Using Econometrics

a beginner's guide

Henry J. Cassidy

USING ECONOMETRICS: A Beginner's Guide

Henry J. Cassidy

Professorial Lecturer
Virginia Polytechnic Institute and State University
Graduate Economics Program in Northern Virginia

and

Director, General Research Division
Office of Policy and Economic Research
Federal Home Loan Bank Board

Reston Publishing Company, Inc.
A Prentice-Hall Company
Reston, Virginia

Library of Congress Cataloging in Publication Data

Cassidy, Henry J
Using econometrics.

Includes index.
1. Econometrics. 2. Regression analysis.
I. Title.
HB139.C37 330′.028 80-27774
ISBN 0-8359-8135-5

Editorial/production supervision and interior design
by Barbara J. Gardetto

10 9 8 7 6 5 4 3 2 1

Printed in the United States of America

Preface

This book is a nontheoretical presentation of the principles of econometric theory. Its objective is to acquaint potential users with the basic concepts of econometrics and, in particular, regression analysis, without the relatively severe mathematical training and sophistication that is usually required. The book is aimed at undergraduates, beginning graduate students, and the economics professional. For many executives today, the utility they derive from modern econometric analysis is severely impaired because of their lack of understanding of the technique of regression analysis.

This text offers a new organization and scope for econometrics. It serves as a "how-to-do-it" manual for regression analysis. The aim is to develop skills and insight sufficiently to permit novice researchers to conduct their own regression analyses and to read and understand regression analyses. The book was designed as an introductory text for students in the Master's Program at Virginia Polytechnic Institute and State University, Graduate Economics Program for Northern Virginia, but it is also intended for undergraduate students.

Even though brief explanations are given, it is assumed that the reader understands the basic propositions of economics, including concepts such as elasticity measures, demand and supply curves, production functions, and some aggregate relationships such as the consumption function. The statistical background is developed as needed in the text in the context of the regression model, the differential operator of calculus is not used, and the text avoids the usual manipulations with calculus and statistical operators. Usually, students who are familiar with the concepts of statistics (and calculus) do much better, other things equal, than the others. If anything should be a prerequisite, it is courses in statistics and algebra.

The text is designed to make the reader a regression *user*, not a theoretical econometrician. To achieve this end efficiently and economically, a number of features characterize and distinguish the approach of this book.

- There are no proofs of the various propositions, although algebraic manipulations of various regression models are presented. An understanding of the properties of the various estimators is stressed, but derivations of estimating formulas and proofs of the properties of estimators need be of no concern to regression users.
- All the required statistical concepts are introduced and explained in the context of the linear regression model.
- Differential and Integral Calculus is not employed.
- The multivariate regression model is introduced in Chapter 1. There is only a short section dealing strictly with the bivariate (one regressor) regression model. The simple model, however, is used throughout the text whenever it is sufficient to demonstrate the points made.
- Learning by doing is highly recommended. Chapter 1 presents an overview of the basic ''mechanical'' elements of multiple regression analysis and suggests a project for students to complete as they work their way through the book.
- Model development is stressed early on, in Chapter 2, and is discussed apart from any discussion of estimating techniques. This early treatment is intended to assist students in formulating their own regression models.
- Monte Carlo simulations are used to introduce notions of expected value, dispersion, inference, and hypothesis testing, in the context of the regression model; and to demonstrate the effects of various specification errors and the consequences of suggested remedies. In a sense, the Monte Carlo simulations are used as a substitute for the mathematical proofs found in most texts.
- The emphasis is on structural equations, since they comprise the cause-and-effect propositions of economics, and systems of equations are emphasized.
- The emphasis is on simplicity in model formation and estimation. The reader is warned that many of the fancier econometric approaches amount to ''curve fitting'' and should be used with caution.

The emphasis of the text is on the application of regression analysis to inference and hypothesis testing. The topics are geared to first-time regression users. However, the book goes beyond the rudimentary mechanics of regression analysis in that it stresses, through the use of Monte Carlo simulations, the consequences for inference and hypothesis testing of making various kinds of specification errors.

