Takeshi Amemiya

ADVANCED ECONOMET

Advanced Econometrics

Takeshi Amemiya

Basil Blackwell

First published in 1985

This edition published in 1986 by Basil Blackwell Ltd,
108 Cowley Road, Oxford OX4 1JF, UK

British Library Cataloguing in Publication Data

Amemiya, Takeshi
Advanced econometrics.
1. Econometrics
I. Title
330'.028 HB139

ISBN 0-631-13345-3

Printed in Great Britain by TJ Press Ltd, Padstow

Preface

This book is intended both as a reference book for professional econometricians and as a graduate textbook. If it is used as a textbook, the material contained in the book can be taught in a year-long course, as I have done at Stanford for many years. The prerequisites for such a course should be one year of calculus, one quarter or semester of matrix analysis, one year of intermediate statistical inference (see list of textbooks in note 1 of Chapter 3), and, preferably, knowledge of introductory or intermediate econometrics (say, at the level of Johnston, 1972). This last requirement is not necessary, but I have found in the past that a majority of economics students who take a graduate course in advanced econometrics do have knowledge of introductory or intermediate econometrics.

The main features of the book are the following: a thorough treatment of classical least squares theory (Chapter 1) and generalized least squares theory (Chapter 6); a rigorous discussion of large sample theory (Chapters 3 and 4); a detailed analysis of qualitative response models (Chapter 9), censored or truncated regression models (Chapter 10), and Markov chain and duration models (Chapter 11); and a discussion of nonlinear simultaneous equations models (Chapter 8).

The book presents only the fundamentals of time series analysis (Chapter 5 and a part of Chapter 6) because there are several excellent textbooks on the subject (see the references cited at the beginning of Chapter 5). In contrast, the models I discuss in the last three chapters have been used extensively in recent econometric applications but have not received in any textbook as complete a treatment as I give them here. Some instructors may wish to supplement my book with a textbook in time series analysis.

My discussion of linear simultaneous equations models (Chapter 7) is also brief. Those who wish to study the subject in greater detail should consult the references given in Chapter 7. I chose to devote more space to the discussion of nonlinear simultaneous equations models, which are still at an early stage of development and consequently have received only scant coverage in most textbooks.

In many parts of the book, and in all of Chapters 3 and 4, I have used the theorem-proof format and have attempted to develop all the mathematical results rigorously. However, it has not been my aim to present theorems in full mathematical generality. Because I intended this as a textbook rather than as a monograph, I chose assumptions that are relatively easy to understand and that lead to simple proofs, even in those instances where they could be relaxed. This will enable readers to understand the basic structure of each theorem and to generalize it for themselves depending on their needs and abilities. Many simple applications of theorems are given either in the form of examples in the text or in the form of exercises at the end of each chapter to bring out the essential points of each theorem.

Although this is a textbook in econometrics methodology, I have included discussions of numerous empirical papers to illustrate the practical use of theoretical results. This is especially conspicuous in the last three chapters of the book.

Too many people have contributed to the making of this book through the many revisions it has undergone to mention all their names. I am especially grateful to Trevor Breusch, Hidehiko Ichimura, Tom MaCurdy, Jim Powell, and Gene Savin for giving me valuable comments on the entire manuscript. I am also indebted to Carl Christ, Art Goldberger, Cheng Hsiao, Roger Koenker, Tony Lancaster, Chuck Manski, and Hal White for their valuable comments on parts of the manuscript. I am grateful to Colin Cameron, Tom Downes, Harry Paarsch, Aaron Han, and Choon Moon for proofreading and to the first three for correcting my English. In addition, Tom Downes and Choon Moon helped me with the preparation of the index. Dzung Pham has typed most of the manuscript through several revisions; her unfailing patience and good nature despite many hours of overtime work are much appreciated. David Criswell, Cathy Shimizu, and Bach-Hong Tran have also helped with the typing. The financial support of the National Science Foundation for the research that produced many of the results presented in the book is gratefully acknowledged. Finally, I am indebted to the editors of the *Journal of Economic Literature* for permission to include in Chapter 9 parts of my article entitled "Qualitative Response Models: A Survey" (*Journal of Economic Literature* 19:1483–1536, 1981) and to North-Holland Publishing Company for permission to use in Chapter 10 the revised version of my article entitled "Tobit Models: A Survey" (*Journal of Econometrics* 24:3–61, 1984).

