Toutenburg, Fieger, Srivastava:

Weighted Modified First Order Regression Procedures for Estimation in Linear Models with Missing X-Observations

Sonderforschungsbereich 386, Paper 127 (1998)

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

Projektpartner

Weighted Modified First Order Regression Procedures for Estimation in Linear Models with Missing X-Observations

H. Toutenburg
Institute of Statistics
University of Munich
80799 Munich, Germany

A. Fieger
Institute of Statistics
University of Munich
80799 Munich, Germany

V.K. Srivastava
Department of Statistics
Lucknow University
Lucknow 226007, India

September 2, 1998

Abstract

This paper considers the estimation of coefficients in a linear regression model with missing observations in the independent variables and introduces a modification of the standard first order regression method for imputation of missing values. The modification provides stochastic values for imputation and, as an extension, makes use of the principle of weighted mixed regression. The proposed procedures are compared with two popular procedures—one which utilizes only the complete observations and the other which employs the standard first order regression imputation method for missing values.

A simulation experiment to evaluate the gain in efficiency and to examine interesting issues like the impact of varying degree of multicollinearity in explanatory variables is proceeded. Some work on the case of discrete regressor variables is in progress and will be reported in a future article to follow.

1 Introduction

It is not uncommon in many applications of the regression analysis that some values of certain explanatory variables are not available due to one reason or the other. A simple solution is then to discard the available values of other variables in the model and to confine attention to the complete data only. Such a solution, it is well known, has often serious statistical consequences and is surely not an efficient strategy. An alternative solution is to plug in imputed values for missing observations and then to carry out the regression analysis. Such imputed values can be obtained in several ways; see, e.g., Little and Rubin (1987) for basic considerations and Little (1992) for a detailed discussion of missing X-values in regression, and Rao and Toutenburg (1995) for a detailed account of MDE-superiority investigations for imputation methods. When these imputed values are non-stochastic, the application of the least squares procedure for the estimation of regression coefficients generally yields biased and inconsistent estimators; see, e.g., Toutenburg, Heumann, Fieger and Park (1995), who have examined the efficiency properties of such procedures with respect to the procedure that uses only the complete observations and provides unbiased estimators of regression coefficients. This raises an interesting issue related to efficiency properties of procedures which employ stochastic values for imputation of missing observations on explanatory variables. This article is a modest attempt in this direction.

We consider the imputation method based on the first order regression. This method and some modifications are discussed in Buck (1960), Afifi and Elashoff (1966) and Dagenais (1973). It essentially amounts to running the auxiliary regressions of each one of explanatory variables (for which some values are missing) on the remaining explanatory variables (for which no value is missing) employing the complete observations only. The estimated equations are then used to formulate predictors for missing values. The thus obtained predicted values are then utilized as substitutes for missing observations on the explanatory variables. This leads to a complete data set and now the regression analysis is performed. As all the observations on the study variable are available, we can easily include the study variable also in the capacity of an additional explanatory variable while running the auxiliary regressions in a bid to utilize all the available information on the variables. This will lead to another imputation method which can be termed as modified first order regression method, and will obviously provide imputed values that are no more non-stochastic. This method was presented in Toutenburg, Srivastava and Fieger (1996) in full detail. Moreover superiority conditions were deduced using large sample asymptotics. Examining the

impact of such imputed values on the estimation of regression coefficients by simulation is the objective of our present investigation.

The plan of this article is as follows. In Section 2, we present the model specification and describe the alternative estimation procedures for the regression coefficients. One is the procedure that discards the incomplete portion of the data while the remaining two employ imputed values obtained from first order regressions. Out of these two, one uses non-stochastic values for imputation while the other uses stochastic values, whereas additionally a weighting procedure is adapted. In Section 3 we discuss efficiency properties and define the risk function. Section 4 describes the model used for simulation and presents efficiency comparisons of the various estimators.

