1 Classical Statistical Procedures

1.1. INTRODUCTION

An alternative title to this book could have been *The Application of Classical Statistical Procedures to Econometrics* or something along these lines. What it purports to do is provide the reader with mathematical tools that facilitate the application of classical statistical procedures to the complicated statistical models that we are confronted with in econometrics. It then demonstrates how these procedures can be applied to a sequence of linear econometric models, each model being more complicated statistically than the previous one. The statistical procedures I have in mind are these centered around the likelihood function: procedures that involve the score vector, the information matrix, and the Cramer–Rao lower bound, together with maximum-likelihood estimation and classical test statistics.

Until recently, such procedures were little used by econometricians. The likelihood function in most econometric models is complicated, and the first-order conditions for maximizing this function usually give rise to a system of nonlinear equations that is not easily solved. As a result, econometricians developed their own class of estimators, instrumental variable estimators, that had the same asymptotic properties as those of maximum-likelihood estimators (MLEs) but were far more tractable mathematically [see Bowden and Turkington (1990)]. Nor did econometricians make much use of the prescribed classical statistical procedures for obtaining test statistics for the hypotheses of interest in econometric models; rather, test statistics were developed on an ad hoc basis.

All that changed in the last couple of decades, when there was renewed interest by econometricians in maximum-likelihood procedures and in developing Lagrangian multiplier test (LMT) statistics. One reason for this change was the advent of large, fast computers. A complicated system of nonlinear equations could now be solved so we would have in hand the maximum-likelihood estimates even though we had no algebraic expression for the underlying estimators. Another more recent explanation for this change in attitude is the

advent of results on zero-one matrices and matrix calculus. Works by Graham (1981), Magnus (1988), Magnus and Neudecker (1988), and Lutkepohl (1996) have shown us the importance of zero-one matrices, their connection to matrix calculus, and the power of matrix calculus particularly with respect to applying classical statistical procedures.

In this introductory chapter, I have a brief and nonrigorous summary of the classical statistical procedures that are used extensively in the latter part of this book.

1.2. THE SCORE VECTOR, THE INFORMATION MATRIX, AND THE CRAMER–RAO LOWER BOUND

Let θ be a $k \times 1$ vector of unknown parameters associated with a statistical model and let $l(\theta)$ be the log-likelihood function that satisfies certain regularity conditions and is twice differentiable. Let $\partial l/\partial\theta$ denote the $k \times 1$ vector of partial derivatives of l. Then $\partial l/\partial\theta$ is called the **score vector**. Let $\partial^2 l/\partial\theta\partial\theta'$ denote the $k \times k$ Hessian matrix of $l(\theta)$. Then the (asymptotic) **information matrix** is defined as

$$I(\theta) = -\lim_{n\to\infty} \frac{1}{n} E(\partial^2 l/\partial\theta\partial\theta'),$$

where n denotes the sample size. Now the limit of the expectation need not be the same as the probability limit. However, for the models we consider in this book, based as they are on the multivariate normal distribution, the two concepts will be the same. As a result it is often more convenient to regard the information matrix as

$$I(\theta) = -p\lim \frac{1}{n} \partial^2 l/\partial\theta\partial\theta'.$$

The inverse of this matrix, $I^{-1}(\theta)$, is called the (asymptotic) **Cramer–Rao** lower bound. Let $\hat{\theta}$ be a consistent estimator of θ such that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V).$$

The matrix V is called the asymptotic covariance matrix of $\hat{\theta}$. Then V exceeds the Cramer–Rao lower bound $I^{-1}(\theta)$ in the sense that $V - I^{-1}(\theta)$ is a positive-semidefinite matrix. If $V = I^{-1}(\theta)$, then $\hat{\theta}$ is called a best asymptotically normally distributed estimator (which is shortened to BAN estimator).