A Teacher's Guide has been developed to accompany the text. It consists of teaching hints as well as questions and answers for each chapter. These questions may constitute homework problems, especially if the students are not assigned a regression project during the course. The Teacher's Guide also contains a suggested syllabus and hints on topics for the regression project. A few proofs are presented that do not appear to be available elsewhere. A separate document contains the documentation for the computer programs, written in Fortran, that were used to conduct the Monte Carlo experiments. Please write Reston Publishing Company for a copy of the Teacher's Guide, and for the program documentation. The program documentation is available to students at a nominal charge.

In all my econometrics courses, I have found vue-graphs or flip-charts of the equations an invaluable teaching aid. It saves my concentration to make the necessary points and emphases, instead of expending my energy writing continuously on the blackboard. It also makes teaching econometrics less of an exhausting chore. I have constructed vue-graphs that essentially outline my classroom discussion, with key words and all the equations, tables, and graphs in the text, put together in a form that aids in the smooth flow of the classroom discussion. I have also video taped the important sections of the text. These video tapes are geared especially for the professional economist who wants to take a self-taught course in econometrics. For information as to the availability and cost of the vue-graphs and the video tapes, and how to get tape copies of the computer programs, write to the author, 3100 N. Oakland Street, Arlington, Virginia 22207.

The list of people who have assisted me along the way is indeed long, and to those people I give my thanks. In particular, however, I would like to thank Professor David Meiselman for his continuing support of the project, financially and morally; Art Bruckheim and Tony Virgilio, for programming and Art for assistance with Chapter 7 (the ARIMA analysis); and to many people for editing and typing assistance, including Michael Bloom, Glenda Brooks, and Kim Brooks. A special note of thanks is due Professor Harry Kelejian and Dr. Richard B. Hoffman, who very thoroughly reviewed earlier drafts; Professor Barry Falk, who contributed to Chapter 7; and Professor P.A.V.B. Swamy, who helped me with the ridge regression analysis of Chapter 11. I especially thank my students, whose suggestions, comments, and corrections over the years are reflected in this book.

Henry J. Cassidy

Contents

1 Overview of Regression Analysis

1.1 Usefulness of Regression Analysis

1.1.1 Usefulness of Economics

Economics is a study of human behavior. Human beings are depicted as economic actors, each of whom plays out one or more roles in the process of attempting to satisfy human wants. The major categories of actors are entrepreneurs, households, and government. Entrepreneurs in a free-enterprise system make decisions about which scarce resources are to be used to produce which goods and services. Households make decisions about which goods to consume, given a set of choices of goods and their prices. Households also offer labor services to entrepreneurs for the purpose of producing goods and services. Governments make decisions affecting various aspects of the production, distribution, and consumption of goods and services.

Economists attempt to understand the whole process of satisfying wants and to understand all the elements of the process. As in the study of any phenomenon, the ultimate purpose of understanding it is to make it work better, to be able to suggest remedies when something goes wrong, or to make it operate differently (viewed as "better" by those making the changes). The suggestions of economists are aimed primarily toward those people whose decisions can make a difference in the way the production, distribution, and consumption process operates, namely, the entrepreneurs and government officials. For example, producers may use economic analysis to decide which products to produce and how to price them; or the government may decide to alter the allocation of wealth, income, or goods and services or to change the levels of aggregate income, employment, and prices by using its powers to tax, spend, and commandeer people and goods.

1.1.2 Usefulness of Econometrics

A very notable characteristic of economists is the disparity of views on how economic processes do or should operate. One of the major purposes of econometrics is to defend or refute alternative theories through the use of quantitative evidence. Once affirmative quantitative evidence has been assembled, it can be used to forecast or predict economic outcomes. The usefulness of accurate forecasting and prediction can be readily appreciated, because the costs of not achieving the decision-maker's desired results can be very high: bankruptcy for the entrepreneur and political upheaval for the government official.

Econometrics attempts to quantify the economic processes. The processes are described in terms of relationships. For example, consumer demand for a particular commodity is actually a relationship between the amount demanded and its price, the prices of complementary and substitute goods, and income. This relationship is called a *function*. The productive process is also a function: output depends on the quantity of factors of production (broadly classified as capital, labor, and raw materials) employed; this particular relationship is called a *production function*.

