Introductory Econometrics: Theory and Applications

R. Leighton Thomas

Introductory Econometrics: Theory and Applications

R. Leighton Thomas

Longman
London and New York

Longman Group Limited
Longman House, Burnt Mill, Harlow
Essex CM20 2JE, England
Associated companies throughout the world

Published in the United States of America
by Longman Inc., New York

First published 1985

British Library Cataloguing in Publication Data
Thomas, R. Leighton
Introductory econometrics (Longman economics series)
I. Title
330'.028
ISBN 0-582-29634-X

Library of Congress Cataloging in Publication Data
Thomas, R. Leighton, 1942–
Introductory econometrics [illegible]
(Longman economics series)
Bibliography: p.
Includes index.
1. Econometrics. I. [illegible]tle. II. Series. [illegible]
HB139.T52 1984 330'.[illegible]8 [illegible]932
ISBN 0-582-29634-X

Printed in Singapore by
Champion Office Supplies Pte Ltd.

Preface

The idea of writing this book arose out of my teaching of the final-year course in applied econometrics at the University of Salford during the late 1970s. Students on this course did not necessarily take a parallel course in introductory econometric theory so the course, although labelled 'applied', had to include a fairly extensive but mainly intuitive theoretical section. The students moreover were composed entirely of non-specialists in econometrics whose background in statistics consisted simply of a second-year course in 'quantitative methods' similar to that run in many British universities and polytechnics.

When searching for a suitable background text for the course, it was immediately clear that no precisely suitable textbook existed. Typical texts either consisted of a long theoretical section with one or two 'applied' topics tagged on at the end, or included a brief theoretical introduction followed by a series of applied chapters each written by a different author. The problem with the latter type of book was, not surprisingly, a lack of integration between the various sections. I have therefore attempted to write an integrated textbook for non-specialist students which includes a fair proportion both of theoretical material and of econometric applications, but in which the range of theoretical topics covered is, to some extent, determined by the applied topics included later. Ideally, students using the book will eventually be studying both theoretical and applied sections at the same time.

The book will, I hope, be of use both to students taking a single final year course in econometrics and to those who take parallel non-specialist courses in both theory and applied. It may also be useful for non-specialist graduate courses in econometrics such as those appearing in the taught Master's programmes at most British universities. When used for a single theory/applied course it will probably be necessary to limit the applied chapters covered and possibly to concentrate on those theoretical topics specifically required for the applied topics chosen.

The typical quantitative methods course taught in the second year of most British university and polytechnic first-degree courses should be a perfectly adequate preparation for tackling this book. Students should have a working knowledge of the two-variable regression model, including its inferential aspects. However, a brief revision of some of the basic concepts of two-variable regression is included in Chapter 2. A limited knowledge of calculus, including partial differentiation, is also assumed as is some familiarity with matrix algebra (e.g. the meaning of an inverse matrix). However, use of matrix algebra is kept to a minimum and largely confined to Chapter 2.

An important feature of the applied econometrics course at Salford is the series of empirical exercises that students are expected to tackle during the year. Any student of econometrics should, at an early stage, get used to handling genuine data and using simple multiple regression packages. Such packages are, of course, now readily available at virtually all British universities and polytechnics. Accordingly, appendices based on the exercises used at Salford have been included at the end of each applied chapter in the book. These appendices are not meant just to be read. Students should use regression programmes to duplicate any estimated equations that are quoted and to follow up any suggestions made for further work. In using the book it is not necessary to wait until the applied section is reached before attempting the exercises. It is quite possible, for example, to tackle the early sections of any of the exercises once Chapter 2 has been read.

I owe a considerable debt to a number of friends and colleagues for making helpful comments on early drafts of various chapters in the book. In particular I must thank George Zis, Mike Sumner and Neil Thompson. For the painstaking job of actually typing both earlier and final drafts, I am most grateful to my wife, Margaret, and to Shirley Wooley. Thanks for typing are also due to Kath Bacon, Susan Mullins and Sharon Machin. None of the above, of course, are responsible for any errors and confusions that remain.

R.L.T. 18 October 1983.