1.3. MAXIMUM LIKELIHOOD ESTIMATORS AND TEST PROCEDURES

Classical statisticians prescribed a procedure for obtaining a BAN estimator, namely the maximum-likelihood procedure. Let $\oplus$ denote the parameter space. Then any value of θ that maximizes $l(\theta)$ over $\oplus$ is called a maximum-likelihood

estimate, and the underlying estimator is called the MLE. The first-order conditions for this maximization are given by

$$\frac{\partial l(\theta)}{\partial \theta} = 0.$$

Let $\tilde{\theta}$ denote the MLE of θ. Then $\tilde{\theta}$ is consistent, and $\tilde{\theta}$ is the BAN estimator so

$$\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} N[0, I^{-1}(\theta)].$$

Let h be a $G \times 1$ vector whose elements are functions of the elements of θ. We denote this by $h(\theta)$. Suppose we are interested in developing test statistics for the null hypothesis

$$H_O : h(\theta) = 0$$

against the alternative

$$H_A : h(\theta) \neq 0.$$

Let $\tilde{\theta}$ denote the MLE of θ and $\bar{\theta}$ denote the constrained MLE of θ; that is, $\bar{\theta}$ is the MLE of θ we obtain after we impose H_O on our statistical model. Let $\partial h(\theta)/\partial \theta$ denote the $k \times G$ matrix whose (ij) element is $\partial h_j/\partial \theta_i$. Then classical statisticians prescribed three competing procedures for obtaining a test statistic for H_O. These are as follows.

LAGRANGIAN MULTIPLIER TEST STATISTIC

$$T_1 = \frac{1}{n}\frac{\partial l(\bar{\theta})'}{\partial \theta} I^{-1}(\bar{\theta}) \frac{\partial l(\bar{\theta})}{\partial \theta}.$$

Note that the LMT statistic uses the constrained MLE of θ. If H_O is true, $\bar{\theta}$ should be close to $\tilde{\theta}$ and as, by the first-order conditions, $\partial l(\tilde{\theta})/\partial \theta = 0$, the derivative $\partial l(\theta)/\partial \theta$ evaluated at $\bar{\theta}$ should also be close to the null vector. The test statistic is a measure of the distance $\partial l(\bar{\theta})/\partial \theta$ is from the null vector.

WALD TEST STATISTIC

$$T_2 = nh(\tilde{\theta})' \left[\frac{\partial h(\tilde{\theta})'}{\partial \theta} I^{-1}(\tilde{\theta}) \frac{\partial h(\tilde{\theta})}{\partial \theta} \right]^{-1} h(\tilde{\theta}).$$

Note that the Wald test statistic uses the (unconstrained) MLE of θ. Essentially it is based on the asymptotic distribution of $\sqrt{n}h(\tilde{\theta})$ under H_O, the statistic itself measuring the distance $h(\tilde{\theta})$ is from the null vector.

LIKELIHOOD RATIO TEST STATISTIC

$$T_3 = 2[l(\tilde{\theta}) - l(\bar{\theta})].$$

Note that the likelihood ratio test (LRT) statistic uses both the unconstrained MLE $\tilde{\theta}$ and the constrained MLE $\bar{\theta}$. If H_O is indeed true, it should not matter whether we impose it or not, so $l(\tilde{\theta})$ should be approximately the same as $l(\bar{\theta})$. The test statistic T_3 measures the difference between $l(\tilde{\theta})$ and $l(\bar{\theta})$.

All three test statistics are asymptotically equivalent in the sense that, under H_O, they all have the same limiting χ^2 distribution and under H_A, with local alternatives, they have the same limiting noncentral χ^2 distribution. Usually imposing the null hypothesis on our model leads to a simpler statistical model, and thus the constrained MLEs $\bar{\theta}$ are more obtainable than the $\tilde{\theta}$ MLEs. For this reason the LMT statistic is often the easiest statistic to form. Certainly it is the one that has been most widely used in econometrics.

1.4. NUISANCE PARAMETERS

Let us now partition θ into $\theta = (\alpha'\beta')'$, where α is a $k_1 \times 1$ vector of parameters of primary interest and β is a $k_2 \times 1$ vector of nuisance parameters, $k_1 + k_2 = k$. The terms used here do not imply that the parameters in β are unimportant to our statistical model. Rather, they indicate that the purpose of our analysis is to make statistical inference about the parameters in α instead of those in β.

