Toutenburg, Srivastava, Fieger:

# Estimation of Parameters in Multiple Regression With Missing X-Observations using Modified First Order Regression Procedure

Sonderforschungsbereich 386, Paper 38 (1996)

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

# Estimation of Parameters in Multiple Regression With Missing $X$-Observations using Modified First Order Regression Procedure

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

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

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

November 4, 1996

**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. Asymptotic properties of the estimators for the regression coefficients arising from the proposed modification are derived when either both the number of complete observations and the number of missing values grow large or only the number of complete observations grows large and the number of missing observations stays fixed. Using these results, the proposed procedure is 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. It is suggested that an elaborate simulation experiment will be helpful to evaluate the gain in efficiency especially in case of discrete regressor variables and to examine some other interesting issues like the impact of varying degree of multicollinearity in explanatory variables. Applications to some concrete data sets may also shed some light on these aspects. Some work on these lines is in progress and will be reported in a future article to follow.

## 1 Introduction

It is not uncommon in many applications of 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 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 MSE-superiority investigations for imputation methods. When these imputed values are nonstochastic, the application of 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 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 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 explanatory variables. This leads to complete data set and now the regression analysis is performed. As all the observations on 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. 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 nonstochastic. Examining the impact of such imputed values on the estimation of regression coefficients is the objective of present investigation.

The plan of this article is as follows. In Section 2, we present the model specification and describe three estimation procedures for the regression coefficients. One is the procedure that discards incomplete portion of data while the remaining two employ imputed values obtained from first order regressions. Out of these two, one uses nonstochastic values for imputation while the other uses stochastic values. Asymptotic properties of these three procedures are discussed in Section 3.

Finally, some summarizing remarks are placed in Section 4. Lastly, the proofs of Theorems are outlined in Appendix.

# 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:

$$\begin{aligned} \mathbf{y}_c &= \mathbf{X}_c\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}_c\,, & (2.1)\\ \mathbf{y}_* &= \mathbf{X}_*\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}_*\,, & (2.2)\end{aligned}$$

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 $\sigma$ 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

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

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 $\alpha$ is the last element.

Thus we can write the model as follows:

$$\begin{aligned} \mathbf{y}_c &= \mathbf{Z}_c\gamma + \alpha\mathbf{x}_c + \sigma\boldsymbol{\epsilon}_c\,, & (2.3)\\ \mathbf{y}_* &= \mathbf{Z}_*\gamma + \alpha\mathbf{x}_* + \sigma\boldsymbol{\epsilon}_*\,. & (2.4)\end{aligned}$$

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

Note: 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 circumstance is to ignore (2.4) and to apply least squares to (2.3). This gives the following estimator of $\boldsymbol{\beta}$:

$$\mathbf{b}_c = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y}_c. \tag{2.6}$$

This estimator $\mathbf{b}_c$ 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 it; see, e.g., Rao and Toutenburg (1995, Chap. 8). Among them, an interesting procedure known as first order regression method 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

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

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

If we include the study variable also as an explanatory variable while running the auxiliary regression of $\mathbf{x}_c$ on $\mathbf{Z}_c$, 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 $\boldsymbol{\beta}$:

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

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

Remark: The estimator $\hat{\boldsymbol{\beta}}$ (2.10) is of OLSE-type ignoring the fact that the disturbances in the filled-up model are

$$
\begin{pmatrix} \mathbf{y}_c \\ \mathbf{y}_* \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_c & \mathbf{x}_c \\ \mathbf{Z}_* & \mathbf{x}_* \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\epsilon} \\ \mathbf{v} \end{pmatrix}
$$

with $\mathbf{v} \sim (\mathbf{0}, \boldsymbol{\Sigma})$ and $\mathrm{E}(\boldsymbol{\epsilon}_c\mathbf{v}') \neq \mathbf{0}$. Using the approximations of the appendix it is easy to prove that

$$
\mathrm{E} \begin{pmatrix} \boldsymbol{\epsilon} \\ \mathbf{v} \end{pmatrix} (\boldsymbol{\epsilon}' \; \mathbf{v}') = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}
$$

with order $O(n^{-3/2})$, so that the use of OLSE instead of GLSE is justified.

