Schaffrin, Toutenburg:

The Impact of Missing Values on the Reliability Measures in a Linear Model

Sonderforschungsbereich 386, Paper 125 (1998)

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

The Impact of Missing Values on the Reliability Measures in a Linear Model

Burkhard Schaffrin [*] Helge Toutenburg [**]

August 13, 1998

Abstract

Reliability measures in linear models are used in geodetic science and elsewhere to quantify the potential to detect outliers and to suppress their impact on the regression estimates. Here we shall study the effect of missing values on these reliability measures with the idea that, under a proper design, they should not change drastically when such a situation occurs.

1 Introduction

Since Baarda (1976) defined reliability measures to quantify the potential to detect outliers in a linear model, this technique has found wide applications in geodesy, photogrammetry, mapping and related areas. Generalizations to include the case of correlated measurements have been proposed quite recently by Schaffrin (1997). It is still unclear, however, what the effect of missing values would be on these reliability measures. This will be studied in more detail in this contribution, thereby relying on the previous studies of Toutenburg, Heumann, Fieger and Park (1995) and Toutenburg, Fieger and Heumann (1999) concerning the missing values problem in regression with particular emphasis on mixed and weighted mixed estimation.

After a brief overview on reliability measures in chapter 1, we shall investigate the situation of missing values in the observation vector as well as in the design matrix in chapter 2 before presenting a short example and drawing some conclusions. The point we want to stress is that not only do we need sufficient reliability in our systems, but also should this reliability not too much be affected by missing values. The results from this study will help to illuminate the situation.

[*] Department of Civil and Environmental Engineering and Geodetic Science, Ohio State University, Columbus, OH 43210-1275, U.S.A.

[**] Institut für Statistik Ludwig-Maximilians-Universität München 80539 München, Germany

2 Reliability Measures in a Linear Model

Let us introduce the Gauß-Markov regression model for uncorrelated observations as

$$y \;=\; \underset{n\times m}{X} \;\; \beta + e, \qquad \mathrm{rk} X = m, \tag{2.1}$$

$$e \sim (0, \sigma_0^2 P^{-1}), \qquad \underset{n\times n}{P} \;:= \mathrm{Diag}(p_j), \tag{2.2}$$

with the unknown $m \times 1$ parameter vector β and the unknown variance component σ_0^2; $Q := P^{-1}$ may be called "cofactor matrix".

Before we consider the case of missing values we give a brief review of Baarda's "reliability measures" as derived in the context of weighted least-squares estimation. The corresponding normal equations are

$$N\hat{\beta} = c \qquad \text{for } [N, c] := X^T P[X, y] \tag{2.3}$$

so that we obtain sequentially

$$\hat{\beta} = N^{-1}c \;\;\sim\;\; (\beta, \sigma_0^2 N^{-1}) \tag{2.4}$$

$$\tilde{e} = y - X\hat{\beta} \;\;\sim\;\; (0, \sigma_0^2(P^{-1} - XN^{-1}X^T)) \tag{2.5}$$

$$\hat{\sigma_0^2} = (n-m)^{-1}\, \tilde{e}^T P \tilde{e} \;\;=\;\; (n-m)^{-1}\, (y^T P y - c^T \hat{\beta}) \tag{2.6}$$

with $\mathrm{E}\{\hat{\sigma}_0^2\} = \sigma_0^2$ for any arbitrary $\sigma_0^2 \in I\!R_+$. E denotes 'expectation', and the symbol

$$Q_{\tilde{e}} := P^{-1} - XN^{-1}X^T \tag{2.7}$$

may denote the cofactor matrix of the residual vector $\tilde{e}$.

