Econometric Foundations

Ron C. Mittelhammer
Washington State University

George G. Judge
University of California, Berkeley

Douglas J. Miller
Purdue University

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK http://www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011-4211, USA http://www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain

First published 2000

Printed in the United States of America

Typeface Times Roman 10.5/13 pt. *System* LaTeX 2_ε [TB]

A catalog record for this book is available from the British Library.

Library of Congress Cataloging in Publication Data

Mittelhammer, Ron.
Econometric foundations / Ron C. Mittelhammer, George G. Judge,
Douglas J. Miller.
p. cm.
ISBN 0-521-62394-4 hb
1. Econometrics. I. Judge, George G. II. Miller, Douglas.
III. Title.
HB139.M575 2000
330′.01′5195 – dc21 99-40040
CIP

ISBN 0 521 62394 4 hardback

Contents

CD-ROM

CHAPTER 1

The Process of Econometric Information Recovery

MOST econometric problems begin with several fundamental questions. One basic question is, How does one develop a plausible basis for reasoning in situations involving partial–incomplete information? Another basic question relates to how one goes about learning from a sample of data.

For the theoretical econometrician, questions tend to be of a nonempirical and hypothetical what-if type: What if a sample of data is described by a particular imagined sampling process? This leads to the question of how one characterizes the sampling process in terms of a probability model that properly identifies the stochastic characteristics of the sampling process as well as the data-restricting constraints, the knowns and unknowns of the problem, and the observable and unobservable components in the model. Given that the data-sampling process can be described by a probability model that expresses the state of knowledge about possible real-world outcomes, another question then arises relating to how one devises effective estimation and hypothesis-testing procedures that will allow the recovery of estimates of the unknowns and provide a basis for making inferences. The theoretical econometrician may, by a process of interpretation, ultimately associate the conceptualized sampling process with a set of *observable economic data*. At this point in the theoretical econometrician's investigation the probability model is interpretable as an econometric model having economic meaning with both real-world economic and statistical implications.

For the applied econometrician, the econometric problem begins with a real-world economic question, perhaps involving the implications of scarcity and choice or perhaps the allocative or distributive impacts, resulting from an action or decision. The next step involves restating the real-world question within a theoretical–conceptual economic model framework in which real-world components are identified to facilitate drawing logic-based conclusions about the question. This step exposes structure and defines the explicit economic model to be used in the empirical analysis of the economic question.

1.1. Introduction

Given a theoretical–conceptual economic playing field, a basis is needed for connecting the real-world data outcomes with their counterparts in the economic model. By visualizing some imagined sampling process by which the outcomes may have evolved and then characterizing this sampling process by a probability model, an econometric model is born. This model then acts as a vehicle for expressing knowledge about real-world outcomes and identifies knowns, unknowns, and observed and unobservable model components. If the applied econometrician is fortunate, the resulting econometric model may be consistent with a probability model that already exists in the literature and for which a well-defined basis for estimation and inference is already available. In this case the applied econometrician will use established statistical procedures to address research questions. On the other hand, the econometric model may not be consistent with a commonly specified and evaluated data-sampling process. Consequently, the applied econometrician must assume the role of the theoretical econometrician in first developing effective estimation and hypothesis-testing procedures and then carrying through the estimation and inference stages needed to answer research questions.

As one reads through this chapter and the chapters ahead, it will at times be necessary to assume the roles of both a theoretical and an applied econometrician to derive maximum benefit from the econometric venture. One goal of the exercises in each chapter is to lead and inspire the reader in this direction. Before going on to consider the question of how to specify a probability–econometric model to provide a basis for learning from a sample of observations, we focus some attention on the real-world component referred to as economic data.

1.2. The Nature of Economic Data

Why do we have books on econometrics? Why not just have books devoted to statistics for economists? What is it that makes economics unique relative to other fields of science?

