Toutenburg, Fieger, Heumann:

Regression modelling with fixed effects - missing values and other problems

Sonderforschungsbereich 386, Paper 123 (1998)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Regression Modelling with Fixed Effects — Missing Values and related Problems

H. Toutenburg A. Fieger C. Heumann

22nd September 1998

Abstract

The paper considers new devices to predict the response variable using a convex target function weighting the response and its expectation. A MDEP-matrix superiority condition is given concerning BLUE, RLSE and mixed estimator where the latter is used in case of imputation for missing values. A small simulation study compares the alternative estimators. Finally the detection of non-MCAR processes in linear regression is discussed.

Some Hypotheses about Statistics for the twenty first century

1. Statistical research in next century will be largely based on artificial intelligence. Appropriate models for any given data set will be searched through computers.
2. We have many empirical studies now. These will be fruitfully employed in developing non-conjugate prior distributions that will describe reality. The availability of realistic prior distributions will lead to high importance to Bayesian inference.
3. Generally we pay attention to one problem at a time. Next century will provide answers to questions that relate to problems which occur simultaneously. For example, consider the traditional linear regression analysis. Such an analysis may not be appropriate due to, for instance, nonlinearity, autocorrelated disturbance and measurement errors. If only one of the problems is present, we know some solution. But if all the three problems are present simultaneously, we have practically no suitable solution. Such issues will be an important aspect of future research.
4. Nonparametric procedures will gain popularity. Considerable efforts will be directed towards the study of performance properties of nonparametric procedures in finite samples.

In general, computer based research will dominate the traditional work.

Problem 1
Predictive Performance of Restricted and Mixed Regression Estimators

1.1 Introduction

Generally predictions from a linear regression model are made either for the actual values of the study variable or for the average values at a time. However, situations may occur in which one may be required to consider the predictions of both the actual and average values simultaneously. For example, consider the installation of an artificial tooth in patients through a specific device. Here a dentist would like to know the life of a restoration, on the average. On the other hand, a patient would be more interested in knowing the actual life of restoration in his/her case. Thus a dentist is interested in the prediction of average value but he may not completely ignore the interest of patients in the prediction of actual value. The dentist may assign higher weightage to prediction of average values in comparison to the prediction of actual values. Similarly, a patient may give more weightage to prediction of actual values in comparison to that of average values.

This section considers the problem of simultaneous prediction of actual and average values of the study variable in a linear regression model when a set of linear restrictions binding the regression coefficients is available, and analyzes the performance properties of predictors arising from the methods of restricted regression and mixed regression besides least squares.

1.2 Specification of Model and Target Function

Let us postulate the following linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{1.1}$$

where $\mathbf{y}$ is a $n \times 1$ vector of n observations on the study variable, $\mathbf{X}$ is a $n \times K$ full column rank matrix of n observations on K explanatory variables, $\boldsymbol{\beta}$ is a column vector of regression coefficients and $\mathbf{u}$ is an $n \times 1$ vector of disturbances.

It is assumed that the elements of $\mathbf{u}$ are independently and identically distributed with mean zero and variance σ^2.

If $\hat{\boldsymbol{\beta}}$ denotes an estimator of $\boldsymbol{\beta}$, then the predictor for the values of study variable within the sample is generally formulated as $\hat{\mathbf{T}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ which is used for predicting either the actual values $\mathbf{y}$ or the average values $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ at a time.

When the situation demands prediction of both the actual and average values together, Toutenburg and Shalabh (1996) defined the following stochastic target function

$$T(\mathbf{y}) = \lambda\mathbf{y} + (1-\lambda)\,\mathrm{E}(\mathbf{y}) = \mathbf{T} \tag{1.2}$$

and use $\hat{\mathbf{T}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ for predicting it where $0 \leq \lambda \leq 1$ is a nonstochastic scalar specifying the weightage to be assigned to the prediction of actual and average values of the study variable; see, e. g. Shalabh (1995).

Remark (i). In case that $\lambda = 0$, we have $\mathbf{T} = \mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and then optimal prediction coincides with optimal estimation of $\boldsymbol{\beta}$, whereas optimality may be defined, e. g., by minimal variance in the class of linear unbiased estimators or by some mean dispersion error criterion if biased estimators are considered. The other extreme case $\lambda = 1$ leads to $\mathbf{T} = \mathbf{y}$. Optimal prediction of $\mathbf{y}$ is then equivalent to optimal estimation of $\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$. If the disturbances are uncorrelated this coincides again with optimal estimation of $\mathbf{X}\boldsymbol{\beta}$, i. e., of $\boldsymbol{\beta}$ itself. If the disturbances are correlated according to $\mathrm{E}(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{W}$, then this information leads to solutions $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}$ (cp. Goldberger, 1962).

Remark (ii). The two alternative prediction problems—the $\mathbf{X}\boldsymbol{\beta}$-superiority and the $\mathbf{y}$-superiority, respectively—are discussed in full detail in Rao and Toutenburg (1995, Chapter 6). As a central result, we have the fact that the superiority (in the Loewner ordering of definite matrices) of one predictor over another predictor can change if the criterion is changed. This was one of the motivations to define a target as in (1.2) that combines these two risks.

In the following we consider this problem but with the nonstochastic scalar λ replaced by a nonstochastic matrix $\boldsymbol{\Lambda}$. The target function is therefore

$$T(\mathbf{y}) = \boldsymbol{\Lambda}\mathbf{y} + (\mathbf{I} - \boldsymbol{\Lambda})\,\mathrm{E}(\mathbf{y}) = \mathbf{T}\,. \tag{1.3}$$

Our derivation of the results makes no assumption about $\boldsymbol{\Lambda}$, but one may have in mind $\boldsymbol{\Lambda}$ as a diagonal matrix with elements $0 \leq \lambda_i \leq 1$, $i = 1, \ldots, n$.

1.3 Exact Linear Restrictions

Let us suppose that we are given a set of J exact linear restrictions binding the regression coefficients:

$$\mathbf{r} = \mathbf{R}\boldsymbol{\beta} \tag{1.4}$$

where $\mathbf{r}$ is a $J \times 1$ vector and $\mathbf{R}$ is a $J \times K$ full row rank matrix.