Should there be a single outlier in the j-th observation we would have to replace the characterization of the observational errors in (2.2) by

$$e \sim (\eta_j\, \beta_0^{(j)}, \sigma_0^2 P^{-1}), \qquad \eta_j^T := [0, \ldots, 1, \ldots 0]\,, \tag{2.8}$$

with η_j as j-th $n \times 1$ unit vector and with $\beta_0^{(j)}$ denoting the size of the outlier. New least-squares estimates can now be taken from the augmented normal equations

$$\begin{bmatrix} \eta_j^T P \eta_j & \eta_j^T P X \\ X^T P \eta_j & N \end{bmatrix} \begin{bmatrix} \hat{\beta}_0^{(j)} \\ \hat{\beta}^{(j)} \end{bmatrix} = \begin{bmatrix} \eta_j^T P y \\ c \end{bmatrix} \tag{2.9}$$

leading to

$$\hat{\beta}_0^{(j)} \;=\; \frac{\eta_j^T P Q_{\tilde{e}} P y}{\eta_j^T P Q_{\tilde{e}} P \eta_j} \tag{2.10}$$

$$\;=\; \frac{\tilde{e}_j}{r_j} \qquad \text{for } r_j := (Q_{\tilde{e}} P)_{jj}, \tag{2.11}$$

and to

$$\begin{aligned}\hat{\xi}^{(j)} &= N^{-1}(c - X^T P \eta_j \, \hat{\beta}_0^{(j)}) = \hat{\xi} - \delta\hat{\xi}^{(j)} && (2.12)\\ \delta\hat{\xi}^{(j)} &= N^{-1} X^T P \eta_j \, (\eta_j^T P Q_{\tilde{e}} P \eta_j)^{-1} \, (\eta_j^T P Q_{\tilde{e}} P y) && (2.13)\end{aligned}$$

where $\delta\hat{\xi}^{(j)}$ describes the effect of the outlier on the estimated parameters. The effect on the corresponding dispersion matrix is given by

$$\begin{aligned}\delta \mathrm{D}\{\hat{\xi}^{(j)}\} &:= \mathrm{D}\{\hat{\xi}\} - \mathrm{D}\{\hat{\xi}^{(j)}\}\\ &= -\sigma_0^2 \, N^{-1} X^T P \eta_j (\eta_j^T P Q_{\tilde{e}} P \eta_j)^{-1} \eta_j^T P X N^{-1} && (2.14)\end{aligned}$$

which seems to indicate a “gain in efficiency” when neglecting the outlier (in spite of its existence). The original residual vector, however, was shifted by

$$\delta\tilde{e} = (Q_{\tilde{e}} P) \, [y - (y - \eta_j \beta_0^{(j)})] = (Q_{\tilde{e}} P)\eta_j \, \xi_0^{(j)}, \qquad (2.15)$$

and in its j-th component by

$$\delta\tilde{e}_j = \eta_j^T (Q_{\tilde{e}} P) \eta_j \, \xi_0^{(j)} = r_j \, \xi_0^{(j)} \qquad (2.16)$$

due to the neglected outlier in the j-th observation. The interpretation of both (2.11) and (2.16) allows us to state that outliers of a given size can be detected more easily by inspecting the corresponding residual when the so-called “redundancy number” r_j is relatively large. (Note that the situation changes completely as soon as correlated observations are involved; see e.g. Schaffrin (1997).)

The redundancy number r_j is the j-th diagonal element of the matrix $Q_{\tilde{e}} P = I_n - X N^{-1} X^T P$, or equivalently of the matrix $P^{1/2} Q_{\tilde{e}} P^{1/2}$, with $P^{1/2} := \mathrm{Diag}(p_j^{1/2})$, since P is diagonal. As idempotent and symmetric matrix $P^{1/2} Q_{\tilde{e}} \, P^{1/2}$ represents an orthogonal projection with bounded diagonal elements

$$0 \leq r_j = \frac{\sigma_0^2}{p_j \, \mathrm{D}\{\hat{\xi}_0^{(j)}\}} \leq 1 \qquad (2.17)$$

with an average size of $\frac{n-m}{n}$ since the trace

$$\mathrm{tr}(Q_{\tilde{e}} P) = r_1 + \ldots + r_n = \mathrm{tr}\, I_n - \mathrm{tr}(N^{-1} X^T P X) = n - m \qquad (2.18)$$

yields the (original) “degrees-of-freedom” of the model. If reliability is of concern, our goal must therefore be to design the model in such a way that the redundancy numbers are not too far away from their average value, and that their values do not change too much in the case of missing values. It is the latter question that we shall further investigate in the following.

