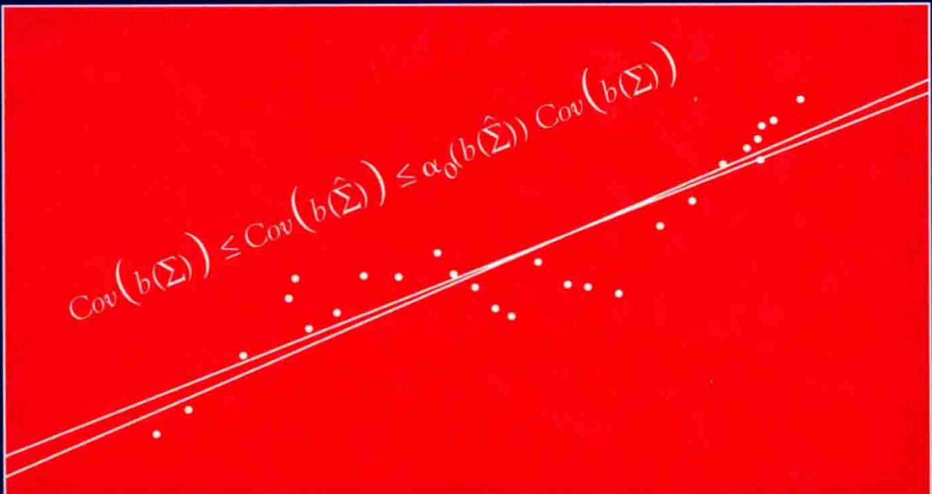
WILEY

Generalized Least Squares



Takeaki Kariya
Hiroshi Kurata

WILEY SERIES IN PROBABILITY AND STATISTICS

Generalized Least Squares

Takeaki Kariya
Kyoto University and Meiji University, Japan

Hiroshi Kurata
University of Tokyo, Japan

John Wiley & Sons, Ltd

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Kariya, Takeaki.
Generalized least squares / Takeaki Kariya, Hiroshi Kurata.
p. cm. – (Wiley series in probability and statistics)
Includes bibliographical references and index.
ISBN 0-470-86697-7 (alk. paper)
1. Least squares. I. Kurata, Hiroshi, 1967-II. Title. III. Series.
QA275.K32 2004
511′.42—dc22

2004047963

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-470-86697-7 (PPC)

Produced from LaTeX files supplied by the author and processed by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by TJ International, Padstow, Cornwall
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Generalized
Least Squares

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

To my late GLS co-worker Yasuyuki Toyooka and to my wife Shizuko
—Takeaki Kariya

To Akiko, Tomoatsu and the memory of my fathers
—Hiroshi Kurata

Preface

Regression analysis has been one of the most widely employed and most important statistical methods in applications and has been continually made more sophisticated from various points of view over the last four decades. Among a number of branches of regression analysis, the method of generalized least squares estimation based on the well-known Gauss–Markov theory has been a principal subject, and is still playing an essential role in many theoretical and practical aspects of statistical inference in a general linear regression model. A general linear regression model is typically of a certain covariance structure for the error term, and the examples are not only univariate linear regression models such as serial correlation models, heteroscedastic models and equi-correlated models but also multivariate models such as seemingly unrelated regression (SUR) models, multivariate analysis of variance (MANOVA) models, growth curve models, and so on.

When the problem of estimating the regression coefficients in such a model is considered and when the covariance matrix of the error term is known, as an efficient estimation procedure, we rely on the Gauss–Markov theorem that the Gauss–Markov estimator (GME) is the best linear unbiased estimator. In practice, however, the covariance matrix of the error term is usually unknown and hence the GME is not feasible. In such cases, a generalized least squares estimator (GLSE), which is defined as the GME with the unknown covariance matrix replaced by an appropriate estimator, is widely used owing to its theoretical and practical virtue.

