Minimum mean square error estimation in linear regression

Erkki P. Liski

University of Tampere, Tampere, Finland

Helge Toutenburg

University of München, München, Germany

Götz Trenkler

University of Dortmund, Dortmund, Germany

Received 30 December 1991; revised manuscript received 27 July 1992

Abstract: Various classes of minimum mean square error (MMSE) estimators are derived in the general linear model. At first the MMSE estimator is derived within the set of all those linear estimators of $\beta$ which are at least as good as a given estimator with respect to dispersion matrix criterion. Thereafter a class of linear biased MMSE estimators, a generalization of optimal ridge estimators, is characterized. Also various techniques of deriving practical variants of MMSE estimators are introduced.

AMS Subject Classification: 62J07.

Key words and phrases: Optimal estimation; admissibility; prior information; biased estimation.

1. The minimum mean square error estimators

Consider the linear regression model

$$y = X\beta + \epsilon,$$  \hspace{1cm} (1.1)

in which $y$ is an observable $n \times 1$ random vector with expectation $E(y) = X\beta$ and with dispersion matrix $D(y) = \sigma^2 I_n$, where $X$ is a known $n \times p$ matrix, while the $p \times 1$ vector $\beta$ and $\sigma^2 > 0$ are unknown parameters. Let $\mathcal{F}$ denote the class of all linear estimators of $\beta$, i.e.,

$$\mathcal{F} = \{ \hat{\beta} = Fy + f \mid F \text{ is a } p \times n \text{ matrix and } f \text{ a } p \times 1 \text{ vector} \}. \hspace{1cm} (1.2)$$

Correspondence to: Prof. E. Liski, University of Tampere, Department of Mathematical Sciences, Statistics Unit, P.O. Box 607, SF 33101 Tampere, Finland.

0378–3758/93/$06.00 \copyright \hspace{0.1cm} 1993$—Elsevier Science Publishers B.V. All rights reserved
A linear parametric functional \( c'\beta \), where \( c \) is a \( p \times 1 \) vector, is said to be estimable in the model (1.1), if it has a linear unbiased estimator. The best linear unbiased estimator (BLUE) of an estimable \( c'\beta \) is \( c'\hat{\beta} \), where \( \hat{\beta} \) is a solution of the normal equation
\[
X'X\hat{\beta} = X'y. \tag{1.3}
\]
A general solution of (1.3) has the form
\[
\hat{\beta} = (X'X)^+ X'y + [I - (X'X)^+ X'X]k,
\]
where \( (X'X)^+ \) is the Moore–Penrose inverse of \( X'X \) and \( k \) is an arbitrary \( p \times 1 \) vector.

The solutions of (1.3) play a central role in the class of linear estimators \( \mathcal{F} \). If \( X \) has not a full rank, then \( \hat{\beta} \) is not unique and there are also linear parametric functionals \( c'\beta \) which are not estimable. Although estimation theory of linear estimable functionals is well developed, there is no unified theory for nonestimable linear functionals. Econometricians use the term multicollinearity to describe the presence of linear relationships or 'near linear relationships' among explanatory variables in linear regression (see, for instance, Judge et al., 1985, Chapter 22). The problems created by this state of affairs are well known. One obvious solution is to obtain and incorporate more information. If new data are not available, then one may supplement sample information with prior information. The most typical forms of prior information are an assumed prior distribution for the parameters or certain constraints on the values of parameters.