Acknowledgements

We are indebted to the following for permission to reproduce copyright material:

Cambridge University Press for Tables 6.1 (Prais & Houthakker, 1955), 10.2 (Courakis, 1978), 10.3–10.5 (Hendry & Mizon, 1978); The Econometric Society for Table 7.1 (Zellner & Geisel. 1970); Elsevier Biomedical Press B.V. Amsterdam, for Table 10.1 from Table 1, p. 227 *Demand for Money in UK* by D. Laidler; International Statistical Institute & the author, Prof. H. Theil for fig. 11.1, Tables 11.2, 11.3 (Theil & Boot, 1962); National Institute of Economic & Social Research for Table 11.4, p. 27 *NIER*, Nov. 1981; Yale University Press for Table 11.1 (L.A. Klein 1950).

Abbreviations

AIDS	almost ideal demand system
AIH	absolute income hypothesis
APC	average propensity to consume
APC	'equilibrium' average propensity to consume
BLUE	best linear unbiased estimator
BB	*National Income and Expenditure Blue Book*
CE	Cambridge Econometrics
CEPG	Cambridge Economic Policy Group
CES	Constant elasticity of substitution
EIU	Economic Intelligence Unit
ETAS	*Economic Trends Annual Supplement*
FIML	full information maximum likelihood
GLS	generalised least squares
ILS	indirect least squares
LBS	London Business School
LCH	life-cycle hypothesis
LIML	limited information maximum likelihood
MLE	maximum likelihood estimation
MPC	marginal propensity to consume
MPS	marginal propensity to save
MRS	marginal rate of substitution
NI	National Institute
OLS	ordinary least squares
PIH	permanent income hypothesis
RIH	relative income hypothesis
SSE	explained sum or squares
SSR	residual sum of squares
SST	total sum of squares
TSLS	two-stage least squares
VES	variable elasticity of substitution

Contents

1 Introduction

Most economic theories have developed out of *a priori* reasoning based on relatively simple assumptions. However, different assumptions will lead to different theories. If we are to provide government with sensible policy prescriptions, we therefore require some way of distinguishing 'good' theories from 'bad' theories. The obvious way is to refer to 'the facts'. In the physical sciences a theory is judged by its ability to make successful predictions. Hypotheses are developed by a combination of *a priori* reasoning and empirical observations and are then used to generate predictions which can be tested against further data. If the predictions are judged 'correct', the hypothesis or theory still stands, while if the predictions are incorrect the hypothesis is either rejected or simply reformulated to take account of the new data. Such traditional 'scientific method' has served the physical sciences well over the past two centuries and one might hope that a similar approach was possible in economics. However, there are problems.

A major problem is that the economist can rarely, if ever, conduct a controlled laboratory experiment. Take two simple examples – one from physics and the other from economics. Suppose we were interested in the effect on the volume, V, of a gas of variations in its temperature, T, and the pressure under which it is kept, P. Specifically, we might wish to test the hypothesis that a given proportionate increase in the temperature of the gas, with pressure held constant, leads to a more than proportionate increase in its volume. That is, resorting to the terminology of economics, we ask the question – is the elasticity of volume with respect to temperature greater than unity? Suppose we assume that a relationship of the form

$$V = AT^{\alpha}P^{\beta} \qquad [1.1]$$

exists where A, α and β are constants. α measures the effect on volume of changes in temperature when pressure, P, is kept constant and β measures the effect on volume of changes in pressure when temperature, T, is held constant.[1]

Equation [1.1] is referred to as a *maintained hypothesis*. In hypothesis-testing situations, typically we make a number of assumptions not all of which are to be tested. Those assumptions we are prepared to accept and do not intend to test constitute the maintained hypothesis. We can never be certain that a maintained hypothesis is valid (e.g., a simpler linear formulation might be preferable to [1.1]), but some such assumptions are always necessary if hypothesis testing is to proceed at all. The form of equation [1.1] is, in fact, very suitable for the purpose at hand since α and β are, of course, elasticities. α is the elasticity of volume with respect to temperature under conditions of constant pressure. To measure α we would set up a laboratory experiment under which pressure is kept constant and we vary the temperature of the gas at will. We then observe, the relationship between temperature and volume under these conditions and come to some conclusion about the size of α. Similarly, β is the elasticity of volume with respect to pressure under conditions of constant temperature. A further controlled experiment would be set up if we wished to come to some conclusion about the size of β.