If these restrictions are ignored, the least squares estimator of $\boldsymbol{\beta}$ is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{1.5}$$

which may not necessarily obey (1.4). Such is, however, not the case with restricted regression estimator given by

$$\mathbf{b}_{\mathrm{R}} = \mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\mathbf{b}) \tag{1.6}$$

which invariably satisfies (1.4).

Employing (1.5) and (1.6), we get the following two predictors for the values of the study variable within the sample:

$$\hat{\mathbf{T}} = \mathbf{X}\mathbf{b}\,, \tag{1.7}$$

$$\hat{\mathbf{T}}_{\mathrm{R}} = \mathbf{X}\mathbf{b}_{\mathrm{R}}\,. \tag{1.8}$$

In the following we compare the estimators $\mathbf{b}$ and $\mathbf{b}_{\mathrm{R}}$ with respect to the predictive mean dispersion error (MDEP) of their corresponding predictions $\hat{\mathbf{T}} = \mathbf{X}\mathbf{b}$ and $\hat{\mathbf{T}}_{\mathrm{R}} = \mathbf{X}\mathbf{b}_{\mathrm{R}}$ for the target function $\mathbf{T}$.

From (1.3), and the fact that the ordinary least squares estimator and the restricted estimator are both unbiased, we see that

$$\begin{aligned} \mathrm{E}_{\mathbf{\Lambda}}(\mathbf{T}) &= \mathrm{E}(\mathbf{y})\,, & (1.9)\\ \mathrm{E}_{\mathbf{\Lambda}}(\hat{\mathbf{T}}) = \mathbf{X}\boldsymbol{\beta} &= \mathrm{E}(\mathbf{y})\,, & (1.10)\\ \mathrm{E}_{\mathbf{\Lambda}}(\hat{\mathbf{T}}_{\mathrm{R}}) = \mathbf{X}\boldsymbol{\beta} &= \mathrm{E}(\mathbf{y})\,, & (1.11) \end{aligned}$$

but

$$\mathrm{E}(\hat{\mathbf{T}}) = \mathrm{E}(\hat{\mathbf{T}}_{\mathrm{R}}) \neq \mathbf{T}\,. \tag{1.12}$$

Equation (1.12) reflects the stochastic nature of the target function $\mathbf{T}$, a problem which differs from the common problem of unbiasedness of a statistic for a fixed but unknown (possibly matrix valued) parameter. Therefore both the predictors are only "weakly unbiased" in the sense that

$$\begin{aligned} \mathrm{E}_{\mathbf{\Lambda}}(\hat{\mathbf{T}} - \mathbf{T}) &= \mathbf{0}\,, & (1.13)\\ \mathrm{E}_{\mathbf{\Lambda}}(\hat{\mathbf{T}}_{\mathrm{R}} - \mathbf{T}) &= \mathbf{0}\,. & (1.14) \end{aligned}$$

1.3.1 MDEP Using Ordinary Least Squares Estimator

To compare alternative predictors, we define the matrix-valued mean-dispersion error for $\tilde{\mathbf{T}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ as follows:

$$\mathrm{MDEP}_{\mathbf{\Lambda}}(\tilde{\mathbf{T}}) = \mathrm{E}(\tilde{\mathbf{T}} - \mathbf{T})(\tilde{\mathbf{T}} - \mathbf{T})'\,. \tag{1.15}$$

First we note that

$$\begin{aligned} \mathbf{T} &= \mathbf{\Lambda}\mathbf{y} + (\mathbf{I} - \mathbf{\Lambda})\,\mathrm{E}(\mathbf{y})\\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{\Lambda}\mathbf{u}\,, & (1.16)\\ \hat{\mathbf{T}} &= \mathbf{X}\mathbf{b}\\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{P}\mathbf{u}\,, & (1.17) \end{aligned}$$

with the symmetric and idempotent projection matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Hence we get

$$\begin{aligned} \mathrm{MDEP}_{\mathbf{\Lambda}}(\hat{\mathbf{T}}) &= \mathrm{E}(\mathbf{P} - \mathbf{\Lambda})\mathbf{u}\mathbf{u}'(\mathbf{P} - \mathbf{\Lambda})'\\ &= \sigma^2(\mathbf{P} - \mathbf{\Lambda})(\mathbf{P} - \mathbf{\Lambda})'\,, & (1.18) \end{aligned}$$

using our previously made assumptions on $\mathbf{u}$.

1.3.2 MDEP Using Restricted Estimator

The problem is now solved by calculation of

$$\mathrm{MDEP}_{\mathbf{\Lambda}}(\hat{\mathbf{T}}_{\mathrm{R}}) = \mathrm{E}(\hat{\mathbf{T}}_{\mathrm{R}} - \mathbf{T})(\hat{\mathbf{T}}_{\mathrm{R}} - \mathbf{T})'\,. \tag{1.19}$$

Using the abbreviation

$$\mathbf{F} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{1.20}$$

and

$$\mathbf{r} - \mathbf{R}\mathbf{b} = -\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\,, \tag{1.21}$$

we get from (1.6), (1.8), (1.16) and (1.17) the following

$$\begin{aligned} \hat{\mathbf{T}}_{\mathrm{R}} - \mathbf{T} &= \mathbf{X}\mathbf{b}_R - \mathbf{T} \\ &= (\mathbf{P} - \mathbf{F} - \mathbf{\Lambda})\mathbf{u}\,. \end{aligned} \tag{1.22}$$

As $\mathbf{F} = \mathbf{F}'$, $\mathbf{P} = \mathbf{P}'$ and $\mathbf{PF} = \mathbf{FP} = \mathbf{F}$, we have

$$\begin{aligned} \mathrm{MDEP}_{\mathbf{\Lambda}}(\hat{\mathbf{T}}_{\mathrm{R}}) &= \sigma^2(\mathbf{P} - \mathbf{F} - \mathbf{\Lambda})(\mathbf{P} - \mathbf{F} - \mathbf{\Lambda})' \\ &= \sigma^2[(\mathbf{P} - \mathbf{\Lambda})(\mathbf{P} - \mathbf{\Lambda})' - (\mathbf{F} - \mathbf{\Lambda}\mathbf{F} - \mathbf{F}\mathbf{\Lambda}')]\,. \end{aligned} \tag{1.23}$$

