Econometrics

Professor Martins - Spring 2025

1 Introduction

The principal objective of any science is the generation of testable hypotheses and conjectures regarding its objects of study. These conjectures and hypotheses are tested by direct confrontation with relevant observed reality. It is this direct confrontation with reality that allows for the establishment of an accepted body of scientific knowledge. The repeated formulation and evaluation of scientific conjectures *via* the observation of relevant phenomena is the essence of scientific inquiry and learning.

Economics would not be a *bona fide* social science if the propositions that emerge from the theoretical models in microeconomics, macroeconomics, industrial organization, development economics, and finance, to name just a few of its specialty branches, were not or could not be evaluated by direct confrontation with relevant observed phenomena.

The observation of economic phenomena for scientific purposes is not a simple task. Economic reality is an incredibly complex web of transactions and relationships among economic agents, and it is impossible to understand it without the use of *theoretical models*. Theoretical models are simplifications of reality which attempt to capture and relate its most salient features. The purpose of theoretical modeling is to create a system of assumptions and relationships, that is a simple enough abstraction of reality to allow for its understanding, and to draw conclusions regarding its evolution. Theoretical models can be built and presented in a variety of ways, but modern economic theory has evolved into a collection of **mathematical** models of economic reality. Hence, economic behavior and interaction among economic agents is normally described by sets of mathematical concepts, equations and inequalities that entwine economic variables. In practice, the observation of economic reality is in essence the observation of the economic concepts and variables that appear in theoretical models.

An important question that arises is how adequately theoretical models represent reality. Theoretical models can be evaluated from different perspectives. At one level, theoretical models can be evaluated based on their internal mathematical consistency. A good theoretical model should be such that its conjectures and hypotheses result deductively from its assumptions. Although internal consistency is a necessary property of any theoretical model, it is not enough to give it scientific validity. As mentioned before, scientific validity of an economic model can result only from testing its conjectures and hypotheses

by observing relevant economic reality, normally in the form of economic variables. Thus, at a more sophisticated level, theoretical models must be contrasted with reality.

As an example, consider a traditional model of consumer behavior. A first course in microeconomic theory normally starts with a mathematical model of consumer behavior based on some assumptions on consumer preferences and constraints. Under these assumptions, a variety of conjectures are drawn about consumer behavior. They are obtained through a set of logical mathematical arguments. One of these conjectures is that there is a relationship between the quantity demanded of a product q, its price p, the prices of other products ρ and the consumer's wealth w. Mathematically, such relationship can be represented by

$$q = f(p, \rho, w), \tag{1}$$

where f is an arbitrary function. The mathematical validity of (1) results from a logical chain of thoughts, however complex it may be, and can be easily verified by consulting a standard microeconomics textbook. However, the scientific validity of (1) depends on answering questions that go beyond the deductive logic used in its derivation. More precisely, the scientific validity of (1) rest on whether or not it is supported by observing and analyzing data on prices, quantity demanded and wealth.

It should be clear that there is nothing special or unique about the consumer behavior example of the previous paragraph. The same distinction between internal mathematical consistency and scientific validity, could have been made using Robert Solow's Macroeconomic Growth Model from a macroeconomic theory textbook, or the Capital Asset Pricing Model that emerges from the theoretical finance literature, to name just a few other theoretical models. Again, the scientific validity of the conjectures that emerge from these models depends on the direct observation and analysis of relevant economic variables they purport to relate.

