Toutenburg, Srivastava:

Estimation of Linear Regression Models with Missingness of Observations on Both the Explanatory and Study Variables-Part I: Theoretical Results

Sonderforschungsbereich 386, Paper 184 (2000)

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

Projektpartner

Estimation of Linear Regression Models with Missingness of Observations on Both the Explanatory and Study Variables–Part I: Theoretical Results

H. Toutenburg
Institut für Statistik
Universität München
80799 München, Germany

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

February 16, 2000

Abstract

This paper discusses the estimation of coefficients in a linear regression model when there are some missing observations on an explanatory variable and the study variable individually as well as simultaneously. The first order regression method of imputation is followed and the least squares procedure is applied. Efficiency properties of estimators are then investigated employing the large sample asymptotic theory.

1 Introduction

Practitioners routinely face the problem of missingness of some observations due to a mariyad of factors on which little control can be exercised. This unavoidable feature of the data set prohibits us from applying the standard statistical procedures fro drawing inferences. There are two popular alternatives to circumvent this problem. One is the amputation strategy which discards the incomplete observations and utilizes only the complete observations for the statistical analysis. Other is the imputation strategy which follows a procedure to find imputed values for missing observations and thus repairs the data so that it looks like a complete data set and permits the application of standard statistical procedures. Both strategies have their own limitations and qualifications.

In the context of the estimation of parameters in linear regression models, considerable attention has been devoted to analyze the comparative performance of amputation and imputation strategies when missingness of observations pertains to either the study variable or some explanatory variables; see, for example, Little (1992), Little and Rubin (1987) and Rao and Toutenburg (1995) for an interesting account. Realistic situations may often necessitate us to assume that there are some cases in which values of some explanatory variables as well as the

study variable are missing simultaneously. Such a framework is considered in this paper and the estimation of regression coefficients in the model is discussed.

The plan of presentation is as follows. In section 2, we describe a linear regression model in which missingness of observations relate to the study variable and only one explanatory variable. The entire set of observations is divided into four parts. The first part consists of complete observations only. Observations on the last explanatory variable are assumed to be missing in the second part while values of the study variable are assumed to be missing in the third part. In the fourth part, observations on both the study variable and the last explanatory variable are missing simultaneously. Under such a framework, first the regression coefficients are estimated by the least squares procedure employing the complete observations in the first part of the data set. A simple imputation procedure is then followed to find the imputed values for the missing observations. These imputed values are substituted for repairing the data set and the least squares procedure is used for estimating the regression coefficients from repaired data. The thus obtained estimators are presented along with those which utilize barely the complete observations. In section 3, we analyze the efficiency properties of these estimators. As general conclusions related to superiority of one estimator over the other are hard to draw, two particular cases of the model are considered. Finally, some concluding remarks are offered in Section 4.

2 Model Specification And The Estimators

Consider a linear regression model with some missing observations. For the sake of clarity in exposition, let us assume that there is barely one explanatory variable on which some observations are not available. Further, it is assumed that some observations on the study variable are also missing.

Corresponding to n_1 complete observations, we have the following regression relationship:

$$y_1 = X_1\beta + \alpha x_1 + \epsilon_1 \tag{2.1}$$

where y_1 is an $n_1 \times 1$ vector of n_1 observations on the study variable, X_1 is a $n_1 \times K$ full column rank matrix of n_1 observations on K explanatory variables, β is a $K \times 1$ vector of regression coefficients, x_1 is a $n_1 \times 1$ vector of n_1 observations on the last explanatory variable, α is the scalar coefficient associated with it and ϵ_1 is an $n_1 \times 1$ vector of disturbances.

Next, suppose that we have a set of n_2 observations such that observations on the last explanatory variable are missing. Thus we can write

$$y_2 = X_2\beta + \alpha x_2^* + \epsilon_2 \tag{2.2}$$

where y_2 is a $n_2 \times 1$ vector of n_2 observations on the study variable, X_2 is a $n_2 \times K$ matrix of n_2 observations on the K explanatory variables, x_2^* denotes the vector of n_2 missing values of the last explanatory variable and ϵ_2 is the vector of disturbances.