Advanced Econometrics

Contents

1 Classical Least Squares Theory 1

2 Recent Developments in Regression Analysis 45

3 Large Sample Theory 81

4 Asymptotic Properties of Extremum Estimators 105

5 Time Series Analysis 159

6 Generalized Least Squares Theory 181

7 Linear Simultaneous Equations Models 228

8 Nonlinear Simultaneous Equations Models 245

9 Qualitative Response Models 267

10 Tobit Models 360

11 Markov Chain and Duration Models 412

Appendix 1 Useful Theorems in Matrix Analysis 459

Appendix 2 Distribution Theory 463

Notes 465

References 475

Name Index 505

Subject Index 511

1 Classical Least Squares Theory

In this chapter we shall consider the basic results of statistical inference in the classical linear regression model—the model in which the regressors are independent of the error term and the error term is serially uncorrelated and has a constant variance. This model is the starting point of the study; the models to be examined in later chapters are modifications of this one.

1.1 Linear Regression Model

In this section let us look at the reasons for studying the linear regression model and the method of specifying it. We shall start by defining Model 1, to be considered throughout the chapter.

1.1.1 Introduction

Consider a sequence of K random variables $(y_t, x_{2t}, x_{3t}, \ldots, x_{Kt})$, $t = 1, 2, \ldots, T$. Define a T-vector $\mathbf{y} = (y_1, y_2, \ldots, y_T)'$, a $(K-1)$-vector $\mathbf{x}_t^* = (x_{2t}, x_{3t}, \ldots, x_{Kt})'$, and a $[(K-1)\times T]$-vector $\mathbf{x}^* = (\mathbf{x}_1^{*\prime}, \mathbf{x}_2^{*\prime}, \ldots, \mathbf{x}_T^{*\prime})'$. Suppose for the sake of exposition that the joint density of the variables is given by $f(\mathbf{y}, \mathbf{x}^*, \theta)$, where θ is a vector of unknown parameters. We are concerned with inference about the parameter vector θ on the basis of the observed vectors $\mathbf{y}$ and $\mathbf{x}^*$.

In econometrics we are often interested in the conditional distribution of one set of random variables given another set of random variables; for example, the conditional distribution of consumption given income and the conditional distribution of quantities demanded given prices. Suppose we want to know the conditional distribution of $\mathbf{y}$ given $\mathbf{x}^*$. We can write the joint density as the product of the conditional density and the marginal density as in

$$f(\mathbf{y}, \mathbf{x}^*, \theta) = f(\mathbf{y}|\mathbf{x}^*, \theta_1) f(\mathbf{x}^*, \theta_2). \tag{1.1.1}$$

Regression analysis can be defined as statistical inferences on θ_1. For this purpose we can ignore $f(\mathbf{x}^*, \theta_2)$, provided there is no relationship between θ_1

and θ_2. The vector $\mathbf{y}$ is called the vector of *dependent* or *endogenous* variables, and the vector $\mathbf{x}^*$ is called the vector of *independent* or *exogenous* variables.

In regression analysis we usually want to estimate only the first and second moments of the conditional distribution, rather than the whole parameter vector θ_1. (In certain special cases the first two moments characterize θ_1 completely.) Thus we can define regression analysis as statistical inference on the conditional mean $E(\mathbf{y}|\mathbf{x}^*)$ and the conditional variance-covariance matrix $V(\mathbf{y}|\mathbf{x}^*)$. Generally, these moments are nonlinear functions of $\mathbf{x}^*$. However, in the present chapter we shall consider the special case in which $E(y_t|\mathbf{x}^*)$ is equal to $E(y_t|\mathbf{x}_t^*)$ and is a linear function of $\mathbf{x}_t^*$, and $V(\mathbf{y}|\mathbf{x}^*)$ is a constant times an identity matrix. Such a model is called the *classical* (or *standard*) linear regression model or the *homoscedastic* (meaning constant variance) linear regression model. Because this is the model to be studied in Chapter 1, let us call it simply Model 1.