In this situation, two approaches can be taken. First, we can derive the information matrix $I(\theta)$ and the Cramer–Rao lower bound $I^{-1}(\theta)$. Let

$$I(\theta) = \begin{bmatrix} I_{\alpha\alpha} & I_{\alpha\beta} \\ I_{\beta\alpha} & I_{\beta\beta} \end{bmatrix},$$

$$I^{-1}(\theta) = \begin{pmatrix} I^{\alpha\alpha} & I^{\alpha\beta} \\ I^{\beta\alpha} & I^{\beta\beta} \end{pmatrix}$$

be these matrices partitioned according to our partition of θ. As far as α is concerned we can now work with $I_{\alpha\alpha}$ and $I^{\alpha\alpha}$ in place of $I(\theta)$ and $I^{-1}(\theta)$, respectively. For example, $I^{\alpha\alpha}$ is the Cramer–Rao lower bound for the asymptotic covariance matrix of a consistent estimator of α. If $\tilde{\alpha}$ is the MLE of α, then

$$\sqrt{n}(\tilde{\alpha} - \alpha) \xrightarrow{d} N(0, I^{\alpha\alpha}),$$

and so on.

A particular null hypothesis that has particular relevance for us is

$$H_O : \alpha = 0$$

against

$$H_A : \alpha \neq 0.$$

Under this first approach, the classical test statistics for this null hypothesis would be the following test statistics.

LAGRANGIAN TEST STATISTIC

$$T_1 = \frac{1}{n} \frac{\partial l(\bar{\theta})'}{\partial \alpha} I^{\alpha\alpha}(\bar{\theta}) \frac{\partial l(\bar{\theta})}{\partial \alpha}.$$

WALD TEST STATISTIC

$$T_2 = n\tilde{\alpha}' I^{\alpha\alpha}(\tilde{\theta})^{-1}\tilde{\alpha}.$$

LIKELIHOOD RATIO TEST STATISTIC

$$T_3 = 2[l(\tilde{\theta}) - l(\bar{\theta})].$$

Under H_O all three test statistics would have a limiting χ^2 distribution with k_1 degrees of freedom, and the nature of the tests insists that we use the upper tail of this distribution to find the appropriate critical region.

The second approach is to work with the concentrated log-likelihood function. Here we undertake a stepwise maximization of the log-likelihood function. We first maximize $l(\theta)$ with respect to the nuisance parameters β to obtain $\bar{\beta} = \bar{\beta}(\alpha)$, say. The vector $\bar{\beta}$ is then placed back in the log-likelihood function to obtain

$$\bar{l}(\alpha) = l[\alpha, \bar{\beta}(\alpha)].$$

The function $\bar{l}(\alpha)$ is called the concentrated likelihood function. Our analysis can now be reworked with $\bar{l}(\alpha)$ in place of $l(\theta)$.

For example, let

$$\bar{I} = -p\lim\frac{1}{n}\frac{\partial \bar{l}}{\partial\alpha\,\partial\alpha'},$$

and let $\hat{\alpha}$ be any consistent estimator of α such that

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} N(0,\ V_\alpha).$$

Then $V_\alpha \geq \bar{I}^{-1}$ in the sense that their difference is a positive-semidefinite matrix. If $\tilde{\alpha}$ is the MLE of α, then $\tilde{\alpha}$ is obtained from

$$\frac{\partial \bar{l}}{\partial\alpha} = 0,$$

$$\sqrt{n}(\tilde{\alpha} - \alpha) \xrightarrow{d} N(0,\ \bar{I}^{-1}),$$

and so on. As far as test procedures go for the null hypothesis $H_O : h(\alpha) = 0$, under this second approach we rewrite the test statistics by using $\bar{I}$ and $\bar{I}$ in place of $I(\theta)$ and $I(\theta)$, respectively. In this book, I largely use the first approach as one of my expressed aims is to achieve the complete information matrix $I(\theta)$ for a sequence of econometric models.

1.5. DIFFERENTIATION AND ASYMPTOTICS

Before we leave this brief chapter, note that classical statistical procedures involve us in much differentiation. The score vector $\partial l/\partial\theta$, the Hessian matrix $\partial^2 l/\partial\theta\partial\theta'$, and $\partial h/\partial\theta$ all involve working out partial derivatives. It is at this stage that difficulties can arise in applying these procedures to econometric

models. As hinted at in Section 1.2, the log-likelihood function $l(\theta)$ for most econometric models is a complicated function, and it is no trivial matter to obtain the derivatives required in our application. Usually it is too great a task for ordinary calculus. Although in some cases it can be done, [see, for example, Rothenberg and Leenders (1964)], what often happens when one attempts to do the differentiation by using ordinary calculus is that one is confronted with a hopeless mess. It is precisely this problem that has motivated the writing of this book. I hope that it will go some way toward alleviating it.