Now consider a situation in economics where we are interested in the effect on a household's consumption expenditure, C, of variations in its disposable income, Y, and its stock of liquid assets, L. Specifically we might be concerned whether, for a given stock of liquid assets, the relationship between consumption and income was one of proportionality. Suppose we specified the maintained hypothesis

$$C = AL^{\alpha}Y^{\beta} \qquad [1.2]$$

where α now measures the elasticity of consumption with respect to liquid assets when income is constant, and β measures the elasticity with respect to income when the liquid asset stock is constant. Given the maintained hypothesis [1.2], testing whether the relationship between C and Y is one of proportionality simply involves testing the hypothesis

$$\beta = 1.$$

Unfortunately, it is very unlikely that we will ever be able to set up controlled experiments in which, for example, we hold a household's liquid asset stock constant and observe the relationship between C and Y. In economics it is very rarely the case that we are able to collect data specifically generated for the purpose in hand. Rather, the economist has to make use of whatever data he can find. Such data can be classified into two kinds – *time series* data and *cross sectional* data.

Time series data on an individual household would consist of observations on the income, consumption and liquid assets of the household for a series of successive periods, e.g. months or years. Although most published time series data on the consumption behaviour of households refers to aggregates (often economy-wide), of very many households, it would be quite feasible to collect data on an individual household. Unfortunately, there would be no way in which we could guarantee that the household's liquid asset stock remained constant while we observed the relationship between C and Y.

Cross-sectional data consist of observations on different households over the same period of time. For example, the Family Expenditure Surveys in the UK and general household surveys in other countries provide such data on many thousands of households. However, there would be no reason to expect household stocks of liquid assets to be constant over the cross-section. Moreover, as we are now dealing with different households, we would also be faced with the problem of variations in size, composition and background.

Clearly, whatever type of data is available for the investigation of [1.2] we are faced with the difficulty that both Y and L will be varying. This situation is, of course, the normal one in economics. Our data is almost invariably such that all variables we are interested in will be non-constant, the controlled experiment not being feasible.

A statistical technique exists that goes some way to overcoming the handicap of being unable to carry out controlled experiments. This technique, known as *multiple regression analysis*, enables us to 'estimate' quantities such as A, α and β in equation [1.2] simultaneously, without the need to hold variables constant artificially. Notice that if we take logarithms of equation [1.2] we obtain the linear equation

$$\log C = \log A + \alpha \log L + \beta \log Y \qquad [1.3]$$

The reader should be familiar with the *least squares technique* of estimating a simple linear relationship between two variables. This technique can, in fact, be extended to the estimation of linear relationships such as [1.3], although in the *multiple regression* case the estimated relationship cannot be depicted in a simple two-dimensional diagram. Multiple regression, then, is the economist's replacement for a controlled laboratory experiment. Often it may not be a very good replacement but it is normally the best we have and much *econometrics* involves its use in one form or another.

Equations [1.2] and [1.3] suggest that an exact or *deterministic* relationship exists between the left-hand side or dependent variable and the two right-hand side or explanatory variables. However, economic relationships are never exact – human beings are unpredictable in their behaviour, and for this reason a random *disturbance* is usually added to such relationships. For example, a household may receive exactly the same income and possess exactly the same liquid asset stock in one week as it does in another. Yet its consumption may well differ for purely random reasons. We therefore rewrite equations such as [1.3] which we wish to estimate as

$$\log C = \log A + \alpha \log L + \beta \log Y + \varepsilon \qquad [1.4]$$

where ε is a random disturbance which may take either a positive or a negative value.[2] This disturbance can also be regarded as reflecting all other factors apart from Y and L which have some (hopefully slight) effect on household consumption. One cannot expect Y and L to encompass all influences on C.[3]