Suppose, for example, that we may augment the model (1.1) by a model
\[
r = R\beta + u, \tag{1.4}
\]
where \( r \) is an \( m \times 1 \) vector, \( R \) is a known \( m \times p \) matrix, \( E(u) = 0 \) and \( D(u) = \sigma^2 I_m \). The model (1.4) can be used to incorporate additional observations or prior information. If \( R \) is such that \( (X', R') \) is of full row rank, then the least squares estimator from the augmented model is
\[
\hat{\beta}_R = (X'X + R'R)^{-1}X'y + (X'X + R'R)^{-1}R'r. \tag{1.5}
\]
The estimator \( \hat{\beta}_R \) is no worse than \( \hat{\beta} \) in the Löwner sense (Löwner, 1934; Chipman, 1964; and Beckenbach and Bellman, 1965), i.e., \( D(\hat{\beta}) \leq D(\hat{\beta}_R) \) is nonnegative definite (n.n.d.). Then we write \( D(\hat{\beta}_R) \leq D(\hat{\beta}) \). If \( r = 0 \) and \( R'R = kI_p \), where \( k \) is a positive number, then formula (1.5) yields a ridge estimator (Hoerl and Kennard, 1970).

The idea of biased estimation is to introduce a small amount of bias to obtain a substantial reduction in variance so that the resulting estimator would have smaller mean square error (MSE) than \( \hat{\beta} \) in the neighbourhood of the true value of \( \beta \). Therefore the biased linear estimators of statistical interest belong to the set of estimators
\[
\mathcal{L}(\hat{\beta}) = \{Fy + f \in \mathcal{F} \mid D(Fy + f) \leq D(\hat{\beta})\}. 
\]
This leads us to consider a set of estimators of the form

\[ \mathcal{E}(Cy + c) = \{ Fy + f \in \mathfrak{F} \mid D(Fy + f) \leq D(Cy + c) \}, \]

where the upper bound of the variance of the estimators in \( \mathcal{E}(Cy + c) \) is the variance of a given estimator \( Cy + c \). The minimum mean square estimators (MMSE) in \( \mathcal{E} \) will be characterized at the end of this section. Well known biased estimators are, for example, ridge estimators (Hoerl and Kennard, 1970), shrinkage estimators (Obenchain, 1975), Marquardt’s estimator (Marquardt, 1970), linearly restricted least squares estimators (Toro-Vizcarrondo and Wallace, 1968), the principal component estimator (Massy, 1965) and the iteration estimator (Trenkler, 1978).

If a linear estimator of type (1.2) is given, its mean square error matrix is defined by

\[
M(\beta) = \sigma^2 F F' + [(FX - I_p)\beta + f][(FX - I_p)\beta + f]
\]

and its mean square error by

\[
MSE(\beta) = E[(\beta - \beta)'(\beta - \beta)] = tr M(\beta) = \text{tr} D(\beta) + \text{bias}(\beta)'\text{bias}(\beta),
\]

(1.7)

where

\[
D(\beta) = \sigma^2 F F' \quad \text{and} \quad \text{bias}(\beta) = (FX - I_p)\beta + f.
\]

**Definition 1.1.** An estimator \( \hat{\beta}^* \in \mathcal{E}(Cy + c) \) is said to be a minimum mean square error (MMSE) estimator in the class \( \mathcal{E}(Cy + c) \) if

\[
MSE(\hat{\beta}^*) \leq MSE(\beta)
\]

for any estimator \( \beta \) in \( \mathcal{E}(Cy + c) \).

A statistical interpretation of MMSE estimators is given in the beginning of the next section. MMSE estimation within the class of all linear estimators under model (1.1), when \( X \) is of full rank, has been considered by Toutenburg (1968), Rao (1971, p. 389), Theil (1971), Swamy and Mehta (1976) and Dwivedi and Srivastava (1978), for example. This approach has also been called as optimal estimation (e.g., Toutenburg, 1982) and as the best linear estimation (Rao, 1973). Rao considers MMSE estimation also in his generalized Gauss-Markov model (Rao, 1973, p. 305). Stahlecker and Trenkler (1985) investigate heterogeneous versions of optimal estimators. Minimizing (1.7) with respect to \( F \) and \( f \) leads to a trivial ‘estimator’ as noted by Toutenburg (1982, p. 46). Therefore \( f \) will be fixed in advance. Choosing an a priori value of \( \beta \), say \( b \), reflects our previous knowledge on \( \beta \). Further, if we accept a priori that \( E(\hat{\beta}) = b \), we have \( FXb + f = b \) and consequently \( f = (I - FX)b \). The usual choice \( b = 0 \) leads to the class of homogeneous linear estimators. Utilizing prior information in the context of MMSE estimation is further discussed in Section 4. Linear admissible estimators are introduced and their possible statistical applications are discussed in Section 3.
2. Derivation of the minimum mean square error estimator