It may be noticed that nonstochastic 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.

# 3 Efficiency Properties

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 squard error matrix as

$$
B(\mathbf{b}_R) = E(\mathbf{b}_R - \boldsymbol{\beta})
$$

$$\begin{aligned}
&= \alpha(\mathbf{X}_c'\mathbf{X}_c + \mathbf{X}_R'\mathbf{X}_R)^{-1}\mathbf{X}_R'\boldsymbol{\theta}\,, && (3.2)\\
M(\mathbf{b}_R) &= E(\mathbf{b}_R - \boldsymbol{\beta})(\mathbf{b}_R - \boldsymbol{\beta})' \\
&= \sigma^2(\mathbf{X}_c'\mathbf{X}_c + \mathbf{X}_R'\mathbf{X}_R)^{-1} + \alpha^2(\mathbf{X}_c'\mathbf{X}_c + \mathbf{X}_R'\mathbf{X}_R)^{-1} \\
&\quad \cdot \mathbf{X}_R'\boldsymbol{\theta}\boldsymbol{\theta}'\mathbf{X}_R(\mathbf{X}_c'\mathbf{X}_c + \mathbf{X}_R'\mathbf{X}_R)^{-1}\,, && (3.3)
\end{aligned}$$

where

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

Toutenburg, Heumann, Fieger and Park (1995) have analyzed 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 squared error criteria.

Deriving the exact distributional properties of the estimator $\hat{\boldsymbol{\beta}}$ arising from our proposed procedure, it can be easily visualized, will be a fairly intricate excercise and may not lead to any meaningful and clear conclusion regarding the efficiency properties of $\hat{\boldsymbol{\beta}}$. Let us therefore consider its asymptotic properties.

If we employ large sample asymptotic theory, we need to assume either $m_* \to \infty$ or $m_c \to \infty$ or both $m_* \to \infty$ and $m_c \to \infty$. The first assumption is obviously not an interesting proposition. It specifies a situation where the number of complete observations is few and the number of missing observations grows large. Just the reverse is true in case of second assumption, that is, the number of missing observations is small while the number of complete observations grows large. In other words, the model has a tendency in which missing observations are given lesser importance as the number of complete observations increases. Such a situation may arise in many practical applications. The third assumption is equally interesting and may be a tenable proposition in some applications because in this case missing observations as well as complete observations both grow large possibly with varying speed. We shall therefore restrict our attention to second and third specifications of asymptotic theory.

### 3.1 Properties When Both $m_c$ And $m_*$ Grows Large

For analyzing the properties of estimators under the specification that both $m_c$ and $m_*$ are large, we introduce a quantity $m$ defined as

$$m = \begin{cases} m_c & \text{if } m_c \le m_* \\ m_* & \text{if } m_c > m_* \end{cases} \tag{3.5}$$

so that $m$ tending to infinity is equivalent to both $m_c$ and $m_*$ tending to infinity.

Next, we assume the asymptotic cooperativeness of explanatory variables, i.e., both $(\frac{1}{m_c}\mathbf{X}_c'\mathbf{X}_c)$ and $(\frac{1}{m_*}\mathbf{X}_*'\mathbf{X}_*)$ tend to finite nonsingular matrices as $m$ tends to infinity.

Under the above specification, it follows from (3.2) that the estimator $\mathbf{b}_R$ is biased and in fact inconsistent. Its mean square matrix to order $O(m^{-1})$ has the same expression as (3.3).

For the estimator $\hat{\boldsymbol{\beta}}$, we have the following results which are derived in Appendix.

