SINET: Ethiop. J. Sci., 22(1):31–40, 1999
 ISSN: 0379–2897

IMPROVEMENT OF MULTIVARIATE STATISTICS BY LINEAR TRANSFORMS

Eshetu Wencheko

Department of Statistics, Faculty of Science,
Addis Ababa University, PO Box 1176, Addis Ababa, Ethiopia

ABSTRACT: Theoretical results about comparison of multivariate estimators and a general linear transform of the same *vis-à-vis* the mean square error criterion are given. Two theorems on admissibility of linear transforms of estimators are introduced. Applications of the theoretical results are demonstrated by considering two linear transforms of the unbiased estimator of the coefficients of the multiple linear regression model.

Key words/phrases: Admissibility, linear transforms, multivariate statistics, strong and weak mean square criteria

INTRODUCTION

Improvement of estimation and prediction by introducing biased estimation procedures had been, and still is a research area of importance. The comparison of risks of the resulting estimators can be shown by using mean square error (MSE). Perlman (1972) gave a necessary and sufficient condition guaranteeing the existence of a scalar $\alpha \in (0,1)$ such that a shrinkage of an unbiased vector-valued estimator would result in an improvement in MSE. Bibby (1972) and Bibby and Toutenburg (1977; 1978) investigated biased estimators and predictors that can be obtained by improving unbiased procedures. A general result due to Perlman was applied by Kleffe (1985) to show that uniformly better estimators can be given through multiplication of an unbiased estimator by some positive constant less than but close to one. The optimal choice of such a multiplier was studied, and specific results for quadratic estimation derived.

This paper studies linear transforms of an unbiased multivariate estimator. Theoretical conditions regarding MSE improvement and admissibility of a linear transform are given. Finally, the general results are applied to two biased estimators of regression coefficients.

COMPARISON OF AN ESTIMATOR WITH ITS LINEAR TRANSFORM

Let $\theta \in \Theta \subset \mathbf{R}^p$ be an unknown vector of parameters. Suppose that $\hat{\theta} \sim D(\mu,\Sigma)$ is an estimator of θ. Consider the linear transformation $A\hat{\theta}$ where $\mathbf{A} \in \mathbf{R}^{p\times p}$ is a constant matrix. The transformation of $\hat{\theta}$ given by $\hat{\theta}_A = A\hat{\theta}$ will be referred to as a linear transform.

Before going into the discussion on comparison of a given estimator of a parameter and its linear transform we will formally introduce our measure of comparison, namely the MSE as well as some terminologies associated with it (Trenkler, 1981). We represent the mean square error matrix of a biased estimator $\tilde{\beta}$ by $M(\tilde{\beta})$ while the unweighted scalar-valued risk is given as $G(\tilde{\beta})$. The weighted scalar-valued risk is $G_H(\tilde{\beta})$, where $\mathbf{H} \in \mathbf{R}^{p\times p}$ is a weight matrix.

Criterion I (strong criterion)

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of θ. $\hat{\theta}_1$ is said to be better than $\hat{\theta}_2$ with respect to criterion I, if $M(\hat{\theta}_2) - M(\hat{\theta}_1)$ is non-negative definite (n.n.d.) for $\theta \in \Theta \subset \mathbf{R}^k$. $\hat{\theta}_1$ is said to be strictly better than $\hat{\theta}_2$ with respect to criterion I if $M(\hat{\theta}_2) - M(\hat{\theta}_1)$ is positive definite (p.d.) for $\theta \in \Theta \subset \mathbf{R}^k$.

Criterion II (weak criterion)

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of θ. Let $\mathbf{H}$ be a non-stochastic n.n.d. matrix. $\hat{\theta}_1$ is said to be H-better than $\hat{\theta}_2$ with respect to criterion II, if $G_H(\hat{\theta}_2) - G_H(\hat{\theta}_1) \geq 0$ for $\theta \in \Theta \subset \mathbf{R}^k$. $\hat{\theta}_1$ is said to be strictly H-better than $\hat{\theta}_2$ with respect to

criterion II, if $G_H(\hat{\theta}_2) - G_H(\hat{\theta}_1) > 0$ for $\theta \in \Theta \subset \mathbf{R}^k$. If the matrix $\mathbf{H} = \mathbf{I}_p$, we simply say $\hat{\theta}_1$ is better (strictly better) than $\hat{\theta}_2$ with respect to criterion II.

Apparently criterion I compares $\hat{\theta}$ and $\hat{\theta}_A$ by studying the MSE matrix difference (Löwner order of matrices), and criterion II enables us to do the same by taking into account the sign of the scalar MSE difference. Since we know that an estimator cannot uniformly dominate all others from the same class, conclusions about improvement pertain to subsets of the space of parameters.

