On the Performance of the Ordinary Least Squares Method under an Error Component Model

P. Mukhopadhyay
Indian Statistical Institute
Calcutta

Rainer Schwabe
University of Technology
Darmstadt

Abstract Starting from the one–dimensional results by Wang et al (1994) we consider the performance of the ordinary least squares estimator in comparison to the best linear unbiased estimator under an error component model with random effects in units and time. Upper bounds are derived for the first–order approximation to the difference between both estimators and for the spectral norm of the difference between their dispersion matrices.

Key words Ordinary least squares estimate. Best linear unbiased estimate. Euclidean norm. Spectral norm. First–order approximation.

1. Introduction.

Throughout the last decades the relative performance has received considerable attention which is attained by the ordinary least squares estimator (OLSE) in comparison to the best linear unbiased estimator (BLUE) of the parameters in a linear model. For example, Bloomfield and Watson (1975) and Knott (1975) evaluated the efficiency of the generalised variance, i. e. the ratio of the determinants of the dispersion matrices of the BLUE and the OLSE. In the presence of autocorrelated errors the efficiency was examined by Krämer (1980). Baksalary and Kala (1980) derived upper bounds for the Euclidean norm of the difference between the OLSE and BLUE. The effect of cluster sampling on the OLSE was investigated by Scott and Holt (1982) and Christensen (1984). More recently, Wang et al (1994) studied the performance of the OLSE by comparing it with the BLUE in the context of two–stage sampling from a finite population. There the observations belonging to the same cluster are assumed to have a constant correlation, the so–called interclass correlation. In this note, starting from the results by Wang et al (1994) we consider the performance of the OLSE relatively to the BLUE in the context of the following error component model with random effects in units and time.

Let $\mathcal{P} = \{1, \cdots, N\}$ be a finite population of a known number N of identifiable units labeled $1, \cdots, N$ and 'y' be a study variable. We are interested in the characteristics of this population over a number of time points $t = 1, \cdots, T$. It is assumed that apart from y we have a set of auxiliary variables 'x_j' $(j = 1, \cdots, p)$, so–called covariates, closely related to y which may also vary in time t. Associated with the unit i we have, therefore, a vector of real numbers $(y_{it}, x_{1it}, x_{2it}, \cdots, x_{pit})$ where y_{it} and x_{jit} are the values of the observations y and the covariates x_j, respectively, on unit i at time t.

For example, we may have N factories belonging to an industrial group, y_{it} being the value of output (value added by manufacture) of the ith factory in the year (at the time point) t. Then $x_1, x_2, \ldots$ may be a set of auxiliary variables, like value of raw materials, the number of workers, $x_{1it}, x_{2it}, \ldots$ being the values of the raw materials (number of workers) for the ith factory in the year t, etc. We consider the following error component model: y_{it} is a realisation of a random variable Y_{it} which follows a superpopulation model distribution for given values of x_{jit} $(j = 1, \cdots, p)$.

$$Y_{it} = \alpha + \sum_{j=1}^{p} \beta_j x_{jit} + \varepsilon_{it} \quad \begin{cases} i = 1, \cdots, N \\ t = 1, \cdots, T \end{cases} \tag{1}$$

where α is a constant (intercept parameter), β_j $(j = 1, \cdots, p)$ is a set of unknown regression coefficients (slope parameters) and ε_{it} is a random error which can be additively decomposed according to

$$\varepsilon_{it} = u_i + v_t + w_{it} \tag{2}$$

into its components u_i, v_t and w_{it} which are associated with the units i, the time points t and the particular observations, respectively. The auxiliary variables are assumed to result in fixed effects $\beta_1, \ldots, \beta_p$ which are the parameters of interest.

The present note provides two–dimensional extensions of the (first–order) approximations obtained by Wang et al (1994) and of their bounds on the spectral norm of the loss in the dispersion matrix when the OLSE is used instead of the BLUE. As a by–product the latter bound is improved also for the one–dimensional situation. Generalisations to higher–order approximations and to more than two variance components are indicated in the concluding Section 4.