1.1.2 Model 1

By writing $\mathbf{x}_t = (1, \mathbf{x}_t^{*\prime})'$, we can define Model 1 as follows. Assume

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + u_t, \qquad t = 1, 2, \ldots, T, \tag{1.1.2}$$

where y_t is a scalar observable random variable, $\boldsymbol{\beta}$ is a K-vector of unknown parameters, $\mathbf{x}_t$ is a K-vector of known constants such that $\Sigma_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'$ is nonsingular, and u_t is a scalar, unobservable, random variable (called the error term or the disturbance) such that $Eu_t = 0$, $Vu_t = \sigma^2$ (another unknown parameter) for all t, and $Eu_tu_s = 0$ for $t \neq s$.

Note that we have assumed $\mathbf{x}^*$ to be a vector of known constants. This is essentially equivalent to stating that we are concerned only with estimating the conditional distribution of $\mathbf{y}$ given $\mathbf{x}^*$. The most important assumption of Model 1 is the linearity of $E(y_t|\mathbf{x}_t^*)$; we therefore shall devote the next subsection to a discussion of the implications of that assumption. We have also made the assumption of homoscedasticity ($Vu_t = \sigma^2$ for all t) and the assumption of no serial correlation ($Eu_tu_s = 0$ for $t \neq s$), not because we believe that they are satisfied in most applications, but because they make a convenient starting point. These assumptions will be removed in later chapters.

We shall sometimes impose additional assumptions on Model 1 to obtain certain specific results. Notably, we shall occasionally make the assumption of serial independence of $\{u_t\}$ or the assumption that u_t is normally distributed. In general, independence is a stronger assumption than no correlation, al-

though under normality the two concepts are equivalent. The additional assumptions will be stated whenever they are introduced into Model 1.

1.1.3 Implications of Linearity

Suppose random variables y_t and $\mathbf{x}_t^*$ have finite second moments and their variance-covariance matrix is denoted by

$$V\begin{bmatrix} y_t \\ \mathbf{x}_t^* \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \boldsymbol{\sigma}_{12}' \\ \boldsymbol{\sigma}_{12} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Then we can always write

$$y_t = \beta_0 + \mathbf{x}_t^{*\prime}\boldsymbol{\beta}_1 + v_t, \tag{1.1.3}$$

where $\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{12}$, $\beta_0 = Ey_t - \boldsymbol{\sigma}_{12}'\boldsymbol{\Sigma}_{22}^{-1}E\mathbf{x}_t^*$, $Ev_t = 0$, $Vv_t = \sigma_1^2 - \boldsymbol{\sigma}_{12}'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{12}$, and $E\mathbf{x}_t^* v_t = \mathbf{0}$. It is important to realize that Model 1 implies certain assumptions that (1.1.3) does not: (1.1.3) does not generally imply linearity of $E(y_t|\mathbf{x}_t^*)$ because $E(v_t|\mathbf{x}_t^*)$ may not generally be zero.

We call $\beta_0 + \mathbf{x}_t^{*\prime}\boldsymbol{\beta}_1$ in (1.1.3) the *best linear predictor* of y_t given $\mathbf{x}_t^*$ because β_0 and $\boldsymbol{\beta}_1$ can be shown to be the values of b_0 and $\mathbf{b}_1$ that minimize $E(y_t - b_0 - \mathbf{x}_t^{*\prime}\mathbf{b}_1)^2$. In contrast, the conditional mean $E(y_t|\mathbf{x}_t^*)$ is called the *best predictor* of y_t given $\mathbf{x}_t^*$ because $E[y_t - E(y_t|\mathbf{x}_t^*)]^2 \leqq E[y_t - g(\mathbf{x}_t^*)]^2$ for any function g.