2 Model Specification And Estimation Procedure

Let us consider the following linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

which is structured as follows:

$$\mathbf{y}_c = \mathbf{X}_c\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}_c\,, \tag{2.1}$$

$$\mathbf{y}_* = \mathbf{X}_*\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}_*\,, \tag{2.2}$$

where $\mathbf{y}_c$ and $\mathbf{y}_*$ denote $m_c \times 1$ and $m_* \times 1$ vectors of observations on the study variable, $\mathbf{X}_c$ and $\mathbf{X}_*$ are $m_c \times K$ and $m_* \times K$ matrices of observations on K explanatory variables, $\boldsymbol{\beta}$ is a $K \times 1$ vector of unknown regression coefficients, $\boldsymbol{\epsilon}_c$ and $\boldsymbol{\epsilon}_*$ are $m_c \times 1$ and $m_* \times 1$ vectors of disturbances and σ is a scalar.

It is assumed that the matrix $\mathbf{X}_*$ is partially observed and contains some missing values. To keep the exposition simple but without any loss of generality, let us assume that the values of the last explanatory variable in $\mathbf{X}_*$ are missing. Thus we can express $\mathbf{X}_*$ as $[\mathbf{Z}_*\,, \mathbf{x}_*]$ where $\mathbf{Z}_*$ is $m_* \times (K-1)$ matrix with no missing values and $\mathbf{x}_*$ is the last column vector with all missing values. Accordingly partitioning $\mathbf{X}_c$ and $\boldsymbol{\beta}$, we write

$$\mathbf{X}_c = [\mathbf{Z}_c\,, \mathbf{x}_c]\,, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\gamma} \\ \alpha \end{pmatrix},$$

where $\mathbf{Z}_c$ comprises first $(K-1)$ column vectors of $\mathbf{X}_c$ and $\mathbf{x}_c$ is the last column vector. Similarly, $\boldsymbol{\gamma}$ denotes a column vector formed by first $(K-1)$ elements of $\boldsymbol{\beta}$ and α is the last element.

Thus we can write the model as follows:

$$\mathbf{y}_c = \mathbf{Z}_c\boldsymbol{\gamma} + \alpha\mathbf{x}_c + \sigma\boldsymbol{\epsilon}_c\,, \tag{2.3}$$

$$\mathbf{y}_* = \mathbf{Z}_*\boldsymbol{\gamma} + \alpha\mathbf{x}_* + \sigma\boldsymbol{\epsilon}_*\,. \tag{2.4}$$

Finally, we assume that the elements of disturbance vectors $\boldsymbol{\epsilon}_c$ and $\boldsymbol{\epsilon}_*$ are independently and identically distributed with mean zero and variance one.

For the following it is assumed that missingness of $\mathbf{x}_*$ depends only on the values of all the explanatory variables but is independent of the study variable $\mathbf{y}$. Using the missing data indicator matrix $\mathbf{R}$ (Rubin, 1976) with (i,j)th element $r_{ij} = 1$ if x_{ij} is observed and $r_{ij} = 0$ if x_{ij} is missing, in our notation $\mathbf{R}$ has the structure

$$\mathbf{R} = \begin{pmatrix} (\mathbf{1}\ldots\mathbf{1}) & \mathbf{1} \\ (\mathbf{1}\ldots\mathbf{1}) & \mathbf{0} \end{pmatrix}$$

corresponding to the dimensions of

$$\begin{pmatrix} \mathbf{Z}_c & \mathbf{x}_c \\ \mathbf{Z}_* & \mathbf{x}_* \end{pmatrix}.$$

Then the assumption on the missing mechanism results in

$$f(\mathbf{y}|\mathbf{R},\mathbf{X}) = \frac{f(\mathbf{y},\mathbf{R}|\mathbf{X})}{f(\mathbf{R}|\mathbf{X})} = f(\mathbf{y}|\mathbf{X}) \tag{2.5}$$

as $f(\mathbf{R}|\mathbf{y},\mathbf{X}) = f(\mathbf{R}|\mathbf{X})$, i.e., regression of $\mathbf{y}$ on $\mathbf{X}$ is independent of $\mathbf{R}$.