2. Assumptions and notations.

Besides the usual requirement of zero mean errors $E(u_i) = E(v_t) = E(w_{it}) = 0$ for the components of ε we will assume that these variance components are homoscedastic

$$\sigma_u^2 = \mathrm{Var}\,(u_i),\ \sigma_v^2 = \mathrm{Var}\,(v_t),\ \sigma_w^2 = \mathrm{Var}\,(w_{it}) \tag{3}$$

and that different components and observations are uncorrelated, i. e. $E(u_i u_j) = E(v_s v_t) = E(w_{is} w_{jt}) = E(u_i v_t) = E(u_i w_{jt}) = E(v_s w_{it}) = 0$. A similar error component model has been considered by Fuller and Battese (1973). The object of our further statistical investigations will be the vector $\beta = (\beta_1, \cdots, \beta_p)'$ of unknown parameters related to the auxiliary variables.

If we denote by

$$Y = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1T} \\ Y_{21} \\ \vdots \\ Y_{NT} \end{bmatrix},\ X = \begin{bmatrix} x_{111} & \cdots & x_{p11} \\ \vdots & & \vdots \\ x_{11T} & \cdots & x_{p1T} \\ x_{121} & \cdots & x_{p21} \\ \vdots & & \vdots \\ x_{1NT} & \cdots & x_{pNT} \end{bmatrix} \text{ and } \varepsilon = \begin{bmatrix} u_1 + v_1 + w_{11} \\ \vdots \\ u_1 + v_T + w_{1T} \\ u_2 + v_1 + w_{21} \\ \vdots \\ u_N + v_T + w_{NT} \end{bmatrix} \tag{4}$$

the vector of observations, the matrix of covariates and the vector of errors, respectively, then the model can be written in matrix notation as

$$Y = \alpha 1_{NT} + X\beta + \varepsilon, \tag{5}$$

where 1_m denotes a m–dimensional vector of ones.

Let

$$A = I_N \otimes J_T \text{ and } B = J_N \otimes I_T, \tag{6}$$

where I_m is a $m \times m$ identity matrix, $J_m = 1_m\, 1'_m$ is a $m \times m$ matrix with all entries equal to one and '$\otimes$' denotes the Kronecker product of matrices. With this notation the dispersion matrix $\mathrm{Cov}\,(\varepsilon) = \mathbf{\Sigma}$ of the error vector ε becomes

$$\mathbf{\Sigma} = \sigma_w^2 I_{NT} + \sigma_u^2 A + \sigma_v^2 B. \tag{7}$$

Note that $\mathbf{\Sigma}$ is regular if $\sigma_w^2 > 0$. Moreover, it can be verified that

$$\mathbf{\Sigma}^{-1} = \frac{1}{\sigma_w^2}(I_{NT} - \phi_1 A - \phi_2 B + \phi_3 J_{NT}) \tag{8}$$

where $\phi_1 = \frac{\sigma_u^2}{\sigma_w^2 + T\sigma_u^2}$, $\phi_2 = \frac{\sigma_v^2}{\sigma_w^2 + N\sigma_v^2}$ and $\phi_3 = \frac{\sigma_u^2\sigma_v^2}{(\sigma_w^2 + T\sigma_u^2)(\sigma_w^2 + N\sigma_v^2)}[1 + \frac{\sigma_w^2}{\sigma_w^2 + N\sigma_v^2 + T\sigma_u^2}]$.

If we define by $\sigma^2 = \mathrm{Var}\,(\varepsilon_{it}) = \sigma_u^2 + \sigma_v^2 + \sigma_w^2$ the total variance of each observation and by $\rho = \frac{\sigma_u^2}{\sigma^2}$ and $\delta = \frac{\sigma_v^2}{\sigma^2}$ the proportions of the variance components associated with the units and the time points, respectively, then the dispersion matrix $\mathbf{\Sigma}$ can be written as $\mathbf{\Sigma} = \sigma^2\Sigma(\rho,\delta)$, where the standardised dispersion $\Sigma(\rho,\delta)$ is given by

$$\Sigma(\rho,\delta) = (1-\rho-\delta)I_{NT} + \rho A + \delta B. \tag{9}$$

We may assume without loss of generality that the auxiliary variables are centered, i. e.

$$\textstyle\sum_i \sum_t x_{jit} = 0 \text{ for each } j(=1,\cdots,p). \tag{10}$$

This may be accomplished by a linear reparametrisation of the model which leaves the parameters $\beta = (\beta_1,\cdots,\beta_p)'$ unchanged (see e. g. Schwabe, 1996, p. 16). Note, however, that this transformation of X may produce an altered value for the overall mean α in the model equation (5) which will not be investigated in the sequel.

