Hindawi Publishing Corporation
Journal of Inequalities and Applications
Volume 2010, Article ID 259672, 5 pages
doi:10.1155/2010/259672

Research Article

From Equivalent Linear Equations to Gauss-Markov Theorem

Czesław Stępniak

Institute of Mathematics, University of Rzeszów, Rejtana 16 A, 35-959 Rzeszów, Poland

Correspondence should be addressed to Czesław Stępniak, stepniak@umcs.lublin.pl

Received 17 December 2009; Revised 20 May 2010; Accepted 27 June 2010

Academic Editor: Andrei Volodin

Gauss-Markov theorem reduces linear unbiased estimation to the Least Squares Solution of inconsistent linear equations while the normal equations reduce the second one to the usual solution of consistent linear equations. It is rather surprising that the second algebraic result is usually derived in a differential way. To avoid this dissonance, we state and use an auxiliary result on equivalence of two systems of linear equations. This places us in a convenient position to attack on the main problems in the Gauss-Markov model in an easy way.

1. Introduction

The Gauss-Markov theorem is the most classical achievement in statistics. Its role in statistics is comparable with that of the Pythagorean theorem in geometry. In fact, there are close relations between both of them.

The Gauss-Markov theorem is presented in many books and derived in many ways. The most popular approaches involve

(i) geometry (cf., Kruskal [1, 2]),

(ii) differential calculus (cf., Scheffé [3], Rao [4]),

(iii) generalized inverse matrices (cf., Rao and Mitra [5], Bapat [6]),

(iv) projection operators (see Seber [7]).

We presume that such a big market has many clients. This paper is intended for some of them. Our consideration is straightforward and self-contained. Moreover, it needs only moderate prerequisites.

The main tool used in this paper is equivalence of two systems of linear equations.

2. Preliminaries

For any matrix $\mathbf{A}$ of $n \times p$ define the sets

$$\begin{aligned} \mathcal{R}(\mathbf{A}) &= \{\mathbf{a} \in R^n : \mathbf{a} = \mathbf{A}\mathbf{x} \ \text{ for some } \mathbf{x} \in R^p\} \ (\text{i.e., the range of } \mathbf{A}), \\ \mathcal{N}(\mathbf{A}) &= \{\mathbf{x} \in R^p : \mathbf{A}\mathbf{x} = \mathbf{0}\} \ (\text{i.e., the kernel of } \mathbf{A}). \end{aligned} \tag{2.1}$$

We note that

$$\mathbf{a}^T\mathbf{b} = \mathbf{0}, \quad \forall \mathbf{a} \in \mathcal{R}(\mathbf{A}),\ \mathbf{b} \in \mathcal{N}\left(\mathbf{A}^T\right). \tag{2.2}$$

It is clear that the range $\mathcal{R}(\mathbf{A})$ constitutes r-dimensional linear space in R^n spanned by the columns of $\mathbf{A}$, where $r = \operatorname{rank}(\mathbf{A})$, while $\mathcal{N}(\mathbf{A}^T)$ constitutes $(n-r)$-dimensional space of all vectors being orthogonal to any vector in $\mathcal{R}(\mathbf{A})$ relative to the usual inner product $(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T\mathbf{b}$. Thus, any vector $\mathbf{y} \in R^n$ may be presented in the form

$$\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2, \quad \text{where } \mathbf{y}_1 \in \mathcal{R}(\mathbf{A}),\ \mathbf{y}_2 \in \mathcal{N}\left(\mathbf{A}^T\right) \text{ are orthogonal.} \tag{2.3}$$

Since $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{0}$ if and only if $\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{0}$ and, hence, $\mathbf{A}\mathbf{x} = \mathbf{0}$, we get $\mathcal{N}(\mathbf{A}\mathbf{A}^T) = \mathcal{N}(\mathbf{A}^T)$.