The fact that a random disturbance is included in economic relationships means that we cannot expect to measure quantities such as A, α and β in [1.4] exactly. This would be the case even if we were able to set up controlled experiments and hold Y and L constant because we would have no control over ε, the random factor. For example, if we held L constant in order to investigate β, then our findings from one experiment might well differ from those in another because of the different random responses of the household, reflected in different values for ε. Similarly, when applying the multiple regression technique, we might obtain one set of estimates for A, α, and β from one sample of observations on Y, L and C and a rather different set from another sample. In other words, the estimators are subject to sampling variability and have sampling distributions.[4] We cannot therefore estimate A, α and β exactly but are reduced to finding, for example, 95 per cent confidence intervals for their values. Similarly, we can never say with certainty that, for example, $\alpha \neq 0$ in equation [1.4], i.e. that liquid assets influence consumption. We can merely test statistically the hypothesis $\alpha = 0$ and reject it or not at, for example, the 5 per cent level of significance. It is when we reject $\alpha = 0$ that we say 'liquid assets are significant at the 5 per cent level'.

At this point it is worth considering the meaning of the term 'level of significance'. For example, to say that a hypothesis is rejected at 'the 5 per cent level of significance' is an admission that there is a probability of 5 per cent that it has been wrongly rejected. That is, that there is a probability of 5 per cent that the hypothesis was true all along and that the characteristics in the data that led to its rejection occurred simply by chance. Hence, if we reject the hypothesis $\alpha = 0$ in [1.4] at the 5 per cent level of significance, we are saying that we believe liquid assets influence consumption but acknowledge a probability of 5 per cent that the statistical association that led us to this conclusion could have occurred by chance.

The fact that the significance level of a test is an admission of the existence of chance has one important implication. Suppose we were interested in other possible determinants of household consumption. Imagine we tried adding, one at a time, twenty different variables to the right-hand side of [1.4]. Remember that, even if such a variable is of no importance in the determination of consumption, there is a 1 in 20, or 5 per cent, probability that it will appear 'significant at the 5 per cent level' purely by chance. Hence, if we try twenty such variables we must expect one of them to appear significant even if *none* of them are of real importance. The danger now is that we might forget the nineteen 'unsuccessful' variables and focus attention on the single 'significant' one, maintaining that we had uncovered evidence that it is an important determinant of consumption. What we would have done, however, is to have confused hopelessly the business of hypothesis *testing* with that of hypothesis *formulation*. The statistical relationship we have uncovered *may* reflect a genuine causal link but it is also possible that it represents a purely spurious relationship that happens to exist just in the data we have observed. We have *formulated* a hypothesis by 'observing' this data. What we cannot do is to *test* this hypothesis using the *same* body of data and it is silly to claim that we have. A hypothesis formulated from one data set obviously needs to be tested on a *new* data set.

The above procedure is an extreme example of what is commonly referred to as 'data-mining'. Unfortunately, such data-mining, albeit in a moderate form, appears to be a fairly common practice in much empirical economic research. One finds impressive-looking regression equations presented in many published papers. What should be realised is that the presented regressions are almost certainly the 'most successful' of a whole series of 'trial' regressions, the vast majority of which do not appear. The presenter may not have tried the twenty variables of the above example, but he will probably have tried two or three and also experimented with different definitions of the one that worked best.[5] Although it is probably an inevitable consequence of the paucity of economic data, there is therefore a tendency for hypothesis formulation and testing to get mixed up in economic research. Because of this it is probably wise to take many of the regression equations reported in the applied chapters of this book with just a slight pinch of salt and mentally downgrade the significance of variables and the overall performance of presented equations. The 'non-statistically minded' reader may have some difficulty in fully understanding what has just been said but, if this is the case, it would be a good idea if he/she returned to this introduction after reading the theoretical Chapters 2–5.

The technique of multiple regression is described in Chapter 2. Because the existence of the random disturbance in relationships such as [1.4] means that this technique will yield only *estimates* of parameters like A, α and β, we are inevitably concerned with the quality of these estimates. *Estimates* are always obtained by the use of some estimating formula or *estimator* and we would obviously like this estimator to be a 'good' one in some sense. Part of Chapter 2 is therefore concerned with defining the properties that we would like our estimators to have. We then consider the conditions under which the least squares method most commonly used in multiple regression analysis will yield estimators possessing these properties. The necessary conditions make up what is frequently referred to as the *classical multiple regression model*. Unfortunately they turn out to be rather restrictive conditions.