Why should we restrict ourselves to the class of estimators \( \mathcal{D}(Cy + c) \)? What would be a statistical interpretation of this restriction? By Lemma 2.4 of Milliken and Akdeniz (1977) a necessary condition for \( Fy + f \in \mathcal{D}(Cy + c) \) is that \( F = CL \), where \( L \) is an \( n \times n \) matrix. Suppose that we have a fixed linear estimator, say \( Cy + c \). Then choosing an estimator \( Fy + f \) from \( \mathcal{D}(Cy + c) \) means, in fact, that \( \beta \) will be estimated from a transformed observation \( z = Ly \), i.e., \( \tilde{\beta} = CLy + f = Cz + f \). Now our aim is the following: Determine such an optimal transformation \( L^* \) which yields the minimum MSE among the estimators of the form \( CLy + f \). If \( C = X^+ \), then the problem reduces to the theory of optimal estimation (cf. Toutenburg, 1982, p. 46).

In many applications of the linear regression model the original data are not available. Nevertheless inference from the transformed data set is still possible provided the information loss is not too heavy.

We assume in the sequel that \( f \) belongs to the column space of \( I_p - FX \), which implies that \( f = (I_p - FX)b \) for some \( b \), and consequently

\[
\text{bias}(\tilde{\beta}) = (FX - I_p)(\beta - b)(\beta - b)'(FX - I_p)'.
\]

If now \( Fy + f \in \mathcal{D}(Cy + c) \), then \( \tilde{\beta} = CLy + (I_p - CLX) \). MSE(\( \tilde{\beta} \))/\( \sigma^2 \) can be written as

\[
\text{MSE}(\tilde{\beta})/\sigma^2 = \text{tr}(L'C'CL) + \text{tr}[(I_p - CLX)\tau'((I_p - CLX)\tau'],
\]

where \( \tau = (\beta - b)/\sigma \). Differentiating MSE(\( \tilde{\beta} \))/\( \sigma^2 \) with respect to \( L \) gives

\[
\frac{\partial}{\partial L} \text{MSE}(\tilde{\beta})/\sigma^2 = 2[C'C'(I_n + X\tau'X\tau') - C'\tau'X\tau'].
\]

Equating the derivative to zero and solving for \( L \) yields the set of solutions

\[
L_* = (C'C)^+C'\tau'X'(I_n + X\tau'X\tau')^{-1} + [I_n - (C'C)^+(C'C)]K,
\]

where \( K \) is an arbitrary \( n \times n \) matrix. Since MSE(\( CLy + f \)) is a convex function of \( L, L_* \) really gives the minimum (Rogers, 1980, p. 92).

The corresponding MMSE estimator is

\[
\tilde{\beta}_e^* = CL*y = P_C\tau'X'(I_n + X\tau'X\tau')^{-1}y = P_C\tau'X'y/(1 + \tau'X'X\tau), \tag{2.1}
\]

where \( P_C = C(C'C)^+C' \) is the orthogonal projector onto \( \mathcal{M}(C) \), the column space of \( C \). Observe that \( \tilde{\beta}_e^* \) does not depend on \( K \), i.e., we may put \( K = 0 \). Note, however, that \( \tilde{\beta}_e^* \) does not necessarily belong to \( \mathcal{D}(Cy + c) \). If \( \tau \in \mathcal{M}(C) \), then

\[
\tilde{\beta}_e^* = \tau'X'(I_n + X\tau'X\tau')^{-1}y = \tau'X'y/(1 + \tau'X'X\tau). \tag{2.2}
\]