The reader might ask why we work with eq. (1.1.2) rather than with (1.1.3). The answer is that (1.1.3) is so general that it does not allow us to obtain interesting results. For example, whereas the natural estimators of β_0 and $\boldsymbol{\beta}_1$ can be defined by replacing the moments of y_t and $\mathbf{x}_t^*$ that characterize β_0 and $\boldsymbol{\beta}_1$ with their corresponding sample moments (they actually coincide with the least squares estimator), the mean of the estimator cannot be evaluated without specifying more about the relationship between $\mathbf{x}_t^*$ and v_t.

How restrictive is the linearity of $E(y_t|\mathbf{x}_t^*)$? It holds if y_t and $\mathbf{x}_t^*$ are jointly normal or if y_t and $\mathbf{x}_t^*$ are both scalar dichotomous (Bernoulli) variables.[1] But the linearity may not hold for many interesting distributions. Nevertheless, the linear assumption is not as restrictive as it may appear at first glance because $\mathbf{x}_t^*$ can be variables obtained by transforming the original independent variables in various ways. For example, if the conditional mean of y_t, the supply of good, is a quadratic function of the price, p_t, we can put $\mathbf{x}_t^* = (p_t, p_t^2)'$, thereby making $E(y_t|\mathbf{x}_t^*)$ linear.

1.1.4 Matrix Notation

To facilitate the subsequent analysis, we shall write (1.1.2) in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{1.1.4}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_T)'$, $\mathbf{u} = (u_1, u_2, \ldots, u_T)'$, and $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)'$. In other words, $\mathbf{X}$ is the $T \times K$ matrix, the tth row of which is $\mathbf{x}_t'$. The elements of the matrix $\mathbf{X}$ are described as

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{T1} & x_{T2} & \cdots & x_{TK} \end{bmatrix}.$$

If we want to focus on the columns of $\mathbf{X}$, we can write $\mathbf{X} = [\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(K)}]$, where each $\mathbf{x}_{(i)}$ is a T-vector. If there is no danger of confusing $\mathbf{x}_{(i)}$ with $\mathbf{x}_t$, we can drop the parentheses and write simply $\mathbf{x}_i$. In matrix notation the assumptions on $\mathbf{X}$ and $\mathbf{u}$ can be stated as follows: $\mathbf{X}'\mathbf{X}$ is nonsingular, which is equivalent to stating rank $(\mathbf{X}) = K$ if $T \geqq K$; $E\mathbf{u} = \mathbf{0}$; and $E\mathbf{u}\mathbf{u}' = \sigma^2\mathbf{I}_T$, where $\mathbf{I}_T$ is the $T \times T$ identity matrix. (Whenever the size of an identity matrix can be inferred from the context, we write it simply as $\mathbf{I}$.)

In the remainder of this chapter we shall no longer use the partition $\boldsymbol{\beta}' = (\beta_0, \boldsymbol{\beta}_1')$; instead, the elements of $\boldsymbol{\beta}$ will be written as $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)'$. Similarly, we shall not necessarily assume that $\mathbf{x}_{(1)}$ is the vector of ones, although in practice this is usually the case. Most of our results will be obtained simply on the assumption that $\mathbf{X}$ is a matrix of constants, without specifying specific values.

1.2 Theory of Least Squares

In this section we shall define the least squares estimator of the parameter $\boldsymbol{\beta}$ in Model 1 and shall show that it is the best linear unbiased estimator. We shall also discuss estimation of the error variance σ^2.

1.2.1 Definition of Least Squares Estimators of $\boldsymbol{\beta}$ and σ^2

The least squares (LS) estimator $\hat{\boldsymbol{\beta}}$ of the regression parameter $\boldsymbol{\beta}$ in Model 1 is defined to be the value of $\boldsymbol{\beta}$ that minimizes the sum of squared residuals[2]

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{1.2.1}$$
$$= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.$$

Putting the derivatives of $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ equal to 0, we have

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \tag{1.2.2}$$

where $\partial S/\partial \boldsymbol{\beta}$ denotes the K-vector the ith element of which is $\partial S/\partial \beta_i$, β_i being the ith element of $\boldsymbol{\beta}$. Solving (1.2.2) for $\boldsymbol{\beta}$ gives

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{1.2.3}$$

Clearly, $S(\boldsymbol{\beta})$ attains the global minimum at $\hat{\boldsymbol{\beta}}$.

