Chapter 1

Introduction

1.1 What is Econometrics?

The term "econometrics" is believed to have been crafted by Ragnar Frisch (1895-1973) of Norway, one of the three principal founders of the Econometric Society, first editor of the journal *Econometrica*, and co-winner of the first Nobel Memorial Prize in Economic Sciences in 1969. It is therefore fitting that we turn to Frisch's own words in the introduction to the first issue of *Econometrica* to describe the discipline.

> A word of explanation regarding the term econometrics may be in order. Its definition is implied in the statement of the scope of the [Econometric] Society, in Section I of the Constitution, which reads: "The Econometric Society is an international society for the advancement of economic theory in its relation to statistics and mathematics.... Its main object shall be to promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems...."
>
> But there are several aspects of the quantitative approach to economics, and no single one of these aspects, taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a defininitely quantitative character. Nor should econometrics be taken as synonomous with the application of mathematics to economics. Experience has shown that each of these three viewpoints, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the *unification* of all three that is powerful. And it is this unification that constitutes econometrics.
>
> Ragnar Frisch, *Econometrica*, (1933), 1, pp. 1-2.

This definition remains valid today, although some terms have evolved somewhat in their usage. Today, we would say that econometrics is the unified study of economic models, mathematical statistics, and economic data.

Within the field of econometrics there are sub-divisions and specializations. **Econometric theory** concerns the development of tools and methods, and the study of the properties of econometric methods. **Applied econometrics** is a term describing the development of quantitative economic models and the application of econometric methods to these models using economic data.

1.2 The Probability Approach to Econometrics

The unifying methodology of modern econometrics was articulated by Trygve Haavelmo (1911-1999) of Norway, winner of the 1989 Nobel Memorial Prize in Economic Sciences, in his seminal paper "The probability approach in econometrics" (1944). Haavelmo argued that quantitative economic models must necessarily be *probability models* (by which today we would mean *stochastic*). Deterministic models are blatently inconsistent with observed economic quantities, and it is incoherent to apply deterministic models to non-deterministic data. Economic models should be explicitly designed to incorporate randomness; stochastic errors should not be simply added to deterministic models to make them random. Once we acknowledge that an economic model is a probability model, it follows naturally that an appropriate tool way to quantify, estimate, and conduct inferences about the economy is through the powerful theory of mathematical statistics. The appropriate method for a quantitative economic analysis follows from the probabilistic construction of the economic model.

Haavelmo's probability approach was quickly embraced by the economics profession. Today no quantitative work in economics shuns its fundamental vision.

While all economists embrace the probability approach, there has been some evolution in its implementation.

The **structural approach** is the closest to Haavelmo's original idea. A probabilistic economic model is specified, and the quantitative analysis performed under the assumption that the economic model is correctly specified. Researchers often describe this as "taking their model seriously". The structural approach typically leads to likelihood-based analysis, including maximum likelihood and Bayesian estimation.

A criticism of the structural approach is that it is misleading to treat an economic model as correctly specified. Rather, it is more accurate to view a model as a useful abstraction or approximation. In this case, how should we interpret structural econometric analysis? The **quasi-structural approach** to inference views a structural economic model as an approximation rather than the truth. This theory has led to the concepts of the pseudo-true value (the parameter value defined by the estimation problem), the quasi-likelihood function, quasi-MLE, and quasi-likelihood inference.

Closely related is the **semiparametric approach**. A probabilistic economic model is partially specified but some features are left unspecified. This approach typically leads to estimation methods such as least squares and the generalized method of moments. The semiparametric approach dominates contemporary econometrics, and is the main focus of this textbook.

Another branch of quantitative structural economics is the **calibration approach.** Similar to the quasi-structural approach, the calibration approach interprets structural models as approximations and hence inherently false. The difference is that the calibrationist literature rejects mathematical statistics (deeming classical theory as inappropriate for approximate models) and instead selects parameters by matching model and data moments using non-statistical *ad hoc*[1] methods.

Trygve Haavelmo

The founding ideas of the field of econometrics are largely due to the Norweigen econometrician Trygve Haavelmo (1911-1999). His advocacy of probability models revolutionized the field, and his use of formal mathematical reasoning laid the foundation for subsequent generations. He was awarded the Nobel Memorial Prize in Economic Sciences in 1989.

[1] *Ad hoc* means "for this purpose" – a method designed for a specific problem – and not based on a generalizable principle.

1.3 Econometric Terms

In a typical application, an econometrician has a set of repeated measurements on a set of variables. For example, in a labor application the variables could include weekly earnings, educational attainment, age, and other descriptive characteristics. We call this information the **data, dataset**, or **sample**.