3 Reliability Measures with Missing Values in the Model

3.1 The case of one missing observation (exogeneous variable)

Let us first investigate the case where one observation, say y_k, is missing and its impact on the reliability measures r_j for $j \neq k$. If x_k^T denotes the k-th row of the matrix X we would have to replace the matrix $N = X^T PX$ by the "reduced" matrix $N - x_k p_k x_k^T$ throughout, with the corresponding inverse

$$(N - x_k p_k x_k^T)^{-1} = N^{-1} + N^{-1} x_k (p_k^{-1} - x_k^T N^{-1} x_k)^{-1} x_k^T N^{-1} \tag{3.1}$$

assuming that it exists for all $k \in \{1, \ldots, n\}$. Thus, for $j \neq k$, the modified reliability measure becomes

$$\begin{aligned} r'_j &:= 1 - x_j^T (N - x_k p_k x_k^T)^{-1} x_j \; p_j \\ &= 1 - (x_j^T N^{-1} x_j) p_j - x_j^T N^{-1} x_k (p_k^{-1} - x_k^T N^{-1} x_k)^{-1} x_k^T N^{-1} x_j \; p_j \\ &= r_j - \frac{(x_j^T N^{-1} x_k)^2 \; p_j}{p_k^{-1} - x_k^T N^{-1} x_k} \\ &= r_j - \frac{p_j p_k \; (x_j^T N^{-1} x_k)^2}{r_k}. \end{aligned} \tag{3.2}$$

Obviously the reliability has deteriorated by an amount which primarily depends on r_k and $x_j^T N^{-1} x_k$. In this sense the deterioration is becoming relatively small if a highly controlled measurement fails to be collected (r_k large), or if the expected correlation between the "adjusted observations" involved, namely $\mathrm{C}\{x_j^T \hat{\beta}, x_k^T \hat{\beta}\}$, turns out to be negligible.

3.2 The case of missing endogeneous variables in one row

In this case we assume that all observations y_j, $j \in \{1, \ldots, n\}$, were taken except that only some values in the k-th row x_k^T of the matrix X are missing. Following one of the techniques as described by Jänner (1993), Toutenburg et al. (1995) or Toutenburg et al. (1999) we may use a substitute row $\bar{x}_k^T$ for x_k^T, generating the possible "bias"

$$\delta_k := (\bar{x}_k^T - x_k^T)\beta. \tag{3.3}$$

Therefore, the matrix $N = X^T PX$ is now to be replaced by $\bar{N} := (N - x_k p_k x_k^T) + \bar{x}_k p_k \bar{x}_k^T$ whose inverse is readily obtained as

$$\begin{aligned} \bar{N}^{-1} &= [(N - x_k p_k x_k^T) + \bar{x}_k p_k \bar{x}_k^T]^{-1} \\ &= (N - x_k p_k x_k^T)^{-1} - (N - x_k p_k x_k^T)^{-1} \bar{x}_k \times \\ &\quad \times [p_k^{-1} + \bar{x}_k^T (N - x_k p_k x_k^T)^{-1} \bar{x}_k]^{-1} \bar{x}_k^T (N - x_k p_k x_k^T)^{-1}. \end{aligned} \tag{3.4}$$

The corresponding reliability measures r'_j consequently change (for $j \neq k$) to

$$\begin{aligned} \bar{r}_j &:= 1 - x_j^T [(N - x_k p_k x_k^T) + \bar{x}_k p_k \bar{x}_k^T]^{-1} x_j \; p_j \\ &= 1 - x_j^T (N - x_k p_k x_k^T)^{-1} x_j p_j + x_j^T (N - x_k p_k x_k^T)^{-1} \bar{x}_k \times \\ &\quad \times [p_k^{-1} + \bar{x}_k^T (N - x_k p_k x_k^T)^{-1} \bar{x}_k]^{-1} \bar{x}_k^T (N - x_k p_k x_k^T)^{-1} x_j p_j \\ &= r'_j + \frac{p_j p_k \; [x_j^T (N - x_k p_k x_k^T)^{-1} \bar{x}_k]^2}{2 - \bar{r}'_k} \end{aligned} \tag{3.5}$$