Economists make forecasts of *variables*. Examples are sales, profits, Gross National Product, taxes, interest rates, and the inflation rate. Some variables are under the direct control of decision-makers. Entrepreneurs can set the prices on the goods they produce, and the government can set the tax rate. These variables that are under the control of decision-makers are called *policy parameters*. Relationships or functions involving policy parameters are thus of special interest to econometricians because it is here that the ultimate payoff is achieved. Prediction is a type of forecasting in which one variable—most likely a policy parameter—is varied to assess how the variable of interest, such as sales or tax revenues, responds, while the other relevant variables are held at specified levels. Sometimes economic theory can be used to predict outcomes, whereas econometrics offers a quantitative estimate of the outcome and provides a technique for holding other variables constant.

The components of econometric analysis are (1) the relationships or functions or, more generally, the economic *models* of human behavior, (2) the data on the variables that are needed to quantify the models, and (3) the quantification of the models. Only the first and third components are explored in this book. In fact, the development of various quantification techniques is what the *science* of econometrics is all about.

1.1.3 Alternative Approaches to Quantification

There are many approaches to quantification, and selecting the appropriate one is the *art* of econometrics. Just as there are alternative economic theories, there are alternative quantification techniques. In general, there are three broad categories of quantification techniques, (1) descriptive statistics, (2) nonparametric techniques, and (3) parametric techniques. Descriptive statistics are various displays of the data, and they attempt to give an intuitive understanding of the relationships without actually quantifying them. Often, they are a useful adjunct to quantified relationships. Nonparametric techniques explore the *degree* of relationship between variables in a function, as contrasted to parametric techniques, which show the amount by which one variable is expected to change if one or more other variables in the relationship change by given amounts. While descriptive statistics and nonparametric techniques are often useful, they do not have the power of the parametric techniques that is so useful for prediction.

1.1.4 Objectives of the Text

There are many parametric techniques, including factor analysis and discriminant analysis. The econometric technique explored in this book is *regression analysis*. The following section presents an overview of regression analysis. The basic—and rather mechanical—elements of regression analysis are presented. The reader is expected to learn how to specify a regression model in terms of the relationships among variables and what the regression technique is supposed to do. The final section of this chapter presents an overview of the steps in applying regression analysis. For a classroom situation, it is suggested that the student specify a regression model at this point; this is the first step of a term project in applied regression analysis.

The major objective of this text is to teach the techniques of regression analysis. Upon completion, the student should be able to formulate, test, evaluate, and use regression models and also be able to read, understand, and evaluate critically articles or reports containing regression analyses. The importance of critical evaluation cannot be stressed too strongly. The *limitations* of regression analysis must be fully perceived and appreciated by anyone attempting to use the findings of regression analysis. Missing data or data inaccurately measured, incorrect formulation of the relationship(s), poor choice of estimating technique, or improper testing procedure—all of these problems imply that predictions from regression analyses need to be viewed with some suspicion. Unfortunately, these kinds of problems are extremely common in applied regression analysis. In many cases, regression analysis is simply not able to provide the proper type of testing of relationships or predictions. This text is designed to highlight the limitations of regression analysis as well as to demonstrate how to use the technique.

This text attempts to provide the minimum set of regression tools and understanding that economists need to perform successful empirical research on a wide range of economic propositions. The emphasis in this text is on the interpretation and application of the techniques of regression analysis. Because the computer accomplishes all the arithmetic and data manipulation required, there is no need for the regression user to be able to prove analytically the mathematical propositions that form the basis of econometrics. An understanding of the regression tools can be achieved without having to investigate the mathematical derivations. An analogy can be drawn to automobile mechanics: they must be able to understand the functioning of the mechanisms of a car in order to diagnose faults and to repair them, but they do not have to be aware of the engineering formulas that were employed in designing the various mechanisms. This text attempts to serve as a guide to building, estimating, evaluating, and using regression models; diagnosis of faults and finding appropriate remedies play an essential role in this process. The reader will learn that, just as with a car, a faulty diagnosis can be costly, and therefore, suggested repairs should be considered very carefully. The purpose of this text is to make the reader a "regression mechanic," one who employs the tools of regression analysis but does not invent or design them—more advanced texts are written for that purpose. But just as car mechanics cannot repair cars unless they understand how they work, neither can regression mechanics build and repair regression models unless they understand the fundamental operations of the technique. This text attempts to supply that understanding without the advanced mathematical detail.