Traditionally, Econometrics has been defined as the set of concepts, methods and procedures used to summarize and analyze economic data that correspond to the economic variables that appear in economic models. As such, Econometrics is the instrument used to evaluate the scientific validity of economic theory. However, as will become clear throughout this class, Econometrics is much more than that. First, economic data has properties that are shared by data observed in other social sciences,

such as Political Science, Sociology, Psychology and other natural sciences such as Meteorology, Biology, and Astronomy. Thus, many of the tools and procedures associated with Econometrics can be used to analyze data and evaluate theoretical models in other sciences. Second, theoretical economists can use the conclusions that emerge from the use of Econometrics to re-specify and fine tune their theoretical models. Hence, Econometrics can be viewed as a tool for building better theoretical models of economic reality. Third, Econometrics can be used, in conjunction with theoretical models, to predict and quantify changes in economic variables. As such, Econometrics can be used as a tool for evaluating proposed economic policy and to aid in forecasting the value of many economic variables of interest. In summary, Econometrics plays important methodological and practical roles in Economics.

2 Economic data

Economic data have two basic characteristics, they are **stochastic**, and they are largely **observational** or **non-experimental** in nature. The stochastic nature of economic data results from the belief that it is impossible to predict with certainty the value of any economic variable. Whether this results from excessive simplicity of theoretical economic models, or inherent chance that permeates economic relationships is for our purposes irrelevant. What is important is that most economic data can be assumed to be realizations of **stochastic phenomena** which exhibit some degree of **stochastic regularity**. Economic variables are therefore said to be **stochastic variables**. Economic data are observational because they are almost always not generated under controlled (laboratory) conditions by economists. They are collected by direct observation of the unconstrained actions of economic agents. Let us take a closer look at these basic characteristics.

2.1 The Stochastic nature of economic data

The defining characteristic of data that emerges from stochastic phenomena is that they exhibit stochastic regularity. By this, we mean that although it is impossible to predict with certainty what a particular realization of a stochastic phenomenon will be, it is possible to identify patterns and draw conclusions about the phenomenon when we observe a **set** of realizations. Consider, for example, Figure 1 which represents quarterly annualized rate of change of the Gross Domestic Product (GDP) of the United States

measured in 2000 (chained) dollars from 1959 until 2007. [1] What is interesting about these data is that

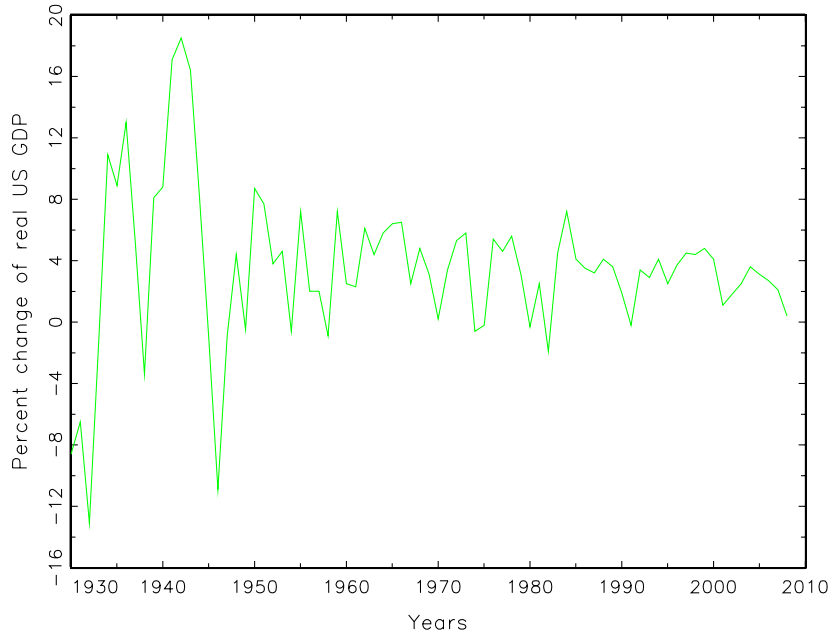

Figure 1: Percent change in US Real GDP from previous period.