where

$$\bar{r}'_k := 1 - \bar{x}_k^T (N - x_k p_k x_k^T)^{-1} \bar{x}_k \; p_k \tag{3.6}$$

in analogy to the definition (3.2). For $j = k$ we have

$$\begin{aligned} \bar{r}_k &:= 1 - \bar{x}_k^T \bar{N}^{-1} \bar{x}_k \; p_k \\ &= 1 - \bar{x}_k^T [(N - x_k p_k x_k^T) + \bar{x}_k p_k \bar{x}_k^T]^{-1} \bar{x}_k \; p_k \\ &= \bar{r}_k' + \frac{(1-\bar{r}_k')^2}{2-\bar{r}_k'} = \frac{1}{2-\bar{r}_k'} \end{aligned}$$

or, conversely,

$$\bar{r}_k' = 2 - \bar{r}_k^{-1} \tag{3.7}$$

and, furthermore, for $j \neq k$ we get

$$r_j' = \bar{r}_j - p_j p_k \; [x_j^T (N - x_k p_k x_k^T)^{-1} \bar{x}_k]^2 \; \bar{r}_k. \tag{3.8}$$

The above formulas (3.5) to (3.8) have been derived on the basis of the matrix $N - x_k p_k x_k^T$ which does not rely on the substitute row $\bar{x}_k^T$ per se. If we decide to use the matrix $\bar{N}$ instead we would, for instance, obtain the following relations, for $(j \neq k)$

$$\begin{aligned} r_j' &= \bar{r}_j - p_j p_k \; \bar{r}_k [x_j^T (\bar{N} - \bar{x}_k p_k \bar{x}_k^T)^{-1} \bar{x}_k]^2 \\ &= \bar{r}_j - p_j p_k \; \bar{r}_k [x_j^T \bar{N}^{-1} \bar{x}_k + x_j^T \bar{N}^{-1} \bar{x}_k p_k (1 - \bar{x}_k^T \bar{N}^{-1} \bar{x}_k \; p_k)^{-1} \bar{x}_k^T \bar{N}^{-1} \bar{x}_k]^2 \\ &= \bar{r}_j - p_j p_k \; \bar{r}_k (x_j^T \bar{N}^{-1} \bar{x}_k)^2 \left(1 + \frac{1-\bar{r}_k}{\bar{r}_k}\right)^2 \\ &= \bar{r}_j - \frac{p_j p_k (x_j^T \bar{N}^{-1} \bar{x}_k)^2}{\bar{r}_k} \end{aligned} \tag{3.9}$$

in formal analogy to (3.2), or by applying (3.7),

$$\bar{r}_j = r_j' + p_j p_k (x_j^T \bar{N}^{-1} \bar{x}_k)^2 (2 - \bar{r}_k') \tag{3.10}$$

with

$$\bar{r}_k' = 1 - \bar{x}_k^T (\bar{N} - \bar{x}_k p_k \bar{x}_k^T)^{-1} \bar{x}_k \; p_k. \tag{3.11}$$

We notice that the above are obviously not just formal relations since they are independent of the row x_k^T with the missing values. However, they do not provide knowledge about the effect of the substitute row $\bar{x}_k^T$ on the original reliability measures r_j (rather than r_j'). Let us now try to tackle this problem by recombining some of the results achieved so far. We start with the fundamental decomposition (for $j \neq k$)

$$\begin{aligned} \bar{r}_j - r_j &= (\bar{r}_j - r_j') - (r_j - r_j') \\ &= p_j p_k \left[\frac{(x_j^T \bar{N}^{-1} \bar{x}_k^T)^2}{\bar{r}_k} - \frac{(x_j^T N^{-1} x_k)^2}{r_k} \right] \\ &= p_j p_k [x_j^T (\bar{N} - \bar{x}_k p_k \bar{x}_k^T)^{-1} \bar{x}_k]^2 \; \bar{r}_k - p_j p_k [x_j^T (N - x_k p_k x_k^T)^{-1} x_k]^2 \; r_k \end{aligned} \tag{3.12}$$