We use the term **observations** to refer to distinct repeated measurements on the variables. An individual observation often corresponds to a specific economic unit, such as a person, household, corporation, firm, organization, country, state, city or other geographical region. An individual observation could also be a measurement at a point in time, such as quarterly GDP or a daily interest rate.

Economists typically denote variables by the italicized roman characters Y, X, and/or Z. The convention in econometrics is to use the character Y to denote the variable to be explained, while the characters X and Z are used to denote the conditioning (explaining) variables. Following mathematical practice, random variables and vectors are denoted by upper case roman characters such as Y and X. We make an exception for equation errors which we typically denote by the lower case letters e, u, or v.

Real numbers (elements of the real line $\mathbb{R}$, also called **scalars**) are written using lower case italics such as x. Vectors (elements of $\mathbb{R}^k$) are typically also written using lower case italics such as x, or using lower case bold italics such as $\boldsymbol{x}$. We use bold in matrix algebraic expressions for compatibility with matrix notation.

Matrices are written using upper case bold italics such as $\boldsymbol{X}$. Our notation will not make a distinction between random and non-random matrices. Typically we use $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ to denote random matrices and use $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{W}$ to denote non-random matrices.

We denote the number of observations by the natural number n, and subscript the variables by the index i to denote the individual observation, e.g. Y_i. In some contexts we use indices other than i, such as in time series applications where the index t is common. In panel studies we typically use the double index it to refer to individual i at a time period t.

We typically use Greek letters such as β, θ, and σ^2 to denote unknown parameters (scalar or vectors). Parameter matrices are written using upper case Latin boldface, e.g. $\boldsymbol{A}$. Estimators are typically denoted by putting a hat "^", tilde "~", or bar "-" over the corresponding letter, e.g. $\widehat{\beta}$ and $\widetilde{\beta}$ are estimators of β, and $\widehat{\boldsymbol{A}}$ is an estimator of $\boldsymbol{A}$.

The covariance matrix of an econometric estimator will typically be written using the upper case boldface $\boldsymbol{V}$, often with a subscript to denote the estimator, e.g. $\boldsymbol{V}_{\widehat{\beta}} = \operatorname{var}\left[\widehat{\beta}\right]$ as the covariance matrix for $\widehat{\beta}$. Hopefully without causing confusion, we will use the notation $\boldsymbol{V}_{\beta} = \operatorname{avar}\left[\widehat{\beta}\right]$ to denote the asymptotic covariance matrix of $\sqrt{n}\left(\widehat{\beta}-\beta\right)$ (the variance of the asymptotic distribution). Covariance matrix estimators will be denoted by appending hats or tildes, e.g. $\widehat{\boldsymbol{V}}_{\beta}$ is an estimator of $\boldsymbol{V}_{\beta}$.

1.4 Observational Data

A common econometric question is to quantify the causal impact of one set of variables on another variable. For example, a concern in labor economics is the returns to schooling – the change in earnings induced by increasing a worker's education, holding other variables constant. Another issue of interest is the earnings gap between men and women.

Ideally, we would use **experimental** data to answer these questions. To measure the returns to schooling, an experiment might randomly divide children into groups, mandate different levels of education to the different groups, and then follow the children's wage path after they mature and enter the labor force. The differences between the groups would be direct measurements of the effects of different levels of education. However, experiments such as this would be widely condemned as immoral! Consequently, in economics experimental data sets are typically narrow in scope.

Instead, most economic data is **observational**. To continue the above example, through data collection we can record the level of a person's education and their wage. With such data we can measure the joint distribution of these variables and assess their joint dependence. But from observational data it is difficult to infer **causality** as we are not able to manipulate one variable to see the direct effect on the other. For example, a person's level of education is (at least partially) determined by that person's choices. These factors are likely to be affected by their personal abilities and attitudes towards work. The fact that a person is highly educated suggests a high level of ability, which suggests a high relative wage. This is an alternative explanation for an observed positive correlation between educational levels and wages. High ability individuals do better in school, and therefore choose to attain higher levels of education, and their high ability is the fundamental reason for their high wages. The point is that multiple explanations are consistent with a positive correlation between schooling levels and education. Knowledge of the joint distribution alone may not be able to distinguish between these explanations.

> Most economic data sets are observational, not experimental. This means that all variables must be treated as random and possibly jointly determined.

This discussion means that it is difficult to infer causality from observational data alone. Causal inference requires identification, and this is based on strong assumptions. We will discuss these issues on occasion throughout the text.