Similarly, there are n_3 observations in which values of the study variable are missing:

$$y_3^* = X_3\beta + \alpha x_3 + \epsilon_3 \tag{2.3}$$

where y_3^* denotes the vector of n_3 missing values of the study variable, the $n_3 \times K$matrix X_3 and the $n_3 \times 1$ vector x_3 contain observations on the explanatory variables and ϵ_3 is a $n_3 \times 1$ vector of disturbances.

Finally, the last part of the data set consists of n_4 observations only on the K explanatory variables so that

$$y_4^* = X_4\beta + \alpha x_4^* + \epsilon_4 \tag{2.4}$$

where y_4^* and x_4^* denote the vectors of missing values of the study variable and the last explanatory variable respectively, X_4 is the $n_4 \times K$ matrix of n_4 available observations on the K explanatory variables and ϵ_4 is a $n_4 \times 1$ vector of disturbances.

We thus have an incomplete data set consisting of $(n_1+n_2+n_3+n_4)$ observations for the estimation of $(K+1)$ regression coefficients.

It is assumed that the elements of $\epsilon_1, \epsilon_2, \epsilon_3$ and ϵ_4 are independently and identically distributed with mean 0 and variance σ^2.

If we delete the incomplete part of data set and use only n_1 complete observations, the least squares estimators of α and β are given by

$$\tilde{\alpha} = \frac{x_1' M y_1}{x_1' M x_1} \tag{2.5}$$

$$\tilde{\beta} = (X_1'X_1)^{-1}X_1'(y_1 - \hat{\alpha}x_1) \tag{2.6}$$

where

$$M = I_{n_1} - X_1(X_1'X_1)^{-1}X_1'\,. \tag{2.7}$$

In order to make full utilization of available observations, we need to find imputed values for missing observations. For this purpose, let us consider the first order regression method of imputation for the missing values of the last explanatory variable. This method consists of running the regresssion of the last explanatory variable on the remaining K explanatory variables using only the n_1 complete observations and then utilizing the estimated relationship for finding the predicted values of the missing observations; see, e. g. , Afifi and Elashoff (1967), Dagenais (1973), Gourieroux and Monfort (1981) and Rao and Toutenburg (1995). This yields the following imputed values for x_2^* and x_4^*:

$$x_2^* = X_2(X_1'X_1)^{-1}X_1'x_1 \tag{2.8}$$

$$x_4^* = X_4(X_1'X_1)^{-1}X_1'x_1 \tag{2.9}$$

In the same spirit, if we run the regression of the study variable on the K explanatory variables utilizing the n_1 complete observations and employ the estimated relationship for finding the predicted values for the missing observations

on the study variable, we obtain the imputed values for y_3^* and y_4^* as follows

$$\hat{y}_3^* = X_3(X_1'X_1)^{-1}X_1'y_1 \tag{2.10}$$
$$\hat{y}_4^* = X_4(X_1'X_1)^{-1}X_1'y_1\,. \tag{2.11}$$

Now let us introduce the following notation

$$\begin{aligned}
\theta &= x_3 - X_3(X_1'X_1)^{-1}X_1'x_1 \\
S &= (X_1'X_1 + X_2'X_2 + X_3'X_3 + X_4'X_4)^{-1} \\
U &= X_1'x_1 + X_2'\hat{x}_2^* + X_3'x_3 + X_4'\hat{x}_4^* \\
&= S^{-1}(X_1'X_1)^{-1}X_1'x_1 + X_3'\theta \\
V &= X_1'y_1 + X_2'y_2 + X_3'\hat{y}_3^* + X_4'\hat{y}_4^* \\
&= S^{-1}(X_1'X_1)^{-1}X - 1'y_1 + X_2'[y_2 - X_2(X_1'X_1)^{-1}X_1'y_1] \\
u &= x_1'x_1 + \hat{x}_2^{*\prime}\hat{x}_2^* + x_3'x_3 + \hat{x}_4{}^{*\prime}\hat{x}_4^* \\
&= x_1'Mx_1 + x_1'X_1(X_1'X_1SX_1'X_1)^{-1}X_1'x_1 + 2x_3'\theta - \theta'\theta \\
v &= x_1'y_1 + \hat{x}_2^{*\prime}y_2 + x_3'\hat{y}_3^* + \hat{x}_4^{*\prime}\hat{y}_4^{*\prime} \\
&= x_1'Mx_1 + x_1'X_1(X_1'X_1)^{-1}(X_1'y_1 + X_2'y_2) \\
&\quad + [x_3'X_3 + x_1'X_1(X_1'X_1)^{-1}X_4'X_4](X_1'X_1)^{-1}X_1'y_1
\end{aligned}$$