where obviously

$$N - x_k p_k x_k^T = \bar{N} - \bar{x}_k p_k \bar{x}_k^T \tag{3.13}$$

holds true, according to the definition of $\bar{N}$. Thus we obtain

$$\bar{r}_j - r_j = p_j p_k \, x_j^T (N - x_k p_k x_k^T)^{-1} (\bar{x}_k \bar{r}_k \bar{x}_k^T - x_k r_k x_k^T)(N - x_k p_k x_k^T)^{-1} x_j \tag{3.14}$$

which shows that the difference between the original reliability measure r_j and its counterpart $\bar{r}_j$ is governed by the difference between the dyadic matrices $\bar{x}_k \bar{r}_k \bar{x}_k^T$ and $x_k r_k x_k^T$. By applying formula (3.7) and its analogon $r_k = (2 - r'_k)^{-1}$ we see that we have to compare the matrix $\bar{x}_k (2 - \bar{r}'_k)^{-1} \bar{x}_k^T$ with the matrix $x_k (2 - r'_k)^{-1} x_k^T$, both of rank 1 while symmetric and positive-semidefinite. Following a theorem by Baksalary and Kala (1983), $\bar{r}_j - r_j$ will turn out nonnegative if and and only if $\bar{x}_k$ is proportionate to x_k with

$$x_k^T (\bar{x}_k \bar{r}_k \bar{x}_k^T)^- x_k r_k \leq 1 \tag{3.15}$$

or

$$x_k^T (\bar{x}_k \bar{x}_k^T)^- x_k \leq \frac{\bar{r}_k}{r_k} = \frac{2 - r'_k}{2 - \bar{r}'_k} \tag{3.16}$$

for any arbitrary choice of the g-inverses. One such choice would certainly be

$$(\bar{x}_k \bar{x}_k^T)^- = \bar{x}_k (\bar{x}_k^T \bar{x}_k)^{-2} \bar{x}_k^T \tag{3.17}$$

in which case we arrive at the inequality

$$r_k \, (x_k^T \bar{x}_k)^2 \leq \bar{r}_k \, (\bar{x}_k^T \bar{x}_k)^2 \tag{3.18}$$

or

$$\frac{(x_k^T \bar{x}_k)^2}{2 - r'_k} \leq \frac{(\bar{x}_k^T \bar{x}_k)^2}{2 - \bar{r}'_k} \tag{3.19}$$

to ensure superiority of $\bar{r}_j$ over r_j for any $j \neq k$. In contrast, for $j = k$ we have the trivial relations

$$\bar{r}_k - r_k = (2 - \bar{r}'_k)^{-1} - (2 - r'_k)^{-1} \geq 0 \tag{3.20}$$

if and only if

$$(2 - r'_k) - (2 - \bar{r}'_k) = \bar{r}'_k - r'_k \geq 0 \, . \tag{3.21}$$

This concludes our analysis of missing endogenous variables in one row only. The case of several rows can be treated along the same lines and is therefore omitted here.

4 A Simple Example

We refer to the simple regression example used by Toutenburg (1992) and again by Rao and Toutenburg (1995). The observations are collected in the 10 × 1 response vector

$$y := [18, 47, 125, 40, 37, 20, 24, 35, 59, 50]^T$$

while the 10×2 coefficient matrix X consists of the following values (row-wise)

$$X^T = [x_1, \ldots, x_{10}] := \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -10 & 19 & 100 & 17 & 13 & 10 & 5 & 22 & 35 & 20 \end{bmatrix}$$

and the weight matrix P is the 10×10 identity matrix $P = I_{10}$.

Consequently, we obtain the cofactor matrix of the residuals by

$$Q_{\tilde{e}} = I_{10} - X(X^TX)^{-1}X^T$$

with

$$(X^TX)^{-1} = \frac{1}{78169} \begin{bmatrix} 13153 & -231 \\ -231 & 10 \end{bmatrix} = N^{-1}$$

and, therefore, the following "redundancy numbers" which may here serve as reliability measures as well:

$r_1 = 0.760$	$r_2 = 0.898$	$r_3 = 0.1435$	$r_4 = 0.895$	$r_5 = 0.897$
$r_6 = 0.878$	$r_7 = 0.858$	$r_8 = 0.900$	$r_9 = 0.882$	$r_{10} = 0.899$

With the exception of the third observation y_3, every other observation appears to be as reliable as expected in view of an expected average value of $8/10 = 0.800$; we call y_3 an observation with "high leverage", or little reliability, due to the small number r_3 which would make the detection of any outlier in this particular measurement most difficult.