1.2 The Regression Analysis Technique

This section presents a bare-bones outline of the general method of performing single-equation regression analysis. Only the major points are discussed in order to give the reader a brief overview of the mechanics. Chapter 2 discusses model development, usually the most difficult exercise in all of regression analysis. If the regression model is correctly and fully specified, estimation becomes a much simpler chore. There are almost always data problems, but these problems are truly secondary to the correct specification of the theoretical model itself. Thus, before the mechanics of estimation are outlined in Chapter 3, the theoretical aspects of model building are explored in Chapter 2.

Chapter 4 presents what is known as the "classical model." The classical model is composed of a standard set of assumptions; Chapters 6 through 9 deal with estimating equations when some of those assumptions are violated. Chapter 10 extends the notions of hypothesis testing, the fundamentals of which must be introduced early (in Chapter 5). Chapter 11 presents a summary of the material of this text (emphasizing the "artistic" elements of applied regression analysis) and presents a few extensions of regression analysis.

1.2.1 Dependent and Independent Variables

Regression analysis is a statistical technique that attempts to "explain" or "predict," in the form of a single empirical equation, movements in one variable, the dependent variable (or "regressand"), as a function of the movements in a set of variables, called the independent variables (or "regressors"). Regression analysis is a very natural statistical tool for economics because most economic propositions can be stated in single-equation functional form: quantity demanded (regressand) is a function of price, income, prices of substitute and complementary goods, etc. (regressors); aggregate consumption (regressand) is a function of disposable income, price expectations, etc. (regressors); output (regressand) is a function of capital and labor inputs (regressors).[1]

---

[1] Often, there are several related economic propositions that, when taken as a group, suggest a system of regression equations. An example is a two-equation model of supply and demand; usually, these two equations must be considered together instead of separately. These models are discussed in Chapters 2 and 9.

The main body of economics is concerned with cause-and-effect propositions: if the price of a good increases by one unit, then the quantity demanded decreases on average by a certain amount, depending on the elasticity of demand (defined in Section 2.2.1); if aggregate disposable income increases by a billion dollars, then aggregate consumption increases by a certain amount, depending on the marginal propensity to consume; if the quantity of capital employed increases by one unit, then output increases by a certain amount, depending on the marginal productivity of capital. All these propositions pose an if-then, or *causal*, relationship that logically postulates a "dependent" variable having movements that are causally determined by movements of a number of specified "explanatory" variables. The *behavior* of a given economic group or entity is explained by the causal propositions of economics: the economic entities are households for the demand and aggregate consumption functions; and whereas production functions are of a technical, or engineering, type, they depict behavior patterns of entrepreneurs, or producers.

The purpose of regression analysis in economics is to quantify and use the behavioral relationships postulated by economic theory: to quantify magnitudes such as the elasticity of demand, the marginal propensity to consume, and the marginal productivity of capital. In many cases, verification or rejection of an economic proposition is the desired end product of the regression analysis, in which case the precise quantification of the magnitude of the coefficients in the relationship is not the goal. In other cases, the purpose of regression analysis is to forecast the value of the dependent variable, given assumed or predicted values of the regressors. The same general statistical approach is used in all cases, but the emphasis on the various steps of regression analysis differs according to the purpose of the regression analysis.

Because regression analysis can be used to quantify the behavioral relationships of economic theory, it forms the basis for most of the empirical research carried out in economics. Although other statistical tools are used and may be more appropriate for certain applications, for most economists, regression analysis is the primary statistical technique that is used.

1.2.2 The Single-Equation, Linear Regression Model

The single-equation, linear regression model, in its simplest form, may be written as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad i = 1, 2 \ldots, n \qquad (1.2.1)$$

where

Y_i = the ith observation on (or of) the *dependent variable* or *regressand*,

X_i = the ith observation on the *explanatory* or *independent* variable, also called the *regressor*,

ϵ_i = the ith observation on the *error* or *disturbance* term,

β_0 and β_1 = the *regression coefficients*, and

n = the number of observations.

