Econometrics I (ECON 371)
Lecture 1
Introduction to Econometrics

Teng Wah Leo

1 What is Econometrics?

Econometrics is the application of Statistical Methods in estimating economic relationships, which are then used to test economic theories and/or evaluating or justifying the implementation of government (social, business and economic) policies, and/or business strategies. Some examples of the application of econometrics are:

1. Forecasting of key macroeconomic variables such as growth or decline in interest rates, inflation rates, Gross Domestic Product (GDP).
2. The study of the impact of immigration policies on the wages of native workers.
3. The study of the impact of divorce laws on marriage and divorce rates.
4. The study of the impact of compulsory education laws on the overall attainment of children.
5. The study of reasons when and why a company may choose to enter into a market.
6. The examination of how the demand and supply side of matching markets find each other, and clears the market optimally.
7. The examination of why consumers make their product choices in both actual markets and online searches.

In no way are these examples exhaustive, and there are at least as many econometric studies as there are economic theories. This is because it is a common practice to first examine an a priori relationship perhaps due to economics or otherwise. If that relationship is significant, we can then seek a rigorous theoretical explanation. You should be asking yourself at least the following questions:

- How do we know when a relationship is significant?
- How do we even know the form of the relationship which would put our initial examination in question?
- What is the best technique of estimating those relationships?
- What are some of the estimation issues?

None of which are trivial questions, and the more you delve deeper into Econometrics and Statistics, the more questions will be born!

The most important input/resource that allows to answer all our questions lie hidden in information, what is commonly known as data, and of which there are several types. The organization/structure, not to mention their abundance in terms of data points/number of observations, and variation will determine the accuracy of our answers to our questions to answer differing types of question. Some of the common types of data are:

1. **Non-experimental/Observational Data**: Some examples would be Canadian Census collected by Statistics Canada, and Survey Data such as Canadian National Longitudinal Survey of Youths (CNLSY). This sorts of data are easily obtainable at some cost or totally free.
2. **Experimental Data**: This sorts of data are collected in controlled environment so as to study behavior and phenomenon accurately. They are often found in natural sciences. Historically, such data are hard to come by in the social sciences, including in Economics. However, with the advent of technology with the increase in computational power, these are becoming the best kind of data used in business. Nonetheless, as you will see, even these sorts of data suffer very similar setbacks the traditional non-experimental data suffers, as they apply in the social sciences.

The reason why mathematics is important to the study of Statistics and Econometrics is that data is structured as matrices. Consequently, all data manipulations in the process of estimation is dependent on our understanding of mathematics. The common structures of data are:

1. **Cross Sectional Data**: Data collected at a fixed point in time. The arrangement of the matrix is such that each observation is arranged in rows, with the variables of each observation indexed on a column. Each observation is the basic unit such as an individual, town, city, state or country.

2. **Time Series Data**: Data collected across time. Here the structure of the matrix is the same as in Cross Sectional, with the exception that the each observation is time, which could be years, quarters in a year, months, weeks, months, days, hours, minutes, seconds, ...

3. **Pooled Cross Sectional Data**: This refers to stacking of cross sectional data one on top of the other either in ascending or descending order. The order of which has no consequence, it is purely aesthetic, after all we are economist. Here, the observations in each period are not the same. That is suppose the observation in each period is an individual, pooled cross section does not involve the same individual in each period. The reason we can still derive meaningful conclusions is from the idea that this observations are typically representative of the general population. However, this in and of itself, is still not sufficient. What allows us to do that? It is Statistical Theory that allows us to extrapolate these facts. This is the reason we will go through some basic statistical theory subsequently.

4. **Panel or Longitudinal Data**: This refers to data collected of a fixed set of observations across time. In other words, if the basic unit of observation is an individual, information on these individuals are collected over a specific duration at fixed intervals. Thus Panel data collects information on a fixed set of individual observations across time, so that we can see how they behave, and thereby determine the reasons for those choices. CNLSY is an example of a panel data.

2 A Simple Example of Empirical Analysis

Empirical Analysis uses data to test a theory or estimate a relationship. We will examine an idea to see the entire process at work. Suppose a priori you think that direct funding to university students while they are completing their degree raises the incentive for them to work hard and do their best. Then if we have data on random application of direct funding that substantially reduces the debt burden, we can write the relationship of this samples Grade Point Average (GPA) as a function of their high school average grade (HiGrade), income of the family equivalized by the number of sibling (EqIncome) and whether they receive additional funding (Funded). (What kind of data should this be?) This could be expressed succinctly as

$$\text{GPA} = f(\text{HiGrade, EqIncome, Funded}) \tag{1}$$

Before we can use our data to examine the veracity of the relationship, there are several considerations,

1. What is the form of the function $f(.)$? In other words, the type of mathemathematical equation the relationship should take? Or is it better not to make that assumption? How do we proceed if it were the latter that we choose?.

2. How do we deal with variables we cannot observe? For example, income of parents tells us the degree of budget constraint on the child's decision. But it says nothing about the natural ability of the child? We could however conjecture that we should then include parental education (PrtEduc) to capture the non-pecuniary aspect of biological inheritance from our parents. (This makes the strong assumption that PrtEduc is a sufficient statistic. An idea we will examine later in the course).

Commonly, at least for this course, we can give the general function $f(.)$ a particular structure, such as an additive model

$$\text{GPA}_i = \beta_0 + \beta_1\text{HiGrade}_i + \beta_2\text{EqIncome}_i + \beta_3\text{Funded}_i + \beta_4\text{PrtEduc}_i + \epsilon_i \tag{2}$$

where the subscript i indexes the observation. We refer to $\beta_0, \beta_1, \ldots, \beta_4$ as the parameters of this econometric model which we are trying to estimate. This is because it tell us the degree of effect the variables have on the sample observations GPA. ϵ is the term for unobserved factors, which may include factors we have no way of observing and summarizing as a variable. It is often referred to as the error term or disturbance term. It is the most important term in econometric analysis which will be revealed as we proceed.

Once we have estimated the model with the data in hand, the significance of the parameters can be tested by examining if it is significantly different from zero, which is equivalent to saying the variable has no effect. With this results we can then examine how such a relationship could be derived from economic theory if what exists, or to develop one if it does not already. If the theory produces a different direction between the relationships of the variables in question, or a different functional relationship altogether, we can then return to the data to reexamine if the results differ when we change the estimating equation. Of course we are being very general here, and the difficulties of the entire process will be revealed as we proceed.

The example we have used presumes that funding has an effect on your GPA. We have eliminated the possibility of the causality from funding to GPA, that is your higher GPA resulted in funding simply because we randomized the allocation of funding. In reality, data we use hardly ever has such pretty structure, hence the value of experimental data. In addition, when performing econometric analysis, we assume that everything else we do not include, is held constant, i.e. *ceteris paribus*. But is anything ever constant? We will address these questions as the course proceeds.