Now let us form the 10×10 matrix R' of modified reliability measures r'_j ($j \neq k$, thus without diagonal elements) where the k-th column refers to the case of y_k missing: $(R')_{jk} := r'_j$ (missing y_k).

If, in addition, we fill in the diagonal elements $(R')_{kk} := r'_k = 2 - r_k^{-1}$ we readily obtain $R' =$

$$\begin{bmatrix} 0.684 & 0.745 & \underline{0.405} & 0.742 & 0.737 & 0.733 & 0.724 & 0.748 & 0.757 & 0.746 \\ 0.880 & 0.886 & 0.873 & 0.886 & 0.886 & 0.885 & 0.884 & 0.887 & 0.888 & 0.886 \\ \underline{0.076} & 0.140 & -\underline{4.969} & 0.141 & 0.143 & 0.142 & 0.136 & 0.134 & \underline{0.089} & 0.138 \\ 0.874 & 0.883 & 0.884 & 0.883 & 0.882 & 0.881 & 0.880 & 0.884 & 0.886 & 0.883 \\ 0.860 & 0.875 & 0.887 & 0.874 & 0.873 & 0.871 & 0.869 & 0.875 & 0.879 & 0.875 \\ 0.846 & 0.865 & 0.872 & 0.864 & 0.863 & 0.861 & 0.858 & 0.866 & 0.871 & 0.866 \\ 0.817 & 0.845 & 0.816 & 0.844 & 0.841 & 0.839 & 0.835 & 0.846 & 0.852 & 0.845 \\ 0.886 & 0.889 & 0.845 & 0.889 & 0.888 & 0.888 & 0.888 & 0.889 & 0.889 & 0.889 \\ 0.879 & 0.872 & \underline{0.554} & 0.873 & 0.874 & 0.875 & 0.876 & 0.871 & 0.866 & 0.872 \\ 0.882 & 0.887 & 0.865 & 0.887 & 0.887 & 0.886 & 0.886 & 0.888 & 0.889 & 0.888 \end{bmatrix}$$

Note that the non-diagonal elements per column must sum up to 7. From this matrix it becomes obvious that the third observation approximately doubles its already high leverage,—respectively cuts its already low reliability by half—if either y_1 or y_9 is missing. Conversely, if y_3 should be missing it would have a deteriorating effect on the reliability of both y_1 and y_9, although critical limits are not yet reached. All the other pairs of observations are mutually unaffected when either one of them should turn out missing.

Now let us consider the case when exactly one value in the second row of X^T is missing and subsequently replaced by the average of the remaining elements.

Thus we obtain for the $\bar{x}_{k\,2}$:

$\bar{x}_{1\,2} = 26.78$	$\bar{x}_{2\,2} = 23.56$	$\bar{x}_{3\,2} = 14.56$	$\bar{x}_{4\,2} = 23.78$	$\bar{x}_{5\,2} = 24.22$
$\bar{x}_{6\,2} = 24.56$	$\bar{x}_{7\,2} = 25.11$	$\bar{x}_{8\,2} = 23.22$	$\bar{x}_{9\,2} = 21.78$	$\bar{x}_{10\,2} = 23.44$