We assume further that $\mathrm{rank}\,(X) = p < NT$, i. e. $[1_{NT}\,X]$ has full column rank $p+1$. Then, under the model specified by (5) and (7), when X satisfies the centering condition (10),

$$\widehat{\beta}^* = (X'\mathbf{\Sigma}^{-1}X)^{-1}X'\mathbf{\Sigma}^{-1}Y \tag{11}$$

is the best linear unbiased estimator (BLUE) of β. As $1_{NT}'\mathbf{\Sigma}^{-1}X = 0$ in view of (8) and (10) the dispersion matrix of $\widehat{\beta}^*$ is

$$\mathrm{Cov}\,(\widehat{\beta}^*) = (X'\mathbf{\Sigma}^{-1}X)^{-1} = \sigma_w^2[X'(I_{NT} - \phi_1 A - \phi_2 B)X]^{-1}. \tag{12}$$

Note that under the above conditions also $\widehat{\alpha} = \overline{Y} = \frac{1}{NT}\sum_i\sum_t Y_{it}$ is the BLUE of α and that the estimators $\widehat{\alpha}$ and $\widehat{\beta}^*$ are uncorrelated.

The calculation of the BLUE $\widehat{\beta}^*$ requires the knowledge of the ratios of the variance components σ_u^2, σ_v^2 and σ_w^2. As an alternative for estimating β one can therefore choose the ordinary least square estimator (OLSE) $\widehat{\beta}_0$ which is given by

$$\widehat{\beta}_0 = (X'X)^{-1}X'Y, \tag{13}$$

when X satisfies the centering condition (10).

To compare the performance of the OLSE $\widehat{\beta}_0$ with the BLUE $\widehat{\beta}^*$ we introduce a whole class of unbiased estimators

$$\widehat{\beta}(\tilde{\rho},\tilde{\delta}) = \left(X'\,\Sigma(\tilde{\rho},\tilde{\delta})^{-1}X\right)^{-1} X'\,\Sigma(\tilde{\rho},\tilde{\delta})^{-1}Y \tag{14}$$

parametrised by $\tilde{\rho}$ and $\tilde{\sigma}$ which covers both estimators of interest as $\widehat{\beta}^* = \widehat{\beta}(\rho, \delta)$ and $\widehat{\beta}_0 = \widehat{\beta}(0,0)$.

3. The OLSE $\widehat{\beta}_0$ as an approximation to the BLUE $\widehat{\beta}^*$.

In the present main section we develop bounds on the deviation of the OLSE from the BLUE.

Theorem 1.

$$\widehat{\beta}^* = \widehat{\beta}_0 - (X'X)^{-1}X'(\rho A + \delta B)(Y - X\widehat{\beta}_0) + O(\rho^2 + \delta^2). \tag{15}$$

Remark. The notation "O" indicates a pointwise approximation of order $O(\rho^2+\delta^2)$, i. e. the remainder term is bounded by $c(\rho^2+\delta^2)$ for some (random) constant c if ρ and δ tend to 0 simultaneously.

Proof of Theorem 1. Expanding $\widehat{\beta}^* = \widehat{\beta}(\rho, \delta)$ in a bivariate Taylor series at $\widehat{\beta}_0 = \widehat{\beta}(0,0)$ we obtain

$$\widehat{\beta}(\rho,\delta) = \widehat{\beta}(0,0) + \rho \left. \frac{\partial \widehat{\beta}(\rho,\delta)}{\partial \rho} \right|_{\rho=0,\delta=0} + \delta \left. \frac{\partial \widehat{\beta}(\rho,\delta)}{\partial \delta} \right|_{\rho=0,\delta=0} + O(\rho^2+\delta^2). \tag{16}$$