Firstly, it is necessary that disturbances such as ε in equation [1.4] should satisfy a whole series of assumptions many of which are unlikely to be met. Secondly, the manner in which we are normally forced to collect our data turns out to be important. Because we are unable to perform controlled experiments we are unable to fix for ourselves the sample values of explanatory variables such as Y and L in [1.4]. Instead, we have to accept any values thrown up by chance by the economic system we are observing. In the jargon, the explanatory variables are *stochastic* or *random* rather than *non-stochastic* or *non-random*. The consequences of this are often serious, particularly if the relationship we are interested in is but one of a simultaneous system of such relationships. Since we represent most economic systems in this way, this is the most common situation.

The consequences of breakdowns in the assumptions that make up the classical regression model and the alternative procedures that are available are considered in detail in Chapters 3 and 4. Indeed, the analysis of such breakdowns and the devising of alternative estimating procedures make up the main subject matter of theoretical econometrics. It is Chapter 4 that is concerned with problems arising out of the simultaneity of most economic relationships. Chapter 5 lists some extensions of normal regression analysis which we will find are used frequently in the applied chapters in the remainder of the book.

Chapters 6–10 each cover an important area of applied work in econometrics. Empirical exercises are included at the end of each of the applied chapters. These involve the use of actual data on the UK economy for estimating regression equations arising out of the material of the preceding chapter. Working with realistic data is a vital part of any course in econometrics since only by actually trying to estimate economic relationships will a student begin to acquire a 'feel' for the difficulties involved. Any of the various statistical packages, such as TSB/ESB, DEMOS and GENSTAT, currently available in UK universities is suitable for tackling these exercises. There is, in fact, no need for the reader to wait until Chapter 6 before turning to the exercises. Early parts of each exercise can and should be attempted once Chapter 2 has been read and understood.

Finally, in Chapter 11, the structure and uses of the major UK macroeconometric models are discussed. We consider how such models are used in forecasting and as an aid to policy formulation. Successful forecasting and the provision of sensible policy prescriptions are two of the ultimate aims of econometrics.

When reading the applied chapters, the reader may be struck by the fact that there appears to be no coherent pattern in the econometric research work performed in the various areas we cover. Unfortunately much empirical work in economics has suffered from the lack of a coherent and constructive research strategy. Work tends to proceed in virtual isolation, taking only token account of previous research in an area. Equations are estimated in a purely *ad hoc* manner with only the most precursory reference to economic theory. The requirements of a constructive research strategy are well summarised in Davidson, *et al.* (1978). Firstly, any new 'model' should only supplant old 'models' if it can account not only for all previously accepted results but also explain some new phenomena that the old models cannot. Secondly, a new model must have a sound basis in economic theory. Thirdly, any new model must be able to account for *all* the properties of the data under consideration. In particular, it should be able to explain the results obtained by previous researchers using the same data set and also explain why their research methods led to such conclusions.

Such a constructive research strategy is now being followed by workers in a number of fields in economics. A rigorous example of the approach is provided by the paper on the UK consumption function by Davidson, *et al.* referred to above. We shall discuss this paper in Chapter 7 but the approach will also be encountered in the chapters on investment and on the demand for money. A characteristic of the approach is to estimate equations involving not the 'levels' of variables but their rates of change. Many economic time series 'trend' either continuously upward or continuously downward (the price level in the postwar UK is an obvious example). Such trend variables will always be highly correlated and there is an obvious danger that such correlations will be at least partly spurious. Working in terms of the rate of change of variables will frequently remove trend elements (until recently the rate of change in the UK price level showed no definite trend either upwards or downwards). Spurious correlations can thus be avoided although there is a danger that unless equations are properly specified vital information relating to the levels of variables will not be made use of.