1.5 Standard Data Structures

There are five major types of economic data sets: cross-sectional, time series, panel, clustered, and spatial. They are distinguished by the dependence structure across observations.

Cross-sectional data sets have one observation per individual. Surveys and administrative records are a typical source for cross-sectional data. In typical applications, the individuals surveyed are persons, households, firms, or other economic agents. In many contemporary econometric cross-section studies the sample size n is quite large. It is conventional to assume that cross-sectional observations are mutually independent. Most of this text is devoted to the study of cross-section data.

Time series data are indexed by time. Typical examples include macroeconomic aggregates, prices, and interest rates. This type of data is characterized by serial dependence. Most aggregate economic data is only available at a low frequency (annual, quarterly, or monthly) so the sample size is typically much smaller than in cross-section studies. An exception is financial data where data are available at a high frequency (daily, hourly, or by transaction) so sample sizes can be quite large.

Panel data combines elements of cross-section and time series. These data sets consist of a set of individuals (typically persons, households, or corporations) measured repeatedly over time. The common modeling assumption is that the individuals are mutually independent of one another, but a given individual's observations are mutually dependent. In some panel data contexts the number of time series observations T per individual is small while the number of individuals n is large. In other panel data contexts (for example when countries or states are taken as the unit of measurement) the number of individuals n can be small while the number of time series observations T can be moderately large. An important issue in econometric panel data is the treatment of error components.

Clustered samples are increasing popular in applied economics and are related to panel data. In clustered sampling the observations are grouped into "clusters" which are treated as mutually independent yet allowed to be dependent within the cluster. The major difference with panel data is that clustered

sampling typically does not explicitly model error component structures, nor the dependence within clusters, but rather is concerned with inference which is robust to arbitrary forms of within-cluster correlation.

Spatial dependence is another model of interdependence. The observations are treated as mutually dependent according to a spatial measure (for example, geographic proximity). Unlike clustering, spatial models allow all observations to be mutually dependent, and typically rely on explicit modeling of the dependence relationships. Spatial dependence can also be viewed as a generalization of time series dependence.

Data Structures

- Cross-section
- Time-series
- Panel
- Clustered
- Spatial

As we mentioned above, most of this text will be devoted to cross-sectional data under the assumption of mutually independent observations. By mutual independence we mean that the i^{th} observation (Y_i, X_i) is independent of the j^{th} observation $\left(Y_j, X_j\right)$ for $i \neq j$. In this case we say that the data are **independently distributed**. (Sometimes the label "independent" is misconstrued. It is a statement about the relationship between observations i and j, not a statement about the relationship between Y_i and X_i.)

Furthermore, if the data is randomly gathered, it is reasonable to model each observation as a draw from the same probability distribution. In this case we say that the data are **identically distributed**. If the observations are mutually independent and identically distributed, we say that the observations are **independent and identically distributed**, **i.i.d.**, or a **random sample**. For most of this text we will assume that our observations come from a random sample.

Definition 1.1 The variables (Y_i, X_i) are a **sample** from the distribution F if they are identically distributed with distribution F.

Definition 1.2 The variables (Y_i, X_i) are a **random sample** if they are mutually independent and identically distributed (i.i.d.) across $i = 1, ..., n$.

In the random sampling framework, we think of an individual observation (Y_i, X_i) as a realization from a joint probability distribution $F\left(y, x\right)$ which we call the **population**. This "population" is infinitely

large. This abstraction can be a source of confusion as it does not correspond to a physical population in the real world. It is an abstraction because the distribution F is unknown, and the goal of statistical inference is to learn about features of F from the sample. The *assumption* of random sampling provides the mathematical foundation for treating economic statistics with the tools of mathematical statistics.

The random sampling framework was a major intellectual breakthrough of the late 19th century, allowing the application of mathematical statistics to the social sciences. Before this conceptual development, methods from mathematical statistics had not been applied to economic data as the latter was viewed as non-random. The random sampling framework enabled economic samples to be treated as random, a necessary precondition for the application of statistical methods.

1.6 Econometric Software

Economists use a variety of econometric, statistical, and programming software.

Stata is a powerful statistical program with a broad set of pre-programmed econometric and statistical tools. It is quite popular among economists, and is continuously being updated with new methods. It is an excellent package for most econometric analysis, but is limited when you want to use new or less-common econometric methods which have not yet been programed. At many points in this textbook specific Stata estimation methods and commands are described. These commands are valid for Stata version 16.

