SECOND EDITION

Elements of Econometrics

Jan Kmenta

Jan Kmenta

PROFESSOR OF ECONOMICS AND STATISTICS
UNIVERSITY OF MICHIGAN

Elements of Econometrics

SECOND EDITION

Macmillan Publishing Company
NEW YORK
Collier Macmillan Publishers
LONDON

Printed in the United States of America

MACMILLAN PUBLISHING COMPANY
866 Third Avenue, New York, New York 10022

Collier Macmillan Canada, Inc.

Library of Congress Cataloging in Publication Data

Kmenta, Jan.
Elements of econometrics.

Includes index.
1. Econometrics. 2. Statistics. I. Title.
HB139.K56 1986 330'.028 86-2740
ISBN 0-02-365070-2

Printing: 3 4 5 6 7 8 Year: 7 8 9 0 1 2 3 4 5

ISBN 0-02-365070-2

Preface

Since the publication of the first edition of this book, the subject of econometrics has expanded considerably and has undergone some marked changes in emphasis. The expansion has involved development of new results in established areas as well as extension of the scope of the subject to new areas. Changes in emphasis have been stimulated by the appearance of new data sources, by the rapid development of computer technology, and by shifts of interest and methodological approaches. This edition represents an effort to incorporate all important new results in the established areas of econometrics, to present most of the new models and methods introduced since 1971, and to reflect the changes in emphasis that have taken place. In addition, rewriting the book has offered a welcome opportunity to simplify some expositions and, I hope, to make them clearer and more succinct.

The motto of my undergraduate alma mater, the University of Sydney in Australia, *sidere mens eadem mutato* ("the same spirit under a different sky"), has also applied in producing this edition. The basic philosophy of making everything as simple and clear as possible, which underlay the first edition, has been maintained. All methods are still explained and discussed within the simplest framework, and generalizations are presented as logical extensions of the simple cases. And while every attempt has been made to preserve a relatively high degree of rigor, every conflict between rigor and clarity of exposition has been resolved in favor of the latter. Finally, in every case the emphasis is on understanding rather than a cookbook type of learning. The consequence of all this is that the organization and style of the second edition are the same as those of the first edition, although much has been added.

The second edition differs from the first in two ways. First, simplifications of exposition and introduction of new topics have been incorporated in the existing sections of the first edition. The largest changes with respect to content concern heteroskedasticity and autocorrelated disturbances in Chapter 8, errors of measurement and grouped data in Chapter 9, multicollinearity and specification errors in Chapter 10, restricted coefficients, nonlinear models, and distributed lags in Chapter 11, pooling of cross-section and time-series data in Chapter 12, and special topics in Chapter 13. Second, new material has been introduced — or old material considerably expanded — by the inclusion of new sections. There is a new section on Bayes theorem in Chapter 3 and on Bayesian inference in Chapter 6 to acquaint the students with the basic ideas of the Bayesian approach to inference. Chapter 8 has a new section on nonnormality and nonzero means of the disturbance, which includes robust estimation. Major additions appear in Chapter 11, which has new sections on models involving qualitative or limited dependent variables, varying coefficients, unobservable variables, disequilibrium models, and model choice. In the Appendix, the section on computational design for least squares estimation has been replaced by a section on asymptotic distributions in regression models with stochastic explanatory variables written by E. P. Howrey and S. H. Hymans.

The incorporation of new results and the inclusion of new sections in the second edition cover most important innovations in econometrics since 1971. There is, however, no section on time-series analysis, even though time-series models do appear in the econometric literature, because these models have no economic content and their use for modeling exogenous variables is not theoretically justified. As pointed out in the text, time-series analysis may be useful for modeling the behavior of the disturbances and thus enriching the dynamic specification of econometric models, but this is a matter for specialists. The fact that time-series models may produce better short-run forecasts than econometric models simply points to the need for improving econometric models rather than replacing them by the ad hoc models of time-series analysis. The only other major omission involves topics usually put under the heading of "longitudinal analysis of labor market data," including duration analysis, intervention analysis, analysis of survey samples of life histories, and others. These topics represent a new and exciting branch of econometrics, but are too specialized to be included in a general textbook. The most important changes in emphasis in econometrics involve increased concern with hypothesis testing as compared to estimation, a burgeoning interest in microeconometrics and less interest in macromodels, the development of specification error tests and tests for model choice to replace more informal procedures, and a shift toward rational expectation models and away from other distributed lag schemes. All of these are — to varying degrees — taken into account in the discussion of different topics.