1.3.3 MDEP Matrix Comparison

Using the results (1.18) and (1.23), the difference of the MDEP-matrices can be written as

$$\begin{aligned} \mathbf{\Delta}_{\mathbf{\Lambda}}(\hat{\mathbf{T}}; \hat{\mathbf{T}}_{\mathrm{R}}) &= \mathrm{MDEP}_{\mathbf{\Lambda}}(\hat{\mathbf{T}}) - \mathrm{MDEP}_{\mathbf{\Lambda}}(\hat{\mathbf{T}}_{\mathrm{R}}) \\ &= \sigma^2(\mathbf{F} - \mathbf{\Lambda}\mathbf{F} - \mathbf{F}\mathbf{\Lambda}') \\ &= \sigma^2\left[(\mathbf{I} - \mathbf{\Lambda})\mathbf{F}(\mathbf{I} - \mathbf{\Lambda})' - \mathbf{\Lambda}\mathbf{F}\mathbf{\Lambda}'\right]\,. \end{aligned} \tag{1.24}$$

Then $\hat{\mathbf{T}}_{\mathrm{R}}$ becomes MDEP-superior to $\hat{\mathbf{T}}$ if $\Delta_{\mathbf{\Lambda}}(\hat{\mathbf{T}}; \hat{\mathbf{T}}_{\mathrm{R}}) \geq 0$.

For $\Delta_{\mathbf{\Lambda}}(\hat{\mathbf{T}}; \hat{\mathbf{T}}_{\mathrm{R}})$ to be non-negative definite, it follows from Baksalary, Schipp and Trenkler (1992) that necessary and sufficient conditions are

$$\begin{aligned} &\text{(i)} \quad \mathcal{R}(\mathbf{\Lambda}\mathbf{F}) \subset ((\mathbf{I} - \mathbf{\Lambda})\mathbf{F} \\ &\text{(ii)} \quad \lambda_1 \leq 1 \end{aligned}$$

where λ_1 denotes the largest characteristic root of the matrix $[(\mathbf{I} - \mathbf{\Lambda})\mathbf{F}(\mathbf{I} - \mathbf{\Lambda}')]^{+}\mathbf{\Lambda}\mathbf{F}\mathbf{\Lambda}'$.

For the simple special case of $\mathbf{\Lambda} = \theta\mathbf{I}$, the conditions reduce to $\theta \leq \frac{1}{2}$.

1.4 Missing values in the $\mathbf{X}$-Matrix and the Mixed Estimator

An interesting problem in all regression models relates to missing data. In general, we may assume the following structure of data:

$$\begin{pmatrix} \mathbf{y}_{\mathrm{obs}} \\ \mathbf{y}_{\mathrm{mis}} \\ \mathbf{y}_{\mathrm{obs}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{\mathrm{obs}} \\ \mathbf{X}_{\mathrm{obs}} \\ \mathbf{X}_{\mathrm{mis}} \end{pmatrix} \boldsymbol{\beta} + \mathbf{u}\,. \tag{1.25}$$

Estimation of $\mathbf{y}_{\mathrm{mis}}$ corresponds to the prediction problem discussed in Chapter 6 of Rao and Toutenburg (1995) in full detail. We may therefore confine ourselves to the structure

$$\mathbf{y}_{\mathrm{obs}} = \begin{pmatrix} \mathbf{X}_{\mathrm{obs}} \\ \mathbf{X}_{\mathrm{mis}} \end{pmatrix} \boldsymbol{\beta} + \mathbf{u} \tag{1.26}$$

and change the notation as follows:

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{y}_* \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{X}_* \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \mathbf{u} \\ \mathbf{u}_* \end{pmatrix} , \quad \begin{pmatrix} \mathbf{u} \\ \mathbf{u}_* \end{pmatrix} \sim (\mathbf{0}, \sigma^2 \mathbf{I}) \ . \tag{1.27}$$

The submodel

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{1.28}$$

presents the completely observed data and should fulfill the standard assumptions (i. e., $\mathbf{X}$ is nonstochastic of full column rank). The other submodel

$$\mathbf{y}_* = \mathbf{X}_*\boldsymbol{\beta} + \mathbf{u}_*$$

is related to the partially observed $\mathbf{X}$-variables. The dimensions of the two models are m_{c} and m_*, respectively, with $n = m_{\mathrm{c}} + m_*$.

Let $\mathbf{M} = (m_{ij})$ define the missing indicator matrix (c. p. Rubin, 1976) with $m_{ij} = 1$ if x_{ij} is not observed and $m_{ij} = 0$ if x_{ij} is observed. Under the assumption that missingness is independent of $\mathbf{y}$, i. e.,

$$f(\mathbf{M}|\mathbf{y}, \mathbf{X}) = f(\mathbf{M}|\mathbf{X})$$

we have

$$f(\mathbf{y}|\mathbf{M}, \mathbf{X}) = \frac{f(\mathbf{y}, \mathbf{M}|\mathbf{X})}{f(\mathbf{M}|\mathbf{X})} = \frac{f(\mathbf{M}, \mathbf{y}|\mathbf{X})}{f(\mathbf{M}|\mathbf{y}, \mathbf{X})} = f(\mathbf{y}|\mathbf{X})$$

which means that the the CC-estimator (complete case)

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{1.29}$$

is consistent for $\boldsymbol{\beta}$.