If we substitute $\hat{x}_2^*$ in place of x_2^* in (2.2), $\hat{y}_3^*$ in place of y_3^* in (2.3), $\hat{x}_4^*$ and $\hat{y}_4^*$ in place of x_4^* and y_4^* respectively in (2.4) and then apply the least squares procedure to the thus obtained equations and (2.1) jointly, the estimators of α and β are to be found to be as follows:

$$\hat{\alpha} = \frac{v - U'SV}{u - U'SU} \tag{2.12}$$
$$= \frac{x_1'My_1 - \theta'X_3SX_2'[y_2 - X_2(X_1'X_1)^{-1}X_1'y_1]}{x_1'Mx_1 + \theta'(I - X_3SX_3')\theta}$$
$$\hat{\beta} = S(V - \hat{\alpha}U) \tag{2.13}$$
$$= (X_1'X_1)^{-1}X_1'y_1 + SX_2'[y_2 - X_2(X_1'X_1)^{-1}X_1'y_1] - [(X_1'X_1)^{-1}X_1'x_1 + SX_3'\theta]\,.$$

When the missingness of the observations on the study variable and the last explanatory variable occur simultaneously so that the model is specified by equations (2.1) and (2.4) only, it is interesting to observe from (2.5), (2.6), (2.12) and (2.13) that $\tilde{\alpha} = \hat{\alpha}$ and $\tilde{\beta} = \hat{\beta}$. This implies that the set of n_4 observations on the K explanatory variables play no role in the least squares estimation of regression coefficients. When the set of n_2 observations is also included so that the model is defined by the equations (2.1), (2.2) and (2.4), we find that $\tilde{\alpha}$ and $\hat{\alpha}$ continue to remain identical but $\tilde{\beta}$ and $\hat{\beta}$ become generally unequal. Now if the set of n_3 incomplete observations is further added so that the model consists of all the four equations (2.1), (2.2), (2.3) and (2.4), we observe that not only $\tilde{\beta}$ and $\hat{\beta}$ are unequal but $\tilde{\alpha}$ and $\hat{\alpha}$ also differ in general.

3 Efficiency Comparisons

Let us first compare the estimators $\tilde{\alpha}$ and $\hat{\alpha}$ of the coefficient α associated with the explanatory variable on which some observations are missing.

It is easy to see that $\tilde{\alpha}$ is an unbiased estimator of α while $\hat{\alpha}$ is not. The bias of $\hat{\alpha}$ is given by

$$\begin{aligned} B(\hat{\alpha}) &= E(\hat{\alpha} - \alpha) \qquad (3.1) \\ &= -\alpha \left[\frac{\theta'(I - X_3 S X_3')\theta + \theta' X_3 S X_2' \{x_2^* - X_2 (X_1' X_1)^{-1} X_1' x_1\}}{x_1' M x_1 + \theta'(I - X_3 S X_3')\theta} \right] . \end{aligned}$$

Further, the variances of $\tilde{\alpha}$ and $\hat{\alpha}$ are

$$\begin{aligned} V(\tilde{\alpha}) &= E(\tilde{\alpha} - \alpha)^2 \qquad (3.2) \\ &= \frac{\sigma^2}{x_1' M x_1} \\ V(\hat{\alpha}) &= E(\hat{\alpha} - E(\alpha))^2 \qquad (3.3) \\ &= \frac{\sigma^2 [x_1' M x_1 + \theta' X_3 S X_2' \{I + X_2 (X_1' X_1)^{-1} X_2'\} X_2 S X_3' \theta]}{[x_1' M x_1 + \theta'(I - X_3 S X_3')\theta]^2} . \end{aligned}$$