The changes incorporated in the second edition have not affected the level of difficulty at which the discussion is set, except perhaps for allowing for a natural growth of sophistication on the part of students of economics. The book is still intended for economists rather than for econometric specialists, and it is expected to be used as a text in first-year graduate or advanced undergraduate economics programs or as a reference book for research workers in economics, business, and other social sciences. The prerequisites are the same as for the first edition, namely, a knowledge of basic economic theory, college algebra, basic calculus, and some descriptive statistics. Chapters 10 – 13 also use matrix algebra, which can be learned from Appendix B. However, the prospective reader should be warned that the book represents a serious and reasonably thorough approach to the subject and is not suitable for audiences desiring only superficial acquaintance with econometric methods.

The first edition contains acknowledgments and expressions of gratitude to my former teachers, colleagues, and students who influenced my thinking and helped me in various ways. Unfortunately, a slip in communication resulted in the omission of my undergraduate professor of statistics, R. S. G. Rutherford of the University of Sydney, who first introduced me to econometrics and who provided me with guidance and support well beyond the call of duty.

Preparation of this revision has greatly benefitted from comments, corrections, and suggestions of colleagues and students too numerous to mention. Those whose input has been particularly extensive and who deserve special credit are D. Asher, E. Berndt, A. Buse, A. Havenner, A. Maeshiro, H.-J. Mittag, A. and M. Nakamura, B. Rafailzadeh, E. Rost, H. Roth, E. Sowey, and V. M. Rao Tummala. Further, I owe a great deal of gratitude to Gerry Musgrave, who helped me with advice and with many calculations presented in the book, to Terry Seaks and Jeff Pliskin, who have gone thoroughly through the entire book and caught many errors and made many useful suggestions, and to Bijan Rafailzadeh, who has carefully read and corrected most of the chapters and worked out the answers to the exercises. (A separate Solutions Manual is available on request from the publisher.) The book has certainly been improved because of their selfless contributions. The suggestions and criticisms concerning the first six chapters offered by reviewers engaged by the publisher—Richard E. Bennett, John F. Chizmar, J. Malcolm Dowlin, Nicholas M. Kiefer, Craig Swan, and David J. Weinschrott—have also been very helpful. The last two chapters were written at the University of Saarland in Germany, whose support is gratefully acknowledged. My thanks go also to Mrs. Morag Nairn for her expert help with managing the manuscript and typing, and to Mrs. Elisabeth Belfer of Macmillan for her assistance in the final stages of the production of the book. The largest thanks, of course, belong to my wife, who has been invaluable in helping to produce the manuscript and who has patiently put up with all the inconveniences that she was subjected to during the lengthy process. Finally, I am indebted to the Literary Executor of the late Sir Ronald A. Fisher, F.R.S., and to Oliver & Boyd Ltd, Edinburgh, for their permission to reprint Table D-3 from their book *Statistical Methods for Research Workers.*

J. K.

Contents

PART ONE Basic Statistical Theory

PART ONE
Basic Statistical Theory

1 Introduction to Statistical Inference

Until the early part of the nineteenth century, statistics was understood to be concerned with characteristics of the state, particularly those related to political and military institutions. Descriptions of these characteristics were at first mainly in verbal terms, but they gradually became more and more numerical. The change to an increasingly more quantitative character was accompanied by an extension to fields other than those concerned with the affairs of the state. Later advances in the theory of probability led to the development of the theory of statistics that permits scientific generalization from incomplete information — in other words, statistical inference.

As a result of this historical development, the subject known as "statistics" consists of two parts: descriptive statistics and statistical inference. Descriptive statistics deals with the collection, organization, and presentation of data, while statistical inference deals with generalizations from a part to the whole. Statistical inference, like any other science, is concerned with the development of methods (statistical theory) as well as with their use (statistical application).

In econometrics we are mainly concerned with statistical inference. Descriptive statistics is relevant only to the extent that measures developed by descriptive statisticians for the purpose of summarizing various characteristics of the data — averages, measures of dispersion, etc. — are also used in statistical inference. But while in the field of descriptive statistics these measures represent ends in themselves, in statistical inference they are only means in the process of inquiry.

1-1 Basic Concepts of Statistical Inference

Before explaining the nature of statistical inference more specifically, we must introduce a few basic concepts. The most crucial concepts in traditional or classical statistics are those of a population and of a sample.