Especially, if \( C \) has full row rank, then \( \tilde{\beta}_e^* \) is of the form (2.2).
Theorem 1.1. The estimator (2.1) is the MMSE estimator in \( \mathcal{D}(Cy + c) \) if and only if

\[
\tau'X'X\tau'(CC')^+ \tau \leq (1 + \tau'X'X\tau)^2.
\] (2.3)

Proof. By the statement (d) of Proposition 1 (Baksalary, Liski and Trenkler, 1989) \( \hat{\beta}_2 \) is in \( \mathcal{D}(Cy + c) \) when \( \lambda_1(L_*^*P_CL_*) \leq 1 \), where \( \lambda_1(\cdot) \) denotes the largest eigenvalue of a matrix. Using the inversion formula for a sum of two matrices (Rao, 1973, p. 33) we have

\[
\tau'X'(I_n + X\tau\tau'X')^{-1} = \tau'X'/(1 + \tau'X'X\tau).
\]

Thus we obtain the identity

\[
L_*^*P_CL_* = (1 + \tau'X'X\tau)^{-2}X\tau\tau'(CC') + \tau\tau'X',
\]

since \( P_C = C^+ \) and \( (C^+)^'C^+ = (CC')^+ \). It follows readily from the above expression that \( \text{rank}(L_*^*P_CL_*) = 1 \) and

\[
\lambda_1(L_*^*P_CL_*) = (1 + \tau'X'X\tau)^{-2}\tau'X'X\tau\tau'(CC')^+\tau,
\]

which is not greater than 1 if and only if the inequality (2.3) holds. \( \square \)

It is of special interest to determine the MMSE estimator in \( \mathcal{D}(\hat{\beta}) \), where \( \hat{\beta} = X^*y \) is the least squares estimator of \( \beta \). The MMSE estimator in \( \mathcal{D}(\hat{\beta}) \) can be written as

\[
\hat{\beta}^* = X^*\tau\tau'X'(I_n + X\tau\tau'X')^{-1}y = X^*\tau\tau'X'y/(1 + \tau'X'X\tau),
\] (2.4)

since now \( P_C = X^*X \). Note that in this case the left-hand side of the inequality (2.3) reduces to \( (\tau'X'X\tau)^2 \), and consequently \( \hat{\beta}^* \) really belongs to \( \mathcal{D}(\hat{\beta}) \). When \( X \) is of full rank, then \( X^*X = I \) and (2.4) takes the form

\[
\hat{\beta}^* = \tau\tau'X'(I_n + X\tau\tau'X')^{-1}y,
\] (2.5)

which is called also as the optimal estimator (cf. Toutenburg, 1982, p. 46).

In fact the MMSE ‘estimators’ are no estimators in true sense, since they depend on the unknown parameters \( \beta \) and \( \sigma^2 \). One way to obtain practical estimators is to choose a prior value for \( \tau = (\beta - b)/\sigma \), say \( t \), based on our previous knowledge, and substitute \( \sigma^2W = \sigma^2tt' \) for \( \tau\tau' \) (cf. Rao, 1973, p. 305). Further, \( \beta \) can be considered also as a random variable having a prior dispersion matrix \( E(\beta - b)(\beta - b)' = \sigma^2W \), or the choice \( W \) in MSE(\( \hat{\beta} \)) can simply represent the relative weight we attach to bias compared to variance (cf. Rao, 1973, p. 305). The matrix \( W \) is n.n.d. and may have rank greater than one, when it is interpreted as a dispersion matrix or as a weight matrix. However, if the unknown value \( \tau \) is replaced by a given \( t \), the estimators (2.1), (2.4) and (2.5) are no more MMSE estimators.
3. A generalization of optimal ridge estimators

Since we search for linear estimators with small MSE, we may appropriately focus our study to the class of admissible linear estimators. Biased estimators like ridge estimators belong to this class.