One thing that tends to make economics and econometrics unique is the nature of economic data and the special characteristics of the sampling processes by which economic data are obtained. In providing an answer to the opening questions of this section, William Barnett, in private correspondence, points us to a classic article by Schumpter in *Econometrica*, Vol. 1, No. 1, 1933, "The Common Sense of Econometrics." In this article Schumpter wrote, "Econometrics is the most quantitative . . . of all sciences, physics not excluded. . . . Every economist is an econometrician whether he wants to be or not." His rationale is that the economy produces and inherently depends upon numbers. Indeed, the very act of transacting in markets depends explicitly upon the numerical values of such variables as quantities and prices. But in physics, for example, the physical world can and will operate without dependence upon numerical measurement of variables. Scientists have to construct devices to measure temperature, pressure, speed, weight, and the like because nature does not "quote" these numbers in markets or even identify a need to know these numbers. So in this sense, as Schumpter observed, economics is inherently more quantitative than any other scientific field.

Economies and markets carry out experiments and produce numerical data through the very nature of their operation. However, they do so in a manner that is not usually in accordance with statistically designed experiments. Although physical scientists can be viewed, as they are in Schumpter's article, as being disadvantaged by the need to measure data, in comparison with economists, who need only record the numerical data that the economy produces, physical scientists have the advantage of being able to run controlled experiments and to generate their data in a manner consistent with established and understood experimental designs. Because the economy is not a statistically designed experiment, economists must in many cases utilize ill-conditioned data. This is a principal reason why econometrics requires special tools for probability model formulation, estimation, and inference and why econometrics is characterized as an experiment in nonexperimental model building. The uncertain nature of economic outcomes goes a long way to explaining why almost everyone, at one time or another, has felt comfortable assuming the role of an economist on certain issues (How often have you heard the phrase "I am not an economist, but. . . ?") and why all economists are in one way or another econometricians.

1.3. The Probability Approach to Economics

The probability theory that you encountered in your courses in theoretical statistics (and that is reviewed in an electronic document on the CD that accompanies this text) has important implications for how one should organize, incorporate, and utilize data and prior information in quantitative economic analyses. In economic problems characterized by incomplete knowledge and uncertainty, this theory, through a process of abstraction and interpretation by analysts, defines a reasoning process for expressing our knowledge about real-world outcomes, for recovering information from data, and for assessing its validity. The calculus of reasoning defined by probability theory facilitates learning and problem resolution and defines a logical basis for evaluating decisions and making choices.

Like Mozart's *Don Giovanni*, conceptual tools such as random variables and stochastic processes, as well as models of economic systems, have been invented or created by theoreticians and empirical analysts rather than discovered. The participants or players in a postulated economic system are presumed to define economic processes that result in measurable outcomes and, by a process of interpretation on the part of the econometrician, these outcomes are viewed in probabilistic form. In most econometric problems, at least a portion of the information available for analysis will be in the form of a sample of data that has been generated as an outcome of some real-world economic process. In addition, the analyst generally has some prior knowledge about the relevant economic processes and institutions that may have conditioned the sample outcomes. If one views the outcomes of the economic system as having come from an imagined sampling process, concepts such as random variables and probability distributions can be used as conceptual tools to characterize the full complement of existing knowledge about these economic observables. In particular, a probability model associated with the imagined sampling process can be defined that serves as a vehicle for describing our state of knowledge relating to how the observable economic data were obtained. Given this information repository, the fundamental econometric problem is concerned with transforming the conceptual probability model and the sample information associated with it into more specific knowledge about the unknown model components and parameters that represent characteristics of real-world economic processes.

1.4. The Process of Searching for Quantitative Economic Knowledge

In searching for quantitative economic knowledge contained in a sample of data, one must begin with some understanding of the economic process to which the data relate as well as some conception of the underlying sampling process by which the data were obtained. Otherwise, a sample of data is merely a collection of numbers with no contextual meaning or information value. Thus, to have a basis for interpreting the observed data, one needs a conceptual model of the process to which the data refer or some basis for specifying a *data-sampling process* (DSP) that links the sample of observations to our state of knowledge about how these observations were obtained.