It may be noticed, that if (2.5) is not valid, i.e. missingness may also depend on $\mathbf{y}$, then we get

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

In this case estimation procedures would depend on the missing data process.

As $\mathbf{x}_*$ is not available, application of least squares to the entire model specified by (2.3) and (2.4) provides although best linear unbiased estimators of regression coefficients but lacks any practical utility. The simplest solution in such circumstances is to ignore (2.4) and to apply least squares to (2.3). This gives the following estimator of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{CC} = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y}_c. \tag{2.6}$$

This estimator $\mathbf{b}_c$ (the complete case estimator, CC) fails to utilize the information contained in m_* observations on the study variable and $(K-1)$ explanatory variables of the model. This kind of complete discard is obviously not always a satisfactory proposition and may often have misleading implications.

An alternative solution is to employ some imputation method so that missing values of the last explanatory variable can be replaced. There are several ways to do this; see, e.g., Rao and Toutenburg (1995, Chap. 8). Among them, an interesting procedure known as first order regression method (FOR) is to run an auxiliary regression of the variable in $\mathbf{x}_c$ on the remaining $(K-1)$ variables in $\mathbf{Z}_c$ and to use the estimated equation for finding the predicted values of missing observations, viz.,

$$\mathbf{x}_R = \mathbf{Z}_*(\mathbf{Z}_c'\mathbf{Z}_c)^{-1}\mathbf{Z}_c'\mathbf{x}_c\ . \tag{2.7}$$

Replacing $\mathbf{x}_*$ in (2.4) by $\mathbf{x}_R$ and then applying least squares to the thus obtained repaired model for estimating $\boldsymbol{\beta}$, we get the following estimator (FOR)

$$\hat{\boldsymbol{\beta}}_{FOR} = (\mathbf{X}_c'\mathbf{X}_c + \mathbf{X}_R'\mathbf{X}_R)^{-1}(\mathbf{X}_c'\mathbf{y}_c + \mathbf{X}_R'\mathbf{y}_*)\ , \tag{2.8}$$

where $\mathbf{X}_R$ is the same as $\mathbf{X}_*$ except that the last column vector $\mathbf{x}_*$ is replaced by $\mathbf{x}_R$.

In order to make full utilization of available information, we may include the study variable also as an explanatory variable while running the auxiliary regression of $\mathbf{x}_c$ on $\mathbf{Z}_c$ so that the imputed values for the elements of $\mathbf{x}_*$ are given by

$$\begin{aligned}
\hat{\mathbf{x}}_* &= [\mathbf{Z}_*\ ,\mathbf{y}_*]\begin{pmatrix} \mathbf{Z}_c'\mathbf{Z}_c & \mathbf{Z}_c'\mathbf{y}_c \\ \mathbf{y}_c'\mathbf{Z}_c & \mathbf{y}_c'\mathbf{y}_c \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{Z}_c'\mathbf{x}_c \\ \mathbf{y}_c'\mathbf{x}_c \end{pmatrix} \\
&= [\mathbf{Z}_*\ ,\mathbf{y}_*]\begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}' & c \end{pmatrix}\begin{pmatrix} \mathbf{Z}_c'\mathbf{x}_c \\ \mathbf{y}_c'\mathbf{x}_c \end{pmatrix} \\
&= \mathbf{Z}_*(\mathbf{Z}_c'\mathbf{Z}_c)^{-1}\mathbf{Z}_c'\mathbf{x}_c + \frac{\mathbf{x}_c'\mathbf{M}\mathbf{y}_c}{\mathbf{y}_c'\mathbf{M}\mathbf{y}_c}\left(\mathbf{y}_* - \mathbf{Z}_*(\mathbf{Z}_c'\mathbf{Z}_c)^{-1}\mathbf{Z}_c'\mathbf{y}_c\right) \\
&= \mathbf{x}_R + \frac{\mathbf{x}_c'\mathbf{M}\mathbf{y}_c}{\mathbf{y}_c'\mathbf{M}\mathbf{y}_c}\left(\mathbf{y}_* - \mathbf{Z}_*(\mathbf{Z}_c'\mathbf{Z}_c)^{-1}\mathbf{Z}_c'\mathbf{y}_c\right)\ ,
\end{aligned} \tag{2.9}$$