**Definition 3.1.** A linear estimator $Ay + a$ is admissible for $\beta$ among the set of linear estimators under the model (1.1) if there does not exist an estimator $Fy + f \in \mathcal{F}$ such that the inequality

$$\text{MSE}(Fy + f) \leq \text{MSE}(Ay + a)$$

holds for all $p \times 1$ vectors $\beta \in \mathbb{R}^p$ and $\sigma^2 > 0$, and is strict for at least one such point $(\beta_0, \sigma_0^2)$.

The set of linear admissible estimators for $\beta$ among $\mathcal{F}$ under the model (1.1) has been originally characterized by Cohen (1966) and Rao (1976). The following lemma is a direct consequence of Theorem 1 in Liski (1988) (c.f. also Theorem 3 in Baksalary, Liski and Trenkler, 1989).

**Lemma 3.1.** Suppose that $X$ in the model (1.1) is of full rank and that

$$X = P\Lambda Q'$$

is the singular value decomposition of $X$ with $P'P = Q'Q = I_p$ and $A = \text{diag}(\lambda_1, ..., \lambda_p)$. Then $Ay + a$ is admissible for $\beta$ if and only if

$$A = X^+ V \Delta V'$$

and

$$a \in \mathcal{M}(I - X^+ V \Delta V'X),$$

(3.1a)

(3.1b)

where $V$ is an $n \times p$ matrix with $V'V = I_p$, $\mathcal{M}(V) = \mathcal{M}(P) = \mathcal{M}(X)$ and $\Delta = \text{diag}(\delta_1, \delta_2, ..., \delta_p)$ is any $p \times p$ diagonal matrix such that $0 \leq \Delta \leq I_p$.

Suppose now that we have fixed an $n \times p$ matrix $V$ that satisfies the conditions of Lemma 3.1. In some cases prior information about $\beta$ can be utilized in determining $V$, but this problem will be discussed in Section 4. Let

$$\mathcal{A}_V = \{X^+ V \Delta V'y + (I - X^+ V \Delta V'X)b \in \mathcal{F} \mid b \in \mathbb{R}^p \text{ and } A \in \mathbb{R}^{p \times p}\},$$

(3.2)

where $\Delta$ is a $p \times p$ diagonal matrix. If $\Delta$ is restricted to the set $\{\Delta \mid 0 \leq \Delta \leq I\}$, then the estimators in $\mathcal{A}_V$ are admissible.

**Theorem 3.1.** Let $X$ be of full rank and let $V$ be a given $n \times p$ matrix such that $\mathcal{M}(V) = \mathcal{M}(X)$ and $V'V = I_p$. Then the MMSE estimator of $\beta$ among $\mathcal{A}_V$ is

$$\hat{\beta}_V^0 = X^+ V \Delta V'y + (I - X^+ V \Delta V'V'y)b,$$
where \( \Delta y = \text{diag}(\delta_1^\gamma, \ldots, \delta_p^\gamma) \) and
\[
(\delta_1^\gamma, \ldots, \delta_p^\gamma) = m' [G \circ (I_p + B)]^{-1}
\] (3.3)
in which \( m' = (v_1^t X \tau \tau^t X^+ v_1, \ldots, v_p^t X \tau \tau^t X^+ v_p) \), \( G = (X^+ V)^t X^+ V \), \( B = V^t X \tau \tau^t X^t V \), and \( G \circ B = [g_{ij} b_{ij}] \) is the Hadamard product of \( G \) and \( B \).

