Statistics and Probability Letters 97 (2015) 16–24

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Some remarks on general linear model with new regressors

Changli Lu [a], Shengjun Gan [a], Yongge Tian [b,*]

[a] College of Economics and Management, Shanghai Maritime University, Shanghai, China

[b] China Economics and Management Academy, Central University of Finance and Economics, Beijing, China

ARTICLE INFO

Article history:
Received 23 May 2014
Received in revised form 23 October 2014
Accepted 23 October 2014
Available online 31 October 2014

MSC:
62J05
62H12

Keywords:
Linear model
Misspecified model
Parametric functions
New regressors
BLUE

ABSTRACT

Assume that an original general linear model is misspecified by adding some new regressors. We investigate in such a case relationships between the best linear unbiased estimators under the two models. In particular, we give necessary and sufficient conditions for the best linear unbiased estimators to be equal under the two models.

1. Introduction

Consider a general linear model

$$\mathscr{M} = \{\mathbf{y},\ \mathbf{X}\boldsymbol{\beta},\ \boldsymbol{\Sigma}\}, \tag{1.1}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a known matrix of arbitrary rank, $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ is a vector of fixed but unknown parameters, $\mathbf{y} \in \mathbb{R}^{n\times 1}$ is an observable random vector with $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $Cov(\mathbf{y}) = \boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ is a known or unknown non-negative definite (nnd) matrix of arbitrary rank.

With the numbers of observations fixed in (1.1), we often face with the task of comparing two models by adding or deleting regressors in the model. If we take (1.1) as a correct model and add regressors $\mathbf{Z}\boldsymbol{\gamma}$, we obtain a misspecified model

$$\mathscr{N} = \{\mathbf{y},\ \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma},\ \boldsymbol{\Sigma}\}, \tag{1.2}$$

where $\mathbf{Z} \in \mathbb{R}^{n\times q}$ is a known matrix of arbitrary rank, $\boldsymbol{\gamma} \in \mathbb{R}^{q\times 1}$ is a vector of fixed but unknown parameters, and $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$ and $Cov(\mathbf{y}) = \boldsymbol{\Sigma}$. Conversely, if taking (1.2) as a correct model with the model matrix $[\mathbf{X},\ \mathbf{Z}]$:

$$\mathscr{N} = \left\{\mathbf{y},\ [\mathbf{X},\ \mathbf{Z}]\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{bmatrix},\ \boldsymbol{\Sigma}\right\}, \tag{1.3}$$

* Corresponding author. Tel.: +86 10 62289253.
E-mail address: yongge.tian@gmail.com (Y. Tian).

http://dx.doi.org/10.1016/j.spl.2014.10.015
0167-7152/

then (1.1) is a new model with a misspecified model matrix $[\mathbf{X},\ \mathbf{0}]$:

$$\mathscr{M} = \left\{ \mathbf{y},\ [\mathbf{X},\ \mathbf{0}] \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix},\ \boldsymbol{\Sigma} \right\}. \tag{1.4}$$

Assume that we want to estimate the vector of parametric functions $\mathbf{K}\boldsymbol{\beta}$, where $\mathbf{K} \in \mathbb{R}^{k\times p}$ is given. The purpose of this note is to establish possible equalities between the best linear unbiased estimators (BLUEs) of $\mathbf{K}\boldsymbol{\beta}$ under (1.1) and (1.2). This kind of work was first proposed and studied by Bhimasankaram and Jammalamadaka (1994), and then by Jammalamadaka and Sengupta (1999, 2007). Under the assumptions in (1.1) or (1.2), the BLUEs of $\mathbf{K}\boldsymbol{\beta}$ are not necessarily the same, and one of which is misspecified. In this note, we take (1.1) as a correct model and give a new investigation to the relations between BLUEs of $\mathbf{K}\boldsymbol{\beta}$ under (1.1) and (1.2) through a variety of matrix rank formulas.