where

$$\begin{aligned}
\mathbf{M} &= \mathbf{I} - \mathbf{Z}_c(\mathbf{Z}_c'\mathbf{Z}_c)^{-1}\mathbf{Z}_c'\ , \\
\mathbf{A} &= (\mathbf{Z}_c'\mathbf{Z}_c)^{-1} + \frac{1}{\mathbf{y}_c'\mathbf{M}\mathbf{y}_c}(\mathbf{Z}_c'\mathbf{Z}_c)^{-1}\mathbf{Z}_c'\mathbf{y}_c\mathbf{y}_c'\mathbf{Z}_c(\mathbf{Z}_c'\mathbf{Z}_c)^{-1}\ , \\
\mathbf{b} &= -\frac{1}{\mathbf{y}_c'\mathbf{M}\mathbf{y}_c}(\mathbf{Z}_c'\mathbf{Z}_c)^{-1}\mathbf{Z}_c'\mathbf{y}_c\ , \\
c &= \frac{1}{\mathbf{y}_c'\mathbf{M}\mathbf{y}_c}\ .
\end{aligned}$$

Substituting $\hat{\mathbf{x}}_*$ for $\mathbf{x}_*$ in (2.4) and then applying least squares to the resulting repaired model, we obtain the following estimator of $\boldsymbol{\beta}$ (modified first order regression estimator, MFOR):

$$\hat{\boldsymbol{\beta}}_{MFOR} = (\mathbf{X}_c'\mathbf{X}_c + \hat{\mathbf{X}}_*'\hat{\mathbf{X}}_*)^{-1}(\mathbf{X}_c'\mathbf{y}_c + \hat{\mathbf{X}}_*'\mathbf{y}_*)\ , \tag{2.10}$$

where $\hat{\mathbf{X}}_*$ is same as $\mathbf{X}_*$ except with $\mathbf{x}_*$ in $\mathbf{X}_*$ being replaced by $\hat{\mathbf{x}}_*$.

It may be noticed that non-stochastic quantities are used to replace the missing values in the traditional first order regression method. In the proposed procedure involving a modification of the first order regression method, we substitute stochastic quantities for missing values. Thus $\mathbf{x}_R$ is a fixed vector while $\hat{\mathbf{x}}_*$ is a random vector.

It is common practice in regression models with missing values to give the completely observed sample matrix a different weight than the sub-matrix containing missing or imputed values (cp. e.g. Rao and Toutenburg, 1995, for the derivation of the weighted mixed regression estimator). As a generalization of the MFOR estimator $\hat{\boldsymbol{\beta}}$ from (2.10) we introduce a weight w, $0 \leq w \leq 1$ and incorporate it in $\hat{\boldsymbol{\beta}}$ in the following manner

$$\hat{\boldsymbol{\beta}}_{wMFOR} = (\mathbf{X}_c'\mathbf{X}_c + w^2\hat{\mathbf{X}}_*'\hat{\mathbf{X}}_*)^{-1}(\mathbf{X}_c'\mathbf{y}_c + w^2\hat{\mathbf{X}}_*'\mathbf{y}_*) \tag{2.11}$$

This estimator is called the weighted MFOR estimator (wMFOR).

The weight w used for the incomplete cases accounts for the increased residual variance for the cases with missing covariate values, yielding weighted least squares estimators. For the setup used in the simulation study, i.e. $\mathbf{X} = (X_1, X_2)$ with missing values in X_2 only, the weight would be

$$w = \sigma_{yy\cdot 12}/\sigma_{yy\cdot 1} = 1 - \rho_{2y\cdot 1}\,,$$

where $\sigma_{yy\cdot 12}$ is the residual variance of y given X_1 and X_2, $\sigma_{yy\cdot 1}$ is the residual variance of y given X_1 only, and $\rho_{2y\cdot 1}$ is the partial correlation coefficient of X_2 and y given X_1 (see Little, 1992).