Denote by $\mathbf{P} = \mathbf{P}_A$ the linear operator from R^n onto $\mathcal{R}(\mathbf{A})$ defined by

$$\mathbf{P}\mathbf{y} = \begin{cases} \mathbf{y}, & \text{if } \mathbf{y} \in \mathcal{R}(\mathbf{A}), \\ \mathbf{0}, & \text{if } \mathbf{y} \in \mathcal{N}(\mathbf{A}^T) \end{cases} \tag{2.4}$$

(i.e., the *orthogonal projector* onto $\mathcal{R}(\mathbf{A})$). It follows from definition (2.4) that $\mathbf{P}\mathbf{P} = \mathbf{P}$. The following lemma (see [8]) will be a key tool in the further consideration.

Lemma 2.1. *For any matrix* $\mathbf{A}$ *and for any vector* $\mathbf{b} \in \mathcal{R}(\mathbf{A})$*, the following are equivalent:*

(i) $\mathbf{A}\mathbf{x} = \mathbf{b}$,

(ii) $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$.

Proof. (i)⇒(ii) is evident (without any condition on $\mathbf{b}$).

(ii)⇒(i). By the assumption that $\mathbf{b} \in \mathcal{R}(\mathbf{A})$, we get $\mathbf{b} = \mathbf{A}\mathbf{c}$ for some $\mathbf{c}$. Thus, (ii) reduces to $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{A}\mathbf{c}$ and its general solution is $\mathbf{x} = \mathbf{c} + \mathbf{x}_0$, where $\mathbf{x}_0 \in \mathcal{N}(\mathbf{A}^T\mathbf{A}) = \mathcal{N}(\mathbf{A})$. Therefore, $\mathbf{x}$ is a solution of (i). □

Remark 2.2. The assumption that $\mathbf{b} \in \mathcal{R}(\mathbf{A})$ in Lemma 2.1 is essential. To see this, let us set

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}. \tag{2.5}$$

Then, $\mathbf{A}^T\mathbf{A} = \left[\begin{smallmatrix} 2 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ and $\mathbf{A}^T\mathbf{b} = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$. Thus, $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$ has a solution $\mathbf{x} = [0, 0]^T$, while $\mathbf{A}\mathbf{x} = \mathbf{b}$ is inconsistent.

3. Least Squares Solution

For any matrix $\mathbf{A}$ of $n \times p$ and for any vector $\mathbf{b} \in R^n$, consider the linear equation

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{3.1}$$

Equation (3.1) may be consistent (if $\mathbf{b} \in \mathcal{R}(\mathbf{A})$) or inconsistent (if not). In the second case we are seeking for such $\mathbf{x}$ that the residual vector $\mathbf{b} - \mathbf{A}\mathbf{x}$ be as small as possible.

Definition 3.1. Any vector $\widehat{\mathbf{x}} \in R^p$ is said to be the Least Squares Solution (LSS) of (3.1) if

$$(\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}})^T(\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}) \leq (\mathbf{b} - \mathbf{A}\mathbf{x})^T(\mathbf{b} - \mathbf{A}\mathbf{x}), \quad \text{for any } \mathbf{x} \in R^p. \tag{3.2}$$

The following theorem shows that this definition is not empty and reduces the LSS of an inconsistent equation (3.1) to the ordinary solution of a consistent one.

Theorem 3.2. *(a) Equation* (3.1) *has at least one LSS.*

(b) Vector $\mathbf{x} \in R^P$ *is an LSS of* (3.1) *if and only if*

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}. \tag{3.3}$$

(c) Condition (3.3) *is equivalent to*

$$\mathbf{A}\mathbf{x} = \mathbf{P}\mathbf{b}, \tag{3.4}$$

where $\mathbf{P} = \mathbf{P}_A$ *is the orthogonal projector onto* $\mathcal{R}(\mathbf{A})$ *defined by* (2.4).

(d) General solution of (3.3) *may be presented in the form* $\mathbf{x} = \mathbf{x}_0 + \mathbf{x}_1$, *where* $\mathbf{x}_0$ *is a particular solution, while* $\mathbf{x}_1 \in \mathcal{N}(\mathbf{A})$.

Remark 3.3. In the statistical literature, (3.3) is said to be normal.