A *population* can be defined as the totality of all possible observations on measurements or outcomes. Examples are incomes of all people in a certain country in a specific period of time, national income of a country over a number of periods of time, and all outcomes of a given experiment such as repeatedly tossing a coin. A population may be either finite or infinite. A *finite population* is one in which the number of all possible observations is less than infinity. However, the distinction between finite and infinite populations is more subtle than may at first appear. For instance, a series of national income figures for the United States for a number of years, e.g., 1948 – 1977, represents a finite collection of thirty observations and thus might seem to be a finite population. But this would be a very narrow interpretation of historical events, since it would imply that the thirty measurements of national income were the only possible ones, i.e., that there is only one course that history might have taken. Now there are obviously not many people who would take such an extremely fatalistic view of the world; most people would admit that it was not impossible for some other, even if only slightly different, values of national income to have occurred. This latter view underlies virtually all policy-oriented research in economics and econometrics and will be used throughout this book. Thus a population of national incomes in a given time interval includes not only the actual history represented by the values that were in fact observed but also the potential history consisting of all the values that might have occurred but did not. The population so defined is obviously an infinite one. Similarly, the population of all possible outcomes of coin tosses is also infinite, since the tossing process can generate an infinite number of outcomes, in this case "heads" and "tails." Most of the populations with which we deal in econometrics are infinite.

Related to the concept of a population is the concept of a *sample,* which is a set of measurements or outcomes selected from the population. The selection can be done by the investigator, in which case we can speak of a sampling experiment, or it may happen independently either by design of others or by nature. In the latter case, the investigator is a mere observer, and this situation is particularly frequent in econometrics. While samples from infinite populations can themselves be infinite, the relevance of such samples is at best only a theoretical one. In practice we deal only with finite samples and, regrettably, quite often only with very small ones. Since samples are obtained by a selection from a given population, the principle of selection clearly plays an important part in determining the composition of the sample. In econometrics our attention is confined to samples drawn in accordance with some specified chance mechanism. Such samples are called *probability samples.* An important type of probability sample is the *random sample.* In finite populations, the principle of selecting a random sample is that of giving every individual in the population an equal chance of being chosen. In the case of infinite populations, a sample is random if each observation (of a measurement or an outcome) is independent of every other observation. The meaning of *independence* will be given in a rigorous way later; at present it is sufficient to note that two events (which can be either measured or counted) are independent if the occurrence of one in no way influences the occurrence of the other.

Both populations and samples can be described by stating their characteristics. Numerical characteristics of a population are called *parameters;* the characteristics of a sample, given in the form of some summary measure, are called *statistics* (a plural of the word "statistic"). Such characteristics may be, for instance, central tendency of measurements (e.g., the mean or the mode), their dispersion (e.g., standard deviation), or, in the case of qualitative phenomena, the proportion of observations of a given kind. Obviously, the parameters of an infinite population are never observed; the parameters of a finite population could be observed in theory but may be impossible to observe in practice.

From our discussion so far it should be clear that statistics deals with phenomena that can be either measured or counted. With respect to a phenomenon that can be measured, we speak of a *variable,* meaning a homogeneous quantity that can assume different values at different points of observation. If a phenomenon can only be counted but not measured (each observation representing one count), we speak of an *attribute.* Thus an attribute is the presence or absence of a given characteristic. An outcome of an event such as the birth of a child leads to an observation of an attribute of sex (i.e., "male" or "not male"); an outcome of a toss of a die may be classified as a presence or an absence of "1," of "2," and so on. In a way the concept of attribute is redundant because we can, and often do, simply assign the value of 1 to the presence, and 0 to the absence, of a given characteristic. In this case we equate "attribute" with the concept of a *qualitative* or *binary variable.* Another and more colorful name, "dummy variable," is also widely used.

The definition of a *variable,* and indeed the name itself, stresses the possibility of variation at different points of observation. On the other hand, a quantity that cannot vary from one observation to another is called a *constant.* If the quantity in question is a variable and not a constant, one may wish to ask about the general source of variation. In particular, it is important to distinguish between those variations that can and those that cannot be fully controlled or predicted. In the case of a variation that cannot be fully controlled or predicted, its existence is due to chance. An obvious example of an uncontrolled variation would be the outcomes of tossing a coin (in the absence of cheating, of course), but many other less obvious instances exist. In fact, as we shall elaborate at length in the rest of this book, most economic variables are always to some extent determined by chance. The variables whose values cannot be fully controlled or determined prior to observation are called *random* or *stochastic variables;* their chief characteristic is that they assume different values (or fall into different value intervals) with some probability other than one. In contrast, a *nonrandom* or *nonstochastic* or *fixed variable* is one that is fully controllable or at least fully predictable. A constant may be regarded as a special case of a fixed variable.