3 Efficiency Properties and MDE-II Superiority

It is easy to see that the estimator $\mathbf{b}_c$ based on complete observations alone is unbiased for $\boldsymbol{\beta}$ with variance covariance matrix as

$$\begin{aligned} V(\mathbf{b}_c) &= E(\mathbf{b}_c - \boldsymbol{\beta})(\mathbf{b}_c - \boldsymbol{\beta})' \\ &= \sigma^2(\mathbf{X}_c'\mathbf{X}_c)^{-1}\,. \end{aligned} \tag{3.1}$$

Next, we observe that the estimator $\mathbf{b}_R$ is biased with bias vector and mean squared error matrix as

$$\begin{aligned} B(\mathbf{b}_R) &= E(\mathbf{b}_R - \boldsymbol{\beta}) \\ &= \alpha(\mathbf{X}_c'\mathbf{X}_c + \mathbf{X}_R'\mathbf{X}_R)^{-1}\mathbf{X}_R'\boldsymbol{\theta}\,, \end{aligned} \tag{3.2}$$

$$\begin{aligned} M(\mathbf{b}_R) &= E(\mathbf{b}_R - \boldsymbol{\beta})(\mathbf{b}_R - \boldsymbol{\beta})' \\ &= \sigma^2(\mathbf{X}_c'\mathbf{X}_c + \mathbf{X}_R'\mathbf{X}_R)^{-1} + \alpha^2(\mathbf{X}_c'\mathbf{X}_c + \mathbf{X}_R'\mathbf{X}_R)^{-1} \\ &\quad \times \mathbf{X}_R'\boldsymbol{\theta}\boldsymbol{\theta}'\mathbf{X}_R(\mathbf{X}_c'\mathbf{X}_c + \mathbf{X}_R'\mathbf{X}_R)^{-1}\,, \end{aligned} \tag{3.3}$$

where

$$\boldsymbol{\theta} = (\mathbf{x}_* - \mathbf{x}_R) = \frac{1}{\alpha}(\mathbf{X}_* - \mathbf{X}_R)\boldsymbol{\beta}\,. \tag{3.4}$$

Toutenburg et al. (1995) have analysed the efficiency properties of $\mathbf{b}_c$ and $\mathbf{b}_R$ in detail and have deduced conditions under which $\mathbf{b}_R$ is superior to $\mathbf{b}_c$ with respect to different weak and strong mean dispersion error criteria.

Deriving the exact distributional properties of the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}(w)$ arising from our proposed procedure, it can be easily visualised, will be a fairly intricate exercise and may not lead to any meaningful and clear conclusion regarding the efficiency properties of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}(w)$. Let us therefore consider its properties using a simulation study.

The mean dispersion error of an estimator $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is defined as

$$\begin{aligned} \mathrm{M}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= \mathrm{E}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})' \\ &= \mathrm{V}(\tilde{\boldsymbol{\beta}}) + \mathrm{Bias}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta})\,\mathrm{Bias}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta})'\,. \end{aligned} \tag{3.5}$$

Comparing two estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ with respect to the MDE-matrix criterion means that $\hat{\boldsymbol{\beta}}_2$ is superior to $\hat{\boldsymbol{\beta}}_1$ if

$$\mathrm{M}(\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}) - \mathrm{M}(\hat{\boldsymbol{\beta}}_2, \boldsymbol{\beta}) = \Delta(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) \geq 0\,.$$