The first step in this search for economic knowledge is for the analyst to identify an *economic process* that the analyst seeks to understand and about which there is incomplete knowledge and uncertainty (henceforth refer to Figure 1.1). By *process* we mean a particular method of doing something that generally involves several steps, operations, and interacting components and that leads to an observable outcome or result. For example, this might be the method by which, given prices and a budget constraint, consumers decide on market purchases (consumer decision process) or the method by which a commodity market leads to a product price (a market price equilibrium process).

The next step is a *process of abstraction* whereby *predata* theories, facts, assumptions, and an imagined sampling process whose outcomes are related to random variables are logically molded together into a *probability–econometric model* of the economic process. In a formal and often idealized way, the econometric model summarizes the analyst's state of knowledge about the mechanisms that are thought to underlie the workings of the economic process under study and the sampling process by which observed data are obtained. The model, which is an abstraction, may be expressed in a variety of ways such as mathematical equations, algorithms, behavioral rules, diagrams, or all of these.

1.4.1. Econometric Model Components

It may be useful to think of an econometric model as being composed of components that include an economic model, a sampling model, and a probability model (Figure 1.1). The *economic model* component distinguishes an *econometric* model from a biological, physical, psychometric, or sociometric model. Models in other disciplines are defined

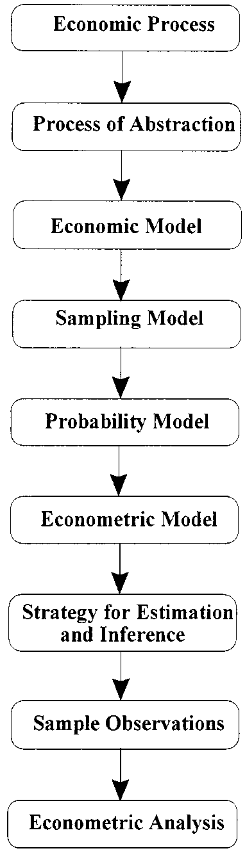

Figure 1.1: Process of Economic Information Recovery.

when appropriate discipline-specific theories, concepts, and knowledge are substituted in place of the economic model component. The economic model is based on a combination of the analyst's understanding of the institutions and mechanisms operating within the economic process being modeled and the economic theory thought to be relevant for explaining data outcomes produced by the economic process.

Once the economic model has been postulated, interpretations and questions relating to the workings of the economic process may be deduced from it. In this way the economic model provides a basis for defining relevant economic variables, forming tentative explanations, and suggesting hypotheses. However, this process of deduction tells us nothing per se about the actual truth or falsity of any explanations, hypotheses, or conclusions. It only ensures that conclusions *deductively* generated from the economic model are internally consistent with the definitions and postulates on which the model is based *provided that* the rules of logic have been applied correctly.

The *sampling model* characterizes a sampling process linkage between observable economic data and the postulated real-world components of the economic model. In particular, the sampling model identifies an imagined sampling process postulating that the observed data are the outcome of a collection of random variables

$\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$. At this point, general assumptions regarding the sampling characteristics of the random variables enter the analysis. For example are the random variables independent and identically distributed (iid), independent or dependent.

Moving farther in the direction of a formal basis for stating specific stochastic characteristics of the random variables in the imagined sampling process, the *probability model* postulates that the economic data are the outcome of some random variable or vector $\mathbf{Y}$ having a joint probability distribution that belongs to some set of potential probability distributions, such as $\{F(\mathbf{y}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Omega\}$. If the elements in the set of probability distributions cannot be identified or indexed by a finite vector of parameter values $\boldsymbol{\theta}$, we may more generally denote the collection of probability distributions as $F(\mathbf{y}) \in \Psi$. By postulating such a probability model, the analyst effectively defines the range of possibilities for the joint probability distribution thought to characterize the behavior of potential sample outcomes. Unknown, uncontrolled, and unobservable components of the probability model are represented by parameters, random variables, or both. Together, the probability and sampling models identify both the candidates for the joint probability distribution of the observed data and the degree of interdependence, or lack thereof, among the individual data observations.