It is assumed that the reader is familiar with standard asymptotic theory. Every attempt has been made to make the rather dull but necessary asymptotic analysis in this book as readable as possible. Only the probability limits of the information matrices that are required in our statistical analysis are worked out in full. The probability limits themselves are assumed to exist – a more formal mathematical analysis would give a list of sufficient conditions needed to ensure this. Finally, as already noted, use is made of the shortcut notation

$$\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, V)$$

rather than the more formally correct notation

$$\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} x \sim N(0, V).$$

2 Elements of Matrix Algebra

2.1. INTRODUCTION

In this chapter, we consider matrix operators that are used throughout the book and special square matrices, namely triangular matrices and band matrices, that will crop up continually in our future work. From the elements of an $m \times n$ matrix, $A = (a_{ij})$ and a $p \times q$ matrix, $B = (b_{ij})$, the Kronecker product forms an $mp \times nq$ matrix. The vec operator forms a column vector out of a given matrix by stacking its columns one underneath the other. The devec operator forms a row vector out of a given matrix by stacking its rows one alongside the other. In like manner, a generalized vec operator forms a new matrix from a given matrix by stacking a certain number of its columns under each other and a generalized devec operator forms a new matrix by stacking a certain number of rows alongside each other. It is well known that the Kronecker product is intimately connected with the vec operator, but we shall see that this connection also holds for the devec and generalized operators as well. Finally we look at special square matrices with zeros above or below the main diagonal or whose nonzero elements form a band surrounded by zeros. The approach I have taken in this chapter, as indeed in several other chapters, is to list, without proof, well-known properties of the mathematical concept, in hand. If, however, I want to present a property in a different light or if I have something new to say about the concept, then I will give a proof.

2.2. KRONECKER PRODUCTS

Let $A = (a_{ij})$ be an $m \times n$ matrix and B a $p \times q$ matrix. The $mp \times nq$ matrix given by

$$\begin{bmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{bmatrix}$$

is called the **Kronecker product** of A and B, denoted by $A \otimes B$. The following useful properties concerning Kronecker products are well known:

$$A \otimes (B \otimes C) = (A \otimes B) \otimes C = A \otimes B \otimes C,$$

$$(A + B) \otimes (C + D) = A \otimes C + A \otimes D + B \otimes C + B \otimes D, \quad \text{if } A + B \text{ and } C + D \text{ exist,}$$

$$(A \otimes B)(C \otimes D) = AC \otimes BD, \quad \text{if } AC \text{ and } BD \text{ exist.}$$

The transpose of a Kronecker product is

$$(A \otimes B)' = A' \otimes B',$$

whereas the rank of a Kronecker product is

$$r(A \otimes B) = r(A)\, r(B).$$

If A is a square $n \times n$ matrix and B is a square $p \times p$ matrix, then the trace of the Kronecker product is

$$\text{tr}(A \otimes B) = \text{tr } A \text{ tr } B,$$

whereas the determinant of the Kronecker product is

$$|A \otimes B| = |A|^p |B|^n,$$

and if A and B are nonsingular, the inverse of the Kronecker product is

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}.$$

Other properties of Kronecker products, although perhaps less well known, are nevertheless useful and are used throughout this book. First note that, in general, Kronecker products do not obey the commutative law, so $A \otimes B \neq B \otimes A$. One exception to this rule is if a and b are two column vectors, not necessarily of the same order; then

$$a' \otimes b = b \otimes a' = ba'. \tag{2.1}$$

This exception allows us to write $A \otimes b$ in an interesting way, where A is an $m \times n$ matrix and b is a $p \times 1$ vector. Partitioning A into its rows, we write

$$A = \begin{pmatrix} a^{1'} \\ \vdots \\ a^{m'} \end{pmatrix},$$

where $a^{i'}$ is the ith row of A. Then clearly from our definition of Kronecker

product

$$A \otimes b = \begin{pmatrix} a^{1'} \otimes b \\ \vdots \\ a^{m'} \otimes b \end{pmatrix} = \begin{pmatrix} b \otimes a^{1'} \\ \vdots \\ b \otimes a^{m'} \end{pmatrix}, \tag{2.2}$$

where we achieve the last equality by using Eq. (2.1).