Using the result

$$\frac{1}{x_1' M x_1 + \theta'(I - X_3 S X_3')\theta} \leq \frac{1}{x_1' M x_1}$$

we observe that

$$\frac{V(\hat{\alpha})}{V(\tilde{\alpha})} \leq \frac{x_1' M x_1 + \theta' X_3 S X_2' [I + X_2 (X_1' X_1)^{-1} X_2'] X_2 S X_3' \theta}{x_1' M x_1 + \theta'(I - X_3 S X_3')\theta} . \qquad (3.4)$$

Thus the estimator $\hat{\alpha}$ has smaller variance in comparison to $\tilde{\alpha}$ so long as the quantity $\theta' A \theta$ is positive where

$$A = I - X_3 S X_3' - X_3 S X_2' X_2 S X_3' - X_3 S X_2' X_2 (X_1' X_1)^{-1} X_2' X_2 S X_3' . \qquad (3.5)$$

As the matrix A does not involve any unknown quantity, the positivity of the characteristic roots of A can be easily checked for any given data set in practice.

If we compare the mean squared error of $\hat{\alpha}$ with the variance of $\tilde{\alpha}$, it is hard to deduce any neat condition for the superiority of $\hat{\alpha}$ over $\tilde{\alpha}$ or vice-versa such that it can be verified in any given application.

Next, let us consider the estimators $\tilde{\beta}$ and $\hat{\beta}$ of β, the vector of regression coefficients associated with the explanatory variables on which no observation is missing.

It can be easily seen that the estimator $\tilde{\beta}$ is unbiased. However, $\hat{\beta}$ is generally biased with bias vector as follows:

$$\begin{aligned} B(\hat{\beta}) &= E(\hat{\beta} - \beta) \qquad (3.6) \\ &= \alpha [S X_2' \{x_2^* - X_2 (X_1' X_1)^{-1} X_1' x_1\} - a S X_3' \theta + (1 - a)(X_1' X_1)^{-1} X_1' x_1] \end{aligned}$$

where

$$a = \frac{x_1' M x_1 - \theta' X_3 S X_2' \{x_2^* - X_2 (X_1' X_1)^{-1} X_1' x_1\}}{x_1' M x_1 + \theta' (I - X_3 S X_3') \theta} \,. \tag{3.7}$$

Similarly, the expression for the variance covariance matrices are as follows:

$$V(\tilde{\beta}) = \sigma^2 \left[(X_1' X_1)^{-1} + \frac{1}{x_1' M x_1} (X_1' X_1)^{-1} X_1' x_1 x_1' X_1 (X_1' X_1)^{-1} \right] \tag{3.8}$$

$$\begin{aligned} V(\hat{\beta}) &= \sigma^2 \left[S X_2' X_2 S + (I - S X_2' X_2)(X_1' X_1)^{-1} (I - X_2' X_2 S) \right. \\ &\quad \left. + (\delta \phi' + \phi \delta') \right] + V(\hat{\alpha}) \delta \delta' \end{aligned} \tag{3.9}$$

where

$$\delta = (X_1' X_1)^{-1} X_1' x_1 + S X_3' \theta \tag{3.10}$$

$$\begin{aligned} \phi &= \frac{1}{x_1' M x_1 + \theta' (I - X_3 S X_3') \theta} [S + S X_2' X_2 - (X_1' X_1)^{-1}] \\ &\quad \cdot X_2' X_2 S X_3' \theta \,. \end{aligned} \tag{3.11}$$

It can be clearly appreciated from the above expression that no inference can be deduced regarding the superiority of one estimator over the other. Same is true when we compare the estimators with respect to the criterion of mean squared error matrix.

Let us now examine two particular cases of our model specification. Case I assumes that the third part of the data set is absent while Case II deletes the second part.

3.1 Case I

Suppose that the model is specified by (2.1), (2.2) and (2.4) only:

$$\begin{aligned} y_1 &= X_1 \beta + \alpha x_1 + \epsilon_1 \\ y_2 &= X_2 \beta + \alpha x_2^* + \epsilon_2 \\ y_4^* &= X_4 \beta + \alpha x_4^* + \epsilon_4 \,. \end{aligned} \tag{3.12}$$

As pointed out earlier, now the estimators $\tilde{\alpha}$ and $\hat{\alpha}$ are identically equal while $\tilde{\beta}$ and $\hat{\beta}$ are generally different.