Another important classification of variables is that which distinguishes between continuous and discrete variables. A *continuous variable* is a variable that can assume any value on the numerical axis or a part of it. Typical examples are time and temperature, but income, expenditure, and similar variables can all be classified as continuous. In fact, most economic variables are continuous or at least approximately so. The last qualification is added to take care of such possible objections as those pointing out that money values of less than a dollar (or possibly a cent) are, in fact, not observable. In contrast to a continuous variable, a *discrete variable* is one that can assume only some specific values on the numerical axis.

These values are usually (but not always) separated by intervals of equal length. Examples are a number of children in a family, a number of dots on a die after a toss, or any binary variable.

The final concept to be introduced at this stage is that of a *distribution.* In the case of a sample we have a frequency distribution, while in the case of a population we speak of a probability distribution. A *frequency distribution* represents an organization of data so as to give the number of observations for each value of the variable (in the case of a discrete variable) or for each interval of values of the variable (in the case of a continuous variable). The number of observations in each class (represented by a point in the case of a discrete variable or by an interval in the case of a continuous variable) is called *absolute frequency.* This can be distinguished from *relative frequency,* which gives the proportion of observations rather than their number for each class. As an example, consider a sample of 64 families being observed with respect to the number of children. The results might be those given in Table 1-1. Another example, this time related to a continuous variable, is given by

Table 1-1

Number of children (=variable)	Number of families (=absolute frequency)	Proportion of families (=relative frequency)
0	4	0.0625
1	12	0.1875
2	20	0.3125
3	16	0.2500
4	8	0.1250
5 and over	4	0.0625
Totals	64	1.0000

family income distribution in the United States in 1978 (Table 1-2). Here, absolute frequencies are not shown, and the relative frequencies are stated in percentages rather than in simple proportions. Sample data in the form of a time series, such as national income figures for a number of years, could also be presented in the form of a frequency distribution, although this is usually not done. The fact that different observations are made at different points of time is relevant only to the extent that the population from which the sample was drawn may have changed through time.

In a population the concept corresponding to a sample frequency distribution is known as a *probability distribution.* Consider, for instance, the population of United States families classified by income received in 1978 as shown in Table 1-2. It is fairly clear that to state that 8.2% of all families received an income of less than \$5000 is equivalent to stating that the probability of selecting (at random) a family with an income of less than \$5000 is 0.082. If the population is infinite, the probabilities can be represented by *limits* of relative frequencies (this will be explained more rigorously in Chapter 3). Picturing, then, the probability distribution of one variable as a population counterpart of the frequency distribution in a sample, we

Table 1-2

Income (=variable)		Percent of families (=relative frequency)
Under $5000		8.2
$5000 to $9999		15.8
$10,000 to $14,999		16.7
$15,000 to $24,999		31.4
$25,000 to $49,999		24.3
$50,000 and over		3.6
	Total	100.0

Source: *Statistical Abstract of the United States,* 1980, p. 451.

can see that it is possible to deal with more than one variable at a time. For example, a distribution giving the probability of death at various ages confines itself to one variable—it is an *univariate distribution.* If, however, we tabulate these probabilities separately for each sex, we are considering two variables and have a *bivariate distribution.* A further classification by other characteristics could produce a *multivariate distribution.*

There also exists another kind of probability distribution for which the probabilities are not relative frequencies but are simply results of a personal judgment. Such probability distributions are particularly relevant in situations where it is impossible to determine probabilities by repeated observations and counting. For instance, in considering the prospects of one's score in a golf game, one can typically form a set of probabilities assigned to various scores. Probability distributions of this kind are called *subjective* or *prior* probability distributions. They play a crucial role in the so-called Bayesian approach to statistical inference, which differs from the traditional sampling theory approach and which will be explained later. (The Reverend Thomas Bayes was an English mathematician who lived in the 18th century.)

1-2 The Nature of Statistical Inference

Having introduced, however briefly, some of the most important concepts of statistical theory, we are now in a position to describe the nature of statistical inference. As indicated earlier, statistical inference is concerned with generalizations about the population on the basis of information provided by a sample. Such a procedure is, of course, frequent in everyday life: we make generalizations about the temperature of our soup on the basis of the first spoonful, or about the life expectancy of a pair of tennis shoes on the basis of past experience. This is precisely what is done in statistical inference, except that we go about it in a somewhat more scientific way. What makes the application of statistical inference scientific is that we take into account the way in which the sample was selected, and that we express our