinguishes a pooled cross section from a panel data set is that in the case of a pooled cross section the members of the cross section are not the same in each year, whereas in the case of a panel data set they are.

- In this unit, we restrict our attention to cross-sectional and time series data structures.

- There are two important differences between cross-sectional and time series data sets.

1.
   - In the case of cross-sectional data, there is no natural ordering of the observations. For example, in the case of a data set consisting of observations on individuals, it doesn't matter whether we record the data for Mr. A before the data for Mr. B or vice versa.
   - Observations on time series data are ordered. There is a natural ordering in time. For example, GDP in the first quarter of the year precedes GDP in the second quarter, which precedes GDP in the third quarter and so on.
2.
   - In the case of cross-sectional data, it is usually reasonable to assume that the observations are independent of each other.
   - For example, in a cross-sectional data set containing observations on individuals, it is usually reasonable to assume that information about Miss A does not contain any information about Miss B and vice versa.

- In the case of a time series data set, it is unreasonable to assume that the observations are independent of each other.
- This is because time series data are usually characterized by some form of **temporal dependence**. For example, because quarterly growth rates of GDP are typically correlated, the growth rate in quarter 1 may contain information about the growth rate in quarter 2.

- Because time series data are characterized by temporal dependence, time series data sets are generally more challenging to work with than cross-sectional data sets.
- Time series data can be used to accomplish two important tasks for which cross-sectional data are inadequate. These tasks are to:

1. Forecast future values of a variable.
   - For example, if we have quarterly observations on Australian GDP from 1900q1 to 2021q2, we can use these observations to forecast the value of Australian GDP in 2021q3.

2. Estimate the **dynamic causal effect** of one variable, x, on another variable y. That is, estimate how a change in the value of x today affects the value of value of y over several time periods.
    - For example, if the Australian Government imposes a tax on alcohol today, the response of consumers will typically be distributed over several time periods, as consumers gradually adjust their consumption of alcohol in response to the new tax. We can use time series data to estimate the effect of the tax on alcohol consumption both today and in future time periods

3 Experimental versus Non-experimental (observational) Data I

- The data available to scientists to conduct empirical analysis can be divided into two broad categories.
- These categories are **experimental data** and **non-experimental data**, also know as **observational data**.
- Experimental data are data that are generated by conducting a controlled experiment.
- As the name suggests, observational data are data that are collected by making observations on outcomes in the real world.
- Experimental data are generated by conducting a **randomized controlled experiment**.
- The key ingredients of a randomized controlled experiment involving individuals are:
    - The treatment

3 Experimental versus Non-experimental (observational) Data II

  - The control group
  - The treatment group
  - Random assignment of the individuals in the experiment to the treatment and the control groups.

- For example, a drug company which wished to test the efficacy of a new drug for cancer might select a random sample of cancer patients and randomly assign half the patients to the treatment group, who receive the drug, and the remaining patients to the control group, who receive a placebo.
- Because the assignment of patients to the control and treatment groups is random, there are no systematic differences between the individuals in the two groups.

3 Experimental versus Non-experimental (observational) Data III

- Since there are on systematic differences between the individuals in each group, if a higher proportion of individuals in the treatment group than in the control group experience a remission in their cancer, we can reasonably infer that this effect has been caused by the cancer drug.
- Unfortunately, in the social sciences, including econometrics, researchers typically have to rely on observational data, as they rarely have access to experimental data.
- As we discuss below, it is much more difficult to estimate the causal effect of one variable on another when using observational rather than experimental data, the estimation of causal effects significantly more difficult in the social sciences than it is in the natural sciences.

- If two variables x and y are **correlated**, that is, informally speaking, they move together, we can use x to **predict** or **forecast**, y.
- In this case, x is called a **predictor** of y.
- Some examples of prediction are:
  - Using past and present rates of unemployment to forecast the future rate of unemployment.
  - Using the proportion of people in the CBD carrying an umbrella in the morning to forecast the probability of rain during the day.
  - Using median house prices in a school district to forecast school quality.
- A critical point to note is that in order for x to be a useful predictor of y, it is not necessary for x "to cause" y.
- For example, median house prices in a school district may be a good predictor of school quality in that district even though high house prices don't "cause" schools in the district to be of high quality.