As an alternative one may impute estimates or fixed values for the missing data so that the partially unknown matrix $\mathbf{X}_*$ is replaced by a known matrix $\mathbf{R}$ resulting in

$$\mathbf{y}_* = \mathbf{R}\boldsymbol{\beta} + (\mathbf{X}_* - \mathbf{R})\boldsymbol{\beta} + \mathbf{u}_* \tag{1.30}$$

or, equivalently written in the shape of stochastic linear restrictions,

$$\mathbf{r} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\delta} + \boldsymbol{\phi} \ , \quad \boldsymbol{\phi} \sim \left(\mathbf{0}, \sigma^2 \mathbf{I}\right) \tag{1.31}$$

with $\boldsymbol{\delta} = (\mathbf{X}_* - \mathbf{R})\boldsymbol{\beta}$ a bias vector. Combining the CC-model (1.28) and the filled-up model (1.31) results in the mixed model (Theil and Goldberger, 1961)

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{r} \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{R} \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\delta} \end{pmatrix} + \begin{pmatrix} \mathbf{u} \\ \boldsymbol{\phi} \end{pmatrix} \ . \tag{1.32}$$

For $\boldsymbol{\delta} = \mathbf{0}$, the BLUE in (1.32) is given by the mixed estimator

$$\tilde{\mathbf{b}}_{\mathrm{R}} = \mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' \left[\mathbf{I} + \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{r} - \mathbf{R}\mathbf{b}) \tag{1.33}$$

with dispersion matrix

$$\begin{aligned} \mathrm{V}(\tilde{\mathbf{b}}_{\mathrm{R}}) &= \mathrm{V}(\mathbf{b}) - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' \left[\mathbf{I} + \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} && (1.34) \\ &= \mathrm{V}(\mathbf{b}) - \mathbf{D} \ , && (1.35) \end{aligned}$$

say, whence it follows that the variance covariance matrix of $\mathbf{b}$ exceeds the variance covariance matrix of $\tilde{\mathbf{b}}_{\mathrm{R}}$ by a non-negative definite matrix and thus $\tilde{\mathbf{b}}_{\mathrm{R}}$ is more efficient than $\mathbf{b}$.

In case that $\boldsymbol{\delta} \neq \mathbf{0}$, the mixed estimator $\tilde{\mathbf{b}}_R$ becomes biased and its bias vector is

$$\mathrm{Bias}(\tilde{\mathbf{b}}_{\mathrm{R}}, \boldsymbol{\beta}) = \mathbf{D}\mathbf{d} \tag{1.36}$$

where

$$\begin{aligned} \mathbf{d} &= (\mathbf{X}'\mathbf{X})\mathbf{R}^{+}\boldsymbol{\delta}\sigma^{-2} && (1.37)\\ \mathbf{R}^{+} &= \mathbf{R}'(\mathbf{R}\mathbf{R}')^{-1}\,. && (1.38) \end{aligned}$$

Therefore $\mathrm{Bias}(\tilde{\mathbf{b}}_{\mathrm{R}}, \boldsymbol{\beta}) \in \mathcal{R}(\mathbf{D})$ and we may apply result A1 given in the Appendix to get the following theorem.

Theorem 1. Let $\mathrm{M}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \mathrm{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'$ define the MDE matrix of an estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Then the biased estimator $\tilde{\mathbf{b}}_{\mathrm{R}}$ is MDE-superior over the OLSE $\mathbf{b}$ in the sense that the variance covariance matrix of $\mathbf{b}$ exceeds the mean squared error matrix of $\tilde{\mathbf{b}}_{\mathrm{R}}$ by a non-negative definite matrix if and only if

$$\rho = \sigma^{-2}\boldsymbol{\delta}'\left[\mathbf{I} + \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1}\boldsymbol{\delta} \leq 1\,. \tag{1.39}$$

If $\mathbf{u}$ and $\boldsymbol{\phi}$ are independently normally distributed, then ρ is the noncentrality parameter of the statistic

$$F = \frac{1}{Js^2}(\mathbf{r} - \mathbf{R}\mathbf{b})'\left[\mathbf{I} + \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1}(\mathbf{r} - \mathbf{R}\mathbf{b}) \tag{1.40}$$

which follows a noncentral $F_{J,n-K}(\rho)$-distribution under $\rho \leq 1$.

1.4.1 MDEP Using Mixed Estimator

Using the mixed estimator $\tilde{\mathbf{b}}_{\mathrm{R}}$, we have $\tilde{\mathbf{T}}_{\mathrm{R}} = \mathbf{X}\tilde{\mathbf{b}}_{\mathrm{R}}$. Hence we have to calculate

$$\mathrm{MDEP}_{\boldsymbol{\Lambda}}\left(\tilde{\mathbf{T}}_{\mathrm{R}}\right) = \mathrm{E}(\tilde{\mathbf{T}}_{\mathrm{R}} - \mathbf{T})(\tilde{\mathbf{T}}_{\mathrm{R}} - \mathbf{T})'\,. \tag{1.41}$$

Using the abbreviation

$$\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\left[\mathbf{I} + \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} \tag{1.42}$$

and taking into account that

$$\begin{aligned} \mathbf{A}\left[\mathbf{I} + \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]\mathbf{A}' &= \mathbf{D} && (1.43)\\ \mathbf{P}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\mathbf{A}' &= \mathbf{D} && (1.44)\\ \mathbf{X}\mathbf{D}\mathbf{X}' &= \mathbf{F} \quad \text{in case that } \boldsymbol{\phi} = \mathbf{0} && (1.45) \end{aligned}$$

we may write

$$\tilde{\mathbf{T}}_{\mathrm{R}} - \mathbf{T} = (\mathbf{P} - \boldsymbol{\Lambda})\mathbf{u} + \mathbf{X}\mathbf{A}\left[\boldsymbol{\phi} - \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\right] + \mathbf{X}\mathbf{A}\boldsymbol{\delta}\,. \tag{1.46}$$

Therefore, using (1.43), (1.44) and (1.45) we get

$$\begin{aligned} \mathrm{MDEP}_{\boldsymbol{\Lambda}}\left(\tilde{\mathbf{T}}_{\mathrm{R}}\right) &= \sigma^2(\mathbf{P} - \boldsymbol{\Lambda})(\mathbf{P} - \boldsymbol{\Lambda})'\\ &\quad - \sigma^2\left(\mathbf{X}\mathbf{D}\mathbf{X}' - \boldsymbol{\Lambda}\mathbf{X}\mathbf{D}\mathbf{X}' - \mathbf{X}\mathbf{D}\mathbf{X}'\boldsymbol{\Lambda}'\right)\\ &\quad + \mathbf{X}\mathbf{A}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{A}'\mathbf{X}' && (1.47) \end{aligned}$$