The sub-regions or conditions for $\hat{\theta}$ improvement over θ according to criterion I and criterion II, respectively, are

$$\mathbf{R}_M = (\mathbf{A}\mu-\theta)\,(\mathbf{A}\mu-\theta)' + \mathbf{A}\Sigma\mathbf{A}' - (\mu-\theta)(\mu-\theta)' - \Sigma \leq \mathbf{0}$$

where " $\leq$ " is the Löwner ordering of matrices, and

$$\mathbf{R}_G = tr(\mathbf{A}\Sigma\mathbf{A}'-\Sigma) + (\mathbf{A}\mu-\theta)'(\mathbf{A}\mu-\theta) - (\mu-\theta)'(\mu-\theta) \leq 0.$$

Matrix differentiation of R_G with respect to A gives

$$\mathbf{A}_{opt} = \theta\mu'\Sigma^{-1}(\mathbf{I}_p + \mu\mu'\Sigma^{-1})^{-1}$$

as the matrix which minimises the squared error risk. Hence, the best estimator of θ in the class of biased estimators

$$C = \{\hat{\theta}_A \mid \hat{\theta}_A = A\hat{\theta},\ E(\hat{\theta}) = \theta\}$$

is $A_{opt}\hat{\theta}$. Because the matrix $\mathbf{A}_{opt}$ is a function of the unknown parameters μ, θ and Σ the best estimator is not operational.

In the following we assume that $\hat{\theta}$ is an unbiased estimator for μ and $\mu = \theta$. Then the matrix-valued and scalar MSE of $\hat{\theta}$, respectively, are

$$M(\hat{\theta}_A) - M(\hat{\theta}) = (A\mu-\theta)(A\mu-\theta)' + A\Sigma A' - \Sigma$$

and

$$G(\hat{\theta}_A) - G(\hat{\theta}) = tr(A\Sigma A' - \Sigma) + (A\mu-\theta)'(A\mu-\theta).$$

The region of improvement of $\hat{\theta}_A$ over $\hat{\theta}$ is

$$(A-I_p)\,\mu\mu'(A-I_p)' + A\Sigma A' - \Sigma \leq 0$$

where "$\leq$" is the Löwner ordering of matrices, and

$$A_{opt} = \mu\mu'\Sigma^{-1}(I_p + \mu\mu'\Sigma^{-1})^{-1}.$$

The following statement gives a condition under which the unbiased estimator $\hat{\theta}$ of θ will be dominated by its linear transform $A\hat{\theta}$.

Theorem 1

Suppose $\hat{\theta}_A = A\hat{\theta}$ is an estimator of the vector of parameters θ, and that $\Sigma' - A\Sigma A'$ is p.d. Then the following two properties are equivalent:

(i) $\hat{\theta}_A$ is strictly better than $\hat{\theta}$ with respect to criterion I.

(ii) $\theta'(A-I_p)'[\Sigma - A\Sigma A']^{-1}(A-I_p)\theta < 1$.

The proof of the above theorem is based on a result by Farebrother (1976) which is restated as follows: Let A be an m x m-matrix, a an m-non-zero vector and let d be a positive scalar. Then $dA - aa'$ is p.d. if and only if $a'A^{-1}a < d$.

Setting $d=1$, $a=(A-I_p)\theta$ and $A=\Sigma - A\Sigma A'$ proves the assertion of the theorem.

ADMISSIBILITY OF A LINEAR TRANSFORM

Adopting the procedure and terminology in Trenkler, we give the following definition of admissibility in terms of the weak criterion.

Definition

Let **H** be an n.n.d. matrix of dimension *pxp*. Consider the class of estimators of $\theta \in \Theta$. An estimator $\hat{\theta}_*$ is said to be (**H**, Θ)-admissible if there exists no other estimator $\tilde{\theta}$ belonging to the same class such that for all $\theta \in \Theta$

$$G_H(\tilde{\theta}) \leq G_H(\hat{\theta}_*)$$

with strict inequality for at least one value of $\theta \in \Theta$. $\hat{\theta}_*$ is called Θ-admissible if it is ($\mathbf{I}_p$,Θ)-admissible.

Rao (1976) showed that ($\mathbf{I}_p$,Θ)-admissibility implies (**H**,Θ)-admissibility. The following restatement of a result due to Rao is used to establish the admissibility of a linear transform of the unbiased estimator $\hat{\theta}$.

Lemma

Let *z* be a *k*-vector random variable such that $\mathbf{E}(z) = \theta$ with dispersion matrix $D(z) = \mathbf{W}$. Assuming that **W** is non-singular the necessary and sufficient conditions for the admissibility of a transform Ly, for a r x k matrix **L**, are

(i) **LWS′** is symmetric, and
(ii) **LWS′ - LWL′** is n.n.d..