A fuller writing of the regression model (equation 1.2.1) is as n equations, one for each observation, as follows:

$$Y_1 = \beta_0 + \beta_1 X_1 + \epsilon_1$$
$$Y_2 = \beta_0 + \beta_1 X_2 + \epsilon_2$$
$$Y_3 = \beta_0 + \beta_1 X_3 + \epsilon_3$$
$$\vdots$$
$$Y_n = \beta_0 + \beta_1 X_n + \epsilon_n$$

That is, the regression model is assumed to hold for each observation.

As an illustration, suppose Y_i is aggregate consumption in year i and X_i is aggregate disposable income in year i; so the regression equation represents the simplest of the (Keynesian) aggregate consumption functions. Figure 1.2.1 is a graphical illustration of this function; the dots represent observed pairs of X and Y, and, in the sample, there are n such observational pairs. A given level of income (X) is postulated to cause or determine a given level of consumption (Y) in a simple, linear manner. "Linearity" refers to the fact that the regression relationship $\text{AVG}(Y_i) = \beta_0 + \beta_1 X_i$ shown in Figure 1.2.1 is a straight line. The notation "$\text{AVG}(Y_i)$" is used to show the distinction between the regression equation 1.2.1 and the straight line shown in the graph. The difference is that the error term ϵ_i is omitted in the straight-line representation of the regression relationship.

The error term ϵ_i is included in the regression relationship (equation 1.2.1) because actual, observed values of Y_i very likely do not all fall on a straight line. For example, if income is $2.0 trillion for the fifth observation in the data set (i.e., $X_5 = 2.0$; in Figure 1.2.1, the observed value X_5 is shown on the horizontal axis) and the

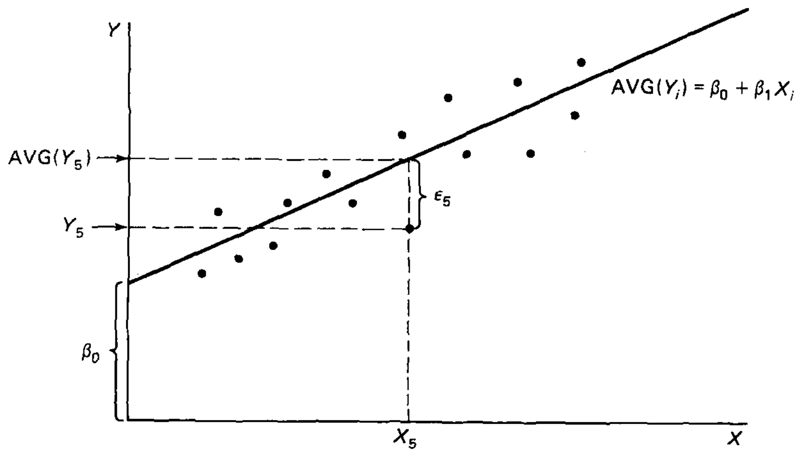

FIGURE 1.2.1 Simple Consumption Function

regression coefficients are $\beta_0 = 0.1$ and $\beta_1 = 0.8$, then the computed value of consumption for the fifth year would be $\beta_0 + \beta_1 X_5 = 0.1 + 0.8\ (2.0) = \$1.7$ trillion. The interpretation of this $1.7 trillion figure is that, *on average*, $1.7 trillion in consumption can be expected each time an income level of $2.0 trillion is observed; hence the use of the AVG notation. In period 5, however, the actual, observed level of consumption (Y_5) may be $1.5 trillion. In another period, such as the tenth year in the sample, income may be $2.0 trillion again, but this time consumption may be $1.8 trillion. Regression relationships are able to show only the *average* or *expected* level of consumption for each level of observed income, which for this fifth observation is shown by the average regression relationship $\text{AVG}(Y_5) = \beta_0 + \beta_1 X_5$. The equation cannot predict without error the actual, or observed, level of consumption in each period. In the example, the error in the fifth period, written as ϵ_5, is the $1.5 trillion of consumption observed less the level of consumption that is expected, $1.7 trillion, or $ − 0.2 trillion.