Let us consider the special case $K = 2$ and $\mathbf{x}_t' = (1, x_{2t})$ and represent each of the T-observations (y_t, x_{2t}) by a point on the plane. Then, geometrically, the least squares estimates are the intercept and the slope of a line drawn in such a way that the sum of squares of the deviations between the points and the line is minimized in the direction of the y-axis. Different estimates result if the sum of squares of deviations is minimized in any other direction.

Given the least squares estimator $\hat{\boldsymbol{\beta}}$, we define

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{1.2.4}$$

and call it the vector of the *least squares residuals*. Using $\hat{\mathbf{u}}$, we can estimate σ^2 by

$$\hat{\sigma}^2 = T^{-1}\hat{\mathbf{u}}'\hat{\mathbf{u}}, \tag{1.2.5}$$

called the least squares estimator of σ^2, although the use of the term *least squares* here is not as compelling as in the estimation of the regression parameters.

Using (1.2.4), we can write

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}} = \mathbf{P}\mathbf{y} + \mathbf{M}\mathbf{y}, \tag{1.2.6}$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{M} = \mathbf{I} - \mathbf{P}$. Because $\hat{\mathbf{u}}$ is orthogonal to $\mathbf{X}$ (that is, $\hat{\mathbf{u}}'\mathbf{X} = \mathbf{0}$), least squares estimation can be regarded as decomposing $\mathbf{y}$ into two orthogonal components: a component that can be written as a linear combination of the column vectors of $\mathbf{X}$ and a component that is orthogonal to $\mathbf{X}$. Alternatively, we can call $\mathbf{P}\mathbf{y}$ the projection of $\mathbf{y}$ onto the space spanned by the column vectors of $\mathbf{X}$ and $\mathbf{M}\mathbf{y}$ the projection of $\mathbf{y}$ onto the space orthogonal to $\mathbf{X}$. Theorem 14 of Appendix 1 gives the properties of a projection matrix such as $\mathbf{P}$ or $\mathbf{M}$. In the special case where both $\mathbf{y}$ and $\mathbf{X}$ are two-dimensional vectors

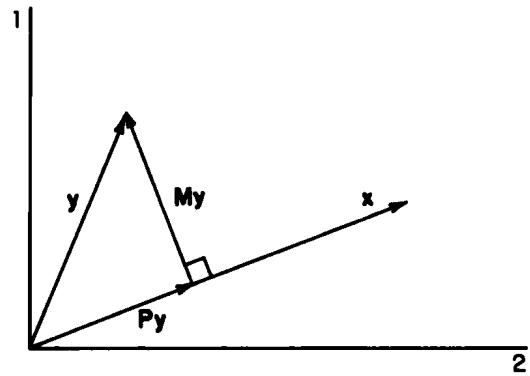

Figure 1.1 Orthogonal decomposition of y

(that is, $K = 1$ and $T = 2$), the decomposition (1.2.6) can be illustrated as in Figure 1.1, where the vertical and horizontal axes represent the first and second observations, respectively, and the arrows represent vectors.

From (1.2.6) we obtain

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{P}\mathbf{y} + \mathbf{y}'\mathbf{M}\mathbf{y}. \tag{1.2.7}$$

The goodness of fit of the regression of y on X can be measured by the ratio $\mathbf{y}'\mathbf{P}\mathbf{y}/\mathbf{y}'\mathbf{y}$, sometimes called R^2. However, it is more common to define R^2 as the square of the sample correlation between y and Py:

$$R^2 = \frac{(\mathbf{y}'\mathbf{L}\mathbf{P}\mathbf{y})^2}{\mathbf{y}'\mathbf{L}\mathbf{y} \cdot \mathbf{y}'\mathbf{P}\mathbf{L}\mathbf{P}\mathbf{y}}, \tag{1.2.8}$$

where $\mathbf{L} = \mathbf{I}_T - T^{-1}\mathbf{l}\mathbf{l}'$ and l denotes the T-vector of ones. If we assume one of the columns of X is l (which is usually the case), we have $\mathbf{L}\mathbf{P} = \mathbf{P}\mathbf{L}$. Then we can rewrite (1.2.8) as

$$R^2 = \frac{\mathbf{y}'\mathbf{L}\mathbf{P}\mathbf{L}\mathbf{y}}{\mathbf{y}'\mathbf{L}\mathbf{y}} = 1 - \frac{\mathbf{y}'\mathbf{M}\mathbf{y}}{\mathbf{y}'\mathbf{L}\mathbf{y}}. \tag{1.2.9}$$