We make use of the differentiation rules $\frac{\partial C(\theta)D(\theta)}{\partial \theta} = \frac{\partial C(\theta)}{\partial \theta} D(\theta) + C(\theta)\frac{\partial D(\theta)}{\partial \theta}$ and $\frac{\partial C(\theta)^{-1}}{\partial \theta} = C(\theta)^{-1}\frac{\partial C(\theta)}{\partial \theta}C(\theta)^{-1}$, etc. for matrix valued functions C and D (see e. g. Magnus and Neudecker, 1988). Then

$$\frac{\partial (X'\Sigma(\rho,\delta)^{-1}X)^{-1}}{\partial \rho} = (X'\Sigma(\rho,\delta)^{-1}X)^{-1}X'H(\rho,\delta)X(X'\Sigma(\rho,\delta)^{-1}X)^{-1}, \tag{17}$$

where

$$H(\rho,\delta) = \frac{\partial \Sigma^{-1}}{\partial \rho} = \Sigma^{-1}\frac{\partial \Sigma}{\partial \rho}\Sigma^{-1} = \Sigma(\rho,\delta)^{-1}(I_{NT} - A)\Sigma(\rho,\delta)^{-1}$$

and, consequently, in view of (14),

$$\begin{aligned} \frac{\partial \widehat{\beta}}{\partial \rho} &= (X'\Sigma^{-1}X)^{-1}(X'\frac{\partial \Sigma^{-1}}{\partial \rho}\, Y) - (X'\Sigma^{-1}X)^{-1}\{\frac{\partial}{\partial \rho}(X'\Sigma^{-1}X)\}\widehat{\beta} \\ &= (X'\Sigma^{-1}X)^{-1}X'H(Y - X\widehat{\beta}). \end{aligned} \tag{18}$$

Similarly, for the derivative with respect to δ we obtain

$$\frac{\partial \widehat{\beta}}{\partial \delta} = (X'\Sigma^{-1}X)^{-1}X'M(Y - X\widehat{\beta}). \tag{19}$$

where

$$M(\rho,\delta) = \frac{\partial \Sigma^{-1}}{\partial \delta} = \Sigma^{-1}\frac{\partial \Sigma}{\partial \delta}\Sigma^{-1} = \Sigma(\rho,\delta)^{-1}(I_{NT} - B)\Sigma(\rho,\delta)^{-1}$$

Note that $\Sigma(0,0) = I_{NT}$ and, hence, $H(0,0) = I_{NT} - A$ and $M(0,0) = I_{NT} - B$. It follows that

$$\begin{aligned}\widehat{\beta}^* &= \widehat{\beta}_0 + \rho(X'X)^{-1}X'(I_{NT} - A)(Y - X\widehat{\beta}_0) \\ &\quad + \delta(X'X)^{-1}X'(I_{NT} - B)(Y - X\widehat{\beta}_0) + O(\rho^2 + \delta^2) \\ &= \widehat{\beta}_0 - (X'X)^{-1}X'(\rho A + \delta B)(Y - X\widehat{\beta}_0) + O(\rho^2 + \delta^2) \qquad (20)\end{aligned}$$

since $X'(Y - X\widehat{\beta}_0) = 0$. □

If we ignore higher order terms, the difference $\widehat{\beta}_0 - \widehat{\beta}^*$ between the OLSE and the BLUE can be approximated by $(X'X)^{-1}X'(\rho A + \delta B)(Y - X\widehat{\beta}_0)$. Note that the OLSE and the BLUE coincide, $\widehat{\beta}_0 = \widehat{\beta}^*$, if $(\rho A + \delta B)X = XS$ for some $p \times p$ matrix S (see e. g. Zyskind, 1967). In particular, this holds if $AX = BX = 0$.

Example. The performance of the OLSE depends on the interactions AX and BX of the auxiliary variable with the dispersion structure generated by the variance components. For illustrative purposes we consider the situation of 4 observations, $N = T = 2$, and a single auxiliary variable, $p = 1$. The auxiliary variable is assumed to be available at levels 1, -1, 0 and 0, thus satisfying condition (10). Now, there are essentially three different assignments of the auxiliary variable possible resulting in the matrices of covariates $X = (1, -1, 0, 0)'$, $X = (1, 0 - 1, 0)'$ and $X = (1, 0, 0, -1)'$. In the first case the difference $\widehat{\beta}_0 - \widehat{\beta}^*$ is approximately equal to $\frac{1}{2}\delta(Y_{21} - Y_{22})$ and the performance depends on the correlation at fixed time points. In the second case the difference equals approximately $\frac{1}{2}\rho(Y_{12} - Y_{22})$ and the within units correlation determines the performance. In the third case both correlations are important and the difference is about $\frac{1}{2}(\rho - \delta)(Y_{12} - Y_{21})$.