If we write

$$S_I = (X_1' X_1 + X_2' X_2 + X_4' X_4)^{-1} \tag{3.13}$$

the expression for the bias vector and the variance covariance matrix of $\hat{\beta}$ can be easily recovered from (3.7) and (3.9). These are as follows:

$$B(\hat{\beta}_I) = \alpha S_I X_2' \{x_2^* - X_2 (X_1' X_1)^{-1} X_1' x_1\} \tag{3.14}$$

$$\begin{aligned} V(\hat{\beta}_I) &= \sigma^2 \left[S_I X_2' X_2 S_I + (I - S_I X_2' X_2)(X_1' X_1)^{-1} (I - X_2' X_2 S_I) \right. \\ &\quad \left. + \frac{1}{x_1' M x_1} (X_1' X_1)^{-1} x_1' x_1 X_1 (X_1' X_1)^{-1} \right] \,. \end{aligned} \tag{3.15}$$

Comparing (3.8) and (3.15), we find

$$V(\tilde{\beta}) - V(\hat{\beta}_I) = \sigma^2 Q \tag{3.16}$$

where

$$Q = S_I X_2' X_2 [(X_1' X_1)^{-1} - S_I] + [(X_1' X_1)^{-1} - S_I] X_2' X_2 S_I \,. \tag{3.17}$$

As the matrix $[(X_1' X_1)^{-1} - S_I]$ is positive definite, we observe that Q is also so. This implies that $\hat{\beta}_I$ is superior to $\tilde{\beta}$ with respect to the criterion of variance covariance matrix.

Next, let us compare the estimators with respect to the criterion of mean squared error matrix.

From (3.14) and (3.15), the difference between the mean squared error matrix of the estimator $\hat{\beta}_I$ and the variance covariance matrix of the estimator $\tilde{\beta}$ can be written as

$$\begin{aligned} \Delta_I &= V(\hat{\beta}_I) + [B(\hat{\beta}_I)][B(\hat{\beta}_I)]' - V(\tilde{\beta}) \\ &= -\sigma^2 Q + \alpha^2 S_I X_2' [x_2^* - X_2 (X_1' X_1)^{-1} X_1' x_1] \\ &\quad \cdot [x_2^* - x_1' X_1 (X_1' X_1)^{-1} X_2'] X_2 S_I \,. \end{aligned} \tag{3.18}$$

Now, using Rao and Toutenburg (1995, Theorem A. 59, p. 304), we find that Δ_I cannot be a nonnegative definite matrix except in the trivial situation $p = 1$. In other words, the estimator $\tilde{\beta}$ cannot be superior to $\hat{\beta}_I$ with respect to the mean squared error matrix criterion with an exception to a trivial case.

If we look at the matrix $(-\Delta_I)$, it follows from Rao and Toutenburg(1995, Theorem A. 57, p.303) that a necessary and sufficient condition for a variance covariance matrix of the unbiased estimator $\tilde{\beta}$ to exceed the mean squared error matrix of the biased estimator $\hat{\beta}_I$ by a nonnegative definite matrix is

$$[x_2^* - x_1' X_1 (X_1' X_1)^{-1} X_2'] X_2 S_I Q^{-1} S_I X_2' [x_2^* - X_2 (X_1' X_1)^{-1} X_1' x_1] \quad \leq \quad \left(\frac{\sigma}{\alpha}\right)^2 . \tag{3.19}$$

Thus the estimator $\hat{\beta}_I$ is superior to $\tilde{\beta}$ when condition (3.19) is satisfied.