Thus R^2 can be interpreted as a measure of the goodness of fit of the regression of the deviations of y from its mean on the deviations of the columns of X from their means. (Section 2.1.4 gives a modification of R^2 suggested by Theil, 1961.)

1.2.2 Least Squares Estimator of a Subset of $\boldsymbol{\beta}$

It is sometimes useful to have an explicit formula for a subset of the least squares estimates $\hat{\boldsymbol{\beta}}$. Suppose we partition $\hat{\boldsymbol{\beta}}' = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2')$, where $\hat{\boldsymbol{\beta}}_1$ is a K_1-vec-

tor and $\hat{\boldsymbol{\beta}}_2$ is a K_2-vector such that $K_1 + K_2 = K$. Partition $\mathbf{X}$ conformably as $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$. Then we can write $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ as

$$\mathbf{X}_1'\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_1'\mathbf{y} \tag{1.2.10}$$

and

$$\mathbf{X}_2'\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{y}. \tag{1.2.11}$$

Solving (1.2.11) for $\hat{\boldsymbol{\beta}}_2$ and inserting it into (1.2.10), we obtain

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y}, \tag{1.2.12}$$

where $\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$. Similarly,

$$\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y}, \tag{1.2.13}$$

where $\mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$.

In Model 1 we assume that $\mathbf{X}$ is of full rank, an assumption that implies that the matrices to be inverted in (1.2.12) and (1.2.13) are both nonsingular. Suppose for a moment that $\mathbf{X}_1$ is of full rank but that $\mathbf{X}_2$ is not. In this case $\boldsymbol{\beta}_2$ cannot be estimated, but $\boldsymbol{\beta}_1$ still can be estimated by modifying (1.2.12) as

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{M}_2^*\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2^*\mathbf{y}, \tag{1.2.14}$$

where $\mathbf{M}_2^* = \mathbf{I} - \mathbf{X}_2^*(\mathbf{X}_2^{*\prime}\mathbf{X}_2^*)^{-1}\mathbf{X}_2^{*\prime}$, where the columns of $\mathbf{X}_2^*$ consist of a maximal number of linearly independent columns of $\mathbf{X}_2$, provided that $\mathbf{X}_1'\mathbf{M}_2^*\mathbf{X}_1$ is nonsingular. (For the more general problem of estimating a linear combination of the elements of $\boldsymbol{\beta}$, see Section 2.2.3.)

1.2.3 The Mean and Variance of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$

Inserting (1.1.4) into (1.2.3), we have

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}.\end{aligned} \tag{1.2.15}$$

Clearly, $E\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ by the assumptions of Model 1. Using the second line of (1.2.15), we can derive the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned}V\hat{\boldsymbol{\beta}} &\equiv E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \\ &= E(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned} \tag{1.2.16}$$

From (1.2.3) and (1.2.4), we have $\hat{\mathbf{u}} = \mathbf{M}\mathbf{u}$, where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Using the properties of the projection matrix given in Theorem 14 of Appendix 1, we obtain

$$\begin{aligned} E\hat{\sigma}^2 &= T^{-1}E\mathbf{u}'\mathbf{M}\mathbf{u} \qquad (1.2.17)\\ &= T^{-1}E \operatorname{tr} \mathbf{M}\mathbf{u}\mathbf{u}' \quad \text{by Theorem 6 of Appendix 1}\\ &= T^{-1}\sigma^2 \operatorname{tr} \mathbf{M}\\ &= T^{-1}(T-K)\sigma^2 \quad \text{by Theorems 7 and 14 of Appendix 1,} \end{aligned}$$

which shows that $\hat{\sigma}^2$ is a biased estimator of σ^2. We define the unbiased estimator of σ^2 by

$$\tilde{\sigma}^2 = (T-K)^{-1}\hat{\mathbf{u}}'\hat{\mathbf{u}}. \qquad (1.2.18)$$

We shall obtain the variance of $\hat{\sigma}^2$ later, in Section 1.3, under the additional assumption that $\mathbf{u}$ is normal.