This book attempts to provide a self-contained treatment of the unified theory of the GLSEs with a focus on their finite sample properties. We have made the content and exposition easy to understand for first-year graduate students in statistics, mathematics, econometrics, biometrics and other related fields. One of the key features of the book is a concise and mathematically rigorous description of the material via the *lower and upper bounds approach*, which enables us to evaluate the finite sample efficiency in a general manner.

In general, the efficiency of a GLSE is measured by relative magnitude of its risk (or covariance) matrix to that of the GME. However, since the GLSE is in general a nonlinear function of observations, it is often very difficult to evaluate the risk matrix in an explicit form. Besides, even if it is derived, it is often impractical to use such a result because of its complication. To overcome this difficulty, our book adopts as a main tool the lower and upper bounds approach, which approaches the problem by deriving a sharp lower bound and an effective upper bound for the risk matrix of a GLSE: for this purpose, we begin by showing that the risk matrix of a GLSE is bounded below by the covariance matrix of the GME (*Nonlinear Version of the Gauss–Markov Theorem*); on the basis of this result, we also derive an effective upper bound for the risk matrix of a GLSE relative to the covariance matrix of the GME (*Upper Bound Problems*). This approach has several important advantages: the upper bound provides information on the finite sample efficiency of a GLSE; it has a much simpler form than the risk matrix itself and hence serves as a tractable efficiency measure; furthermore, in some cases, we can obtain the optimal GLSE that has the minimum upper bound among an appropriate class of GLSEs. This book systematically develops the theory with various examples.

The book can be divided into three parts, corresponding respectively to Chapters 1 and 2, Chapters 3 to 6, and Chapters 7 to 9. The first part (Chapters 1 and 2) provides the basics for general linear regression models and GLSEs. In particular, we first give a fairly general definition of a GLSE, and establish its fundamental properties including conditions for unbiasedness and finiteness of second moments. The second part (Chapters 3–6), the main part of this book, is devoted to the detailed description of the lower and upper bounds approach stated above and its applications to serial correlation models, heteroscedastic models and SUR models. First, in Chapter 3, a nonlinear version of the Gauss–Markov theorem is established under fairly mild conditions on the distribution of the error term. Next, in Chapters 4 and 5, we derive several types of effective upper bounds for the risk matrix of a GLSE. Further, in Chapter 6, a uniform bound for the normal approximation to the distribution of a GLSE is obtained. The last part (Chapters 7–9) provides further developments (including mathematical extensions) of the results in the second part. Chapter 7 is devoted to making a further extension of the Gauss–Markov theorem, which is a maximal extension in a sense and leads to a further generalization of the nonlinear Gauss–Markov theorem proved in Chapter 3. In the last two chapters, some complementary topics are discussed. These include concentration inequalities, efficiency under elliptical symmetry, degeneracy of the distribution of a GLSE, and estimation of growth curves.

This book is not intended to be exhaustive, and there are many topics that are not even mentioned. Instead, we have done our best to give a systematic and unified presentation. We believe that reading this book leads to quite a solid understanding of this attractive subject, and hope that it will stimulate further research on the problems that remain.

The authors are indebted to many people who have helped us with this work. Among others, I, Takeaki Kariya, am first of all grateful to Professor Morris L. Eaton, who was my PhD thesis advisor and helped us get in touch with the publishers. I am also grateful to my late coauthor Yasuyuki Toyooka with whom

I published some important results contained in this book. Both of us are thankful to Dr. Hiroshi Tsuda and Professor Yoshihiro Usami for providing some tables and graphs and Ms Yuko Nakamura for arranging our writing procedure. We are also grateful to John Wiley & Sons for support throughout this project. Kariya's portion of this work was partially supported by the COE fund of Institute of Economic Research, Kyoto University.