In the following we will study the relationship between $\widehat{\beta}^*$ and $\widehat{\beta}_0$ by providing an upper bound for $||\widehat{\beta}^* - \widehat{\beta}_0||$, where $||\cdot||$ is the Euclidean norm. Let $\lambda_1(G) \geq \cdots \geq \lambda_m(G)$ denote the eigenvalues of a $m \times m$ symmetric matrix G.

Lemma 1. *If D is a symmetric positive semidefinite $m \times m$ matrix and C an arbitrary $m \times n$ matrix, then $\lambda_1(C'DC) \leq \lambda_1(D)\lambda_1(C'C)$.*

Proof. By Rayleigh's formula for the largest eigenvalue of a symmetric positive semidefinite matrix (see e. g. Rao, 1973, p. 62) we have

$$\begin{aligned}\lambda_1(C'DC) &= \sup_z \frac{z'C'DCz}{z'z} = \sup_z \frac{z'C'DCz}{z'C'Cz}\frac{z'C'Cz}{z'z} \\ &\leq \sup_z \frac{z'C'DCz}{z'C'Cz} \cdot \sup_z \frac{z'C'Cz}{z'z} \leq \lambda_1(D)\lambda_1(C'C).\end{aligned}$$

□

Theorem 2.

$$||\widehat{\beta}^* - \widehat{\beta}_0|| \leq \frac{(T\rho + N\delta)||Y - X\widehat{\beta}_0||}{\lambda_p^{1/2}(X'X)} + O(\rho^2 + \delta^2) \qquad (21)$$

Proof. By Theorem 1 we have

$$||\widehat{\beta}^* - \widehat{\beta}_0|| = ||(X'X)^{-1}X'F\eta|| + R = (\eta'F'X(X'X)^{-2}X'F\eta)^{1/2} + R,$$

where $F = \rho A + \delta B$, $\eta = Y - X\widehat{\beta}_0$ and $R = O(\rho^2 + \delta^2)$. By Rayleigh's inequality and Lemma 1 we obtain

$$\begin{aligned} \eta'F'X(X'X)^{-2}X'F\eta &\leq \lambda_1\big[F'X(X'X)^{-2}X'F\big]\,||\eta||^2 \\ &\leq \lambda_1\big[X(X'X)^{-2}X'\big]\,\lambda_1(F'F)||\eta||^2. \end{aligned}$$

Now, F is symmetric and positive semidefinite and, hence, $\lambda_1(F'F) = \lambda_1^2(F)$.

Next, note that for every matrix C the positive eigenvalues of $C'C$ and CC' coincide, since for every eigenvector z of $C'C$ corresponding to an eigenvalue $\lambda > 0$ the vector $Cz \neq 0$ is an eigenvector of CC' with $CC'Cz = \lambda Cz$. By letting $C = (X'X)^{-1}X'$ we, thus, obtain

$$\lambda_1\big[X(X'X)^{-2}X'\big] \;=\; \lambda_1\big[(X'X)^{-1}\big] \;=\; \frac{1}{\lambda_p(X'X)}.$$

The largest eigenvalue of F is $T\rho + N\delta$ which completes the proof. □

We note that the upper bound in (21) may become large, when strong multicollinearity is present in X, i. e. $\lambda_p(X'X)$ is small and X is ill–conditioned. However, for balanced X the smallest eigenvalue $\lambda_p(X'X)$ grows approximately like NT for N and T large. Hence, the bound of Theorem 2 becomes $c(\rho\sqrt{T/N} + \delta\sqrt{N/T})$.