Throughout this note, $\mathbb{R}^{m\times n}$ stands for the collection of all $m \times n$ real matrices. The symbols $\mathbf{A}'$, $r(\mathbf{A})$ and $\mathscr{R}(\mathbf{A})$ stand for the transpose, rank and range (column space) of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, respectively; $\mathbf{I}_m$ stands for the identity matrix of order m. The Moore–Penrose inverse of $\mathbf{A} \in \mathbb{R}^{m\times n}$, denoted by $\mathbf{A}^+$, is defined to be the unique solution $\mathbf{X}$ satisfying the four matrix equations $\mathbf{AXA} = \mathbf{A}$, $\mathbf{XAX} = \mathbf{X}$, $(\mathbf{AX})' = \mathbf{AX}$ and $(\mathbf{XA})' = \mathbf{XA}$. Further, let $\mathbf{P_A}$, $\mathbf{E_A}$ and $\mathbf{F_A}$ stand for the three orthogonal projectors (symmetric idempotent matrices) $\mathbf{P_A} = \mathbf{AA}^+$, $\mathbf{E_A} = \mathbf{I}_m - \mathbf{AA}^+$ and $\mathbf{F_A} = \mathbf{I}_n - \mathbf{A}^+\mathbf{A}$, where $\mathbf{E_A}$ and $\mathbf{F_A}$ satisfy $\mathbf{E_A} = \mathbf{F_{A'}}$ and $\mathbf{F_A} = \mathbf{E_{A'}}$, and the ranks of $\mathbf{E_A}$ and $\mathbf{F_A}$ are $r(\mathbf{E_A}) = m - r(\mathbf{A})$ and $r(\mathbf{F_A}) = n - r(\mathbf{A})$, respectively. The symbols $i_+(\mathbf{A})$ and $i_-(\mathbf{A})$ stand for the positive and negative inertias of $\mathbf{A} = \mathbf{A}' \in \mathbb{R}^{m\times m}$, which are defined to be the number of the positive and negative eigenvalues of $\mathbf{A}$ counted with multiplicities, respectively. For two symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ of the same size, $\mathbf{A} \preccurlyeq \mathbf{B}$ means that $\mathbf{B} - \mathbf{A}$ is nnd.

2. Preliminaries

The linear regression model is one of the most widely-used models in statistical analysis, which is also the foundation for more advanced methods. In the estimation theory of regression models, the BLUEs of $\mathbf{K}\boldsymbol{\beta}$ under (1.1) have been main objects of study due to their simple and optimal statistical properties. Assume that $\mathbf{K} \in \mathbb{R}^{k\times p}$ is a given matrix. Then the vector of parametric functions $\mathbf{K}\boldsymbol{\beta}$ is said to be estimable under (1.1) if there exists a matrix $\mathbf{L} \in \mathbb{R}^{k\times n}$ such that the expectation $E(\mathbf{Ly}) = \mathbf{K}\boldsymbol{\beta}$ holds; see, e.g., Alalouf and Styan (1979).

Assume that $\mathbf{K}\boldsymbol{\beta}$ is estimable under (1.1), and let $\mathscr{S}$ be the collection of all $\mathbf{Ly} - \mathbf{K}\boldsymbol{\beta}$ with $E(\mathbf{Ly}) = \mathbf{K}\boldsymbol{\beta}$, i.e.,

$$\mathscr{S} = \{\mathbf{Ly} - \mathbf{K}\boldsymbol{\beta} \mid E(\mathbf{Ly} - \mathbf{K}\boldsymbol{\beta}) = \mathbf{0},\ \mathbf{L} \in \mathbb{R}^{k\times n},\ \mathbf{K} \in \mathbb{R}^{k\times p}\}. \tag{2.1}$$

Note that $\mathbf{Ly} - \mathbf{K}\boldsymbol{\beta}$ in $\mathscr{S}$ is not necessarily unique. Thus it is natural to seek an element of $\mathscr{S}$ according to a certain optimal criterion as the best choice for all unbiased estimators of $\mathbf{K}\boldsymbol{\beta}$. A well-known case is to find an element of $\mathscr{S}$ that minimizes the covariance matrix of $\mathbf{Ly} - \mathbf{K}\boldsymbol{\beta} \in \mathscr{S}$ in the Löwner partial ordering, i.e., to find $\mathbf{L}_0\mathbf{y} - \mathbf{K}\boldsymbol{\beta} \in \mathscr{S}$ such that

$$Cov(\mathbf{L}_0\mathbf{y} - \mathbf{K}\boldsymbol{\beta}) \preccurlyeq Cov(\mathbf{Ly} - \mathbf{K}\boldsymbol{\beta}) \quad \text{s.t. } \mathbf{Ly} - \mathbf{K}\boldsymbol{\beta} \in \mathscr{S}. \tag{2.2}$$