although it seems impossible to predict at any particular year what the rate of growth will be, it is possible to discern various regularities for the entire set of observations. First, it seems that the arithmetic average of growth (taken over time) seems to be constant. Second, the variation around the average seems to be declining over time. Third, the histogram for growth rates on Figure 2 seems to exhibit a certain symmetry around the average. We are able to identify these patterns by simple graphical examination of the data. Although we are still unable to predict precisely next quarter's annualized GDP growth, it seems warranted to make statements such as: (a) there is a small chance the GDP growth next year will exceed 8 percent, or (b) there is a small chance of observing three consecutive negative growth rates of GDP (recessions). These are examples of *probabilistic* statements which can be made only because the data seems to exhibit some stochastic regularity. Discovering the stochastic regularities of economic data and making precise probabilistic statements about the stochastic phenomenon under study is therefore an important part of Econometrics.

[1] Source: US Department of Commerce, Bureau of Economic Analysis, http://www.bea.gov/

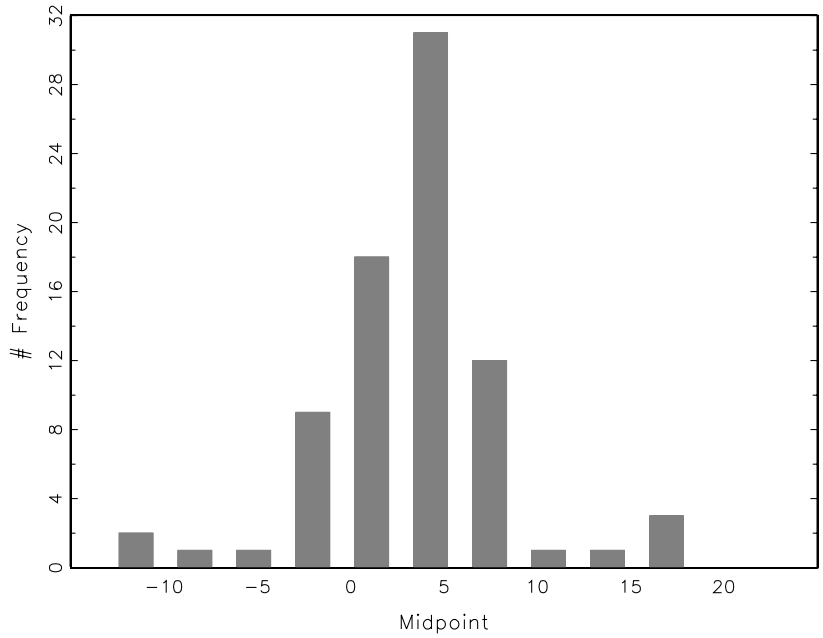

Figure 2: Frequency histogram for GDP change

2.2 The Observational Nature of Economic Data

Regarding its stochastic nature, the GDP growth data discussed above is not very different from the data shown in Figure 3. It shows the results of 100 tosses of a fair coin, with 0 representing the observation of Heads and 1 representing the observation of Tails. As in the case of GDP growth, it is impossible to predict what the will be the next outcome of a toss. However, a very clear stochastic regularity emerges when we observe the histogram in Figure 4. As in the case of the GDP growth rate, a variety of probabilistic statements can be made regarding the outcomes of the coin toss. However, there are fundamental conceptual differences between the stochastic phenomenon that produced the coin data and that which produced the GDP growth data. First, the coin data was generated under the auspices of the observer and under her control, whereas the GDP growth data resulted from a stochastic mechanism completely outside of the observer's control. In fact, the user of economic data on GDP growth is in most cases even different from those that have collected and measured the data. Second, with the help of a computer we can repeat the coin experiment as many time as we want under ostensibly similar circumstances. This is clearly not the case for the GDP growth data. Not only are we limited on the number of observable data points, but we are also unable to guarantee that the observed data emerge