The combination of the probability, sampling, and economic models results in an *econometric model* that links a specified sampling process to the data. The adjective *econometric* arises from the realization, identification, and incorporation of an economic component into the formation and interpretation of the model. The econometric model represents our knowledge of the sampling of economic data in terms of a collection of random variables, $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$, that have a certain economic interpretation, a certain dependence structure, and a joint probability distribution that belongs to some set of probability distributions $\{F(\mathbf{y}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Omega\}$ or $F(\mathbf{y}) \in \Psi$. Having defined the econometric model, the analyst has effectively specified a complete model of the sampling of the economic data under investigation. This means that, if values for the unknown and unobservable components of the perceived econometric model were known or assumed, the analyst would expect, or hope, that data consistent with the economic process being analyzed could be simulated from the econometric model. To the extent that the econometric model represents an accurate depiction of the true data-sampling process, the simulated outcomes could be used to produce additional samples of economic data relating to the economic process under study.

1.4.2. Econometric Analysis

Given a fully specified econometric model, the analyst has created a complete probabilistic and economic description and interpretation of the imagined sampling process for the economic data being analyzed. The model thus provides a complete picture of the analyst's state of knowledge about a set of economic outcomes and identifies *what is assumed* and *what is left to be discovered in the research process.*

An analyst's econometric model of sample outcomes for the random variables $\mathbf{Y}$ is usually specified in terms of a *systematic* or *signal component* and an unobservable random *error*, *disturbance*, or *noise component* ε. The two components are assumed to combine in a way that determines the *exact* values of observed sample outcomes.

In particular, the extent to which the value of $\mathbf{Y}$ cannot be functionally represented in terms of the systematic component is accounted for by some function of ε, the content of which in some sense reflects the development of economics as a science. An example is the additive formulation $\mathbf{Y} = \mathbf{g}(\mathbf{X}, \boldsymbol{\theta}) + \varepsilon$ in which the sum of the systematic and noise components represents the sample outcome, where $\boldsymbol{\theta}$ is a vector of unknown and unobservable parameters and $\mathbf{X}$ is a set of conditioning variables. More specific examples of the characterization of $\mathbf{Y}$ in terms of systematic and noise components will be examined in Chapter 2.

Once the econometric model has been specified, the applied econometrician's objective is to proceed to the econometric analysis of the model. A necessary ingredient for such an analysis is the collection of *sample observations* of economic data relating to the economic process under study. The analyst must then also devise a *strategy for estimation and inference* in which appropriate statistical procedures for information discovery and recovery are identified within the context of the model being analyzed.

Given the sample observations and identified statistical procedures, the analyst then conducts an *econometric analysis* by applying the statistical procedures to the sample data and generating estimates and inferences. The analyst then provides a statistical and economic interpretation of the results obtained to complete the econometric analysis of the economic process.

1.5. The Inverse Problem

A challenge in econometric analyses is that unknown and uncontrolled components of the econometric model cannot generally be observed directly, and thus the analyst must use indirect observations based on observable data to recover information on these components. This challenge is associated with a concept in systems and information theory called the *inverse problem*, which is the problem of recovering information about unknown and uncontrolled components of a model from indirect observations on these components. The adjective *indirect* refers to the fact that, although the observed data are considered to be directly influenced by the values of model components, the observations are not themselves the direct values of these components but only indirectly reflect the influence of the components. Thus, the relationship characterizing the effect of unobservable components on the observed data must be somehow inverted to recover information about the unobservable model components from the data observations. Because econometric relations generally contain a systematic and a noise component, the problem of recovering information about unknowns and unobservables $(\boldsymbol{\theta}, \varepsilon)$ from sample observations $(\mathbf{y}, \mathbf{x})$ within the context of an econometric model $\mathbf{Y} = \eta(\mathbf{X}, \varepsilon, \boldsymbol{\theta})$ is referred to as an *inverse problem with noise*. A solution to this inverse problem is of the general form $(\mathbf{y}, \mathbf{x}) \Rightarrow (\boldsymbol{\theta}, \varepsilon)$.