1.4.2 MDEP-Matrix Comparison

The difference of the MDEP-matrices of $\hat{\mathbf{T}}$ and $\tilde{\mathbf{T}}_{\mathrm{R}}$ can be written as

$$\boldsymbol{\Delta}_{\boldsymbol{\Lambda}}(\hat{\mathbf{T}};\tilde{\mathbf{T}}_{\mathrm{R}}) = \sigma^2\left[(\mathbf{I}-\boldsymbol{\Lambda})\mathbf{X}\mathbf{D}\mathbf{X}'(\mathbf{I}-\boldsymbol{\Lambda})' - \boldsymbol{\Lambda}\mathbf{X}\mathbf{D}\mathbf{X}'\boldsymbol{\Lambda}'\right] - \mathbf{X}\mathbf{A}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{A}'\mathbf{X}' \quad (1.48)$$

Then using Baksalary et al. (1992) and the result A1 of the Apendix, we have

$$\boldsymbol{\Delta}_{\boldsymbol{\Lambda}}(\hat{\mathbf{T}};\tilde{\mathbf{T}}_{\mathrm{R}}) \geq \mathbf{0}$$

if and only if

$$\text{(i)} \quad \left[(\mathbf{I}-\boldsymbol{\Lambda})\mathbf{X}\mathbf{D}\mathbf{X}'(\mathbf{I}-\boldsymbol{\Lambda})' - \boldsymbol{\Lambda}\mathbf{X}\mathbf{D}\mathbf{X}'\boldsymbol{\Lambda}'\right] \geq 0 \quad (1.49)$$

$$\text{(ii)} \quad \sigma^{-2}\boldsymbol{\delta}'\mathbf{A}'\mathbf{X}'\left[(\mathbf{I}-\boldsymbol{\Lambda})\mathbf{X}\mathbf{D}\mathbf{X}'(\mathbf{I}-\boldsymbol{\Lambda})' - \boldsymbol{\Lambda}\mathbf{X}\mathbf{D}\mathbf{X}'\boldsymbol{\Lambda}'\right]^{-}\mathbf{X}\mathbf{A}\boldsymbol{\delta} \leq 1\,. \quad (1.50)$$

Problem 2
Missing values in the $\mathbf{X}$-matrix and the weighted mixed regression estimator

In the following we again assume the situation given in equation (1.26), that is missing values in $\mathbf{X}$ only. Filling in replacement values for the missing values leads to the setup of biased mixed estimation as in equations (1.31) and (1.32). Since the additional information is biased, it seems pertinent to use a weight lower than one for this part of the model. This can be achieved by rewriting the target function to be minimized from

$$S(\boldsymbol{\beta}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + (\mathbf{r}-\mathbf{R}\boldsymbol{\beta})'(\mathbf{r}-\mathbf{R}\boldsymbol{\beta})$$

to

$$S(\boldsymbol{\beta},\lambda) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda(\mathbf{r}-\mathbf{R}\boldsymbol{\beta})'(\mathbf{r}-\mathbf{R}\boldsymbol{\beta})\ ,$$

with $0 \leq \lambda \leq 1$. The solution given by

$$\mathbf{b}(\lambda) = (\mathbf{X}'\mathbf{X}+\lambda\mathbf{R}'\mathbf{R})^{-1}(\mathbf{X}'\mathbf{y}+\lambda\mathbf{R}'\mathbf{r})$$

may be called the weighted mixed regression estimator (WMRE). This estimator may be interpreted as the familiar mixed estimator in the model

$$\begin{pmatrix}\mathbf{y}\\ \sqrt{\lambda}\mathbf{r}\end{pmatrix} = \begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{R}\end{pmatrix}\boldsymbol{\beta} + \begin{pmatrix}\mathbf{u}\\ \sqrt{\lambda}\boldsymbol{\phi}\end{pmatrix}\ .$$

Using $\mathbf{Z}_\lambda = (\mathbf{X}'\mathbf{X}+\lambda\mathbf{R}'\mathbf{R})$, we have the alternative representation

$$\begin{aligned}\mathbf{b}(\lambda) &= \mathbf{Z}_\lambda^{-1}(\mathbf{X}'\mathbf{X}\boldsymbol{\beta}+\mathbf{X}'\mathbf{u}+\lambda\mathbf{R}'\mathbf{X}_*\boldsymbol{\beta}+\lambda\mathbf{R}'\boldsymbol{\phi})\\ &= \boldsymbol{\beta}+\lambda\mathbf{Z}_\lambda^{-1}\mathbf{R}'(\mathbf{X}_*-\mathbf{R})\boldsymbol{\beta}+\mathbf{Z}_\lambda^{-1}(\mathbf{X}'\mathbf{u}+\lambda\mathbf{R}'\boldsymbol{\phi})\end{aligned}$$

from which it follows that the WMRE is biased and its bias vector is given by

$$\mathrm{Bias}\,\mathbf{b}(\lambda) = \lambda\mathbf{Z}_\lambda^{-1}\mathbf{R}'\boldsymbol{\delta}\ ,$$

with covariance matrix as

$$\mathrm{V}(\mathbf{b}(\lambda)) = \sigma^2\mathbf{Z}_\lambda^{-1}(\mathbf{X}'\mathbf{X}+\lambda^2\mathbf{R}'\mathbf{R})\mathbf{Z}_\lambda^{-1}\ .$$