The eigenvalues of the dispersion matrix $\mathbf{\Sigma}$ are

$$\begin{array}{l} \lambda_1 = T\sigma_u^2 + N\sigma_v^2 + \sigma_w^2 = \sigma^2[1 + (T-1)\rho + (N-1)\delta] \\ \lambda_i = N\sigma_v^2 + \sigma_w^2 = \sigma^2[1 + (N-1)\delta - \rho] \quad \text{with multiplicity } T-1 \\ \lambda_j = T\sigma_u^2 + \sigma_w^2 = \sigma^2[1 + (T-1)\rho - \delta] \quad \text{with multiplicity } N-1 \\ \lambda_{N+T} = \cdots = \lambda_{NT} = \sigma_w^2 = \sigma^2(1 - \rho - \delta) \end{array} \tag{22}$$

According to the results of Magness and McGuire (1962) the following inequality holds for the efficiency when a linear function $c'\beta$ of the parameter β is estimated by the BLUE property of $\widehat{\beta}^*$ and by the Kantorovich inequality (see e. g. Rao (1973), p. 74)

$$1 \geq \frac{\mathrm{Var}\,(c'\widehat{\beta}^*)}{\mathrm{Var}\,(c'\widehat{\beta}_0)} \geq \frac{4\lambda_1(\mathbf{\Sigma})\lambda_{NT}(\mathbf{\Sigma})}{(\lambda_1(\mathbf{\Sigma}) + \lambda_{NT}(\mathbf{\Sigma}))^2} = \frac{4(T\sigma_u^2 + N\sigma_v^2 + \sigma_w^2)\sigma_w^2}{(T\sigma_u^2 + N\sigma_v^2 + 2\sigma_w^2)^2} \tag{23}$$

If we assume $X'X = I_p$. This resukt can be also found in Hannan (1970, p. 422). Obviously, the minimum efficiency $\inf_{c\in R^p} \mathrm{Var}\,(c'\widehat{\beta}^*)/\mathrm{Var}\,(c'\widehat{\beta}_0)$ is bounded from below by $1 - O(\rho^2 + \delta^2)$.

Alternatively, as in Wang et al (1994), we consider the spectral norm $d = ||\mathrm{Cov}\,(\widehat{\beta}_0) - \mathrm{Cov}\,(\widehat{\beta}^*)||_2$ of the difference between the dispersion matrix $\mathrm{Cov}\,(\widehat{\beta}_0)$ of the OLSE and the dispersion matrix $\mathrm{Cov}\,(\widehat{\beta}^*)$ of the BLUE.

Note that the spectral norm $||G||_2$ of a matrix G is defined as $||G||_2 = \sup_{||z||=1} ||Gz||$ and, hence, $d = \lambda_1(\mathrm{Cov}\,(\widehat{\beta}_0) - \mathrm{Cov}\,(\widehat{\beta}^*))$ as $\mathrm{Cov}\,(\widehat{\beta}_0) - \mathrm{Cov}\,(\widehat{\beta}^*)$ is positive semidefinite and symmetric. We note that the Euclidean norm $||\widehat{\beta}^* - \widehat{\beta}_0||$ measures the distance between the two estimated values of β while the spectral norm of $\mathrm{Cov}\,(\widehat{\beta}_0) - \mathrm{Cov}\,(\widehat{\beta}^*)$ measures how far the dispersion matrices of the estimates differ.

Theorem 3.

$$d \le \frac{T\rho + N\delta}{\lambda_p(X'X)}$$

Remark. Let $\tau = \lambda_1(\mathbf{\Sigma})/\lambda_{NT}(\mathbf{\Sigma}) = \frac{1+(T-1)\rho+(N-1)\delta}{1-\rho-\delta}$ denote the condition number of the dispersion matrix $\mathbf{\Sigma}$. Then we can reformulate the result of Theorem 3 as

$$d \le \frac{\{1 + (T-1)\rho + (N-1)\delta\}\{1 - \frac{1}{\tau}\}}{\lambda_p(X'X)}.$$

By letting $T = 1$ we obtain $d \le \{1 + (N-1)\delta\}\{1 - \frac{1}{\tau}\}\lambda_p^{-1}(X'X)$ which improves upon the inequality of Theorem 4 by Wang et al (1994) as $\tau > 1$.