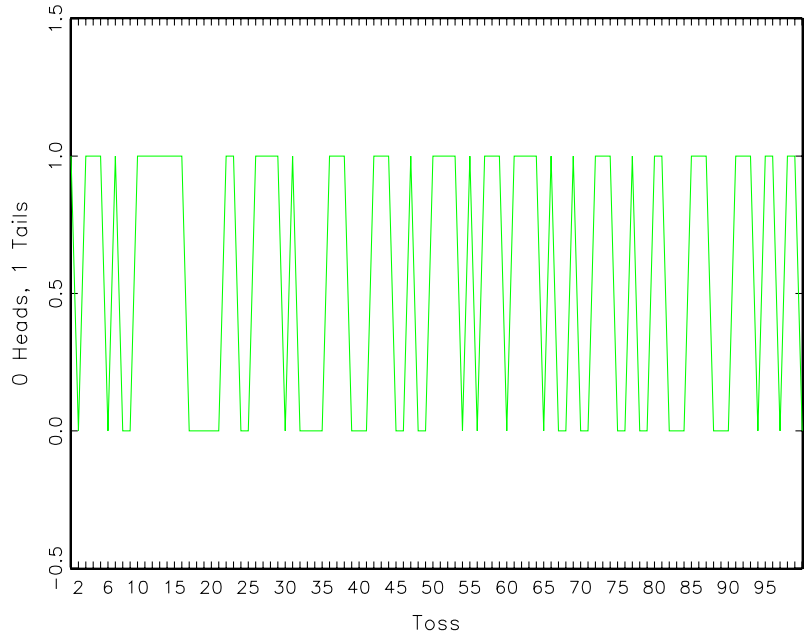

Figure 3: Heads/Tails in 100 tosses

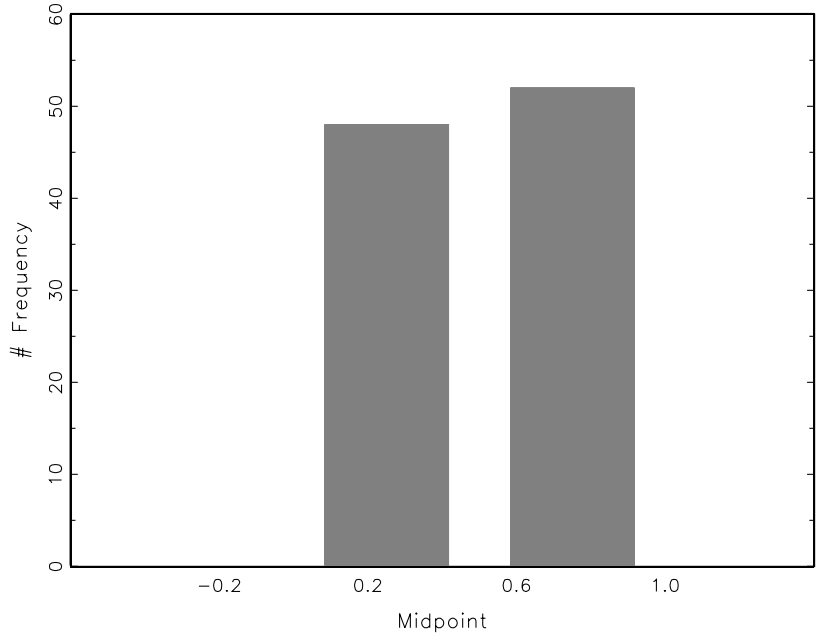

Figure 4: Histogram for 100 tosses of a fair coin

from the same stochastic phenomena. This inability to get *repeated* observations is a distinguishing characteristic of many economic data. The observational nature of the GDP growth data is common to most economic data and it has a profound effect on how we analyze them.

