Srivastava, Toutenburg:

On the First Order Regression Procedure of Estimation for Incomplete Regression Models

Sonderforschungsbereich 386, Paper 175 (1999)

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

Projektpartner

On the First Order Regression Procedure of Estimation for Incomplete Regression Models

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

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

November 11, 1999

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.

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 (1987) 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 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:

$$y_c = X_c\beta + \alpha x_c + \sigma\epsilon_c \tag{2.1}$$

$$y_* = X_*\beta + \alpha x_{mis} + \sigma\epsilon_* \tag{2.2}$$

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, ϵ_c and ϵ_* are column vectors of disturbances assumed to be independently and identically distributed with zero mean and unit variance and σ 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 α and βrespectively are given by the solution of following equations:

$$\begin{aligned} X_c'X_c\hat{\beta}_c + \hat{\alpha}_c X_c'x_c &= X_c'y_c \qquad (2.3) \\ x_c'X_c\hat{\beta}_c + \hat{\alpha}_c x_c'x_c &= x_c'y_c\,. \qquad (2.4) \end{aligned}$$

Premultiplying (2.3) by $x_c'X_c(X_c'X_c)^{-1}$ and then subtracting from (2.4), we get

$$\hat{\alpha}_c = \frac{x_c'My_c}{x_c'Mx_c} \qquad (2.5)$$

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

Substituting it in (2.3), we find

$$\hat{\beta}_c = (X_c'X_c)^{-1}X_c'y_c - \frac{x_c'My_c}{x_c'Mx_c}(X_c'X_c)^{-1}X_c'x_c\,. \qquad (2.6)$$

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 employing the estimated relationship to find the predicted values corresponding to the rows of X_*. This yields the following vector of imputed values:

$$\hat{x}_{mis} = X_*(X_c'X_c)^{-1}X_c'x_c\,. \qquad (2.7)$$

Using it to replace x_{mis} in (2.2) and then applying the least squares procedure to the thus repaired model, we obtain the following equations specifying the estimators $\hat{\alpha}$ and $\hat{\beta}$:

$$\begin{aligned} (X_c'X_c + X_*'X_*)\hat{\beta} + \hat{\alpha}(X_c'x_c + X_*'\hat{x}_{mis}) &= X_c'y_c + X_*'y_* \\ (x_c'X_c + \hat{x}_{mis}'X_*)\hat{\beta} + \hat{\alpha}(x_c'x_c + \hat{x}_{mis}'\hat{x}_{mis}) &= x_c'y_c + \hat{x}_{mis}'y_*\,. \end{aligned}$$

Substituting the expression (2.7), we obtain

$$\begin{aligned} (X_c'X_c + X_*'X_*')[\hat{\beta} + \hat{\alpha}(X_c'X_c)^{-1}x_c] &= X_c'y_c + X_*'y_* \\ x_c'X_c(X_c'X_c)^{-1}(X_c'X_c + X_*'X_*)[\hat{\beta} + \hat{\alpha}(X_c'X_c)^{-1}X_c'x_c] + \hat{\alpha}x_c'Mx_c & \\ = x_c'y_c + x_c'X_c(X_c'X_c)^{-1}X_*'y_*\,. & \end{aligned}$$

Premultiplying the first equation by $x_c'X_c(X_c'X_c)^{-1}$ and then substracting from the second equation, we get

$$\hat{\alpha} = \frac{x_c'My_c}{x_c'Mx_c} \qquad (2.8)$$

whence it follows that

$$\hat{\beta} = (X_c'X_c + X_*'X_*)^{-1}(X_c'y_c + X_*'y_*) - \frac{x_c'My_c}{x_c'Mx_c}(X_c'X_c)^{-1}X_c'x_c\,. \tag{2.9}$$

It is interesting to observe from (2.5) and (2.8) that the estimator of α, the regression coefficient associated with the explanatory variable on which some observations are missing, remains same whether we amputate the incomplete part of data or impute the missing observations by the first order regression method. This point has been noted by Afifi and Elashoff (1967) in the context of bivariate regression models. In fact, the equality of estimators can be easily seen when there are two or more explanatory variables on which some observations are missing.

If we write

$$\tilde{\beta}_c = (X_c'X_c)^{-1}X_c'y_c \tag{2.10}$$

$$\tilde{\beta} = (X_c'X_c + X_*'X_*)^{-1}(X_c'y_c + X_*'y_*) \tag{2.11}$$

then these are the estimators of β when $\alpha = 0$, i. e. , the last explanatory variable is dropped from the model. When it is retained, a sort of correction is to be applied to these estimators and this correction is same for both the estimators; see (2.6) and (2.9).

These observations suggest a practical rule as follows. First, test the hypothesis $H : \alpha = 0$ against the alternative hypothesis $A : \alpha \neq 0$ using the statistic

$$t = \frac{(m_c - K - 1)^{\frac{1}{2}}x_c'My_c}{[(y_c'My_c)(x_c'Mx_c) - (x_c'My_c)^2]^{\frac{1}{2}}} \tag{2.12}$$

which follows a t-distribution with $(m_c - K - 1)$ degrees of freedom under H.

If the hypothesis $H : \alpha = 0$ is retained, we may estimate β by either $\tilde{\beta}_c$ or $\tilde{\beta}$. A choice between $\tilde{\beta}_c$ and $\tilde{\beta}$ can be exercised on the basis of their performance properties which are easy to analyze following Toutenburg et al. (1995).

On the other hand, if the hypothesis $H : \alpha = 0$ is rejected, we may choose $\hat{\beta}_c$ or $\hat{\beta}$ for the estimation of β. Again, the properties of $\hat{\beta}_c$ and $\hat{\beta}$ can be easily studied on the lines of Toutenburg et al. (1995).

From the above proposition, we can present the following preliminary test estimators of β:

$$b_c = \begin{cases} \tilde{\beta}_c & \text{if } |t| < c \\ \hat{\beta}_c & \text{if } |t| \geq c \end{cases} \tag{2.13}$$

$$b = \begin{cases} \tilde{\beta} & \text{if } |t| < c \\ \hat{\beta} & \text{if } |t| \geq c \end{cases} \tag{2.14}$$

where c denotes the tabulated values obtained from the t-distribution for a two-sided test with a preassigned level of significance.

It may be observed that b_c and b are preliminary test estimators arising from amputation and imputation strategies respectively. Further, their performance properties can be studied on the lines of Giles and Srivastava (1993); see also Judge and Bock (1978).

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.

Giles, D. E. A. and Srivastava, V. K. (1993). The exact distribution of a least squares regression coefficient estimator after a preliminary t-test, *Statistics and Probability Letters* **16**: 59–64.

Judge, G. G. and Bock, M. E. (1978). *The Statistical Implications of Pre-test and Stein-Rule Estimators in Econometrics*, North Holland, Amsterdam.

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.

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.