Proof of Theorem 3. Let $(X'X)^{1/2}$ denote a square root of $X'X$, i.e. $X'X = (X'X)^{1/2}(X'X)^{1/2}$, and $(X'X)^{-1/2}$ its inverse. Then we have

$$\begin{aligned}\mathrm{Cov}\,(\widehat{\beta}_0) - \mathrm{Cov}\,(\widehat{\beta}^*) \;&=\; (X'X)^{-1/2}[(X'X)^{-1/2}X'\mathbf{\Sigma}X(X'X)^{-1/2} \\ &\quad - (X'X)^{1/2}(X'\mathbf{\Sigma}^{-1}X)^{-1}(X'X)^{1/2}](X'X)^{-1/2}\end{aligned}$$

and, consequently, by Lemma 1

$$\begin{aligned}d \;\le\; &\lambda_1\big[(X'X)^{-1}\big] \\ &\big\{\lambda_1\big[(X'X)^{-1/2}X'\mathbf{\Sigma}X(X'X)^{-1/2}\big] - \lambda_p\big[(X'X)^{1/2}(X'\mathbf{\Sigma}^{-1}X)^{-1}(X'X)^{1/2}\big]\big\}.\end{aligned}$$

Again, by Lemma 1

$$\begin{aligned}&\lambda_1\big[(X'X)^{-1/2}X'\mathbf{\Sigma}X(X'X)^{-1/2}\big] \\ &\quad\le\; \lambda_1(\mathbf{\Sigma})\lambda_1\big[(X'X)^{-1/2}X'X(X'X)^{-1/2}\big] \;=\; \lambda_1(\mathbf{\Sigma})\end{aligned}$$

and, similarly,

$$\begin{aligned}&\lambda_p\big[(X'X)^{1/2}(X'\mathbf{\Sigma}^{-1}X)^{-1}(X'X)^{1/2}\big] \\ &\quad=\; \lambda_1^{-1}\big[(X'X)^{-1/2}X'\mathbf{\Sigma}^{-1}X(X'X)^{-1/2}\big] \;\ge\; \lambda_1^{-1}(\mathbf{\Sigma}^{-1}) \;=\; \lambda_{NT}(\mathbf{\Sigma}).\end{aligned}$$

Combining these results we obtain

$$d \le \frac{\lambda_1(\mathbf{\Sigma}) - \lambda_{NT}(\mathbf{\Sigma})}{\lambda_p(X'X)}$$

which completes the proof in view of (22). (Note also that $\lambda_1(\mathbf{\Sigma}) - \lambda_{NT}(\mathbf{\Sigma}) = \lambda_1(\rho A + \delta B)$.) $\square$

4. Additional remarks.

The results of Section 3 can be extended to situations in which more than two additional variance components are present. The additive structure (2) of the random errors expands to

$$\varepsilon_{j_1,\dots,j_k} = u_{1,j_1} + \dots + u_{k,j_k} + w_{j_1,\dots,j_k}.$$

Consequently, the dispersion matrix $\mathbf{\Sigma}$ can be written as $\mathbf{\Sigma} = \sigma^2\Sigma(\boldsymbol{\rho})$ where the standardised dispersion

$$\Sigma(\boldsymbol{\rho}) = \left(1 - \sum_{j=1}^{k} \rho_j\right) I + \sum_{j=1}^{k} \rho_j A_j$$

decomposes additively, $\boldsymbol{\rho} = (\rho_1, \dots, \rho_k)'$, $\rho_j \geq 0$ and $\sum_{j=1}^{k} \rho_j < 1$. I and A_j are $N \times N$ matrices where N is the number of different observations, e. g. $N = \prod_{j=1}^{k} N_j$ for a full factorial or, alternatively, $N = N_1^2$ for a Latin or Graeco–Latin square. Then in generalisation of Theorem 1 we obtain

$$\widehat{\beta}^* = \widehat{\beta}_0 - (X'X)^{-1}X' \textstyle\sum_{j=1}^{k} \rho_j A_j (Y - X\widehat{\beta}_0) + O(||\boldsymbol{\rho}||^2). \tag{24}$$