An error term is always present for regression equations. In the consumption function example, consumption may have been less than expected in period 5 because of uncertainty over the future course of the economy, causing consumers to save more and to consume less than they on average would if the uncertainty did not exist. Another reason for the error in period 5 may be that the consumption function attempts to portray the behavior of people, and there is an element of unpredictability in human behavior; that is, just by chance households consumed less than they were expected to on average. A third reason for the error may be that the *observed* amount of consumption in period 5 may have been different from the *actual* level of consumption due to some error (such as sampling error) in the measurement of consumption in the National Income Accounts. A fourth reason for the error may be that the true relationship is not exactly linear as was assumed. Any of these reasons can explain the error between the observed values of Y_i and the values expected from the hypothesized linear relationship, $\text{AVG}(Y_i)$.

The expression $\beta_0 + \beta_1 X_i$ is called the *deterministic* part of the regression equation. It indicates what the expected level of consumption is for any given value of X_i. Notationally, the deterministic part of the equation is written as in Figure 1.2.1:

$$\text{AVG}(Y_i) = \beta_0 + \beta_1 X_i \qquad (1.2.2)$$

The part of equation 1.2.1 that distinguishes it from the deterministic part (equation 1.2.2) is the error term, ϵ_i. This part of the regression equation is called the *stochastic* or *random* part. It is always included because the observed value of Y_i is very unlikely to be the same as the regression-line value $\text{AVG}(Y_i)$.

The coefficient β_0 is called the "constant term" or the "Y-intercept." In Figure 1.2.1, it is the expected value of Y_i when X_i is zero. The coefficient β_1 is called the "slope" coefficient. It shows the increase in the average or expected value of Y_i caused by an increase in X_i of one unit (e.g., one trillion dollars). In other words, it shows the response of the expected value of Y_i to *changes* in X_i. Since being able to predict changes in the dependent variable if X_i changes is the essence of quantifying behavioral relationships, most of the emphasis in regression analysis is on slope coefficients such as β_1.

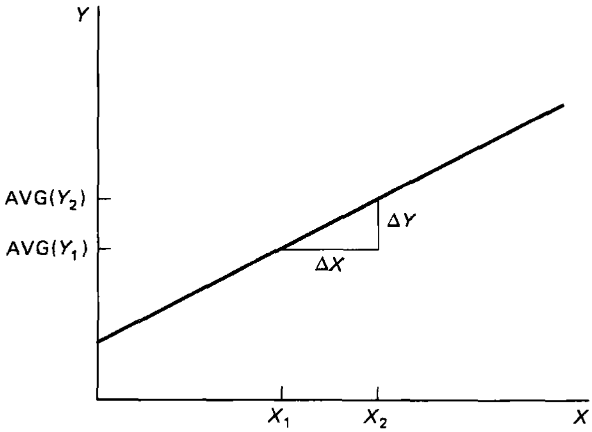

FIGURE 1.2.2 Graphic Representation of the Slope of the Regression Line

Figure 1.2.2 is a graphic interpretation of the slope of the regression line. If X were postulated to increase from X_1 to X_2, the value of Y predicted from the regression line would increase from Y_1, the level associated with X_1, to Y_2, the level associated with X_2. For linear (i.e., straight-line) models, the response in the predicted value of Y due to the change in X, i.e., the slope, is measured as:

$$\frac{\text{AVG}(Y_2) - \text{AVG}(Y_1)}{X_2 - X_1}$$

When Δ is used to denote the change in the variables (and Y is used instead of AVG(Y)), the slope coefficient can be represented as:

$$\beta_1 = \frac{\Delta Y}{\Delta X}$$

Some readers may recognize this as the "rise" (ΔY) divided by the "run" (ΔX), which refers to the directions in the graph by which the changes in Y and X are measured.

Variables other than income are likely to affect the level of consumption. If the interest rate is relatively high, people may save more and consume less, or if uncertainty about recession is rampant, consumption may be curtailed. If we define

X_{i1} = the ith observation on disposable income,

X_{i2} = the ith observation on the interest rate, and

X_{i3} = the ith observation on an index of consumer uncertainty over recession,

then all three variables can be expressed as determinants of Y in a *multiple linear regression model*:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \qquad (1.2.3)$$