Second, it is clear from the definition of the Kronecker product that if A is partitioned into submatrices, say

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1K} \\ \vdots & & \vdots \\ A_{l1} & \cdots & A_{lK} \end{bmatrix}$$

then

$$A \otimes B = \begin{bmatrix} A_{11} \otimes B & \cdots & A_{1K} \otimes B \\ \vdots & & \vdots \\ A_{l1} \otimes B & \cdots & A_{lK} \otimes B \end{bmatrix}.$$

Suppose we now partition B into an arbitrary number of submatrices, say

$$B = \begin{bmatrix} B_{11} & \cdots & B_{1r} \\ \vdots & & \vdots \\ B_{s1} & \cdots & B_{sr} \end{bmatrix}.$$

Then, in general,

$$A \otimes B \neq \begin{bmatrix} A \otimes B_{11} & \ldots & A \otimes B_{1r} \\ \vdots & & \vdots \\ A \otimes B_{s1} & \ldots & A \otimes B_{sr} \end{bmatrix}.$$

One exception to this rule is given by the following theorem.

Theorem 2.1. Let a be an $m \times 1$ vector and B be a $p \times q$ matrix. Write $B = (B_1 \cdots B_r)$, where each submatrix of B has p rows. Then

$$a \otimes B = (a \otimes B_1 \cdots a \otimes B_r).$$

Proof of Theorem 2.1. Clearly

$$a \otimes B = \begin{pmatrix} a_1 B \\ \vdots \\ a_m B \end{pmatrix} = \begin{bmatrix} a_1(B_1 & \cdots & B_r) \\ & \vdots & \\ a_m(B_1 & \cdots & B_r) \end{bmatrix} = \begin{bmatrix} a_1 B_1 & \cdots & a_1 B_r \\ \vdots & & \vdots \\ a_m B_1 & \cdots & a_m B_r \end{bmatrix}$$
$$= (a \otimes B_1 \cdots a \otimes B_r). \qquad \square$$

Now consider A as an $m \times n$ matrix partitioned into its columns $A = (a_1 \cdots a_n)$ and a partitioned matrix $B = (B_1 \cdots B_r)$. Then, by using Theorem 2.1, it is clear that we can write

$$A \otimes B = (a_1 \otimes B_1 \cdots a_1 \otimes B_r \cdots a_n \otimes B_1 \cdots a_n \otimes B_r).$$

This property of Kronecker products allows us to write $A \otimes B$ in a useful way. Partitioning A and B into their columns, we write

$$A = (a_1 \cdots a_n), \quad B = (b_1 \cdots b_q).$$

Then

$$A \otimes B = (a_1 \otimes b_1 \cdots a_1 \otimes b_q \cdots a_n \otimes b_1 \cdots a_n \otimes b_q).$$

Third, note that if A and B are $m \times n$ and $p \times q$ matrices, respectively, and x is any column vector, then

$$A(I_n \otimes x') = (A \otimes 1)(I_n \otimes x') = A \otimes x',$$
$$(x \otimes I_p)B = (x \otimes I_p)(1 \otimes B) = x \otimes B.$$

This property, coupled with the Kronecker product of $A \otimes B$, where A is partitioned, affords us another useful way of writing $A \otimes B$. Partitioning A into its columns, we obtain

$$A \otimes B = (a_1 \otimes B \cdots a_n \otimes B) = [(a_1 \otimes I_p)B \cdots (a_n \otimes I_p)B].$$

Finally, note that for A $m \times n$, and B $p \times q$

$$A \otimes B = (A \otimes I_p)(I_n \otimes B) = (I_m \otimes B)(A \otimes I_q).$$

2.3. THE VEC AND THE DEVEC OPERATORS

2.3.1. Basic Definitions

Let A be an $m \times n$ matrix and a_j be its jth column. Then **vec** A is the $mn \times 1$ vector

$$\text{vec } A = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix},$$