Consequently, in extension of Theorem 2, we obtain

$$||\widehat{\beta}^* - \widehat{\beta}_0|| \leq \frac{\lambda_1\left(\sum_{j=1}^{k} \rho_j A_j\right)}{\lambda_p^{1/2}(X'X)} ||Y - X\widehat{\beta}_0|| + O(||\boldsymbol{\rho}||^2), \tag{25}$$

where $\lambda_1(\sum_{j=1}^{k} \rho_j A_j) = \sum_{j=1}^{k} \frac{N}{N_j}\rho_j$ in case of a full factorial, or $\lambda_1(\sum_{j=1}^{k} \rho_j A_j)$ $= N_1 \sum_{j=1}^{k} \rho_j$ for a Latin or a Graeco–Latin square. Similarly, for Theorem 3 we have

$$d \leq \frac{\lambda_1\left(\sum_{j=1}^{k} \rho_j A_j\right)}{\lambda_p(X'X)}. \tag{26}$$

Furthermore, also second order approximations are available in the spirit of Wang et al (1994) by using multivariate Taylor expansions of second degree

$$\begin{aligned}\widehat{\beta}^* \; = \; & \widehat{\beta}_0 - (X'X)^{-1}X'F\eta \qquad (27)\\ & - (X'X)^{-1}X'\left\{\textstyle\sum_{j=1}^{k} \rho_j F - F(I_N - X(X'X)^{-1}X')F\right\} \eta + O(||\boldsymbol{\rho}||^3),\end{aligned}$$

where $F = \sum_{j=1}^{k} \rho_j A_j$ and $\eta = Y - X\widehat{\beta}_0$. In particular, for the two–dimensional situation of the previous sections with units and time points we obtain

$$\begin{aligned}\widehat{\beta}^* \; = \; & \widehat{\beta}_0 - (X'X)^{-1}X'(\rho A + \delta B)\eta \qquad (28)\\ & + (X'X)^{-1}X'\{(T-1)\rho^2 A + (N-1)\delta^2 B - \rho\delta(A+B)\\ & \quad - (\rho A + \delta B)X(X'X)^{-1}X'(\rho A + \delta B)\}\eta + O(|\rho|^3 + |\delta|^3).\end{aligned}$$

References

Baksalary, J.K. and Kala, R. (1980): A new bound for the Euclidean norm of the difference between the least squares and the best linear unbiased estimators. *Ann. Statist.* **8**, 679–681.

Bloomfield, P. and Watson, G.S. (1975): The inefficiency of least squares. *Biometrika* **62**, 121–128.

Christensen, R. (1984): A note on ordinary least squares methods for two–stage sampling. *J. Amer. Statist. Assoc.* **79**, 720–721.

Fuller, W.A. and Battese, G.E. (1973): Transformations for estimation of linear models with nested–error structure. *J. Amer. Statist. Assoc.* **68**, 626–632.

Hannan, R.J. (1970): *Multiple Time Series.* Wiley, New York.

Knott, M. (1975): On the minimum efficiency of least squares. *Biometrika* **62**, 129–132.

Krämer, W. (1980): Finite sample efficiency of ordinary least squares in the linear regression model with autocorrelated errors. *J. Amer. Statist. Assoc.* **75**, 1005–1009.

Magness, T.A. and McGuire, J.B. (1962): Comparison of least squares and minimum variance estimates of regression parameters. *Ann. Math. Statist.* **33**, 462–470.

Magnus, J.R. and Neudecker, H. (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Wiley, Chichester.

Rao, C.R. (1973): *Linear Statistical Inference and Its Applications. 2nd ed.* Wiley, New York.

Schwabe, R. (1996): *Optimum Designs for Multi–Factor Models. Lecture Notes in Statistics* **113**. Springer, New York.

Scott, A.J. and Holt, D. (1982): The effect of two–stage sampling on ordinary least squares methods. *J. Amer. Statist. Assoc.* **77**, 848–854.

Wang, S.G., Chow, S.C. and Tse, S.K. (1994): On ordinary least–squares methods for sample surveys. *Statist. Probab. Letters* bf 20, 173–182.

Zyskind, G. (1967): On canonical forms, non–negative covariance matrices and best and simple least squares linear estimators in linear models. *Ann. Math. Staitst.* **38**, 1092–1109.