Motivation for viewing a problem in econometric analysis as an inverse problem can be provided by a familiar illustration involving the theory of the firm. Firm managers need to make decisions concerning the profit-maximizing mix of inputs and levels of outputs under fixed prices. To make these optimizing decisions, information is needed on unknown components of the real-world production process such as the marginal products of inputs. The marginal products are not directly observable, but we can observe the levels of outputs that result when various levels of inputs are used. These observable outcomes of the production process are *indirect* observations on the marginal products that, although not equal to the values of the marginal products themselves, are influenced by their values. Thus, we are confronted with an inverse problem: How can we best use the observed levels of input and output to recover information about the unobservable marginal products? At this point, in the absence of effective methods of information recovery, it must be clear that few, if any, rational or informed bets could be devised relative to the values of unknown, unobserved, and unobservable components of our econometric models. The principal objective of this book is to provide a foundation for the development of effective information recovery methods for this and other inverse problems.

1.6. A Comment

We view an economic–probability–econometric model as a starting point that lets us state, for all to see, what we are maintaining, or willing to assume, is known and what we consider unknown and seek to discover relative to an economic process under investigation. One of our econometric friends once remarked that he would rather be asked by a curious 3-year-old where babies come from than to try to answer the question, Where do econometric models come from? Perhaps it is less important where they come from than what the models represent, which is a starting place – a postulational base that leads to questions, experimentation, data collection, estimation, and finally inference and conclusions. In other words, they are the basis for a research process in which the model, the data, and the method of information recovery are interdependent links in the knowledge search and recovery chain. In Chapter 2, we review some interesting econometric questions and begin to examine the process of progressing from an economic model to a probability–econometric model. Our intention is to start the reader thinking about the estimation and inference methods for solving inverse problems with noise.

1.7. Notation

Before moving to Chapter 2 to begin our conceptualization of some alternative econometric models, we review here some notational conventions that give meaning to Figure 1.1 and the formulations we use in this book. A scalar random variable is denoted by a capital letter such as X or Y. A multivariate random variable in the form of a vector or matrix is denoted by a bold capital letter such as **X** or **Y**. A subscripted index distinguishes between different random variables. For example, we will use Y_i to indicate one representative of a collection of random variables, $(Y_1, Y_2, \ldots, Y_n)$. Random variables whose outcomes we seek to explain will be referred to as *dependent variables*. We will also be interested in *explanatory variables,* whose values are used to help explain the values of dependent variables.

It is most often the case that an econometric model will contain more than one explanatory variable. An index is then needed if we wish to distinguish the explanatory variables from one another for a given observation. Depending on the circumstances, explanatory variables may either be fixed or random. In either case we will use a double subscript, where the first subscripted index denotes the observation number and the second specifies the particular explanatory variable number. We normally use **X** or **x** for a matrix of random or fixed explanatory variables, respectively. We emphasize that in either the random or fixed case, boldface denotes a vector or a matrix, whereas a nonboldfaced symbol will denote a scalar. Thus, if the explanatory variables are random, then X_{ij} represents the ith potential observation on the jth random explanatory variable. If the explanatory variables are fixed, then x_{ij} is the ith value of the jth fixed explanatory variable. An alternative notation for indicating the (i, j)th element of the explanatory variable matrix will be the standard matrix element notation $\mathrm{X}[i, j]$ or $\mathrm{x}[i, j]$. We will also represent the ith row of the explanatory variable matrix by $\mathbf{X}_{i.}$ or $\mathbf{x}_{i.}$, or in standard matrix notation, by $\mathbf{X}[i, .]$ or $\mathbf{x}[i, .]$. The corresponding notation for designating the jth column of the explanatory variable matrix will be $\mathbf{X}_{.j}$ or $\mathbf{x}_{.j}$, and in standard matrix notation $\mathbf{X}[., j]$ or $\mathbf{x}[., j]$.