3.2 Case II

Let us be given the following model:

$$\begin{aligned} y_1 &= X_1 \beta + \alpha x_1 + \epsilon_1 \\ y_3^* &= X_3 \beta + \alpha x_3 + \epsilon_3 \\ y_4^* &= X_4 \beta + \alpha x_4^* + \epsilon_4 \,. \end{aligned} \tag{3.20}$$

In this case, the estimators $\hat{\alpha}$ and $\hat{\beta}$ reduce to the following:

$$\hat{\alpha}_{II} = \frac{x_1' M y_1}{x_1' M x_1 + \theta' (I - X_3 S_{II} X_3') \theta} \tag{3.21}$$

$$\hat{\beta}_{II} = (X_1' X_1)^{-1} X_1' y_1 - \hat{\alpha}_{II} \delta_{II} \tag{3.22}$$

where

$$S_{II} = (X_1'X_1 + X_3'X_3 + X_4'X_4)^{-1} \tag{3.23}$$

$$\delta_{II} = (X_1'X_1)^{-1}X_1'x_1 + S_{II}X_3'\theta \tag{3.24}$$

Comparing (3.21) with (2.5), it is obvious that $\tilde{\alpha}$ and $\hat{\alpha}_{II}$ are generally different. In fact, the magnitude of $\hat{\alpha}_{II}$ is smaller than that of
$tilde\alpha$. Similarly, if we compare (3.22) with (2.6), the estimators $\tilde{\beta}$ and $\hat{\beta}_{II}$ are seen to be generally different.

From (3.1) we observe that

$$B(\hat{\alpha}_{II}) = -\frac{\alpha\theta'(I - X_3S_{II}X_3')\theta}{x_1'Mx_1 + \theta'(I - X_3S_{II}X_3')\theta} \tag{3.25}$$

so that the bias of $\hat{\alpha}_{II}$ has a sign opposite to that of α. Further, the magnitude of bias is always smaller than the absolute value of α.

Similarly, from (3.3), we have

$$V(\hat{\alpha}_{II}) = \frac{\sigma^2x_1'Mx_1}{[x_1'Mx_1 + \theta'(I - X_3S_{II}X_3')\theta]^2}\,. \tag{3.26}$$

Comparing with (3.2), we find that $\hat{\alpha}_{II}$ has invariabily smaller variance than $\tilde{\alpha}$.

Further, it is found that the mean squared error of $\hat{\alpha}_{II}$ is less than the variance of $\tilde{\alpha}$ provided that

$$\left(\frac{\sigma}{\alpha}\right)^2\left[\frac{1}{x_1'Mx_1} + \frac{2}{\theta'(I - X_3S_{II}X_3')\theta}\right] > 1 \tag{3.27}$$

which is indeed necessary and sufficient condition for the superiority of $\hat{\alpha}_{II}$ over $\tilde{\alpha}$ according to the mean squared error criterion.

Just the reverse is true, i. e. , $\tilde{\alpha}$ is superior to $\hat{\alpha}_{II}$ when the inequality (3.27) holds with an opposite sign.

Similarly, from (3.6) and (3.9), the bias vector and the variance covariance matrix are

$$B(\hat{\beta}_{II}) = \alpha\left[(X_1'X_1)^{-1}X_1'x_1 - \frac{x_1'Mx_1}{x_1'Mx_1 + \theta'[I - X_3S_{II}X_3']\theta}\right] \tag{3.28}$$

$$V(\hat{\beta}_{II}) = \sigma^2(X_1'X_1)^{-1} + V(\hat{\alpha}_{II})\delta_{II}\delta_{II}'\,. \tag{3.29}$$

Comparing (3.29) with (3.8), we find that

$$V(\tilde{\beta}) - V(\hat{\beta}_{II}) = \frac{\sigma^2w}{x_1'Mx_1}\left[w\delta_{II}\delta_{II}' - \frac{1}{w}(X_1'X_1)^{-1}X_1'x_1x_1'X_1(X - 1'X_1)^{-1}\right] = \tag{3.30}$$

where

$$w = \frac{x_1'Mx_1}{x_1'Mx_1 + \theta'(I - X_3S_{II}X_3')\theta}\,. \tag{3.31}$$

The expression on the right hand side of (3.30) is obviously a semi-definite matrix but no comment can be made regarding its positiveness or negativeness.