The quantity $V\hat{\boldsymbol{\beta}}$ can be estimated by substituting either $\hat{\sigma}^2$ or $\tilde{\sigma}^2$ (defined above) for the σ^2 that appears in the right-hand side of (1.2.16).

1.2.4 Definition of Best

Before we prove that the least squares estimator is best linear unbiased, we must define the term *best*. First we shall define it for scalar estimators, then for vector estimators.

DEFINITION 1.2.1. Let $\hat{\theta}$ and θ^* be scalar estimators of a scalar parameter θ. The estimator $\hat{\theta}$ is said to be *at least as good as* (or *at least as efficient as*) the estimator θ^* if $E(\hat{\theta}-\theta)^2 \leqq E(\theta^*-\theta)^2$ for all parameter values. The estimator $\hat{\theta}$ is said to be *better* (or *more efficient*) than the estimator θ^* if $\hat{\theta}$ is at least as good as θ^* and $E(\hat{\theta}-\theta)^2 < E(\theta^*-\theta)^2$ for at least one parameter value. An estimator is said to be *best* (or *efficient*) in a class if it is better than any other estimator in the class.

The mean squared error is a reasonable criterion in many situations and is mathematically convenient. So, following the convention of the statistical literature, we have defined "better" to mean "having a smaller mean squared error." However, there may be situations in which a researcher wishes to use other criteria, such as the mean absolute error.

DEFINITION 1.2.2. Let $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ be estimators of a vector parameter $\boldsymbol{\theta}$. Let $\mathbf{A}$ and $\mathbf{B}$ be their respective mean squared error matrices; that is, $\mathbf{A} = E(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})'$ and $\mathbf{B} = E(\boldsymbol{\theta}^*-\boldsymbol{\theta})(\boldsymbol{\theta}^*-\boldsymbol{\theta})'$. Then we say $\hat{\boldsymbol{\theta}}$ is *better* (or

more efficient) than θ^* if

$$\mathbf{c}'(\mathbf{B} - \mathbf{A})\mathbf{c} \geqq 0 \quad \text{for every vector} \quad \mathbf{c} \quad \text{and every parameter value} \tag{1.2.19}$$

and

$$\mathbf{c}'(\mathbf{B} - \mathbf{A})\mathbf{c} > 0 \quad \text{for at least one value of} \quad \mathbf{c} \quad \text{and at least one value of the parameter.} \tag{1.2.20}$$

This definition of *better* clearly coincides with Definition 1.2.1 if θ is a scalar.

In view of Definition 1.2.1, equivalent forms of statements (1.2.19) and (1.2.20) are statements (1.2.21) and (1.2.22):

$$\mathbf{c}'\hat{\theta} \quad \text{is at least as good as} \quad \mathbf{c}'\theta^* \quad \text{for every vector} \quad \mathbf{c} \tag{1.2.21}$$

and

$$\mathbf{c}'\hat{\theta} \quad \text{is better than} \quad \mathbf{c}'\theta^* \quad \text{for at least one value of} \quad \mathbf{c}. \tag{1.2.22}$$

Using Theorem 4 of Appendix 1, they also can be written as

$$\mathbf{B} \geqq \mathbf{A} \quad \text{for every parameter value} \tag{1.2.23}$$

and

$$\mathbf{B} \neq \mathbf{A} \quad \text{for at least one parameter value.} \tag{1.2.24}$$

(Note that $\mathbf{B} \geqq \mathbf{A}$ means $\mathbf{B} - \mathbf{A}$ is nonnegative definite and $\mathbf{B} > \mathbf{A}$ means $\mathbf{B} - \mathbf{A}$ is positive definite.)