2.1 Ways of finding an optimal λ

One strategy to find an optimal λ is to minimize the MDEP. Let $\tilde{y} = \tilde{\mathbf{x}}'\boldsymbol{\beta} + \sigma\tilde{\epsilon}$ be a nonobserved future realisation of the regression model that is to be predicted by $p = \tilde{\mathbf{x}}'\mathbf{b}(\lambda)$. Minimizing the MDEP of p given by $\mathrm{E}(p - \tilde{y})^2$ with respect to λ leads to the relation (Rao and Toutenburg, 1995)

$$\begin{aligned} \lambda &= \frac{1}{1 + \sigma^{-2}\rho_1(\lambda)\rho_2^{-1}(\lambda)} \\ \rho_1(\lambda) &= \mathrm{tr}[\mathbf{Z}_\lambda^{-1}\mathbf{S}\mathbf{Z}_\lambda^{-1}\mathbf{R}^{-1}\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{R}\mathbf{Z}_\lambda^{-1}] \\ \rho_2^{-1}(\lambda) &= \mathrm{tr}[\mathbf{Z}_\lambda^{-1}\mathbf{S}_\mathrm{R}\mathbf{Z}^{-1}\mathbf{S}\mathbf{Z}_\lambda^{-1}] \ , \end{aligned}$$

with $\mathbf{S} = \mathbf{X}'\mathbf{X}$ and $\mathbf{S}_\mathrm{R} = \mathbf{R}'\mathbf{R}$. In general, the solution has to be found iteratively while σ^2 and $\boldsymbol{\delta}$ have to be estimated by some procedure, e.g., σ^2 may be estimated from the complete cases. For the special case that only one observation is missing (i.e., $\mathbf{r}$ and $\boldsymbol{\delta}$ are scalars), an explicit but unknown solution is available as

$$\lambda = \frac{1}{1 + \sigma^{-2}\delta^2} \ . \tag{2.1}$$

A second strategy is to minimize the trace of the MDE matrix with respect to λ, which is given by

$$\mathrm{tr}\,\mathrm{MDE}(\mathbf{b}(\lambda), \boldsymbol{\beta}) = \mathrm{tr}\,[\sigma^2\mathbf{Z}_\lambda^{-1}(\mathbf{S} + \lambda^2\mathbf{S}_\mathrm{R})\mathbf{Z}_\lambda^{-1}] + \lambda^2\mathbf{Z}_\lambda^{-1}\mathbf{R}'\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{R}\mathbf{Z}_\lambda^{-1}]$$

Note that the solution λ_tr has to be found iteratively.

A third way is to compare $\mathbf{b}(\lambda)$ and $\mathbf{b}$ with respect to the MDE criterion. This results in the condition that $\mathbf{b}(\lambda)$ is MDE better than $\mathbf{b}$, if

$$\rho_\lambda = \sigma^{-2}\boldsymbol{\delta}'[(2\lambda^{-1} - 1)\mathbf{I} + \mathbf{R}\mathbf{S}^{-1}\mathbf{R}']^{-1}\boldsymbol{\delta} \leq 1 \, .$$

It can be shown, that the generalized version of (2.1),

$$\lambda_\mathrm{e} = \frac{1}{1 + \sigma^{-2}\boldsymbol{\delta}'\boldsymbol{\delta}}$$

always fulfills this condition. Alternatively λ_max could be chosen such that $\rho_\lambda = 1$ holds. Again, λ_max has to be found iteratively.

2.2 A small simulation study

In a small simulation study we compared the estimators $\mathbf{b}$ (complete case estimator, which is the same as $\mathbf{b}(\lambda)$ with $\lambda = 0$), $\mathbf{b}(1) = \mathbf{b}_\mathrm{R}$ and $\mathbf{b}(\lambda)$ (with $0 < \lambda < 1$). The comparison of the respective estimators was conducted using the scalar risk function

$$R(\mathbf{I}) = \mathrm{E}\left(\mathbf{b}(\lambda) - \boldsymbol{\beta}\right)'\left(\mathbf{b}(\lambda) - \boldsymbol{\beta}\right)$$

estimated by its empirical version

$$\hat{R}(\mathbf{b}(\lambda), \boldsymbol{\beta}) = \frac{1}{\#\mathrm{rep}} \sum_{i=1}^{\#\mathrm{rep}} \left(\mathbf{b}(\lambda)_i - \boldsymbol{\beta}\right)'\left(\mathbf{b}(\lambda)_i - \boldsymbol{\beta}\right)$$

where #rep means the number of repeated simulations of the error terms applied to one specific covariate data set. The details of the setup can be obtained from the authors on request.

Using the weights computed from $\boldsymbol{\delta}$ and σ^2 which are known in the simulation study (generating 100 different covariate data sets), all weighted estimators were found to be better than the complete case estimator $\mathbf{b}$, as expected from the theory. Comparing the weighted estimators using the different λ-values previously mentioned with the estimator $\mathbf{b}_{\mathrm{R}}$ shows that $\mathbf{b}_{\lambda_{\mathrm{tr}}}$ performs best in this comparison ($\mathbf{b}_{\lambda_{\mathrm{tr}}}$ was better than $\mathbf{b}_{\mathrm{R}}$ in 91 of 100 runs, while $\mathbf{b}_{\lambda_{\mathrm{e}}}$ was better than $\mathbf{b}_{\mathrm{R}}$ in only 50 of 100 runs). On the other hand, using the weights computed from estimated $\hat{\boldsymbol{\delta}} = \mathbf{r} - \mathbf{R}\mathbf{b}$ and $\hat{\sigma}^2$ (from the complete data), we observed that $\mathbf{b}_{\hat{\lambda}_{\mathrm{e}}}$ was better than $\mathbf{b}$ in 99 of 100 runs, $\mathbf{b}_{\hat{\lambda}_{\mathrm{tr}}}$ was better than $\mathbf{b}$ in 97 of 100 runs, while $\mathbf{b}_{\mathrm{R}}$ was better than $\mathbf{b}$ in only 79 of 100 runs, but also that, e. g. $\mathbf{b}_{\hat{\lambda}_{\mathrm{e}}}$ was better than $\mathbf{b}_{\mathrm{R}}$ in only 32 of 100 runs and $\mathbf{b}_{\hat{\lambda}_{\mathrm{tr}}}$ was better than $\mathbf{b}_{\mathrm{R}}$ in only 43 of 100 runs. These results yield no transitive ordering of the estimators.

One possible reason for these results could be that the true λ_{e} is typically underestimated by $\hat{\lambda}_{\mathrm{e}} = 1/(1 + \hat{\sigma}^{-2}\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\delta}})$ (the degree is also depending on σ^2 and the covariance structure of $\mathbf{X}$), since it can be shown that $\mathrm{E}(\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\delta}}) = \boldsymbol{\delta}'\boldsymbol{\delta} + \sigma^2(J + \sum_{j=1}^{J}\mu_j)$, where μ_j are the eigenvalues of $\mathbf{R}\mathbf{S}^{-1}\mathbf{R}'$.