**Theorem 1:** *The bias vector of $\hat{\boldsymbol{\beta}}$ up to order $O(m^{-\frac{1}{2}})$ is null. Its mean squared error matrix up to order $O(m^{-1})$ is given by*

$$M(\hat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{\Omega} \mathbf{X}_c' \mathbf{X}_c \boldsymbol{\Omega} \,, \tag{3.6}$$

*where $\boldsymbol{\Omega} = (\mathbf{X}_c'\mathbf{X}_c + \mathbf{X}_*'\mathbf{X}_* + \frac{\sigma^2 m_*}{\alpha^2}\mathbf{e}\mathbf{e}')^{-1}$ with $\mathbf{e}$ as $K \times 1$ vector having first $(K-1)$ elements zero and last element one.*

Now comparing $\mathbf{b}_c$, $\mathbf{b}_R$ and $\hat{\boldsymbol{\beta}}$, we observe that $\mathbf{b}_c$ is unbiased and consistent for $\boldsymbol{\beta}$ but it ignores the available set of $m_*$ observations. The estimator $\mathbf{b}_R$ is neither unbiased nor consistent although it utilizes the additional $m_*$ observations on $(K-1)$ explanatory variables only. The estimator $\hat{\boldsymbol{\beta}}$ not only utilizes the additional $m_*$ observations on $(K-1)$ explanatory variables but also uses the $m_*$ observations on study variable. This improves the performance of estimators. Thus our proposed estimator is consistent but biased.

Next, let us compare the estimators with respect to the criterion of mean squared error matrix to order $O(m^{-1})$. Restricting attention to consistent estimators $\mathbf{b}_c$ and $\hat{\boldsymbol{\beta}}$, we observe from (3.1) and (3.6) that

$$V(\mathbf{b}_c) - M(\hat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{\Omega} \left[ \boldsymbol{\Omega}^{-1} (\mathbf{X}_c'\mathbf{X}_c)^{-1} \boldsymbol{\Omega}^{-1} - \mathbf{X}_c'\mathbf{X}_c \right] \boldsymbol{\Omega} \,. \tag{3.7}$$

Substituting the expression for $\boldsymbol{\Omega}^{-1}$ inside the square brackets, it is seen that the expression for matrix difference is positive definite. This implies that the biased estimator $\hat{\boldsymbol{\beta}}$ is more efficient than the unbiased estimator $\mathbf{b}_c$.

So far as the comparison of $\mathbf{b}_R$ with $\mathbf{b}_c$ and $\hat{\boldsymbol{\beta}}$ is concerned, we feel that it is rather inappropriate and improper to compare the asymptotic approximations for bias vectors and mean squared error matrices of an inconsistent estimator with a consistent estimator. We have therefore not done it. At this point, it may be mentioned that the comparative study of $\mathbf{b}_c$ and $\mathbf{b}_R$ conducted by Toutenburg, Heumann, Fieger and Park (1995) remains meaningful and valid because they have employed exact, and not asymptotic, expressions.

## 3.2 Properties When $m_c$ Grows Large While $m_*$ Stays Fixed

In order to study the asymptotic properties under the present assumption, we assume that explanatory variables in (2.1) or (2.3) are asymptotically cooperative in the sense that $\mathbf{V}_c = (\frac{1}{m_c}\mathbf{X}_c'\mathbf{X}_c)^{-1}$ tends to a finite nonsingular matrix as $m_c$ tends to infinity.

Using (2.3), (2.4) and (2.9), it can be easily seen from (2.10) that

$$m_c^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sigma m_c^{-\frac{1}{2}} \mathbf{V}_c \mathbf{X}_c' \boldsymbol{\epsilon}_c + O_p(m_c^{-\frac{1}{2}}) \,, \tag{3.8}$$

whence it follows that the asymptotic distribution of $m_c^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is multivariate normal with mean vector $\mathbf{0}$ and variance covariance matrix $\sigma^2\mathbf{V}_c$ which is the same as the asymptotic distribution of $m_c^{-\frac{1}{2}}(\mathbf{b}_c - \boldsymbol{\beta})$ and $m_c^{\frac{1}{2}}(\mathbf{b}_R - \boldsymbol{\beta})$.