**Proof.** The MSE-function of an estimator \( \tilde{\theta} \in \mathcal{S} \) can be written as
\[
\text{MSE}(\tilde{\theta})/\sigma^2 = \text{tr}(AGA) + \text{tr} \left\{ (I - X^+ V \Delta V^t X) \tau \tau^t (I - X^+ V \Delta V^t X)^t \right\}.
\] (3.4)
Differentiating (3.4) with respect to \( \delta_i \) yields
\[
\frac{\partial}{\partial \delta_i} \text{MSE}(\tilde{\theta})/\sigma^2 = 2g_{ii} \delta_i + 2g_{ij} \Delta b_{ij} - 2v_i^t X \tau \tau^t X^+ v_i, \quad i = 1, \ldots, p,
\]
in which \( g_i \) and \( b_i \) are the \( i \)-th columns of \( G \) and \( B \), respectively. The conditions \((\partial/\partial \delta_i)\text{MSE}(\tilde{\theta})/\sigma^2 = 0, i = 1, \ldots, p,\) imply the equations
\[
g_{11} \delta_1^\gamma + \sum_{j=1}^p g_{1j} b_{1j} \delta_j^\gamma = v_1^t X \tau \tau^t X^+ v_1,
\]
\[
g_{22} \delta_2^\gamma + \sum_{j=1}^p g_{2j} b_{2j} \delta_j^\gamma = v_2^t X \tau \tau^t X^+ v_2,
\]
\[
\vdots
\]
\[
g_{pp} \delta_p^\gamma + \sum_{j=1}^p g_{pj} b_{pj} \delta_j^\gamma = v_p^t X \tau \tau^t X^+ v_p,
\]
which can be written in the matrix form
\[
[G \circ I_p + G \circ B] \delta_v = m,
\]
where \( \delta_v^\gamma = (\delta_1^\gamma, \ldots, \delta_p^\gamma) \), \( G = [g_{ij}] = [v_i^t (X X^t)^+ v_j] \) and \( B = [b_{ij}] = [v_i^t X \tau \tau^t X^t v_j] \).

Clearly both \( G \) and \( B \) are nonnegative definite. Use the singular value decomposition \( X = P \Lambda Q^t \) to write \( G = V' P \Lambda^{-2} P' V \). Since \( \mathcal{M}(V) = \mathcal{M}(X) = \mathcal{M}(P), \) \( V' P \) is nonsingular. Hence \( G \) is positive definite. By the Schur product theorem (Horn and Johnson, 1985, p. 458) \( G \circ I_p \) is positive definite and \( G \circ B \) is nonnegative definite. Therefore \( G \circ I_p + G \circ B = G \circ (I_p + B) \) is positive definite, and so is also \([G \circ (I_p + B)]^{-1} \). Thus the solution (3.3) is established. Applying a reasoning similar to that used in Section 2 we see that the minimum is attained at the point \( \delta_v^\gamma \).

In many applications of the model (1.1) only transformed observations \( z = Ly \), instead of the original data \( y \), are available. Suppose, for example, that \( z \) consists of certain mean values \( Jy \) calculated from the original \( y \). Then
is a block diagonal matrix, where \( n_i J_i \) is an \( n_i \times n_i \) matrix of one's, \( n_1 + n_2 + \cdots + n_k = n \) and \( p < k \). Let \( L = J \) and note that \( J \) has a decomposition \( VV' \), where \( V \) is such an \( n \times k \) matrix that \( V'V = I_k \). If we have observed only the vector \( z = Jy = VV'y \), then \( V'z = V'y \). Instead of the original projector \( VV'y \) we may now consider the predictor

\[
X\hat{\beta}_V^0 = XX^+V\Delta_0 V'y = V\Delta_0 V'y
\]

corresponding to the estimator \( \hat{\beta}_V^0 = X^+V\Delta_0 V'y \), where the diagonal matrix \( \Delta_0 \) can be obtained from Theorem 3.1. Consequently, this is a natural application for the estimators \( \hat{\beta}_V^0 \).

**Corollary 3.1.** Let \( X = PAQ' \) be the singular value decomposition of \( X \). If \( V = P \), then

\[
\delta_i^P = \frac{\lambda_i^2 \alpha_i^2}{1 + \lambda_i^2 \alpha_i^2}, \quad i = 1, \ldots, p,
\]

in which \( \tau' Q = (\alpha_1, \ldots, \alpha_p) \) and \( A = \text{diag}(\lambda_1, \ldots, \lambda_p) \).