These observations suggest the construction of a bias corrected version of the estimators. An interesting direction is to use bootstrapping techniques to obtain a bias correction (using different resampling techniques). The results of this approach indicate that the estimates can be improved concerning the bias. But there is still noticeable underestimation.

2.3 Some concluding remarks

For handling the problem of missing values of some explanatory variables, weighted mixed estimation seems to be a promising approach. However, the determination of the weighing scalar requires careful attention. It will also be interesting to develop suitable procedures for confidence intervals and hypothesis testing. So far the results hold only for $J < p$, i. e., the number of restrictions is smaller than the number of variables. For the missing value context, we also need to investigate the case when $J > p$.

Problem 3
Detection of non-MCAR processes in linear regression models

Missing data values in $\mathbf{X}$ are said to be missing completely at random (MCAR) if

$$f(\mathbf{M}|\mathbf{y}, \mathbf{X}, \boldsymbol{\phi}) = f(\mathbf{M}|\boldsymbol{\phi}) \quad \forall \mathbf{y}, \mathbf{X},$$

using the indicator matrix $\mathbf{M}$, defined in section 1.4.

For a mixed model with missing values in $\mathbf{X}_1$, we have

$$\begin{aligned} \mathrm{E}(y_i|X_{i1},\ldots,X_{ip}) &= \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \\ \mathrm{E}(y_i|X_{i2},\ldots,X_{ip}) &= \beta_0 + \beta_1 \tilde{X}_{i1} + \sum_{j=1}^{p} \beta_j X_{ij} \end{aligned}$$

with $\tilde{X}_{i1} = \mathrm{E}(X_{i1}|X_{i2},\ldots,X_{ip})$. This means that imputing conditional means $\tilde{X}_{i1}$ and applying least squares on the completed data produce consistent estimates assuming MCAR (Little, 1992).

MCAR Diagnosis

There are several approaches to detect missing data, which are non-MCAR. These include

- comparison of the means of $\mathbf{y}$ in the complete subsample (CC-data) and in the partially observed subsample,
- diagnostic plots, as introduced by Simon and Simonoff (1986), or
- the usage of diagnostic measures originally intended for the detection of outliers.

We will discuss the latter ideas in more detail.

Possible diagnostics include

- Cook's distance,
- the change in the residual sum of squares, or
- the change in the determinant of $\mathbf{X}'\mathbf{X}$

where originally the comparison is between the data sets $\mathbf{X}$ and $\mathbf{X}_{(i)}$, the data without case number i.

In the context of detecting a non-MCAR mechanism, the CC-data $\mathbf{X}$ and the partially observed data $\mathbf{X}_*$ are compared. Cook's distance now compares the (weighted) difference of the CC-estimator $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and the mixed-estimator $\tilde{\mathbf{b}}_\mathrm{R}$ from (1.33)

$$\frac{(\tilde{\mathbf{b}}_\mathrm{R} - \mathbf{b})'(\mathbf{X}'\mathbf{X} + \mathbf{R}'\mathbf{R})(\tilde{\mathbf{b}}_\mathrm{R} - \mathbf{b})}{p s^2_\mathrm{R}} .$$

Analogouesly, the change in the residual sum of squares (DRSS)

$$\frac{(\mathrm{RSS}_\mathrm{R} - \mathrm{RSS}_\mathrm{c})/m_*}{\mathrm{RSS}_\mathrm{c}/(m_\mathrm{c} - m_* - K + 1)}$$

and the change in the determinant (DXX)

$$\frac{\det(\mathbf{X}'\mathbf{X})}{\det(\mathbf{X}'\mathbf{X} + \mathbf{R}'\mathbf{R})}$$

are used to gain information on the nature of the missing data mechanism.

Idea

The basic idea is to compare CC and 'valid imputation assuming MCAR' (Simonoff, 1988). If MCAR does not hold, then a MCAR-imputation for the missing values in $\mathbf{X}_*$ is not adequate. If we compare the diagnostic measure to it's distribution under H_0: "MCAR is valid", we should be able to detect a possible non-MCAR process. This is more general than comparing group means (see above), as this procedure can also detect non-MCAR with $E(y) = E(y_*)$.

Distribution under H_0

The distribution of the diagnostic measures under H_0 can be investigated using a Monte-Carlo method. The algorithm is as follows

- compute $\mathbf{b}$
- replace $\mathbf{X}_*$ by 'valid imputation assuming MCAR' $\mathbf{R}$
- replace $\mathbf{y}_*$ by $\hat{\mathbf{y}}_* = \mathbf{R}\mathbf{b}_c + \boldsymbol{\epsilon}$, $\epsilon \sim N(0, \hat{\sigma}_c^2)$
- produce MCAR samples from the filled-in data repeatedly to generate a Null-distribution

The basic idea here is that, no matter what the true missing data mechanism is, the generated data will always have unobserved values that are MCAR. A basic underlying assumption that has to be fullfilled to keep type-I error under control is that the relationship between the missing values in $\mathbf{X}_*$ and the observed values can adequately be fitted by a linear regression model.

Simulation Study

A simulation study was conducted to investigate the properties of the above approach for different imputation methods and different correlation structures of the data matrix X. The structure was as follows.

- Generate $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2)$ with missing values only in $\mathbf{x}_2$.
- Repeat this step for varying $\rho = \text{corr}(x_1, x_2)$, and
- varying amount of cases with missing values.
- Consider different non-MCAR processes.

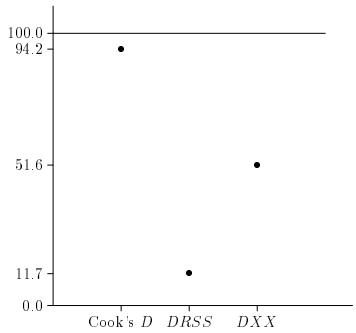
The processes generating missing values were a mean split and a variance split process. The mean split process selects a value x_{i2} as missing value with probability p_1 if $(x_{i2} - \bar{x}_2)$ exceeds a specified constant c. If $(x_{i2} - \bar{x}_2) \leq c$ the value is selected as missing value with probability p_2.