It is thus observed that all the three estimators, viz., $\mathbf{b}_c$, $\mathbf{b}_R$ and $\hat{\boldsymbol{\beta}}$ are asymptotically equivalent in the sense of possessing same asymptotic distribution. Let us therefore consider higher order approximations.

It is easy to see from (3.2) and (3.3) that the bias vector of $\mathbf{b}_R$ to order $O(m_c^{-1})$ is

$$B(\mathbf{b}_R) = \frac{\alpha}{m_c}\mathbf{V}_c\mathbf{X}_R'\boldsymbol{\theta}\,, \tag{3.9}$$

while the mean squared error matrix to order $O(m_c^{-2})$ is given by

$$\begin{aligned} M(\mathbf{b}_R) &= \frac{\sigma^2}{m_c}\mathbf{V}_c + \frac{1}{m_c^2}\mathbf{V}_c\left(\sigma^2\mathbf{X}_R'\mathbf{X}_R + \alpha^2\mathbf{X}_R'\boldsymbol{\theta}\boldsymbol{\theta}'\mathbf{X}_R\right)\mathbf{V}_c \\ &= \frac{\sigma^2}{m_c}\mathbf{V}_c + \frac{1}{m_c^2}\mathbf{V}_c\big[\sigma^2\mathbf{X}_R'\mathbf{X}_R \\ &\quad + \mathbf{X}_R'(\mathbf{X}_* - \mathbf{X}_R)\boldsymbol{\beta}\boldsymbol{\beta}'(\mathbf{X}_* - \mathbf{X}_R)'\mathbf{X}_R\big]\mathbf{V}_c \end{aligned} \tag{3.10}$$

Similar results for the estimator $\hat{\boldsymbol{\beta}}$ are derived in Appendix and presented below.

**Theorem 2:** *The estimator $\hat{\boldsymbol{\beta}}$ is unbiased up to order $O(m_c^{-1})$. An asymptotic approximation for its mean squared error matrix to order $O(m_c^{-2})$ is given by*

$$M(\hat{\boldsymbol{\beta}}) = \frac{\sigma^2}{m_c}\mathbf{V}_c + \frac{\sigma^2}{m_c^2}\mathbf{V}_c\left[\frac{2\sigma^2 m_*(\lambda - \sigma^2)}{\lambda\alpha}\mathbf{e}\mathbf{e}' + \mathbf{W} + \mathbf{W}'\right]\mathbf{V}_c\,, \tag{3.11}$$

*where*

$$\begin{aligned} \mathbf{W} &= \mathbf{X}_*'\mathbf{Z}_*\mathbf{V}_c\left(\frac{1}{m_c}\mathbf{Z}_c'\mathbf{X}_c\right) + \frac{\alpha(\lambda - \sigma^2)}{\lambda}\mathbf{X}_*'\boldsymbol{\theta}\mathbf{e}' \\ &= \mathbf{X}_*'\mathbf{X}_R + \frac{(\lambda - \sigma^2)}{\lambda}\mathbf{X}_*'(\mathbf{X}_* - \mathbf{X}_R)\boldsymbol{\beta}\mathbf{e}'\,, \end{aligned} \tag{3.12}$$

*with $\lambda = (\sigma^2 + \frac{\alpha^2}{m}\mathbf{x}_c'\mathbf{M}\mathbf{x}_c)$ and $\mathbf{e}$ as $K \times 1$ vector with first $(K-1)$ elements zero and last element one.*

Comparing the estimators $\mathbf{b}_c$, $\mathbf{b}_R$ and $\hat{\boldsymbol{\beta}}$ with respect to the criterion of bias to order $O(m_c^{-1})$, we observe that $\mathbf{b}_c$ and $\hat{\boldsymbol{\beta}}$ are unbiased while $\mathbf{b}_R$ is not unless $\mathbf{x}_*$ and $\mathbf{x}_R$ turn out to be numerically same.