**Proof.** It can be easily verified by direct calculation that \( g_{ii} = \lambda_i^2 \) and \( g_{ij} = 0 \) when \( i \neq j \). Similarly we obtain \( b_{ii} = \alpha_i^2 \lambda_i^2, p_i'X\tau'X^+p_i - \alpha_i^2 \) and \( b_{ij} = 0 \) (when \( i \neq j \)), where \( p_i \) is the \( i \)-th column of \( P \). \( \square \)

If we choose \( V = P \), then the resulting set \( \mathcal{A}_P \) is the class of optimal ridge estimators introduced by Obenchain (1975) and formula (3.5) yields the optimal ridge factors (Obenchain, 1978, Theorem 2).

In some economic applications subjective expectations are used as prior information for parameters. We have consulted a group of experts, for example, and each of them has expressed a forecast on the value of \( y \) for given values of explanatory variables (Bates and Granger, 1969). Suppose that linear restrictions on the expectation \( E(y/\sigma) = X\beta/\sigma = X\tau \) can be introduced so that

\[
H'X\tau = d,
\]

where \( H = (h_1, \ldots, h_p) \) is an \( n \times p \) matrix with \( \mathcal{M}(H) = \mathcal{M}(X) \) and \( d' = (d_1, \ldots, d_r) \). Let \( H = V\Omega Z' \) be the singular value decomposition of \( H \) and \( V = PU \) with \( UU' = I_p \). Then we have

\[
V'X\tau = U'P'X\tau = U'\Omega Z' \tau = U' \Lambda \alpha = \Omega^{-1}Z'd,
\]
and consequently $a = A^{-1}U\Omega^{-1}Z'd$. Now this prior estimate can be used in constructing the optimal ridge factors given in Corollary 3.1.

Also Trenkler (1984) considered MMSE ridge estimation. In fact all the biased estimators mentioned in Section 1, except the restricted least squares estimator, are included in $A_p$. Note that by (3.2) the generalized ridge estimators can also be expressed in the form

$$A_p = \left\{ Q\Delta_Q'\hat{\beta} + (I - Q\Delta_Q')b \mid b \in \mathbb{R}^p \text{ and } 0 \leq \Delta \leq I \right\}.$$ 

For example the usual ridge estimator is obtained by choosing $A = A^2(A^2 + kI)^{-1}$, where $k$ is a nonnegative scalar. Practical estimators are again obtained by substituting $\tau = t$ into (3.3). Goodness of the resulting estimator, measured with MSE, depends on the choice of $t$ reflecting the precision of our prior knowledge. However, the admissibility of the estimator is independent of the choice $t$. In fact, we can apply the formula (3.3) also in the case when $X$ is not of full rank. Then $A^{-1}$ should be replaced with $A^+$. Estimation of nonestimable linear functions of $\beta$ is considered in the next section.

4. Further considerations

Let us now investigate the linear estimators $\tilde{\beta} = CLy + f$ of $K\beta$, where $K$ is a given $q \times p$ matrix, $C$ a $q \times n$ matrix and $f$ a $q \times 1$ vector. We compare estimators with respect to the weighted mean square error

$$\text{MSE}(\tilde{\beta}; W) = E[(\tilde{\beta} - K\beta)'W(\tilde{\beta} - K\beta)],$$

where $W = T'T$ is a nonnegative matrix. Therefore, it does not mean an essential restriction to use the ordinary mean square error

$$\text{MSE}(\tilde{\beta})/\sigma^2 = \text{tr}(L'C'CL) + \text{tr}[(K - CLX)\tau\tau'(K - CLX)'],$$

since weighting can always be taken into account by substituting $C := TC$, $f := Tf$ and $K := TK$. Note that in (4.2) $\tau = (K\beta - k)/\sigma$, where $k$ is a prior value of $K\beta$. Since $X$ is not necessarily of full rank, $K\beta$ may be nonestimable. We may further generalize the situation by assuming the generalized Gauss-Markov model (GMM) introduced by Rao (cf. Rao, 1973, p. 294). In this model $D(y) = \sigma^2D$, where $D$ can also be singular. In this general case the MMSE estimator of $K\beta$ is (cf. Rao, 1973, p. 306)

$$\hat{\beta}_s = PC\tau\tau'X'(D + X\tau\tau'X)^{-1}y.$$ 

The MMSE estimator exists whether $K\beta$ is estimable or not. Estimation of non-estimable parametric functions of $\beta$ have been considered by Chipman (1964), by Rao (1973, p. 306) and by Majumdar and Mitra (1978), for example. The most well
known method here is the minimum bias estimation due to Chipman (1964). Other approaches to estimation of nonestimable parametric functions utilizes usually certain restrictions introduced in the parametric space (cf. Majumdar and Mitra, 1978).