Proof. By properties of the projector $\mathbf{P}$, we get

$$\begin{aligned}(\mathbf{b} - \mathbf{A}\mathbf{x})^T(\mathbf{b} - \mathbf{A}\mathbf{x}) &= [\mathbf{P}\mathbf{b} + (\mathbf{I} - \mathbf{P})\mathbf{b} - \mathbf{A}\mathbf{x}]^T[\mathbf{P}\mathbf{b} + (\mathbf{I} - \mathbf{P})\mathbf{b} - \mathbf{A}\mathbf{x}] \\ &= (\mathbf{P}\mathbf{b} - \mathbf{A}\mathbf{x})^T(\mathbf{P}\mathbf{b} - \mathbf{A}\mathbf{x}) + [(\mathbf{I} - \mathbf{P})\mathbf{b}]^T[(\mathbf{I} - \mathbf{P})\mathbf{b}] \\ &= (\mathbf{P}\mathbf{b} - \mathbf{A}\mathbf{x})^T(\mathbf{P}\mathbf{b} - \mathbf{A}\mathbf{x}) + \mathbf{b}^T(\mathbf{I} - \mathbf{P})\mathbf{b} \\ &\geq \mathbf{b}^T(\mathbf{I} - \mathbf{P})\mathbf{b}\end{aligned} \tag{3.5}$$

with the equality if and only if (3.4) holds. Moreover, by definition of $\mathbf{P}$, (3.4) is consistent and, by Lemma 2.1, it is equivalent to (3.3).

Statement (d) follows directly from definition of kernel. □

4. Gauss-Markov Model and Gauss-Markov Theorem

Let $\mathbf{y}$ be an arbitrary random vector in R^n with finite second moment $E(\mathbf{y}^T\mathbf{y})$. Then there exist a unique vector $\boldsymbol{\mu} \in R^n$ and a unique symmetric nonnegative definite matrix $\mathbf{V}$ of $n \times n$ such that

$$\begin{aligned} E\left(\mathbf{a}^T\mathbf{y}\right) &= \mathbf{a}^T\boldsymbol{\mu}, \\ \operatorname{Cov}\left(\mathbf{A}^T\mathbf{y}, \mathbf{B}^T\mathbf{y}\right) &= \mathbf{A}^T\mathbf{V}\mathbf{B} \end{aligned} \tag{4.1}$$

for all vectors $\mathbf{a} \in R^n$ and all matrices $\mathbf{A}$ and $\mathbf{B}$ of n rows. Traditionally, such $\boldsymbol{\mu}$ and $\mathbf{V}$ are called the expectation and the dispersion of the random vector $\mathbf{y}$.

As usual, we will assume that $\boldsymbol{\mu}$ and $\mathbf{V}$ have the representations

$$\begin{aligned} \boldsymbol{\mu} &= \mathbf{X}\boldsymbol{\beta}, \\ \mathbf{V} &= \sigma^2\mathbf{I}_n, \end{aligned} \tag{4.2}$$

where $\mathbf{X}$ is a given matrix of $n \times p$ while $\boldsymbol{\beta} \in R^p$ and $\sigma^2 > 0$ are unknown parameters. We will refer to the structure $(\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ as to the standard Gauss-Markov model. In the context of the model we will consider unbiased estimation of the parametric vector $\boldsymbol{\Psi} = \mathbf{C}^T\boldsymbol{\beta}$, where $\mathbf{C}$ is of $p \times q$, by estimators of the form $\widehat{\boldsymbol{\Psi}} = \mathbf{D}^T\mathbf{y}$, where $\mathbf{D}$ is of $n \times p$ matrix. Since $\mathbf{D}^T\mathbf{y}$ is unbiased if and only if $\mathbf{D}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{C}^T\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$, $\mathbf{C}^T\boldsymbol{\beta}$ is estimable if and only if

$$\mathbf{C}^T = \mathbf{D}^T\mathbf{X}, \quad \text{for some } \mathbf{D}. \tag{4.3}$$

Without loss of generality, we may and will assume that $\mathcal{R}(\mathbf{D}) \subseteq \mathcal{R}(\mathbf{X})$. We note that such a matrix $\mathbf{D}$ is uniquely determined by $\mathbf{C}$.