As an example of the preceding notation, we may develop a model of individual incomes Y_i using observations on explanatory variables representing scores on intelligence tests, years of schooling, highest degree obtained, grade point average (GPA), gender, race, and geographical region. In general functional notation we may then specify that $\mathrm{Y}_i = \mathrm{g}(\mathbf{X}_{i.})$ or $\mathrm{Y}_i = \mathrm{g}(\mathbf{X}[i, .])$, for $i = 1, \ldots, n$.

In representing general functions of variables, we will on occasion need to distinguish between *scalar* and *vector* functions of variables. The general notation will be $\mathrm{g}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$, respectively, where again, boldface denotes a vector. Furthermore, it is often the case in representing the systematic part of econometric models that the same functional definition is applied to each data observation. In this case, we will use the notation $\mathrm{g}(\mathbf{x}_{i.})$ to denote the function applied to the observation $\mathbf{x}_{i.}$ and then continue to use $\mathbf{g}(\mathbf{x})$ to denote the vector of all of the observations, that is, $\mathbf{g}(\mathbf{x}) = [\mathrm{g}(\mathbf{x}_{1.}), \mathrm{g}(\mathbf{x}_{2.}), \ldots, \mathrm{g}(\mathbf{x}_{n.})]'$.

In some circumstances, we will want to characterize the values of dependent variables over time. In cases where we want to emphasize the temporal nature of the observations we will use a t subscript to denote the time index. For example, one may be interested in the values of a dependent and associated explanatory variables at n distinct time periods. The dependent and explanatory variables at time t will be denoted by Y_t and $\mathbf{X}_{t.}$ or $\mathbf{X}[t, .]$, respectively. A data set of n observations over time would then consist of y_t and $\mathbf{x}_{t.}$ or $\mathbf{x}[t, .]$ for $t = 1, \ldots, n$, where y_t is the observed value of the random variable Y_t, and $\mathbf{x}_{t.}$ or $\mathbf{x}[t, \cdot]$ is the observed value of $\mathbf{X}_{t.}$ at time t.

Note that there will be a few exceptions to the conventions introduced above when precedent in the literature is so strong as to warrant an exception. An exception already encountered in the text is the use of ε to denote the *random* variable representing the noise component of an econometric model. Because we will later use **e** to denote an outcome of the noise component, as is very often done in the literature, we avoid confusion with the letter **E**, which is the conventional notation for mathematical expectation, and instead choose ε to be the random variable whose outcome is **e**. We will be careful to identify notational exceptions when they are first introduced in the text, and, regarding exceptions to the capital letter–random variable convention, we will endeavor to use Greek-letter alternatives.

Now that we have a context for discussing econometric models and the notation to represent them formally, in Chapter 2 we identify and classify a range of econometric models that will be of major interest as we work our way through the chapters to come.

1.8. Idea Checklist – Knowledge Guides

1. Assume you are a theoretical econometrician. Identify a general format that you might use in developing a research project or reporting a working paper or journal article.

2. Assume you are an applied econometrician. Identify a general format that you might use in developing a research project or reporting a working paper or journal article.

3. To use later as a basis of comparison for how much your understanding of econometric analysis has matured, write a short essay on the topic: Where do econometric models come from?

4. To use later as a basis of comparison for how much your understanding of econometric analysis has matured, write a short essay on the topic: Is econometrics necessary?

5. Test your ability to specify a simple linear statistical model that involves a set of data from which you want to recover information on a mean-location level and a variance-scale parameter.