Wooldridge, Introductory Econometrics, 4th ed.

Chapter 1: Nature of Econometrics and Economic Data

What do we mean by *econometrics*? Econometrics is the field of economics in which statistical methods are developed and applied to estimate economic relationships, test economic theories, and evaluate plans and policies implemented by private industry, government, and supranational organizations. Econometrics encompasses forecasting—not only the high-profile forecasts of macroeconomic and financial variables, but also forecasts of demand for a product, likely effects of a tax package, or the interaction between the demand for health services and welfare reform.

Why is econometrics separate from mathematical statistics? Because most applications of

statistics in economics and finance are related to the use of non-experimental data, or observational data. The fundamental techniques of statistics have been developed for use on experimental data: that gathered from controlled experiments, where the design of the experiment and the reliability of measurements from its outcomes are primary foci. In relying on observational data, economists are more like astronomers, able to collect and analyse increasingly complete measures on the world (or universe) around them, but unable to influence the outcomes.

This distinction is not absolute; some economic policies are in the nature of experiments, and economists have been heavily involved in both their design and implementation. A good example within the last five years is the implementation of welfare reform, limiting individuals' tenure on the welfare rolls to five years

of lifetime experience. Many doubted that this would be successful in addressing the needs of those welfare recipients who have low job skills; but the reforms have been surprisingly successful, as a recent article in *The Economist* states, at raising the employment rate among this cohort. Economists are also able to carefully examine the economic consequences of massive regime changes in an economy, such as the transition from a planned economy to a capitalist system in the former Soviet bloc. But fundamentally applied econometricians observe the data, and use sophisticated techniques to evaluate their meaning.

We speak of this work as empirical analysis, or empirical research. The first step is the careful formulation of the question of interest. This will often involve the application or development of an economic model, which may be as simple as noting that normal goods have negative price elasticities, or exceedingly complex,

involving a full-fledged description of many aspects of a set of interrelated markets and the supply/demand relationships for the products traded (as would, for instance, an econometric analysis of an antitrust issue, such as *U.S. v Microsoft).* Economists are often attacked for their imperialistic tendencies–applying economic logic to consider such diverse topics as criminal behavior, fertility, or environmental issues–but where there is an economic dimension, the application of economic logic and empirical research based on econometric practice may yield valuable insights. Gary Becker, who has made a career of applying economics to non-economic realms, won a Nobel Prize for his efforts.Crime, after all, is yet another career choice, and for high school dropouts who don't see much future in flipping burgers at minimum wage, it is hardly surprising that there are ample applicants for positions in a drug dealer's distribution network. In risk-adjusted terms (gauging the risk of getting shot, or arrested and

successfully prosecuted...) the risk-adjusted hourly wage is many times the minimum wage. Should we be surprised by the outcome?

Irregardless of whether empirical research is based on a formal economic model or economic intuition, the hypotheses about economic behavior must be transformed into an econometric model that can be applied to the data. In an economic model, we can speak of functions such as $Q = Q(P, Y)$; but if we are to estimate the parameters of that relationship, we must have an explicit functional form for the Q function, and determine that it is an appropriate form for the model we have in mind. For instance, if we were trying to predict the efficiency of an automobile in terms of its engine size (displacement, in in^3 or liters), Americans would likely rely on a measure like mpg – miles per gallon. But the engineering relationship is not linear between mpg and displacement; it

is much closer to being a linear function if we relate gallons per mile ($gpm = 1/mpg$) to engine size. The relationship will be curvilinear in mpg terms, requiring a more complex model, but nearly linear in gpm vs displacement. An econometric model will spell out the role of each of its variables: for instance,

$$gpm_i = \beta_0 + \beta_1 displ_i + \epsilon_i$$

would express the relationship between the fuel consumption of the i^{th} automobile to its engine size, or displacement, as a linear function, with an additive error term ϵ_i which encompasses all factors not included in the model. The $parameters$ of the model are the β terms, which must be estimated via statistical methods. Once that estimation has been done, we may test specific $hypotheses$ on their values: for instance, that β_1 is positive (larger engines use more fuel), or that β_1 takes on a certain value. Estimating this relationship for Stata's

$auto.dta$ dataset of 74 automobiles, the predicted relationship is

$$\widehat{gpm}_i = 0.029 + 0.011 displ_i$$

where displacement is measured in hundreds of in^3. This estimated relationship has an "R^2" value of 0.59, indicating that 59% of the variation of gpm around its mean is "explained" by displacement, and a root-mean-squared error of 0.008 (which can be compared to gpm's mean of 0.050, corresponding to about 21 mpg).