- Indeed, causation is likely to run in the opposite direction, with high school quality attracting large numbers of home buyers to the district, leading to high house prices.
- A sufficient condition for x to be a useful predictor of y is that x and y are correlated.
- The higher the correlation between x and y, the more useful x is likely to be as a predictor of y.
- Forecasting is sometimes referred to as **predictive analytics**.

- In principle, econometric methods can also be used to determine whether or not a change in a variable x causes a change in another variable y and, if such a causal relationship exists, to estimate by how much y changes in response to a given change in x.
- Such causal models are frequently of great interest to policy makers.
- However, given that econometricians almost always work with observational data, causal analysis is fraught with difficulty and must be undertaken with great care.
- For example, suppose that we are interested in whether or not obtaining a university degree **causes** individuals to earn higher wages.

- Assume that we collect data on wages and educational status for a random sample of 1000 individuals and specify the following linear regression model

$$wage_i = \beta_0 + \beta_1 uni_i + u_i, \tag{5}$$

  where uni_i is a dummy variable which assumes a value of 1 if individual i has a university degree and a value of zero if they do not. (Dummy variables are extensively discussed later in the unit)
- If we estimate (5) and find that $\widehat{\beta}_1$ is positive and statistically significant, can we conclude that having a university degree **causes** an individual to earn higher wages?
- To answer this question, recall our discussion above of the elements of a randomized controlled experiment.
- In the current example:
  - the treatment is obtaining a university degree

  - the treatment group consists of those individuals in the sample who have a university degree.
  - the control group consists of those individuals in the sample who do not have a university degree.

- However, a key feature of a randomized controlled experiment is missing.
- The individuals in the sample have not been randomly assigned to the treatment and control groups.
- Some individuals in the sample have **chosen** to be in the treatment group (by choosing to obtain a university degree) and some have **chosen** to be in the control group (by choosing not to obtain a university degree).

- Therefore, we have to carefully consider whether or not there are systematic differences between the people who have chosen to obtain a university degree (the treatment group) and those who have chosen not to obtain a university degree (the control group), and if there are systematic differences between the individuals in the two groups, whether these differences rather than differences in educational status may explain the observed differences in wages.
- Can you think of any systematic differences between the individuals in each group that might be important in explaining the differences in wages between those with and those without a university degree?
- When working with observational data, one must always carefully consider the possibility of systematic differences between those who have chosen to be in the treatment group and those who have chosen to be in the control group, and the impact of such differences on one's ability to make causal inferences.

- If we are satisfied that we have established a causal relationship between the dependent variable and the regressor(s) of interest, the estimated magnitude of the causal effect may be used to inform economic policy.
- For example, if we can show that increasing the minimum wage causes a decrease in poverty, that finding may inform government policy on the appropriate level for the minimum wage.
- Empirical analysis which is used to inform economic or social policy is sometimes called **prescriptive analytics**.

‘

- A third purpose for which econometric modelling is used is to test economic theories (and theories from other disciplines).
- For example, the "quantity theory of money", first proposed by the Nobel Prize winning economist Milton Friedman, states that, at least in the long run, the rate of inflation in an economy will be equal to the rate of monetary growth.
- A crude way of testing this theory (much more sophisticated tests are available) would be to estimate the following linear regression equation

$$\pi_i = \beta_0 + \beta_1 m_i + u_i, \;\; i = 1, 2, ..., n, \tag{6}$$

where

$$\begin{aligned} \pi_i &= \text{rate of inflation in country i,} \\ m_i &= \text{rate of monetary growth in country i,} \end{aligned}$$

and test the null hypothesis that

$$\beta_1 = 1.$$

- If we reject the null hypothesis, this would be evidence against the quantity theory of money.