Looking at the expressions (3.10) and (3.11), it is difficult to draw any clear inference from the comparison of $\mathbf{b}_c$, $\mathbf{b}_R$ and $\hat{\boldsymbol{\beta}}$ with respect to the criterion of mean squared error matrix to the order of our approximation.

# 4 Some Remarks

We have considered the problem of estimating the coefficients in a linear regression model with some missing observations on some of the explanatory variables. To keep the exposition simple, we have assumed that missing observations pertain to one explanatory variable only. Our investigations can, however, be easily extended to the case when there are more than one explanatory variable with possibly varying numbers of missing observations.

Recognizing that the application of least squares procedure to the entire model does not lead to feasible estimators for the regression coefficients, two estimation strategies are chiefly considered. The first strategy ignores the incomplete observations fully and involves application of least squares to complete data set only. This kind of complete discard of information contained in the incomplete data set may not always be a satisfactory proposition. The second strategy comprises substituting imputed values for the missing observations and then applying least squares to completed or repaired data set. For finding the imputed values, the method of first order regression is considered and two alternatives are presented. One provides nonstochastic values for imputation through the conventional method while the other yields stochastic values through a modification in the conventional method.

We have investigated the asymptotic properties of the three estimation procedures mentioned above. When both the number of complete observations and the number of incomplete observations are large, it is observed that the conventional first order regression method yields inconsistent and biased estimators of regression coefficients while the simple strategy of ignoring the incomplete data set completely produces consistent and unbiased estimators. On the other hand, the modified first order regression method gives consistent and biased estimators but it is asymptotically more efficient than the method which uses only complete data set. It is interesting to note that the proposed modification overcomes the problem of inconsistency of estimators arising from the conventional first order regression method.

When only the number of complete observations grows large while the number of incomplete observations stays fixed, our investigations have revealed that all the three estimation procedures are asymptotically equivalent in the sense that they share the same distributional properties asymptotically and thus do not permit us to prefer one over the other. We have therefore considered higher order asymptotic approximations.

These approximations do not provide us any clear inference regarding the superiority of one estimation procedure over the other. Perhaps an elaborate simulation experiment may be helpful to examine this aspect. It will also be interesting to investigate the impact of the presence of multicollinearity of varying degree on the lines of Hill and Ziemer (1983). Applications to some concrete data sets involving possibly some discrete variables may also shed some light on the efficiency properties. Some work in these directions is in progress and will be reported in future.

# A Appendix

In order to derive the results stated in Theorem 1, we first observe that

$$\begin{aligned}
&\left(\mathbf{y}_* - \mathbf{Z}_*(\mathbf{Z}_c'\mathbf{Z}_c)^{-1}\mathbf{Z}_c'\mathbf{y}_c\right) \\
&\quad = \alpha\left(\mathbf{x}_* - \mathbf{Z}_*(\mathbf{Z}_c'\mathbf{Z}_c)^{-1}\mathbf{Z}_c'\mathbf{x}_c\right) + \sigma\boldsymbol{\epsilon}_* - \sigma\mathbf{Z}_*(\mathbf{Z}_c'\mathbf{Z}_c)^{-1}\mathbf{Z}_c'\boldsymbol{\epsilon}_c \\
&\quad = \alpha(\mathbf{x}_* - \mathbf{x}_R) + \sigma\boldsymbol{\epsilon}_* - \sigma\mathbf{Z}_*(\mathbf{Z}_c'\mathbf{Z}_c)^{-1}\mathbf{Z}_c'\boldsymbol{\epsilon}_c\,. \qquad (A.1)
\end{aligned}$$

Notice that the last term on the right hand side is $O_p(m_c^{-\frac{1}{2}})$ while the other terms are of order $O_p(1)$.