Another distinctive characteristic of the approach is to start with a very general formulation and then use the data evidence to simplify the estimating equation along lines consistent with economic theory. This contrasts with the more conventional approach adopted by many investigators where economic theory is used to specify an initial simple form for estimating equations which is then modified according to the characteristics of the data. The problem with the new approach is that such economic theory as exists frequently provides little more information than that certain variables bear some proportionate relationship to one another when in steady state. For example, the equilibrium relationship between consumption and income or capital stock and output can be taken as one of proportionality. This leaves so much scope for data-based simplification of the general equations, by, for example, experimenting with various lag structures, that some economists would regard the approach as verging on the 'data-mining' discussed earlier. However, the data-based approach at the very least provides a standard by which the explanatory power of more conventionally estimated equations can be judged. Possibly a judicious combination of the two approaches will prove the most fruitful in future research efforts.

At this stage the reader may find some difficulty in fully appreciating some of the issues just discussed. However, if this is the case then it should prove helpful to re-read this introduction once the book as a whole has been studied.

Notes

1. Those familiar with Charles' Law will recognise that the actual relationship between V, P and T is $PV/T = \text{constant}$. Hence, since this implies $V = (\text{constant})\ TP^{-1}$, experimentation should yield $\alpha = 1$ and $\beta = -1$.
2. This implies that the original equation [1.2] should be rewritten as $C = AL^{\alpha}Y^{\beta}\theta$ where θ is the 'antilog' of ε, i.e. $\theta = e^{\varepsilon}$. θ must be assumed always greater than zero, otherwise consumption would be negative and $\varepsilon = \log\theta$ would not be defined.

3. ε may also reflect the fact that we cannot always measure variables with perfect accuracy. While a relationship such as [1.3] could hold for the true values of C, L and Y it may well not hold exactly for the data we obtain. Possible errors in the measurement of C, L and Y are thus another reason for adding a random disturbance to [1.3].
4. The situation is analogous to that when we attempt to estimate a population mean, μ, by the mean of a random sample, $\bar{x}$. The reader should be familiar with the fact that $\bar{x}$ is subject to sampling variability, i.e. that different random samples will yield different values for $\bar{x}$.
5. For example, the percentage annual rate of price inflation can be calculated in several ways. Similarly, there are alternative definitions for the 'broad money stock'.

2 Multiple regression analysis

In this chapter we shall be concerned with the so-called 'classical linear regression model'. A working knowledge of the two-variable regression model will be assumed and we shall deal mainly with what is known as multiple regression. However, we first revise some crucial aspects of simple two-variable regression analysis.

2.1 Revision of some important concepts in two-variable regression

In simple regression analysis a linear relationship is assumed between a *dependent* variable Y and an *explanatory* variable X

$$Y = \beta_1 + \beta_2 X + \varepsilon \qquad [2.1]$$

For example, Y might be the weekly consumption expenditure of a household of given size and composition and X the weekly dispossable income of such a household. β_1 and β_2 are fixed constants and ε is a random *disturbance*. The disturbance reflects, firstly, all factors other than disposable income which influence the consumption expenditure of this type of household, e.g. its tastes, social and educational background, the size of its bank balance, etc. ε may therefore be positive or negative. It might be positive for a household which because of past savings has a large positive bank balance and may be negative for a household which has incurred large debts. Secondly, the disturbance reflects the basic unpredictable or random nature of human behaviour. We do not expect two households with the same disposable income and identical in other respects to, necessarily, make exactly the same consumption expenditure. Neither can we expect a given household to make exactly the same consumption expenditure in two successive weeks even when the conditions under which it operates remain unchanged.

The disturbance ε may be regarded as a random variable with its own probability distribution and it is convenient to assume for the moment that its average or expected value is zero, i.e. $E(\varepsilon) = 0$. Taking expectations over equation [2.1], we then have for a household of given income X

$$E(Y) = \beta_1 + \beta_2 X \qquad [2.2]$$

Equation [2.2] is sometimes referred to as the *population regression line* and β_1 and β_2 are population parameters. $E(Y)$ may be regarded as the average or expected consumption expenditure of households with the given disposable income X. The parameter β_1, of course, represents the expected expenditure of a household with zero income while β_2 measures the change in expected expenditure per unit change in disposable income X. The population regression line is