MATLAB, GAUSS, and OxMetrics are high-level matrix programming languages with a wide variety of built-in statistical functions. Many econometric methods have been programed in these languages and are available on the web. The advantage of these packages is that you are in complete control of your analysis, and it is easier to program new methods than in Stata. Some disadvantages are that you have to do much of the programming yourself, programming complicated procedures takes significant time, and programming errors are hard to prevent and difficult to detect and eliminate. Of these languages, GAUSS used to be quite popular among econometricians, but currently MATLAB is more popular.

An intermediate choice is R. R has the capabilities of the above high-level matrix programming languages, but also has many built-in statistical environments which can replicate much of the functionality of Stata. R is the dominant programming language in the statistics field, so methods developed in that arena are most commonly available in R. Uniquely, R is open-source, user-contributed, and best of all, completely free! A growing group of econometricians are enthusiastic fans of R.

For highly-intensive computational tasks, some economists write their programs in a standard programming language such as Fortran or C. This can lead to major gains in computational speed, at the cost of increased time in programming and debugging.

There are many other packages which are used by econometricians, include Eviews, Gretl, PcGive, Python, Julia, RATS, and SAS.

As the packages described above have distinct advantages many empirical economists use multiple packages. As a student of econometrics you will learn at least one of these packages and probably more than one. My advice is that all students of econometrics should develop a basic level of familiarity with Stata, MATLAB, and R.

1.7 Replication

Scientific research needs to be documented and replicable. For social science research using observational data this requires careful documentation and archiving of the research methods, data manipulations, and coding.

The best practice is as follows. Accompanying each published paper an author should create a complete replication package (set of data files, documentation, and program code files). This package should contain the source (raw) data used for analysis, and code which executes the empirical analysis and other numerical work reported in the paper. In most cases this is a set of programs which may need to be executed sequentially. (For example, there may be an initial program which "cleans" and manipulates the data, and then a second set of programs which estimate the reported models.) The ideal is full documentation and clarity. This package should be posted on the author(s) website, and posted at the journal website when that is an option.

A complicating factor is that many current economic data sets have restricted access and cannot be shared without permission. In these cases the data cannot be posted nor shared. The computed code, however, can and should be posted.

Most journals in economics require authors of published papers to make their datasets generally available. For example:

Econometrica states:

> *Econometrica* has the policy that all empirical, experimental and simulation results must be replicable. Therefore, authors of accepted papers must submit data sets, programs, and information on empirical analysis, experiments and simulations that are needed for replication and some limited sensitivity analysis.

The *American Economic Review* states:

> It is the policy of the American Economic Association to publish papers only if the data and code used in the analysis are clearly and precisely documented and access to the data and code is non-exclusive to the authors. Authors of accepted papers that contain empirical work, simulations, or experimental work must provide, prior to acceptance, information about the data, programs, and other details of the computations sufficient to permit replication, as well as information about access to data and programs.

The *Journal of Political Economy* states:

> It is the policy of the *Journal of Political Economy* to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication.

If you are interested in using the data from a published paper, first check the journal's website, as many journals archive data and replication programs online. Second, check the website(s) of the paper's author(s). Most academic economists maintain webpages, and some make available replication files complete with data and programs. If these investigations fail, email the author(s), politely requesting the data. You may need to be persistent.

As a matter of professional etiquette, all authors absolutely have the obligation to make their data and programs available. Unfortunately, many fail to do so, and typically for poor reasons. The irony of the situation is that it is typically in the best interests of a scholar to make as much of their work (including all data and programs) freely available, as this only increases the likelihood of their work being cited and having an impact.

Keep this in mind as you start your own empirical project. Remember that as part of your end product, you will need (and want) to provide all data and programs to the community of scholars. The greatest form of flattery is to learn that another scholar has read your paper, wants to extend your work, or wants to use your empirical methods. In addition, public openness provides a healthy incentive for transparency and integrity in empirical analysis.

1.8 Data Files for Textbook

On the textbook webpage http://www.ssc.wisc.edu/~bhansen/econometrics/ there are posted a number of files containing data sets which are used in this textbook both for illustration and for end-of-chapter empirical exercises. For most of the data sets there are four files: (1) Description (pdf format); (2) Excel data file; (3) Text data file; (4) Stata data file. The three data files are identical in content: the observations and variables are listed in the same order in each, and all have variable labels.

For example, the text makes frequent reference to a wage data set extracted from the Current Population Survey. This data set is named `cps09mar`, and is represented by the files `cps09mar_description.pdf`, `cps09mar.xlsx`, `cps09mar.txt`, and `cps09mar.dta`.