We will use the weakened MDE-II superiority criterion for investigating the efficiency properties of the alternative estimators. Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ two competing estimators. Then $\hat{\boldsymbol{\beta}}_2$ is said to be MDE-II better than $\hat{\boldsymbol{\beta}}_2$ if

$$\mathrm{tr}(\Delta(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)) \geq 0\,, \tag{3.6}$$

or, equivalent if

$$\frac{\mathrm{tr}(\mathrm{M}(\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}))}{\mathrm{tr}(\mathrm{M}(\hat{\boldsymbol{\beta}}_2, \boldsymbol{\beta}))} \geq 1 \tag{3.7}$$

4 Model Specification and some Simulation Results

To investigate the properties of the considered estimators in case of small sample sizes ($n = 30$), a simulation study was conducted. We considered data $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2)$, where $\mathbf{x}_1$ was always observed and $\mathbf{x}_2$ contained some missing values. The structure of the data is as described in (2.3) and (2.4), with $\mathbf{Z} = \mathbf{x}_1$. Varying degrees of multicollinearity are governed by $\rho = \mathrm{corr}(x_1, x_2)$. The simulation study investigates the considered estimators for different data sets $(\mathbf{y}, \mathbf{X})$.

The algorithm for the creation of data $(\mathbf{y}, \mathbf{X})$ was as follows.

(1) The independent variables $(\mathbf{x}_1, \mathbf{x}_2)$ were generated as i.i.d. multivariate normal with mean $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} = \sigma_X^2 \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$ for a given σ_X^2.

(2) The regression coefficients $\boldsymbol{\beta}$ were set to $\boldsymbol{\beta} = (1, 1, 1)'$, yielding response values $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with a random error $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$. Using this setup, the overall model fit (measured by R^2) is driven by the ratio $\sigma_\epsilon^2 / \sigma_X^2$, where smaller values signalize a better fit.

(3) Having created data as described in (1) and (2), missing values in $\mathbf{x}_2$ were generated randomly for a specified percentage of missing values.

Using the data set generated by the above algorithm, the complete-case estimator $\hat{\boldsymbol{\beta}}_{\mathrm{CC}}$, the FOR estimator $\hat{\boldsymbol{\beta}}_{\mathrm{FOR}}$, the MFOR estimator $\hat{\boldsymbol{\beta}}_{\mathrm{MFOR}}$ and a weighted version of the MFOR estimator $\hat{\boldsymbol{\beta}}_{\mathrm{wMFOR}}$ were evaluated.

By repeating step (2) of the algorithm $R = 200$ times, the empirical variances of the considered estimators were computed, using (component wise, $j = 1, 2$)

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_j) = \frac{1}{R-1} \sum_{i=1}^{R} (\hat{\boldsymbol{\beta}}_{ij} - \bar{\hat{\boldsymbol{\beta}}}_j)^2\,.$$

The bias of the considered estimators was estimated by

$$\widehat{\mathrm{Bias}}(\hat{\boldsymbol{\beta}}_j) = \frac{1}{R} \sum_{i=1}^{R} (\hat{\boldsymbol{\beta}}_{ij} - \boldsymbol{\beta}_j)\,,$$

yielding $\widehat{\mathrm{MDE}}(\hat{\boldsymbol{\beta}}_j) = \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_j) + \widehat{\mathrm{Bias}}(\hat{\boldsymbol{\beta}}_j)^2$. In order to compare the efficiency of the estimators in relation to the complete case method, the ratio of the MDE-II risks, i.e. the ratio of the traces of the MDE matrices of the CC estimator and the respective estimator is used:

$$\mathrm{eff} = \frac{\mathrm{tr}(\mathrm{MDE(CC)})}{\mathrm{tr}(\mathrm{MDE(Estimator)})}.$$

A value greater than one indicates superiority to CC.

The algorithm was repeated, creating different $\mathbf{X}$ matrices in step (1). The so obtained results were averaged, which is taking the expectation over $\mathbf{X}$. Figure 4.1 shows the ratios for the considered estimators for varying degrees of multicollinearity, indexed by ρ. Figure 4.2 shows the corresponding bias terms, and figure 4.3 shows the corresponding variance components.

