On the first order regression procedure of estimation for incomplete regression models

V. K. Srivastava¹, H. Toutenburg²

¹ Department of Statistics, University of Lucknow, Lucknow 226007, India
² Institute of Statistics, University of Munich, Germany (e-mail: toutenb@stat.uni-muenchen.de)

Received: July 30, 2002; revised: April 15, 2003

Abstract

This article discusses some properties of the first order regression method for imputation of missing values on an explanatory variable in linear regression model and presents an estimation strategy based on hypothesis testing.

Keywords: first order regression, missing values, linear regression model, imputation, estimation strategy, hypothesis testing

1 Introduction

When some observations on some of the explanatory variables in a linear regression model are missing, there are several imputation procedures to obtain their substitutes; see, e.g., Little and Rubin (2002) and Rao and Toutenburg (1995) for an interesting account. Among them, a popular procedure is the method of first order regression. It essentially amounts to running an auxiliary regression of each explanatory variable (on which the observations are missing) on the remaining explanatory variables (on which no observation is missing). The estimated equation is then used to find the predicted values for the missing

¹ This work was carried out before Professor V.K. Srivastava passed away in 2001.
² The author is grateful to the referees for their illuminating comments on an earlier draft of this paper.
observations on that explanatory variable. These predicted values are then used to complete the data set which, in turn, is employed for the estimation of regression coefficients by the method of least squares. Properties of the resulting estimators have been systematically analyzed by Toutenburg, Heumann, Fieger and Park (1995). This note mentions some additional points and presents two preliminary test estimators.

2 The First Order Regression Method

Consider the following linear regression model with some missing observations:

\[ \begin{align*} y_c &= X_c \beta + \alpha x_c + \sigma \epsilon_c \\ y* &= X* \beta + \alpha x_{mis} + \sigma \epsilon* \end{align*} \]

(2.1) \hspace{1cm} (2.2)

where \( y_c \) and \( y* \) are the column vectors of \( m_c \) and \( m* \) observations respectively on the study variable, \( X_c \) and \( X* \) are matrices of order \( m_c \times K \) respectively of the observations on \( K \) explanatory variables, \( x_c \) is a column vector of \( m_c \) observations on an additional explanatory variable while \( x_{mis} \) denotes the column vector of \( m* \) missing observations, \( \epsilon_c \) and \( \epsilon* \) are column vectors of disturbances assumed to be independently and identically distributed with zero mean and unit variance and \( \sigma \) is an unknown positive scalar.

It is assumed that the matrix \( X_c \) has full column rank while this may not be necessarily the case with \( X* \).

For the estimation of regression coefficients, if we amputate the incomplete part of data and accordingly apply the least squares procedure to (2.1), the estimators \( \hat{\alpha}_c \) and \( \hat{\beta}_c \) of \( \alpha \) and \( \beta \) respectively are given by the solution of the following equations:

\[ \begin{align*} X_c'X_c\hat{\beta}_c + \hat{\alpha}_cX_c'x_c &= X_c'y_c \\ x_c'X_c\hat{\beta}_c + \hat{\alpha}_cX_c'x_c &= x_c'y_c. \end{align*} \]

(2.3) \hspace{1cm} (2.4)

Premultiplying (2.3) by \( x_c'Mx_c \) and then subtracting from (2.4), we get

\[ \hat{\alpha}_c = \frac{x_c'My_c}{x_c'Mx_c} \]

(2.5)

where \( M = I - X_c(X_c'X_c)^{-1}X_c \).

Substituting it in (2.3), we find

\[ \hat{\beta}_c = (X_c'X_c)^{-1}X_c'y_c - \frac{x_c'My_c}{x_c'Mx_c}(X_c'X_c)^{-1}X_c'x_c. \]

(2.6)

If we do not delete the incomplete part of data, the first order regression method may be employed for finding the imputed values of the missing observations. This procedure consists of running the regression of \( x_c \) on \( X_c \) and...