We have shown in Section 2 that the MMSE estimator deviates from the optimal estimator (2.5) only when \( C \) is not of full row rank and \( \tau \notin \mathcal{M}(C) \). When we have the principal component (PC) estimator (see Gunst and Mason, 1977, p. 618), for example, \( C \) is not of full row rank. Thus the MMSE estimator for the PC estimator is not equal to the usual optimal estimator. As was stressed in the end of Section 2, the MMSE estimator is not a true estimator and it cannot be directly applied in practice. On the other hand, the derived practical estimators do not have any minimum MSE properties and they are not linear estimators.

To make the optimal estimator operational, Farebrother (1975) proposed to replace the parameters \( \beta \) and \( \sigma^2 \) by their estimates \( \hat{\beta} \) and \( \hat{\sigma}^2 \). As soon as \( b \) is fixed, we have an estimate \( \hat{\tau} \) of \( \tau \). Following Farebrother, the adaptive version of the MMSE estimator can be defined as

\[
\hat{\beta}_a = aP_C \hat{\tau}.
\]

where \( a = \hat{\tau}'X'(I_n + X\hat{\tau}'X')^{-1}y \). Now \( P_C \hat{\tau} \) is the orthogonal projection of \( \hat{\tau} \) onto \( \mathcal{M}(C) \). In the case of the PC estimator \( P_C \) is the orthogonal projector onto the subspace spanned by the eigenvectors of \( X'X \) corresponding to the \( p_1 \) (\( p_1 < p \)) largest eigenvalues of \( X'X \). The \( p - p_1 \) smallest eigenvalues are then usually considered to be close to zero.

Dwivedi and Srivastava (1978) derived the large sample asymptotic approximations for the bias vector and for the MSE matrix of the Farebrother estimator. They also compared its performance with that of \( \hat{\beta} \). Vinod (1976) generalized Farebrother’s idea by introducing two iterative adaptive estimation procedures, which can also be applied to the MMSE estimation. According to Vinod’s simulation results his estimator \( FPV^* \) (and Stein’s positive part rule) did well over a wide range of experimental settings. It would be interesting to study Vinod’s estimator more closely in the context of the MMSE estimation. Note however, that the superiority of an estimator over another depends, in general, on the structure of \( X'X \) and on the values of the unknown parameters \( \beta \) and \( \sigma \).

The successful application of an adaptive MMSE estimator requires some a priori information on parameters. Therefore the blind use of an estimator cannot be recommended. Fortunately, in most practical applications a priori knowledge is available, although it is not always easy to incorporate this information into the estimation procedure. An extensive and excellent discussion on the use of a priori information in econometric applications can be found in Stahlecker (1987, Chapters 3 and 4). However, there exists no comprehensive and easily interpretable comparative study on the practical estimators based on the MMSE estimation. In Section 3 we gave an example, where a priori knowledge motivated the use of the
generalized optimal ridge estimator. Also this estimator can be made practical by estimating the parameter vector \((\delta^*_1, \ldots, \delta^*_p)\). Hemmerle (1975) studied the operational generalized optimal ridge type estimator in the special case when \(V = P\).

**Acknowledgements**

The authors are grateful to the Editor and to the referees for their helpful suggestions. We also thank H. Knautz for his critical comments. The third author was supported by the Deutsche Forschungsgemeinschaft (DFG), grant number TR 253/1-2, which is gratefully acknowledged.

**References**