Writing $v = (\frac{1}{m_c}\boldsymbol{\epsilon}_c'\boldsymbol{\epsilon}_c - 1)$ and noticing that $v$ has order $O_p(m_c^{-\frac{1}{2}})$, we can express

$$\begin{aligned}
\frac{\mathbf{x}_c'\mathbf{M}\mathbf{y}_c}{\mathbf{y}_c'\mathbf{M}\mathbf{y}_c} &= \frac{1}{\alpha} - \frac{\mathbf{y}_c'\mathbf{M}(\mathbf{y}_c - \alpha\mathbf{x}_c)}{\alpha\mathbf{y}_c'\mathbf{M}\mathbf{y}_c} \\
&= \frac{1}{\alpha} - \frac{\sigma\alpha\mathbf{x}_c'\mathbf{M}\boldsymbol{\epsilon}_c}{(\alpha^2\mathbf{x}_c'\mathbf{M}\mathbf{x}_c + \sigma^2 m_c) + (2\alpha\sigma\mathbf{x}_c'\mathbf{M}\boldsymbol{\epsilon}_c + \sigma^2 m_c v) - \sigma^2\boldsymbol{\epsilon}_c'\mathbf{Z}_c(\mathbf{Z}_c'\mathbf{Z}_c)^{-1}\mathbf{Z}_c'\boldsymbol{\epsilon}_c} \\
&= \frac{1}{\alpha} - \frac{\sigma\alpha}{m_c\lambda}\mathbf{x}_c'\mathbf{M}\mathbf{x}_c\left[1 + \frac{2\alpha\sigma\mathbf{x}_c'\mathbf{M}\boldsymbol{\epsilon}_c + \sigma^2 m_c v}{m_c\lambda} - \frac{\sigma^2\boldsymbol{\epsilon}_c'\mathbf{Z}_c(\mathbf{Z}_c'\mathbf{Z}_c)^{-1}\mathbf{Z}_c'\boldsymbol{\epsilon}_c}{m_c\lambda}\right]^{-1},
\end{aligned}$$

where $\lambda = (\sigma^2 + \frac{\alpha^2}{m_c}\mathbf{x}_c'\mathbf{M}\mathbf{x}_c)$.

Expanding the expression inside the square brackets, we find

$$\frac{\mathbf{x}_c'\mathbf{M}\mathbf{y}_c}{\mathbf{y}_c'\mathbf{M}\mathbf{y}_c} = \frac{1}{\alpha} - \frac{\sigma\alpha}{m_c\lambda}\mathbf{x}_c'\mathbf{M}\boldsymbol{\epsilon}_c + O_p(m_c^{-1})\,. \tag{A.2}$$

Using (A.1) and (A.2) in (2.9), we get

$$\begin{aligned}
\hat{\mathbf{x}}_* &= (\mathbf{x}_* + \frac{\sigma}{\alpha}\boldsymbol{\epsilon}_*) \\
&\quad - \frac{\sigma}{m_c}\left[\frac{1}{\alpha}\mathbf{Z}_*\mathbf{V}_c\mathbf{Z}_c'\boldsymbol{\epsilon}_c + \frac{\alpha^2}{\lambda}\mathbf{x}_c'\mathbf{M}\boldsymbol{\epsilon}_c(\boldsymbol{\theta} + \frac{\sigma}{\alpha}\boldsymbol{\epsilon}_*)\right] + O_p(m_c^{-1})
\end{aligned}$$

from which we can express

$$\hat{\mathbf{X}}_* = (\mathbf{X}_* + \frac{\sigma}{\alpha}\mathbf{U}_0) - \frac{\sigma}{m_c}\mathbf{U} + O_p(m_c^{-1})\,, \tag{A.3}$$

where

$$\begin{aligned}
\mathbf{U}_0 &= [\mathbf{0}\ \ \boldsymbol{\epsilon}_*] \\
\mathbf{U} &= \left[\mathbf{0}\quad \frac{1}{\alpha}\mathbf{Z}_*\mathbf{V}_c\mathbf{Z}_c'\boldsymbol{\epsilon}_c + \frac{\alpha^2}{\lambda}\mathbf{x}_c'\mathbf{M}\boldsymbol{\epsilon}_c(\boldsymbol{\theta} + \frac{\sigma}{\alpha}\boldsymbol{\epsilon}_*)\right],
\end{aligned}$$

$\mathbf{0}$ denoting a null matrix of order $m_* \times (K-1)$.