The matrix **S** of dimension *r* x *k* is given in the theorem preceding the above result in Rao, and it was used to indicate that **Lz** satisfies (**S**,Θ)-admissibility.

In the notations of this paper $\mathbf{A} = \mathbf{L}$, $\hat{\theta} = z$ and $\Sigma = \mathbf{W}$. Since we assumed that **A** is a square matrix, we take the identity matrix of dimension *p* in place of **S**. The application of the Lemma leads to the following result.

Theorem 2

Let $\hat{\theta}_A = A\hat{\theta}$ be a linear transform of $\hat{\theta}$. Then the following conditions are necessary and sufficient for the admissibility of $\hat{\theta}_A$:

(i) **A** is symmetric and it commutes with Σ, and
(ii) $A\Sigma - A\Sigma A'$ is n.n.d.

Proof: The proof in (ii) uses the Lemma. The requirement (i) is satisfied if the transformation matrix **A** and the covariance matrix Σ commute, that is $A\Sigma = \Sigma A$. Then $A\Sigma = \Sigma' A'$ proving that $A\Sigma$ is symmetric.

APPLICATION TO THE LINEAR REGRESSION MODEL

Let us consider the multiple linear regression model $M(y, X\beta, \sigma^2 I_n)$, where β is a fixed but unknown p-vector of coefficients and σ^2 is the common variance of the error terms. We assume that the *nxp* non-stochastic regression matrix **X** has full column rank. In the subsequent discussion we will use expressions related to the spectral decomposition of the regression matrix **X**. The matrices **X** can be decomposed as $X = Q\Lambda^{1/2}P'$, where **P** and **Q** are matrices of dimensions *pxp* and *nxp*, respectively. These two matrices satisfy $P'P = Q'Q = I_p$, and the columns of **P** are the orthonormal eigenvectors corresponding to the eigenvalues λ_j, j = 1,2, ..., p of **X'X**. We represent the diagonal matrix of eigenvalues by Λ.

We limit our interest to linear transforms **Ab** of the ordinary least squares estimator (OLSE), $b = (X'X)^{-1}X'y$. The linear transform **Ab** is superior to b according to criterion I in the region

$$(A\beta-\beta)(A\beta-\beta)' + \sigma^2 A(X'X)^{-1}A' - \sigma^2(X'X)^{-1} \leq 0$$

and the best estimator of β is $\hat{\beta}_{opt} = A_{opt}b$ where

$$A_{opt} = \beta\beta'X'X[\sigma^2 I_p + \beta\beta'X'X]^{-1}.$$

Since $\mathbf{A}_{opt}$ involves β and σ^2 the best estimator is unknown, and hence not operational.

Example 1 - Comparison of the ridge estimator and OLSE

The generalised ridge estimator (Hoerl and Kennard, 1970a,b) is given as

$$b(\mathbf{K})=(\mathbf{X}'\mathbf{X}+\mathbf{K})^{-1}\mathbf{X}'y, \quad \mathbf{K}=\text{diag}(k_j)\neq 0,\ k_j \geq 0.$$

Note that $b(\mathbf{K}) = \mathbf{A}b$, where $\mathbf{A} = \mathbf{P}(\Lambda+\mathbf{K})^{-1}\Lambda\mathbf{P}'$, and thus a linear transform of the ordinary least squares estimator of the vector of coefficients.

The improvement region for the generalised ridge estimator over OLSE, according to criterion I, is

$$\sigma^2[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\ \mathbf{A}'-(\mathbf{X}'\mathbf{X})^{-1}]+(\mathbf{A}-\mathbf{I}_p)\beta\beta'(\mathbf{A}-\mathbf{I}_p)' \leq 0$$

which is the same as

$$\sigma^2\mathbf{P}[(\Lambda+\mathbf{K})^{-2}\Lambda-\Lambda^{-1}]\mathbf{P}' + \mathbf{P}[(\Lambda+\mathbf{K})^{-1}\Lambda - \mathbf{I}_p]\gamma\gamma'[(\Lambda+\mathbf{K})^{-1}\Lambda-\mathbf{I}_p]\mathbf{P}' \leq 0,$$

where $\gamma=\mathbf{P}\beta$. In both expressions above "$\leq$" is the Löwner ordering of matrices.