First note that the FOR estimator is unbiased (for random $\mathbf{X}$), whereas it is generally biased assuming a fixed $\mathbf{X}$. Second, for the investigated setup with missing values in only one variable, the FOR estimator coincides with the CC estimator for this component.

The gain in efficiency is determined both by the ratio $\sigma_\epsilon^2/\sigma_X^2$, i.e., the overall model fit, and the amount of missing values. Generally speaking, the better the model fit achieved by the CC estimator, the smaller the possible gain by using a filled-up model. For increasing amount of missing data the gain in efficiency increases. Both conclusions seem natural, since having a nearly perfect fit with CC, there is not much room for further improvement, and having nearly no missing data, the estimators tend to be identical.

More interesting is the effect of multicollinearity on the behaviour of the considered estimators. If x_1 and x_2 are not correlated, i.e. if $\rho = 0$, then the bias for $\hat{\beta}_1$ (the covariate with no missing values) is zero for all considered estimation methods, as $\hat{\beta}_1$ and $\hat{\beta}_2$ are estimated independently (for the normal model). Generally, the bias of the MFOR (weighted and unweighted) estimator increases with increasing ρ (see Figure 4.2). Weighted MFOR has a smaller bias than unweighted MFOR – the bias is reduced by introducing a weight.

The efficiency of the MFOR procedures compared to the CC method decreases for increasing $\rho = \text{corr}(x_1, x_2)$ (see Figure 4.1). Only for values of $\rho > 0.8$, i.e. if multicollinearity is severe, the wMFOR procedure is less efficient than the FOR procedure. In these situations, however the model itself is questionable, as the variances (see Figure 4.3) increase exponentially.

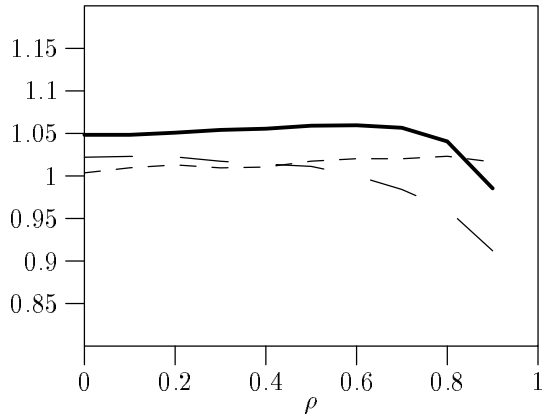

Figure 4.1: Ratios of tr(MDE) for different degrees of multicollinearity (indexed by ρ). MDE(CC)/ MDE(FOR) dashed line (short), MDE(CC)/ MDE(MFOR) dashed line (long), MDE(CC)/ MDE(wMFOR) thick line. ($\sigma_X^2 = \sigma_\epsilon^2 = 1.0$, 10% cases with missing values).