2.3 Representing Data

Since we will be working with economic data, we require a convenient way to represent observations on variables that are of interest. We represent n observations on the economic variable X as a sequence $\{x_k\}_{k=1}^n$. Hence, X may represent quarterly annualized GDP growth and x_k will represent the observed value of this variable at time period k. When k is an index representing time, we say that $\{x_k\}_{k=1}^n$ is a **time series** on X. When k represents something other than time, e.g., a specific economic agent, a region of the country, etc. we say that $\{x_k\}_{k=1}^n$ is a cross-section on X. For example, X may be the year 2000 GDP and x_k will represent the observed value for country k. Whenever $k = (i, t)$ with $i = 1, 2, ...N$ and $t = 1, 2, ...T$ indexing sections and time respectively, we say that $\{x_{it}\}_{i=1,t=1}^{N,T}$ is a panel on X. For example, X may represent GDP and x_{it} will represent the observed GDP for country i in time period t.

Measurement scale and ordering also provide a useful terminology for the description of economic data. The variable X is said to be measured in **ratio** scale if for any two observation x_k and x_s: (a) $\frac{x_k}{x_s}$ and (b) $x_k - x_s$ are meaningful quantities, and (c) there exists an order relation between x_k and x_s, i.e., $x_k \geq x_s$ or $x_k \leq x_s$ or both. Many economic data are measured in the ratio scale. These include costs, prices, quantities demanded and supplied, etc. The variable X is said to be measured in **interval scale** if (b) and (c) are valid, but (a) is not. A typical example is time. X is said to be measured in **ordinal scale** if (c) is valid, but not (a) and (b). Examples of such variables include income level (upper, middle, low), educational level (High School degree, Bachelor's degree, Graduate degree), etc. X is said to be measured in **nominal scale** if neither (a), (b) or (c) are valid. Variables such as marital status, gender and employment status are in this category. Variables measured in ordinal or nominal scales are normally referred to as **categorical** data.

3 Statistical Models

The stochastic and observational nature of economic data makes **statistical models** the primary tool for their analysis. A statistical model is a set of general assumptions regarding the stochastic nature of the

phenomenon which produced the observed data. Ideally, these assumptions are general enough to account for the stochastic regularity patterns exhibited by the data. A precise description of a statistical model requires the definition of concepts that we will study later, such as **probability**, **stochastic variables**, **probability distributions**, and **statistical independence**. However, at this stage, it is important to gain a general appreciation for what constitutes a statistical model. We provide two examples that illustrate some of the basic characteristics of these models.

Example 1: Suppose that the stochastic phenomenon of interest is the toss of a coin. The toss produces a result R which takes the value H for heads or T for tails. We have data on R that results from n consecutive tosses. They are represented by the time series $\{r_t\}_{t=1}^n$. If $n = 3$ such time series could be $\{H, T, T\}$ or $\{T, T, H\}$, etc. A statistical model for R can be described by the following assumptions,

1. $r_t = H$ with probability p and $r_t = T$ with probability $1 - p$ for all $t = 1, ..., n$. p is an unknown **parameter** of the model (in this case it represent a probability) that takes values in the interval $[0, 1]$.

2. r_t and r_s are statistically independent for all $t \neq s$.

Abstracting from the statistical terminology, assumption 1 states that there is a chance p of observing heads and $1 - p$ of observing tails at any particular toss. Assumption 2 suggests that the fact that a particular result from toss t conveys no information about future or past results.

This very simple statistical model contains the basic elements that are present in many (parametric) statistical models. (i) A parameter (p) that takes value in a pre specified set $[0, 1]$ that defines a class of probabilities (all pairs $(p, 1 - p)$ for $p \in [0, 1]$) associated with the occurrence of the variable of interest (R). (ii) an assumption about how the data is obtained. In this case the time series is assumed to come from n independent tosses with the same chance of observing T and H in each toss. Statistical models can be more complicated, but they will in many cases retain these features.

Let us now look at an example closer to economics.

Example 2: One of the most enduring and studied questions in the field of finance is whether or not financial asset prices are predictable. Suppose we observe a time series $\{p_t\}_{t=1}^n$ on the price P of a stock. One of the earliest statistical models proposed to capture the stochastic regularities of such data is that of Bachelier (1900). His model can be generally described by the following assumptions,

1. $\log p_t - \log p_{t-1} = \mu + \epsilon_t$ for all $t = 1, ..., n$ and μ is an unknown parameter taking values in the set

M.