Takeaki Kariya
Hiroshi Kurata

Contents

1

Preliminaries

1.1 Overview

This chapter deals with some basic notions that play indispensable roles in the theory of generalized least squares estimation and should be discussed in this preliminary chapter. Our selection here includes three basic notions: multivariate normal distribution, elliptically symmetric distributions and group invariance. First, in Section 1.2, some fundamental properties shared by the normal distributions are described without proofs. A brief treatment of Wishart distributions is also given. Next, in Section 1.3, we discuss the classes of spherically and elliptically symmetric distributions. These classes can be viewed as an extension of multivariate normal distribution and include various heavier-tailed distributions such as multivariate t and Cauchy distributions as special elements. Section 1.4 provides a minimum collection of notions on the theory of group invariance, which facilitates our unified treatment of generalized least squares estimators (GLSEs). In fact, the theory of spherically and elliptically symmetric distributions is principally based on the notion of group invariance. Moreover, as will be seen in the main body of this book, a GLSE itself possesses various group invariance properties.

1.2 Multivariate Normal and Wishart Distributions

This section provides without proofs some requisite distributional results on the multivariate normal and Wishart distributions.

Multivariate normal distribution. For an n-dimensional random vector y, let $\mathcal{L}(y)$ denote the distribution of y. Let

$$\mu = (\mu_1, \dots, \mu_n)' \in R^n \quad \text{and} \quad \Sigma = (\sigma_{ij}) \in \mathcal{S}(n),$$

Generalized Least Squares Takeaki Kariya and Hiroshi Kurata
 ISBN: 0-470-86697-7 (PPC)

where $\mathcal{S}(n)$ denotes the set of $n \times n$ positive definite matrices and a' the transposition of vector a or matrix a. We say that y is distributed as an n-dimensional *multivariate normal distribution* $N_n(\mu, \Sigma)$, and express the relation as

$$\mathcal{L}(y) = N_n(\mu, \Sigma), \tag{1.1}$$

if the probability density function (pdf) $f(y)$ of y with respect to the Lebesgue measure on R^n is given by

$$f(y) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right) \quad (y \in R^n). \tag{1.2}$$

When $\mathcal{L}(y) = N_n(\mu, \Sigma)$, the mean vector $E(y)$ and the covariance matrix $\mathrm{Cov}(y)$ are respectively given by

$$E(y) = \mu \quad \text{and} \quad \mathrm{Cov}(y) = \Sigma, \tag{1.3}$$

where

$$\mathrm{Cov}(y) = E\{(y-\mu)(y-\mu)'\}.$$

Hence, we often refer to $N_n(\mu, \Sigma)$ as the normal distribution with mean μ and covariance matrix Σ.

Multivariate normality and linear transformations. Normality is preserved under linear transformations, which is a prominent property of the multivariate normal distribution. More precisely,

Proposition 1.1 *Suppose that $\mathcal{L}(y) = N_n(\mu, \Sigma)$. Let A be any $m \times n$ matrix such that $\mathrm{rank}\, A = m$ and let b be any $m \times 1$ vector. Then*

$$\mathcal{L}(Ay + b) = N_m(A\mu + b, A\Sigma A'). \tag{1.4}$$

Thus, when $\mathcal{L}(y) = N_n(\mu, \Sigma)$, all the marginal distributions of y are normal. In particular, partition y as

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad \text{with} \quad y_j : n_j \times 1 \quad \text{and} \quad n = n_1 + n_2,$$

and let μ and Σ be correspondingly partitioned as

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \tag{1.5}$$

Then it follows by setting $A = (I_{n_1}, 0) : n_1 \times n$ in Proposition 1.1 that

$$\mathcal{L}(y_1) = N_{n_1}(\mu_1, \Sigma_{11}).$$

Clearly, a similar argument yields $\mathcal{L}(y_2) = N_{n_2}(\mu_2, \Sigma_{22})$. Note here that y_j's are not necessarily independent. In fact,

Proposition 1.2 *If $\mathcal{L}(y) = N_n(\mu, \Sigma)$, then the conditional distribution $\mathcal{L}(y_1|y_2)$ of y_1 given y_2 is given by*

$$\mathcal{L}(y_1|y_2) = N_{n_1}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2), \Sigma_{11.2}) \tag{1.6}$$

with

$$\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

It is important to notice that there is a one-to-one correspondence between $(\Sigma_{11}, \Sigma_{12}, \Sigma_{22})$ and $(\Sigma_{11.2}, \Theta, \Sigma_{22})$ with $\Theta = \Sigma_{12}\Sigma_{22}^{-1}$. The matrix Θ is often called the *linear regression coefficient* of y_1 on y_2.