Note that $Cov(\mathbf{Ly} - \mathbf{K}\boldsymbol{\beta}) = E[(\mathbf{Ly} - \mathbf{K}\boldsymbol{\beta})(\mathbf{Ly} - \mathbf{K}\boldsymbol{\beta})']$ under (2.1), where $(\mathbf{Ly} - \mathbf{K}\boldsymbol{\beta})(\mathbf{Ly} - \mathbf{K}\boldsymbol{\beta})'$ is the well-known matrix loss function of $\mathbf{Ly}$ with respect to $\mathbf{K}\boldsymbol{\beta}$; see Rao (1976). Thus (2.2) is also equivalent to

$$E[(\mathbf{L}_0\mathbf{y} - \mathbf{K}\boldsymbol{\beta})(\mathbf{L}_0\mathbf{y} - \mathbf{K}\boldsymbol{\beta})'] \preccurlyeq E[(\mathbf{Ly} - \mathbf{K}\boldsymbol{\beta})(\mathbf{Ly} - \mathbf{K}\boldsymbol{\beta})'] \quad \text{s.t. } \mathbf{Ly} - \mathbf{K}\boldsymbol{\beta} \in \mathscr{S}. \tag{2.3}$$

The estimator $\mathbf{L}_0\mathbf{y}$ satisfying this inequality is the well-known BLUE of the parameter vector $\mathbf{K}\boldsymbol{\beta}$ under (1.1) and is denoted by $\mathrm{BLUE}_{\mathscr{M}}(\mathbf{K}\boldsymbol{\beta})$. If $\mathbf{K} = \mathbf{X}$, the estimator $\mathbf{L}_0\mathbf{y}$ satisfying (2.2) is the BLUE of the mean vector $\mathbf{X}\boldsymbol{\beta}$ in (1.1) and is denoted by $\mathrm{BLUE}_{\mathscr{M}}(\mathbf{X}\boldsymbol{\beta})$. Eqs. (2.1) and (2.2) show that the BLUE under (1.1) is in fact a quadratic matrix-valued function optimization problem over a given matrix set. This kind of problems on the minimization of covariance matrices of estimators occurs everywhere in statistics, which can be regarded as some special cases of optimization problems on quadratic matrix-valued functions in the Löwner sense. In this case, matrix theory plays an important role in solving this kind of estimation problems in regression analysis.

Some new and useful theory on optimization problems of quadratic matrix-valued functions was developed in recent years. Thus, many minimization problems on covariance matrices of estimators in statistics can be solved analytically. A recent result on constrained quadratic matrix-valued function optimization problem related to (2.1) and (2.2) is given below; see Tian (2012).

Lemma 2.1. *Let $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{B} \in \mathbb{R}^{m\times p}$ be given, and $\mathbf{P} = \mathbf{P}' \in \mathbb{R}^{n\times n}$ be nnd. Also assume that there exists $\mathbf{X}_0 \in \mathbb{R}^{m\times n}$ such that $\mathbf{X}_0\mathbf{A} = \mathbf{B}$. Then the maximal positive inertia of $\mathbf{X}_0\mathbf{P}\mathbf{X}_0' - \mathbf{XPX}'$ subject to all solutions of $\mathbf{XA} = \mathbf{B}$ is*

$$\max_{\mathbf{XA}=\mathbf{B}} i_+(\mathbf{X}_0\mathbf{P}\mathbf{X}_0' - \mathbf{XPX}') = r\begin{bmatrix} \mathbf{X}_0\mathbf{P} \\ \mathbf{A}' \end{bmatrix} - r(\mathbf{A}) = r(\mathbf{X}_0\mathbf{P}\mathbf{E_A}). \tag{2.4}$$

Hence there exists solution $\mathbf{X}_0$ of $\mathbf{X}_0\mathbf{A} = \mathbf{B}$ such that

$$\mathbf{X}_0\mathbf{P}\mathbf{X}_0' \preccurlyeq \mathbf{XPX}' \tag{2.5}$$