The structure of economic data

We must acquaint ourselves with some terminology to describe the several forms in which economic and financial data may appear. A great deal of the work we will do in this course will relate to **cross-sectional** data: a sample of units (individuals, families, firms, industries, countries...) taken at a given point

in time, or in a particular time frame. The sample is often considered to be a random sample of some sort when applied to micro-data such as that gathered from individuals or households. For instance, the official estimates of the U.S. unemployment rate are gathered from a monthly survey of individuals, in which each is asked about their employment status. It is not a count, or census, of those out of work. Of course, some cross sections are not samples, but may represent the population: e.g. data from the 50 states do not represent a random sample of states. A cross-sectional dataset can be conceptualized as a spreadsheet, with variables in the columns and observations in the rows. Each row is uniquely identified by an observation number, but in a cross-sectional datasets the ordering of the observations is immaterial. Different variables may correspond to different time periods; we might have a dataset containing municipalities,

their employment rates, and their population in the 1990 and 2000 censuses.

The other major form of data considered in econometrics is the **time series**: a series of evenly spaced measurements on a variable. A time-series dataset may contain a number of measures, each measured at the same frequency, including measures derived from the originals such as lagged values, differences, and the like. Time series are innately more difficult to handle in an econometric context because their observations almost surely are interdependent across time. Most economic and financial time series exhibit some degree of persistence. Although we may be able to derive some measures which should not, in theory, be explainable from earlier observations (such as tomorrow's stock return in an efficient market), most economic time series are both interrelated and autocorrelated—that is, related to themselves

across time periods. In a spreadsheet context, the variables would be placed in the columns, and the rows labelled with dates or times. The order of the observations in a time-series dataset matters, since it denotes the passage of equal increments of time. We will discuss time-series data and some of the special techniques that have been developed for its analysis in the latter part of the course.

Two combinations of these data schemes are also widely used: **pooled cross-section/time series** (CS/TS) datasets and **panel**, or **longitudinal**, data sets. The former (CS/TS) arise in the context of a repeated survey–such as a presidential popularity poll–where the respondents are randomly chosen. It is advantageous to analyse multiple cross-sections, but not possible to link observations across the cross-sections. Much more useful are panel data sets, in which we have timeseries of observations on the same unit: for instance, $C_{i,t}$

might be the consumption level of the i^{th} household at time t. Many of the datasets we commonly utilize in economic and financial research are of this nature: for instance, a great deal of research in corporate finance is carried out with Standard and Poor's COMPUSTAT, a panel data set containing 20 years of annual financial statements for thousands of major U.S. corporations. There is a wide array of specialized econometric techniques that have been developed to analyse panel data; we will not touch upon them in this course.

Causality and ceteris paribus

The hypotheses tested in applied econometric analysis are often posed to make inferences about the possible causal effects of one or more factors on a response variable: that is, do changes in consumers' incomes "cause" changes in their consumption of beer? At some level,

of course, we can never establish causation–unlike the physical sciences, where the interrelations of molecules may follow well-established physical laws, our observed phenomena represent innately unpredictable human behavior. In economic theory, we generally hold that individuals exhibit rational behavior; but since the econometrician does not observe all of the factors that might influence behavior, we cannot always make sensible inferences about potentially causal factors. Whenever we "operationalize" an econometric model, we implicitly acknowledge that it can only capture a few key details of the behavioral relationship, and is leaving many additional factors (which may or may not be observable) in the "pound of **ceteris paribus**." ceteris paribus–literally, other things equal–always underlies our inferences from empirical research. Our best hope is that we might control for many of the factors, and be able to use our empirical findings

to ascertain whether systematic factors have been omitted. Any econometric model should be subjected to diagnostic testing to determine whether it contains obvious flaws. For instance, the relationship between mpg and $displ$ in the automobile data is strictly dominated by a model containing both $displ$ and $displ^2$, given the curvilinear relation between mpg and $displ$. Thus the original linear model can be viewed as unacceptable in comparison to the polynomial model; this conclusion could be drawn from analysis of the model's residuals, coupled with an understanding of the engineering relationship that posits a nonlinear function between mpg and $displ$.