As is well known, the condition $\Sigma_{12} = 0$ is equivalent to the independence between y_1 and y_2. In fact, if $\Sigma_{12} = 0$, then we can see from Proposition 1.2 that

$$\mathcal{L}(y_1) = \mathcal{L}(y_1|y_2) \ (= N_{n_1}(\mu_1, \Sigma_{11})),$$

proving the independence between y_1 and y_2. The converse is obvious.

Orthogonal transformations. Consider a class of normal distributions of the form $N_n(0, \sigma^2 I_n)$ with $\sigma^2 > 0$, and suppose that the distribution of a random vector y belongs to this class:

$$\mathcal{L}(y) \in \{N_n(0, \sigma^2 I_n) \mid \sigma^2 > 0\}. \tag{1.7}$$

Let $\mathcal{O}(n)$ be the group of $n \times n$ orthogonal matrices (see Section 1.4). By using Proposition 1.1, it is shown that the distribution of y remains the same under orthogonal transformations as long as the condition (1.7) is satisfied. Namely, we have

Proposition 1.3 *If $\mathcal{L}(y) = N_n(0, \sigma^2 I_n)$ $(\sigma^2 > 0)$, then*

$$\mathcal{L}(\Gamma y) = \mathcal{L}(y) \quad for\ any\ \Gamma \in \mathcal{O}(n). \tag{1.8}$$

It is noted that the orthogonal transformation $a \to \Gamma a$ is geometrically either the rotation of a or the reflection of a in R^n. A distribution that satisfies (1.8) will be called a *spherically symmetric distribution* (see Section 1.3). Proposition 1.3 states that $\{N_n(0, \sigma^2 I_n) \mid \sigma^2 > 0\}$ is a subclass of the class of spherically symmetric distributions.

Let $\|A\|$ denote the Euclidean norm of matrix A with

$$\|A\|^2 = tr(A'A),$$

where $tr(\cdot)$ denotes the trace of a matrix $\cdot$. In particular,

$$\|a\|^2 = a'a$$

for a vector a.

Proposition 1.4 *Suppose that $\mathcal{L}(y) \in \{N_n(0, \sigma^2 I_n) \mid \sigma^2 > 0\}$, and let*

$$x \equiv \|y\| \text{ and } z \equiv y/\|y\| \text{ with } \|y\|^2 = y'y. \tag{1.9}$$

Then the following three statements hold:

(1) $\mathcal{L}\left(x^2/\sigma^2\right) = \chi^2_n$, *where χ^2_n denotes the χ^2 (chi-square) distribution with degrees of freedom n;*

(2) *The vector z is distributed as the uniform distribution on the unit sphere $\mathcal{U}(n)$ in R^n, where*

$$\mathcal{U}(n) = \{u \in R^n \mid \|u\| = 1\};$$

(3) *The quantities x and z are independent.*

To understand this proposition, several relevant definitions follow. A random variable w is said to be distributed as χ^2_n, if a pdf of w is given by

$$f(w) = \frac{1}{2^{n/2}\Gamma(n/2)} w^{\frac{n}{2}-1} \exp\left(-w/2\right) \quad (w > 0), \tag{1.10}$$

where $\Gamma(a)$ is the Gamma function defined by

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt \quad (a > 0). \tag{1.11}$$