In a similar way, if we consider the difference between the mean squared error matrix of $\hat{\beta}_{II}$ and the variance covariance matrix of $\tilde{\beta}$, we get

$$\begin{aligned}
\Delta_{II} &= V(\hat{\beta}_{II}) + [B(\hat{\beta}_{II})][B(\hat{\beta}_{II})]' - V(\tilde{\beta}) \qquad (3.32)\\
&= (1-w)^2\left[\alpha^2 - \frac{\sigma^2}{x_1'Mx_1}\left(\frac{1+w}{1-w}\right)\right](X_1'X_1)^{-1}X_1'x_1x_1'X_1(X_1'X_1)^{-1}\\
&\quad + w^2\left[\alpha^2 + \frac{\sigma^2}{x_1'Mx_1}\right]S_{II}X_3'\theta\theta'X_3S_{II}\\
&\quad - w(1-w)\left[\alpha^2 - \frac{\sigma^2}{x_1'Mx_1}\left(\frac{w}{1-w}\right)\right]\left[(X_1'X_1)^{-1}X_1'x_1\theta'X_3S_{II} + \right.\\
&\quad \left. S_{II}X_3'\theta x_1'X_1(X_1'X_1)^{-1}\right]
\end{aligned}$$

which is clearly a matrix of rank 1.

It is difficult to determine whether Δ_{II} is positive semi-definite or not.

4 Some Remarks

We have considered a linear regression model under a general framework for missingness of some observations. The entire set of observations consists of four parts. The first part has complete observations on all variables in the model while the remaining three parts refer to incomplete observations. Out of these three parts, observations on an explanatory variable are missing in the second part, while observations on the study variable are misisng in the third part. In the fourth part, observations on the study variable as well as the explanatory variable are missing simultaneously.

Two strategies for the estimation of regression coefficients have been considered. The first strategy consists of amputing the incomplete observations and applying least squares procedure using the first part of data. The second strategy follows a simple imputation procedure in which separate regressions of the study variable and the explanatory variable (on which some observations are missing) on the remaining explanatory variables (on which no observation is missing) are run employing the first part of data set end the estimated regression equations are used to find predicted values for the imputation of missing observations on the study and explanatory variables. After substituting these imputed values for the missing observations, the least squares procedure is applied.

It is seen that amputation strategy provides unbiased estimators of the regression coefficients while the imputation strategy gives generally biased estimators. Comparing the variance covariance matrices and mean squared error matrices of the estimators arising from the two strategies, no clear conclusion is found regarding superiority of one strategy over the other. However, some interesting observations are made in particular cases.

If the missigness of observations relate to both the study variable and explanatory variable simultaneously but not individually, i. e. , the data set consists

of first and fourth parts only, the amputation and imputation strategies yield identical estimators, and thus imputation is not worthwhile.

When we add the second part of data in which only the values of the explanatory variable are missing, the amputation and imputation strategies continue to provide identical estimator for the coefficient associated with the explanatory variable. However, they give generally different estimators for the coefficients associated with the remaining explanatory variables on which no observation is missing in the data set. The estimators arising from the amputation strategy are unbiased but less efficient, with respect to the criterion of variance covariance matrix, than the estimators stemming from the imputation strategy which are generally biased. If we take the criterion as mean squared error matrix, no uniform superiority of any strategy over the other is observed.

Instead of the second part when the third part of data containing missing values on the study variable is added, the scenario changes completely and the estimators based on amputation and imputation strategies are different whether we consider the estimator of the coefficient of the explanatory variable with some missing values or the coefficient of the remaining explanatory variables without any missing value. As mentioned earlier, the estimators found from the strategy of amputation are always unbiased but the estimators obtained from the strategy of imputation are generally not so. It is observed that the bias of the estimator of the coefficient of the explanatory with some missing values has a sign opposite to that of coefficient and the bias is always smaller in magnitude of coefficient itself. Further, if we compare the variances, it is interesting to note that the imputation strategy is rated uniformly superior to the amputation strategy. Such a result does not remain true when the estimators are compared with respect to the criterion of mean squared error. So far as the estimators of the coefficients associated with the explanatory variables with no missing values are concerned, no definite comment can be made regarding the bias vector, the variance covariance matrix or the mean squared error matrix.

References

Afifi, A. A. and Elashoff, R. M. (1967). Missing observations in multivariate statistics: Part II: Point estimation in simple linear regression, *Journal of the American Statistical Association* **62**: 10–29.

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

Gourieroux, G. and Monfort, A. (1981). On the problem of missing data in linear models, *Review of Economic Studies* **48**: 579–586.

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.