and, moreover, the following modified inverses

$$
\begin{aligned}
(X'X - x_1 x_1^T + \bar{x}_1 \bar{x}_1^T)^{-1} &= \begin{bmatrix} 10 & 267.78 \\ 267.78 & 13770.17 \end{bmatrix}^{-1} \\
&= \frac{1}{65995.56} \begin{bmatrix} 13770.17 & -267.78 \\ -267.78 & 10 \end{bmatrix} \\
(X'X - x_2 x_2^T + \bar{x}_2 \bar{x}_2^T)^{-1} &= \begin{bmatrix} 10 & 235.56 \\ 235.56 & 13347.07 \end{bmatrix}^{-1} \\
&= \frac{1}{77982.22} \begin{bmatrix} 13347.07 & -235.56 \\ -235.56 & 10 \end{bmatrix} \\
(X'X - x_3 x_3^T + \bar{x}_3 \bar{x}_3^T)^{-1} &= \begin{bmatrix} 10 & 145.56 \\ 145.56 & 3364.99 \end{bmatrix}^{-1} \\
&= \frac{1}{12462.22} \begin{bmatrix} 3364.99 & -145.56 \\ -145.56 & 10 \end{bmatrix} \\
(X'X - x_4 x_4^T + \bar{x}_4 \bar{x}_4^T)^{-1} &= \begin{bmatrix} 10 & 237.78 \\ 237.78 & 13429.49 \end{bmatrix}^{-1} \\
&= \frac{1}{77755.56} \begin{bmatrix} 13429.49 & -237.78 \\ -237.78 & 10 \end{bmatrix} \\
(X'X - x_5 x_5^T + \bar{x}_5 \bar{x}_5^T)^{-1} &= \begin{bmatrix} 10 & 242.22 \\ 242.22 & 13570.61 \end{bmatrix}^{-1} \\
&= \frac{1}{77035.56} \begin{bmatrix} 13570.61 & -242.22 \\ -242.22 & 10 \end{bmatrix} \\
(X'X - x_6 x_6^T + \bar{x}_6 \bar{x}_6^T)^{-1} &= \begin{bmatrix} 10 & 245.56 \\ 245.56 & 13656.19 \end{bmatrix}^{-1} \\
&= \frac{1}{76528.89} \begin{bmatrix} 13656.19 & -245.56 \\ -245.56 & 10 \end{bmatrix} \\
(X'X - x_7 x_7^T + \bar{x}_7 \bar{x}_7^T)^{-1} &= \begin{bmatrix} 10 & 251.11 \\ 251.11 & 13758.51 \end{bmatrix}^{-1} \\
&= \frac{1}{74528.89} \begin{bmatrix} 13758.51 & -251.11 \\ -251.11 & 10 \end{bmatrix} \\
(X'X - x_8 x_8^T + \bar{x}_8 \bar{x}_8^T)^{-1} &= \begin{bmatrix} 10 & 232.22 \\ 232.22 & 13208.17 \end{bmatrix}^{-1} \\
&= \frac{1}{78155.56} \begin{bmatrix} 13208.17 & -232.22 \\ -232.22 & 10 \end{bmatrix} \\
(X'X - x_9 x_9^T + \bar{x}_9 \bar{x}_9^T)^{-1} &= \begin{bmatrix} 10 & 217.78 \\ 217.78 & 12402.37 \end{bmatrix}^{-1} \\
&= \frac{1}{76595.56} \begin{bmatrix} 12402.37 & -217.78 \\ -217.78 & 10 \end{bmatrix}
\end{aligned}
$$

$$\begin{aligned}(X'X - x_{10}x_{10}^T + \bar{x}_{10}\bar{x}_{10}^T)^{-1} &= \begin{bmatrix} 10 & 234.44 \\ 234.44 & 13302.43 \end{bmatrix}^{-1} \\ &= \frac{1}{78062.22}\begin{bmatrix} 13302.43 & -234.44 \\ -234.44 & 10 \end{bmatrix}\end{aligned}$$

Next, we form the 10×10 matrix $\bar{R}$ which shows the substitute reliability measures $\bar{r}_j$ ($j \neq k$, thus without diagonal elements again) where the k-th column indicates which element in the second row of X^T was missing:

$$(\bar{R})_{jk} := \bar{r}_j \text{ (missing } x_{k\,2}).$$

In addition, we may fill in the diagonal elements

$$(\bar{R})_{kk} := \bar{r}_k = \frac{1}{2 - \bar{r}'_k}$$

with

$$\bar{r}'_k := 1 - \bar{x}_k^T(N - x_k x_k^T)^{-1}\bar{x}_k$$

and arrive at $\bar{R} =$

$$\begin{bmatrix}
\underline{0.900} & 0.756 & \underline{0.416} & 0.753 & 0.748 & 0.743 & 0.735 & 0.759 & 0.768 & 0.757 \\
0.891 & 0.900 & 0.884 & 0.897 & 0.896 & 0.896 & 0.895 & 0.898 & 0.899 & 0.897 \\
\underline{0.088} & 0.151 & \underline{0.900} & 0.152 & 0.155 & 0.154 & 0.148 & 0.146 & \underline{0.101} & 0.149 \\
0.886 & 0.894 & 0.895 & 0.900 & 0.893 & 0.892 & 0.891 & 0.895 & 0.897 & 0.895 \\
0.871 & 0.885 & 0.898 & 0.885 & 0.900 & 0.882 & 0.880 & 0.887 & 0.890 & 0.886 \\
0.857 & 0.876 & 0.883 & 0.876 & 0.874 & 0.900 & 0.869 & 0.878 & 0.882 & 0.877 \\
0.828 & 0.856 & 0.827 & 0.855 & 0.852 & 0.850 & 0.900 & 0.858 & 0.863 & 0.856 \\
0.896 & 0.900 & 0.856 & 0.900 & 0.899 & 0.899 & 0.899 & 0.900 & 0.900 & 0.900 \\
0.890 & 0.883 & \underline{0.564} & 0.884 & 0.885 & 0.886 & 0.887 & 0.882 & 0.900 & 0.883 \\
0.893 & 0.898 & 0.876 & 0.898 & 0.898 & 0.898 & 0.896 & 0.899 & 0.900 & 0.900
\end{bmatrix}$$

Note that now all the elements per column must sum up to 8, including the diagonal elements which equally show the value of 0.900 due to the way we constructed the substitute elements $\bar{x}_{k\,2}$.

All the elements in the matrix $\bar{R}$ turn necessarily out larger than their corresponding entries in the matrix R' which shows the positive impact, in general, of the imputation method used. Specifically, however, the third observation can no longer be identified as of "high leverage" so that it may be treated as an outlier more often than not. In this case the reliability measures of both y_1 and y_9 drop considerably, with or without imputation. (More sophisticated imputation techniques can be considered along the same lines.)

5 Conclusions

Traditionally reliability measures are computed for every observation in a linear model in order to quantify their potential to detect ouliers. These reliability measures may be more or less drastically affected by the occurrence of the “missing value” situation. We have studied some of the typical cases and provided analytical formulas for them so that counter-measures can be taken in time to avoid problems associated with the lack of control among some of the data, assuming they are uncorrelated. The case of correlated observations will be treated elsewhere.

Acknowledgement

This study was completed while the first author, Burkhard Schaffrin, was visiting faculty of the Sonderforschungsbereich 386: “Statistische Analyse diskreter Strukturen, Modellierung und Anwendung in Biometrie und Ökonometrie” at the Ludwig-Maximilians University in Munich, Germany, with Helge Toutenburg as his host. This and the financial support through the German Science Foundation (DFG) is most gratefully acknowledged.

References

Baarda, W. (1976). Reliability and precision of networks, *Proceedings of the VII. Intl. Course for Surveying Engineering*, Darmstadt.

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.

Jänner, M. (1993). *Neue Ansätze zur Lösung des Problems fehlender Werte im linearen Regressionsmodell*, Peter Lang Europäischer Verlag der Wissenschaften, Frankfurt am Main.

Rao, C. R. and Toutenburg, H. (1995). *Linear Models: Least Squares and Alternatives (corrected second printing, 1997)*, Springer, New York.

Schaffrin, B. (1997). Reliability measures for correlated observations, *Journal of Surveying Engineering* **123**: 126–133.

Toutenburg, H. (1992). *Lineare Modelle*, Physica, Heidelberg.

Toutenburg, H., Fieger, A. and Heumann, C. (1999). Regression modelling with fixed effects – missing values and other problems, *in* C. R. Rao (ed.), *Statistics of the 21st Century*, Springer, New York.

Toutenburg, H., Heumann, C., Fieger, A. and Park, S. H. (1995). Missing values in regression: Mixed and weighted mixed estimation, *in* V. Mammitzsch and H. Schneeweiß (eds), *Gauss Symposium*, de Gruyter, Berlin, pp. 289–301.