Criterion II gives the condition of superiority of $b(\mathbf{K})$ as

$$\sigma^2 tr[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'-(\mathbf{X}'\mathbf{X})^{-1}]+\beta'(\mathbf{A}-\mathbf{I}_p)'(\mathbf{A}-\mathbf{I}_p)\beta \leq 0$$

$$\leftrightarrow \sigma^2 tr\mathbf{P}[(\Lambda+\mathbf{K})^{-2}\Lambda-\Lambda^{-1})\mathbf{P}']+\beta'\mathbf{P}[(\Lambda+\mathbf{K})^{-1}\Lambda-\mathbf{I}_p]^2\ \mathbf{P}'\beta \leq 0$$

$$\leftrightarrow -\sigma^2\sum_{j=1}^{p} k_j(k_j+2\lambda_j)/\lambda_j(\lambda_j+k_j)^2+\gamma'\text{diag}(\lambda_j/(\lambda_j+k_j)-1)^2\gamma \leq 0.$$

The ridge estimator is admissible because $A\sum=\sigma^2[P(\Lambda+K)^{-3}\Lambda^2]p'=\sum'A'$ and in addition $A\sum-A\sum A'=\sigma^2P[(\Lambda+K)^{-3}\Lambda^2-(\Lambda+K)^{-4}\Lambda^3]P'=\sigma^2P\Delta P'$ is n.n.d. because $\Delta=\text{diag}(k_j\lambda_j^2/(\lambda_j+k_j)^4)$, j=1, 2, ..., p is p.d.

The above conditions can be rewritten for the ordinary ridge estimator $b(k)=(\mathbf{X}'\mathbf{X}+k\mathbf{I}_p)^{-1}\mathbf{X}'y$, $k>0$, if k_js are replaced with k, that is $\mathbf{K} = k\mathbf{I}_p$.

Example 2 - Comparison of the shrunken estimator and OLSE

The shrunken estimator (Mayer and Willke, 1973) is given by

$$b_c = cb \quad c \in (0,1].$$

Note that $b_c = \mathbf{A}b$ with $\mathbf{A} = c\mathbf{I}_p$.

The shrunken estimator is an improvement over the least squares estimator according to criterion I if

$$(c+1)\sigma^2(\mathbf{X}'\mathbf{X})^{-1}+(c-1)\beta\beta'\geq 0,$$

and according to criterion II when

$$(1+c)/(1-c)\sum_{j=1}^{p}\lambda_j^{-1} \geq \beta'\beta/\sigma^2$$

With regard to the two criteria for admissibility, we have $\mathbf{A\Sigma}=c^3\sigma^2\mathbf{P}\Lambda^{-1}\mathbf{P}'=\mathbf{\Sigma}'\mathbf{A}'$. Furthermore, $\mathbf{A\Sigma}$ - $\mathbf{A\Sigma A}'$ = $c^3\sigma^2\mathbf{P}\Delta\mathbf{P}'$ is n.n.d. for all $c \in (0,1]$, where Δ = diag$((1-c)\lambda_j^{-1})$, j=1, 2, ..., p. Thus, we have shown that the shrunken estimator is also an admissible linear transform of b.

SUMMARY

Comparisons of the MSE risks of a vector-valued $\hat{\theta}$ of a parameter θ and a linear transform $A\hat{\theta}$ were considered. The sub-region of the parameter space where the linear transform shows superiority over the unbiased estimator $\hat{\theta}$ was derived, and also the optimal transformation matrix A given. The condition for the dominance of the linear transform was stated, and its admissibility established. Finally, two linear transforms of the OLSE, namely the shrunken estimator and the ridge estimator were taken to illustrate how the theoretical considerations in the second section of the paper can be utilised.

REFERENCES

1. Bibby, J. (1972). Minimum mean square error estimation, ridge regression and some unanswered questions. *Progress in Statistics* 1:107–121.
2. Bibby, J. and Toutenburg, H. (1977). *Prediction and Improved Estimation in Linear Models.* Wiley, Chichester.
3. Bibby, J. and Toutenburg, H. (1978). Improved estimation and prediction. *Zeitschrift für angewandte Mathematik und Mechanik* 58:45–49.
4. Farebrother, R.W. (1976). Further results on the mean square error of ridge regression. *Journal of the Royal Statistical Society* B38:248–250.
5. Hoerl, A.E. and Kennard, R.W. (1970a). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics* 12:55–67.
6. Hoerl, A.E. and Kennard, R.W. (1970b). Ridge regression: Applications to non-orthogonal problems. *Technometrics* 12:69–82.
7. Kleffe, J. (1985). Some remarks on improving unbiased estimators by multiplication with a constant. In: *Linear Statistical Inference*, pp. 150–161 (Calinski, T. and Klonecki, W., eds). Cambridge, Massachussets.
8. Mayer, L.S. and Willke, T.A. (1973). On biased estimation in linear models. *Technometrics* 15:497–508.
9. Perlman, M.D. (1972). Reduced mean square error estimation for several parameters. *Sankhya* B34:89–92.

10. Rao, C.R. (1976). Estimation of parameters in a linear model. (The Wald Memorial Lectures). *Annals of Statistics* 6:1023–1037.

11. Trenkler, G. (1981). Biased estimators in the linear regression model. In: *Mathematical Systems in Economics* No. 58. Oelgeschlager, Gunn und Hain. Cambridge, Massachussets.