It may be recalled that both $m_c$ and $m_*$ are large. It is equivalent to saying that $m$ is large where $m$ denotes the minimum of $m_c$ and $m_*$, i.e.,

$$m = \begin{cases} m_c & \text{if } m_c < m_* \\ m_* & \text{if } m_c > m_* \end{cases}$$

Now from (A.3), it is easy to see that

$$(\mathbf{X}_c'\mathbf{X}_c + \hat{\mathbf{X}}_*'\hat{\mathbf{X}}_*)^{-1} = \boldsymbol{\Omega} + O_p(m^{-\frac{3}{2}})\,,$$

so that we can express

$$\begin{aligned}
(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= (\mathbf{X}_c'\mathbf{X}_c + \hat{\mathbf{X}}_*'\hat{\mathbf{X}}_*)^{-1}\left[\sigma(\mathbf{X}_c'\boldsymbol{\epsilon}_c + \hat{\mathbf{X}}_*'\boldsymbol{\epsilon}_*) + \hat{\mathbf{X}}_*'(\mathbf{X}_* - \hat{\mathbf{X}}_*)\boldsymbol{\beta}\right] \\
&= \sigma\boldsymbol{\Omega}\mathbf{X}_c'\boldsymbol{\epsilon}_c + O_p(m^{-1})\,.
\end{aligned}$$

Thus we have bias vector up to order $O(m^{-\frac{1}{2}})$ as null and mean squared error matrix up to order $O(m^{-1})$ as $\sigma^2 \boldsymbol{\Omega} \mathbf{X}_c' \mathbf{X}_c \boldsymbol{\Omega}$.

This establishes the results of Theorem 1.

Using (A.3) in (2.10), we have

$$\begin{aligned}
(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= (\mathbf{X}_c'\mathbf{X}_c + \hat{\mathbf{X}}_*'\hat{\mathbf{X}}_*)^{-1}(\mathbf{X}_c'\mathbf{y}_c + \hat{\mathbf{X}}_*'\mathbf{y}_*) - \boldsymbol{\beta} \\
&= (\mathbf{X}_c'\mathbf{X}_c + \hat{\mathbf{X}}_*'\hat{\mathbf{X}}_*)^{-1}\left[\sigma(\mathbf{X}_c'\boldsymbol{\epsilon}_c + \hat{\mathbf{X}}_*'\boldsymbol{\epsilon}_*) + \hat{\mathbf{X}}_*'(\mathbf{X}_* - \hat{\mathbf{X}}_*)\boldsymbol{\beta}\right] \\
&= \boldsymbol{\eta}_{-\frac{1}{2}} + \boldsymbol{\eta}_{-\frac{3}{2}} + O_p(m_c^{-2}), \qquad \text{(A.4)}
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\eta}_{-\frac{1}{2}} &= \frac{\sigma}{m_c}\mathbf{V}_c\mathbf{X}_c'\boldsymbol{\epsilon}_c \\
\boldsymbol{\eta}_{-\frac{3}{2}} &= \frac{\sigma}{m_c^2}\mathbf{V}_c(\mathbf{X}_*' + \frac{\sigma}{\alpha}\mathbf{U}_0')\left[\mathbf{Z}_*\mathbf{V}_c\mathbf{Z}_c'\boldsymbol{\epsilon}_c + \frac{\alpha^3}{\lambda}\mathbf{x}_c'\mathbf{M}\boldsymbol{\epsilon}_c(\boldsymbol{\theta} + \frac{\sigma}{\alpha}\boldsymbol{\epsilon}_*)\right].
\end{aligned}$$

Here the suffixes of $\boldsymbol{\eta}$ indicate the order of magnitude in probability.