We shall now prove the equivalence of (1.2.20) and (1.2.24). Because the phrase "for at least one parameter value" is common to both statements, we shall ignore it in the following proof. First, suppose (1.2.24) is not true. Then $\mathbf{B} = \mathbf{A}$. Therefore $\mathbf{c}'(\mathbf{B} - \mathbf{A})\mathbf{c} = 0$ for every $\mathbf{c}$, a condition that implies that (1.2.20) is not true. Second, suppose (1.2.20) is not true. Then $\mathbf{c}'(\mathbf{B} - \mathbf{A})\mathbf{c} = 0$ for every $\mathbf{c}$ and every diagonal element of $\mathbf{B} - \mathbf{A}$ must be 0 (choose $\mathbf{c}$ to be the zero vector, except for 1 in the ith position). Also, the i,jth element of $\mathbf{B} - \mathbf{A}$ is 0 (choose $\mathbf{c}$ to be the zero vector, except for 1 in the ith and jth positions, and note that $\mathbf{B} - \mathbf{A}$ is symmetric). Thus (1.2.24) is not true. This completes the proof.

Note that replacing $\mathbf{B} \neq \mathbf{A}$ in (1.2.24) with $\mathbf{B} > \mathbf{A}$—or making the corresponding change in (1.2.20) or (1.2.22)— is unwise because we could not then rank the estimator with the mean squared error matrix

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

higher than the estimator with the mean squared error matrix

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$

A problem with Definition 1.2.2 (more precisely, a problem inherent in the comparison of vector estimates rather than in this definition) is that often it does not allow us to say one estimator is either better or worse than the other. For example, consider

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}. \tag{1.2.25}$$

Clearly, neither $\mathbf{A} \geqq \mathbf{B}$ nor $\mathbf{B} \geqq \mathbf{A}$. In such a case one might compare the trace and conclude that $\hat{\theta}$ is better than θ^* because $\operatorname{tr} \mathbf{A} < \operatorname{tr} \mathbf{B}$. Another example is

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}. \tag{1.2.26}$$

Again, neither $\mathbf{A} \geqq \mathbf{B}$ nor $\mathbf{B} \geqq \mathbf{A}$. If one were using the determinant as the criterion, one would prefer $\hat{\theta}$ over θ^* because $\det \mathbf{A} < \det \mathbf{B}$.

Note that $\mathbf{B} \geqq \mathbf{A}$ implies both $\operatorname{tr} \mathbf{B} \geqq \operatorname{tr} \mathbf{A}$ and $\det \mathbf{B} \geqq \det \mathbf{A}$. The first follows from Theorem 7 and the second from Theorem 11 of Appendix 1. As these two examples show, neither $\operatorname{tr} \mathbf{B} \geqq \operatorname{tr} \mathbf{A}$ nor $\det \mathbf{B} \geqq \det \mathbf{A}$ implies $\mathbf{B} \geqq \mathbf{A}$.

Use of the trace as a criterion has an obvious intuitive appeal, inasmuch as it is the sum of the individual variances. Justification for the use of the determinant involves more complicated reasoning. Suppose $\hat{\theta} \sim N(\theta, \mathbf{V})$, where $\mathbf{V}$ is the variance-covariance matrix of $\hat{\theta}$. Then, by Theorem 1 of Appendix 2, $(\hat{\theta} - \theta)'\mathbf{V}^{-1}(\hat{\theta} - \theta) \sim \chi^2_K$, the chi-square distribution with K degrees of freedom, K being the number of elements of θ. Therefore the $(1 - \alpha)$% confidence ellipsoid for θ is defined by

$$\{\theta | (\hat{\theta} - \theta)'\mathbf{V}^{-1}(\hat{\theta} - \theta) < \chi^2_K(\alpha)\}, \tag{1.2.27}$$

where $\chi^2_K(\alpha)$ is the number such that $P[\chi^2_K \geqq \chi^2_K(\alpha)] = \alpha$. Then the volume of the ellipsoid (1.2.27) is proportional to the determinant of $\mathbf{V}$, as shown by Anderson (1958, p. 170).

A more intuitive justification for the determinant criterion is possible for the case in which θ is a two-dimensional vector. Let the mean squared error matrix of an estimator $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)'$ be