The data sets currently included are

- `AB1991`
  - Data file from Arellano and Bond (1991)
- `AJR2001`
  - Data file from Acemoglu, Johnson, and Robinson (2001)
- `AK1991`
  - Data file from Angrist and Krueger (1991)
- `AL1999`
  - Data file from Angrist and Lavy (1999)
- `BMN2016`
  - Data file from Bernheim, Meer and Novarro (2016)
- `cps09mar`
  - household survey data extracted from the March 2009 Current Population Survey
- `Card1995`
  - Data file from Card (1995)
- `CHJ2004`
  - Data file from Cox, B. E. Hansen and Jimenez (2004)
- `CK1994`
  - Data file from Card and Krueger (1994)
- `CMR2008`
  - Date file from Card, Mas, and Rothstein (2008)

- `DDK2011`
  - Data file from Duflo, Dupas, and Kremer (2011)
- `DS2004`
  - Data file from DiTella and Schargrodsky (2004)
- `FRED-MD` and `FRED-QD`
  - U.S. monthly and quarterly macroeconomic databases from McCracken and Ng (2015)
- `Invest1993`
  - Data file from Hall and Hall (1993)
- `LM2007`
  - Data file from Ludwig and Miller (2007) and Cattaneo, Titiunik, and Vazquez-Bare (2017)
- `Kilian2009`
  - Data file from Kilian (2009)
- `Koppelman`
  - Data file from Forinash and Koppelman (1993), Koppelman and Wen (2000) and Wen and Koppelman (2001)
- `MRW1992`
  - Data file from Mankiw, Romer, and Weil (1992)
- `Nerlove1963`
  - Data file from Nerlov (1963)
- `PSS2017`
  - Data file from Papageorgiou, Saam, and Schulte (2017)
- `RR2010`
  - Data file from Reinhard and Rogoff (2010)

1.9 Reading the Manuscript

I have endeavored to use a unified notation and nomenclature. The development of the material is cumulative, with later chapters building on the earlier ones. Nevertheless, every attempt has been made to make each chapter self-contained so readers can pick and choose topics according to their interests.

To fully understand econometric methods it is necessary to have a mathematical understanding of its mechanics, and this includes the mathematical proofs of the main results. Consequently, this text is self-contained with nearly all results proved with full mathematical rigor. The mathematical development and proofs aim at brevity and conciseness (sometimes described as mathematical elegance), but also at pedagogy. To understand a mathematical proof it is not sufficient to simply *read* the proof, you need to follow it and re-create it for yourself.

Nevertheless, many readers will not be interested in each mathematical detail, explanation, or proof. This is okay. To use a method it may not be necessary to understand the mathematical details. Accordingly I have placed the more technical mathematical proofs and details in chapter appendices. These appendices and other technical sections are marked with an asterisk (*). These sections can be skipped without any loss in exposition.

Key concepts in matrix algebra and a set of useful inequalities are reviewed in Appendices A & B. It may be useful to read or review Appendix A.1-A.11 before starting Chapter 3, and review Appendix B before Chapter 6. It is not necessary to understand all the material in the appendices. They are intended to be reference material and some of the results are not used in this textbook.

Part I

Regression

Chapter 2

Conditional Expectation and Projection

2.1 Introduction

The most commonly applied econometric tool is least squares estimation, also known as **regression**. Least squares is a tool to estimate the conditional mean of one variable (the **dependent variable**) given another set of variables (the **regressors**, **conditioning variables**, or **covariates**).

In this chapter we abstract from estimation and focus on the probabilistic foundation of the conditional expectation model and its projection approximation. This includes a review of probability theory. For a background in intermediate probability theory see Chapters 1-5 of *Probability and Statistics for Economists*.

2.2 The Distribution of Wages

Suppose that we are interested in wage rates in the United States. Since wage rates vary across workers we cannot describe wage rates by a single number. Instead, we can describe wages using a probability distribution. Formally, we view the wage of an individual worker as a random variable *wage* with the **probability distribution**

$$F(y) = \mathbb{P}\left[wage \leq y\right].$$

When we say that a person's wage is random we mean that we do not know their wage before it is measured, and we treat observed wage rates as realizations from the distribution F. Treating unobserved wages as random variables and observed wages as realizations is a powerful mathematical abstraction which allows us to use the tools of mathematical probability.

A useful thought experiment is to imagine dialing a telephone number selected at random, and then asking the person who responds to tell us their wage rate. (Assume for simplicity that all workers have equal access to telephones and that the person who answers your call will answer honestly.) In this thought experiment, the wage of the person you have called is a single draw from the distribution F of wages in the population. By making many such phone calls we can learn the full distribution.

When a distribution function F is differentiable we define the **probability density function**

$$f(y) = \frac{d}{dy}F(y).$$

The density contains the same information as the distribution function, but the density is typically easier to visually interpret.