As $E(\boldsymbol{\eta}_{-\frac{1}{2}}) = \mathbf{0}$, the bias vector to order $O(m_c^{-1})$ is equal to a null vector. Further, the mean squared error matrix to order $O(m_c^{-2})$ is given by

$$M(\hat{\boldsymbol{\beta}}) = E(\boldsymbol{\eta}_{-\frac{1}{2}}\boldsymbol{\eta}_{-\frac{1}{2}}') + E(\boldsymbol{\eta}_{-\frac{3}{2}}\boldsymbol{\eta}_{-\frac{1}{2}}' + \boldsymbol{\eta}_{-\frac{1}{2}}\boldsymbol{\eta}_{-\frac{3}{2}}'). \qquad \text{(A.5)}$$

It is easy to see that

$$E(\boldsymbol{\eta}_{-\frac{1}{2}}\boldsymbol{\eta}_{-\frac{1}{2}}') = \frac{\sigma^2}{m_c}\mathbf{V}_c.$$

For the second term on the right hand side, we observe that

$$\begin{aligned}
&E\mathbf{V}_c\mathbf{X}_*'\left[\mathbf{Z}_*\mathbf{V}_c\mathbf{Z}_c'\boldsymbol{\epsilon}_c + \frac{\alpha^3}{\lambda}\mathbf{x}_c'\mathbf{M}\boldsymbol{\epsilon}_c(\boldsymbol{\theta} + \frac{\sigma}{\alpha}\boldsymbol{\epsilon}_*)\right]\boldsymbol{\epsilon}_c'\mathbf{X}_c\mathbf{V}_c \\
&= \mathbf{V}_c\mathbf{X}_*'\left[\mathbf{Z}_*\mathbf{V}_c\mathbf{Z}_c' + \frac{\alpha^3}{\lambda}\boldsymbol{\theta}\mathbf{x}_c'\mathbf{M}\right]\mathbf{X}_c\mathbf{V}_c, \\
&E\mathbf{V}_c\mathbf{U}_0'\left[\mathbf{Z}_*\mathbf{V}_c\mathbf{Z}_c'\boldsymbol{\epsilon}_c + \frac{\alpha^3}{\lambda}\mathbf{x}_c'\mathbf{M}\boldsymbol{\epsilon}_c(\boldsymbol{\theta} + \frac{\sigma}{\alpha}\boldsymbol{\epsilon}_*)\right]\boldsymbol{\epsilon}_c'\mathbf{X}_c\mathbf{V}_c \\
&= \frac{\alpha^2\sigma m_*}{\lambda}\mathbf{V}_c\mathbf{e}\mathbf{x}_c'\mathbf{M}\mathbf{X}_c\mathbf{V}_c.
\end{aligned}$$

Using these, we obtain $E(\boldsymbol{\eta}_{-\frac{3}{2}}\boldsymbol{\eta}_{-\frac{1}{2}}')$. The last term on the right hand side of (A.5) is its transpose. Substituting these and using

$$\begin{aligned}
\mathbf{x}_c'\mathbf{M}\mathbf{X}_c &= (\mathbf{x}_c'\mathbf{M}\mathbf{Z}_c \quad \mathbf{x}_c'\mathbf{M}\mathbf{x}_c) \\
&= \mathbf{x}_c'\mathbf{M}\mathbf{x}_c\mathbf{e}' \\
&= \frac{(\lambda - \sigma^2)m_c}{\alpha^2}\mathbf{e}'
\end{aligned}$$

we obtain the result (3.11) stated in Theorem 2.

## 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, *Journal of Econometrics* **1**: 317–328.

Hill, R. C. and Ziemer, R. F. (1983). Missing regressor values under conditions of multicollinearity, *Communications in Statistics, Part A–Theory and Methods* **12**: 2557–2573.

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.