The well-known Gauss-Markov theorem provides a constructive way for estimation of the function $\boldsymbol{\Psi}$. It is based on a solution of the normal equation $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$ which plays the role of the estimator for $\boldsymbol{\beta}$.

Theorem 4.1. *For any estimable $\boldsymbol{\Psi} = \mathbf{C}^T\boldsymbol{\beta}$ in the standard Gauss-Markov model $(\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, there exists a unique linear unbiased estimator with minimal dispersion. This estimator, called the Least Squares Estimator (LSE) of $\boldsymbol{\Psi}$, may be presented in the form $\mathbf{C}^T\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is an arbitrary LSS of $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$ or, equivalently, it is a solution of the normal equation*

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}. \tag{4.4}$$

Proof. By Theorem 3.2 the condition $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$ is equivalent to $\mathbf{P}_X\mathbf{y} = \mathbf{X}\boldsymbol{\beta} = E\mathbf{y}$. Therefore, by (4.3), for any estimable $\boldsymbol{\Psi} = \mathbf{C}^T\boldsymbol{\beta}$ and for any solution $\widehat{\boldsymbol{\beta}}$ of (4.4), the statistic $\mathbf{C}^T\widehat{\boldsymbol{\beta}} = \mathbf{D}^T\mathbf{X}\widehat{\boldsymbol{\beta}}$ is unbiased. On the other hand,

$$E\left(\mathbf{D}_1^T\mathbf{y}\right) = E\left(\mathbf{D}_2^T\mathbf{y}\right), \quad \text{iff } \mathcal{R}(\mathbf{D}_1 - \mathbf{D}_2) \subseteq \mathcal{N}\left(\mathbf{X}^T\right). \tag{4.5}$$

Hence, any unbiased estimator of $\boldsymbol{\Psi}$ may be presented in the form $\mathbf{D}^T\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{B}^T\mathbf{y}$, where the first component is the LSE of $\boldsymbol{\Psi}$ while $\mathbf{B}^T\mathbf{D} = \mathbf{B}^T\mathbf{X} = \mathbf{0}$. In particular the components are not correlated. Therefore, the variance of the sum is greater than the variance of the LSE $\mathbf{D}^T\mathbf{X}\widehat{\boldsymbol{\beta}}$, unless $\mathbf{B}^T\mathbf{y} \neq \mathbf{0}$. Moreover, by Theorem 3.2(d) this estimator is invariant with respect to the choice of the LSS $\widehat{\boldsymbol{\beta}}$. In consequence, the LSE of the function $\boldsymbol{\Psi}$ is unique. □

Acknowledgment

Thanks are due to a reviewer for his comments leading to the improvement in the presentation of this paper.

References

[1] W. Kruskal, "The coordinate-free approach to Gauss-Markov estimation, and its application to missing and extra observations," in *Proceedings of 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 435–451, University of California Press, Berkeley, Calif, USA, 1961.
[2] W. Kruskal, "When are Gauss-Markov and least squares estimators identical? A coordinate-free approach," *Annals of Mathematical Statistics*, vol. 39, pp. 70–75, 1968.
[3] H. Scheffé, *The Analysis of Variance*, John Wiley & Sons, New York, NY, USA, 1959.
[4] C. R. Rao, *Linear Statistical Inference and Its Applications*, John Wiley & Sons, New York, NY, USA, 2nd edition, 1973.
[5] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and Its Applications*, John Wiley & Sons, New York, NY, USA, 1971.
[6] R. B. Bapat, *Linear Algebra and Linear Models*, Universitext, Springer, New York, NY, USA, 2nd edition, 2000.
[7] G. A. F. Seber, *Linear Regression Analysis*, John Wiley & Sons, New York, NY, USA, 1977.
[8] C. Stępniak, "Through a generalized inverse," *Demonstratio Mathematica*, vol. 41, no. 2, pp. 291–296, 2008.