2. ϵ_t are unobserved values of a stochastic variable ϵ which takes values in $(-\infty, +\infty)$ and has probabilities of occurrence given by a function F_θ that depends on an unknown parameter θ that takes values in the set Θ. It is also assumed that realizations ϵ_t and ϵ_s are statistically independent.

Assumption 1 tells us that the difference between the log price of the stock from one period to another is a constant - called the *drift*, plus an increment ϵ_t. Assumption 2 makes a series of assumptions regarding the stochastic nature of the increment. Since, μ is a parameter of the statistical model, Assumption 2 is in effect making various statements about the observed differences $\log p_t - \log p_{t-1}$ for $t = 1, 2, ..., n$. As in the case of the first example, this model depends on parameters μ and θ that take values in pre specified sets. These parameters define a class of probabilities (given by the function F_θ) associated with the occurrence of the variable of interest (P). There is also an assumption about how the data is obtained. It is assumed that price differences between any two adjacent periods ($\log p_t - \log p_{t-1}$) contain no information about future or past price differences and that the probabilities associated with these price differences do not change with time.

The adequacy of a postulated statistical model depends how well it captures the stochastic regularities of the relevant observed data. Assessing the adequacy of statistical models depends on two separate but interrelated procedures: **estimation** and **specification testing**. Estimation is a set of procedures, normally the result of an analysis of the data, that produce values (estimates) for the unknown parameters of the model. With estimated parameters we can make predictions, for example, about the future price of a stock or assess the probability that heads will show on the next toss of a coin. Specification testing is a set of data based procedures that permit the evaluation of some of the underlying assumptions of the model. Estimation and specification testing are interrelated activities. First, specification testing cannot be done without prior estimation. However, testing can reveal characteristics of the stochastic phenomena that produced the data that call for new or modified assumptions on the statistical model. This new set of assumptions in turn normally call for different estimation procedures. Figure 1.5 gives a condensed diagrammatic representation of the interdependency of estimation and specification testing.

Statistical Model + Observed Data $\rightarrow$ *Estimation* $\rightarrow$ *Testing* $\rightarrow$

Respecification- New Statistical Model + Observed Data $\rightarrow$ *Estimation* $\rightarrow$ *Testing.*

Figure 1.5

The combined use of observed data, statistical models, estimation and testing procedures to capture the regularities that characterize a stochastic phenomenon is known as **Empirical Modeling.** There is an underlying characteristic of the discussion above that should be emphasized. The adequacy of a statistical model depends only on how effective it is in capturing the stochastic regularity of the observed data, and should be in no way dependent on the theoretical model it will be used to evaluate. In this sense, empirical modeling is *(economic) theory free* .

Although empirical modeling is very much a *data driven* process, it depends on some important ways on the theoretical models we wish to evaluate. First, in most instances, the variables that appear in the statistical model are suggested by the theoretical model. Second, the mathematical structure that relates variables in the theoretical model is in many cases incorporated into the statistical model by representing chracteristics of the postulated probabilistic structure. Therefore, theoretical models are an indispensable input in the specification of statistical models. However, once these basic specification steps are taken, empirical modeling should proceed free of any theoretical dictates. This separation allows for the revelation of empirical regularities among variables that theory may not had suggested. This can be used by theorists to refine and re-specify theoretical models.

There is a major issues with the evaluation of theoretical models *via*empirical modeling. It is often the case that variables that are collected and used in empirical modeling do not correspond precisely to the variables a theorist envisioned when constructing the theoretical model. In this case, empirical modeling is of little, if any, use in weeding out bad from good theoretical models. To avoid this problem, it is important to: a) know how the data was collected and measured, b) assert, to the highest degree possible, the correspondence between observed variables and theoretical variables.