The variance split process is alike the mean split, but the absolute difference $|x_{i2} - \bar{x}_2| > c$ is used to decide if a value is selected as suitable for the missing value.

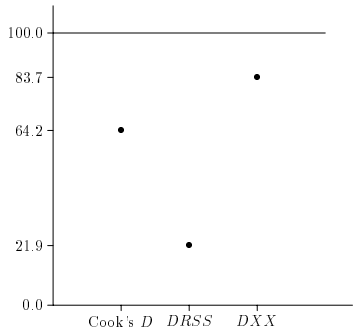
In brief, the simulation studies suggested that Cook's distance performes good for mean split while DRSS and DXX for variance split. Interestingly enough,

performance also depended on ρ. For low absolute ρ, the useage of DRSS seems to perform better, whereas for high ρ, DXX gives better results.

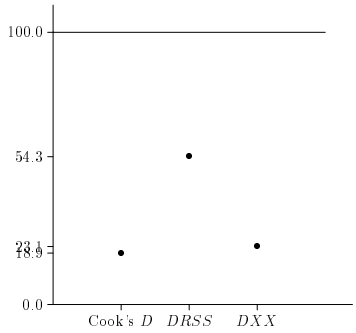
In contrast to the simulation study, the missing data mechanism is unknown in real applications so that there is no general ranking of the diagnostic measures, concerning their ability for the detection of non-MCAR processes.



Mean Split, $p_1 = 0.005$, cutoff= 0.25, $p_2 = 0.74$, FOR, $\alpha = 0.05$, $\rho = 0.7$



Variance Split, $p_1 = 0.006$, cutoff= 0.8, $p_2 = 0.7$, FOR, $\alpha = 0.05$, $\rho = 0.7$



Variance Split, $p_1 = 0.006$, cutoff= 0.8, $p_2 = 0.7$, FOR, $\alpha = 0.05$, $\rho = 0.3$

Appendix

Result A1 (Baksalary and Kala, 1983). Let $\mathbf{A}$ be a non-negative definite matrix and let $\mathbf{a}$ be a column vector. Then $\mathbf{A} - \mathbf{a}\mathbf{a}' \geq \mathbf{0} \Leftrightarrow$

$$\mathbf{a} \in \mathcal{R}(\mathbf{A}) \quad \text{and} \quad \mathbf{a}'\mathbf{A}^{-}\mathbf{a} \leq 1 \ ,$$

where $\mathbf{A}^{-}$ is any g-inverse of $\mathbf{A}$, that is, $\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$.

Result A2 (Baksalary, Liski and Trenkler, 1989). Let $\mathbf{A} = \mathbf{C}_1\mathbf{C}_1' - \mathbf{C}_2\mathbf{C}_2'$. Then $\mathbf{A} \geq 0 \Leftrightarrow$

$$\begin{aligned} &\text{(i)} \quad \mathcal{R}(\mathbf{C}_2) \subset \mathcal{R}(\mathbf{C}_1) \\ &\text{(ii)} \quad \lambda_1(\mathbf{C}_2'(\mathbf{C}_1\mathbf{C}_1')^-\mathbf{C}_2) \leq 1\,. \end{aligned}$$

Theorem A3 (Baksalary et al., 1992). Let $\mathbf{F}$ be a symmetric non-negative definite $n \times n$-matrix. Then

$$(\mathbf{I} - \mathbf{\Lambda})'\mathbf{F}(\mathbf{I} - \mathbf{\Lambda})' - \mathbf{\Lambda}'\mathbf{F}\mathbf{\Lambda} \geq \mathbf{0} \qquad \Leftrightarrow$$

1. $\mathcal{R}(\mathbf{\Lambda}'\mathbf{F}) \subset \mathcal{R}((\mathbf{I} - \mathbf{\Lambda})'\mathbf{F})$
2. $\lambda_1[\{(\mathbf{I} - \mathbf{\Lambda})'\mathbf{F}(\mathbf{I} - \mathbf{\Lambda})\}^+\mathbf{\Lambda}'\mathbf{F}\mathbf{\Lambda}] \leq 1$.

References

Baksalary, J. K. and Kala, R. (1983). Partial orderings between matrices one of which is of rank one, *Bulletin of the Polish Academy of Science, Mathematics* **31**: 5–7.

Baksalary, J. K., Liski, E. P. and Trenkler, G. (1989). Mean square error matrix improvements and admissibility of linear estimators, *Journal of Statistical Planning and Inference* **23**: 312–325.

Baksalary, J., Schipp, B. and Trenkler, G. (1992). Some further results on hermitian matrix inequalities, *Linear Algebra and its Applications* **160**: 119–129.

Goldberger, A. S. (1962). Best linear unbiased prediction in the generalized regression model, *Journal of the American Statistical Association* **57**: 369–375.

Little, R. J. A. (1992). Regression with missing X's: a review, *Journal of the American Statistical Association* **87**: 1227–1237.

Puntanen, S. and Styan, G. (1989). On the equality of the ordinary least squares estimator and the best linear unbiased estimator, *The American Statistician* **43**: 153–164.

Rao, C. R. and Toutenburg, H. (1995). *Linear Models: Least Squares and Alternatives*, Springer, New York.

Rubin, D. B. (1976). Inference and missing data, *Biometrika* **63**: 581–592.

Shalabh (1995). Performance of Stein-rule procedure for simultaneous prediction of actual and average values of study variable in linear regression model, *Bulletin of the International Statistical Institute* **56**: 1375–1390.

Simon, G. A. and Simonoff, J. S. (1986). Diagnostic plots for missing data in least squares regression, *Journal of the American Statistical Association* **81**: 501–509.

Simonoff, J. S. (1988). Regression diagnostics to detect nonrandom missingness in linear regression, *Technometrics* **30**: 205–214.

Theil, H. and Goldberger, A. S. (1961). On pure and mixed estimation in econometrics, *International Economic Review* **2**: 65–78.

Toutenburg, H. and Shalabh (1996). Predictive performance of the methods of restricted and mixed regression estimators, *Biometrical Journal* **38**: 951–959.