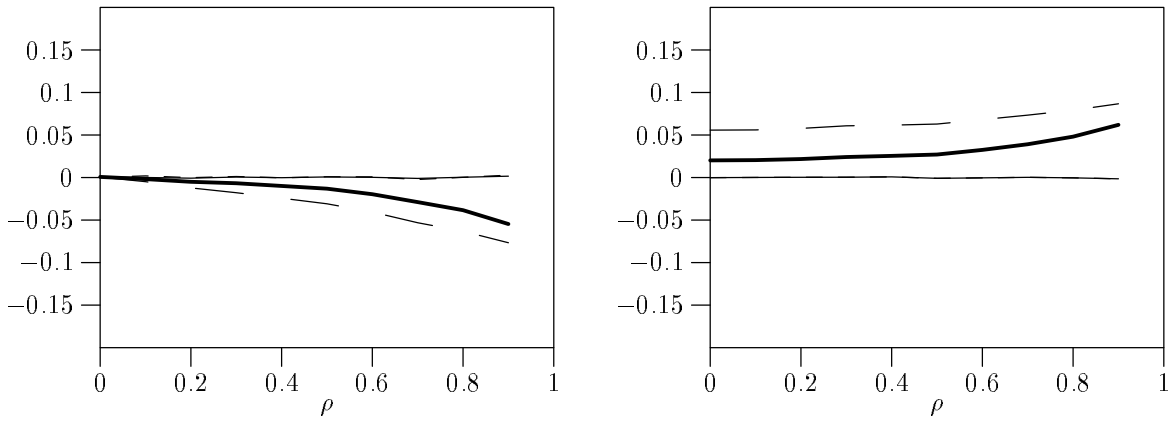

Figure 4.2: Bias terms for different degrees of multicollinearity (indexed by ρ), $\hat{\boldsymbol{\beta}}_1$ left panel, and $\hat{\boldsymbol{\beta}}_2$ right panel. Bias(CC) solid line, Bias(FOR) dashed line (short), Bias(MFOR) dashed line (long), Bias(wMFOR) thick line. ($\sigma_X^2 = \sigma_\epsilon^2 = 1.0$, 10% cases with missing values). Note: Bias(CC) and Bias(FOR) coincide for β_2, as x_2 is the only variable affected by missing values.

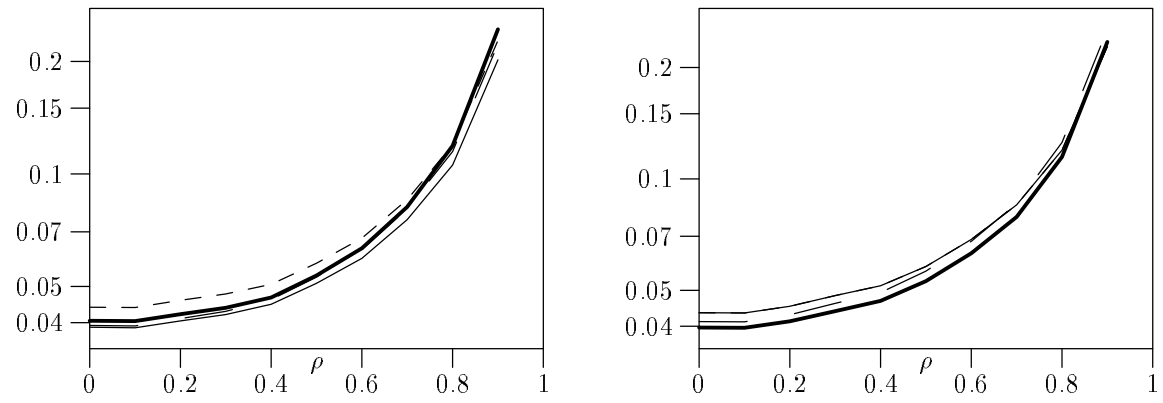

Figure 4.3: Variance terms for different degrees of multicollinearity (indexed by ρ, using a log scale), $\hat{\boldsymbol{\beta}}_1$ left panel, and $\hat{\boldsymbol{\beta}}_2$ right panel. Var(CC) solid line, Var(FOR) dashed line (short), Var(MFOR) dashed line (long), Var(wMFOR) thick line. ($\sigma_X^2 = \sigma_\epsilon^2 = 1.0$, 10% cases with missing values).

References

Afifi, A. A. and Elashoff, R. M. (1966). Missing observations in multivariate statistics: I: Review of the literature, *Journal of the American Statistical Association* **61**: 595–604.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer, *Journal of the Royal Statistical Society, Series B* **22**: 302–307.

Dagenais, M. G. (1973). The use of incomplete observations in multiple regression analysis: A generalized least squares approach, *Journal of Econometrics* **1**: 317–328.

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

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*, Wiley, New York.

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.

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.

Toutenburg, H., Srivastava, V. K. and Fieger, A. (1996). Estimation of parameters in multiple regression with missing X-observations using first order regression procedure, *SFB386 – Discussion Paper 38*, Ludwig-Maximilians-Universität München.