A random vector z such that $z \in \mathcal{U}(n)$ is said to have a uniform distribution on $\mathcal{U}(n)$ if the distribution $\mathcal{L}(z)$ of z satisfies

$$\mathcal{L}(\Gamma z) = \mathcal{L}(z) \text{ for any } \Gamma \in \mathcal{O}(n). \tag{1.12}$$

As will be seen in the next section, statements (2) and (3) of Proposition 1.4 remain valid as long as the distribution of y is spherically symmetric. That is, if y satisfies $\mathcal{L}(\Gamma y) = \mathcal{L}(y)$ for all $\Gamma \in \mathcal{O}(n)$ and if $P(y = 0) = 0$, then $z \equiv y/\|y\|$ is distributed as the uniform distribution on the unit sphere $\mathcal{U}(n)$, and is independent of $x \equiv \|y\|$.

Wishart distribution. Next, we introduce the Wishart distribution, which plays a central role in estimation of the covariance matrix Σ of the multivariate normal distribution $N_n(\mu, \Sigma)$. In this book, the Wishart distribution will appear in the context of estimating a seemingly unrelated regression (SUR) model (see Example 2.4) and a growth curve model (see Chapter 9).

Suppose that p-dimensional random vectors $y_1, \ldots, y_n$ are independently and identically distributed as the normal distribution $N_p(0, \Sigma)$ with $\Sigma \in \mathcal{S}(p)$. We call the distribution of the matrix

$$W = \sum_{j=1}^n y_j y_j'$$

the *Wishart distribution* with parameter matrix Σ and degrees of freedom n, and express it as

$$\mathcal{L}(W) = W_p(\Sigma, n). \tag{1.13}$$

When $n \geq p$, the distribution $W_p(\Sigma, n)$ has a pdf of the form

$$f(W) = \frac{1}{2^{np/2}\Gamma_p(n/2)|\Sigma|^{n/2}}|W|^{\frac{n-p-1}{2}} \exp\left(-\frac{tr(W\Sigma^{-1})}{2}\right), \tag{1.14}$$

which is positive on the set $\mathcal{S}(p)$ of $p \times p$ positive definite matrices. Here $\Gamma_p(a)$ is the multivariate Gamma function defined by

$$\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\left(a - \frac{j-1}{2}\right) \quad \left(a > \frac{p-1}{2}\right). \tag{1.15}$$

When $p = 1$, the multivariate Gamma function reduces to the (usual) Gamma function:

$$\Gamma_1(a) = \Gamma(a).$$

If W is distributed as $W_p(\Sigma, n)$, then the mean matrix is given by

$$E(W) = n\Sigma.$$

Hence, we often call $W_p(\Sigma, n)$ the Wishart distribution with mean $n\Sigma$ and degrees of freedom n. Note that when $p = 1$ and $\Sigma = 1$, the pdf $f(W)$ in (1.14) reduces to that of the χ^2 distribution χ_n^2, that is, $W_1(1, n) = \chi_n^2$. More generally, if $\mathcal{L}(w) = W_1(\sigma^2, n)$, then

$$\mathcal{L}(w/\sigma^2) = \chi_n^2. \tag{1.16}$$

(See Problem 1.2.2.)

Wishart-ness and linear transformations. As the normality is preserved under linear transformations, so is the Wishart-ness. To see this, suppose that $\mathcal{L}(W) = W_p(\Sigma, n)$. Then we have

$$\mathcal{L}(W) = \mathcal{L}\left(\sum_{j=1}^{n} y_j y_j'\right),$$

where y_j's are independently and identically distributed as the normal distribution $N_p(0, \Sigma)$. Here, by Proposition 1.1, for an $m \times p$ matrix A such that $rank A = m$, the random vectors $Ay_1, \ldots, Ay_n$ are independent and each Ay_j has $N_p(0, A\Sigma A')$. Hence, the distribution of

$$\sum_{j=1}^{n} Ay_j(Ay_j)' = A\left(\sum_{j=1}^{n} y_j y_j'\right)A'$$