Economic and Social Review Vol. 9 No. 4

Substituting Means for Missing Observations in Regression

D. CONNIFFE
An Foras Talúntais

I INTRODUCTION

Econometric textbooks usually discuss the procedures to be adopted when missing values occur in data intended for regression analysis. There is a large statistical literature on this topic. A variety of methods, differing both in their computational complexity and in the assumptions required, have been suggested, compared or simulated by, among others, Buck (1960), Afifi and Elashoff (1966), Haitovsky (1968), and Hocking and Smith (1968). More recently, there have been the papers of Dagenais (1973), and Beale and Little (1975). Yet textbooks sometimes give a not dishonourable mention to the simplest *ad-hoc* procedure of all: that of replacing missing values by sample means. Apart from the obvious computational ease of the procedure, advantages over ordinary least squares (on the complete observations only) are claimed for it in some circumstances. To quote Kmenta (1971): 'One redeeming feature of the estimators . . . is the fact that when the correlation between X and Y is low, the mean square error of these estimators is less than that of ordinary least squares'.

This note examines why the simple method of substituting means seems to give good results in certain circumstances. The objective is not to promote this simple approach instead of more sophisticated methods – no one has ever claimed that it could be generally superior – but to explain when and why its mean square error properties are superior to ordinary least squares or even to other methods.

II MISSING VALUES OF THE DEPENDENT VARIABLE

Consider the case of simple regression when there are r_1 observations on x and r_2 ($< r_1$) complete pairs of both x and y. The estimator of the regression coefficient obtained by replacing missing y values by the simple mean and performing the usual calculations is

$$\hat{\beta} = \frac{S'xy}{S'x^2} = \frac{Sxy}{S'x^2} = \beta^* \frac{Sx^2}{S'x^2} \qquad (1)$$

where S' implies summation over all r_1 pairs including those in which the sample mean has been substituted, and β^* is the usual estimator from the r_2 complete observations. Afifi and Elashoff (1967) gave comparisons of the mean-square error of $\hat{\beta}$ relative to that of β^*, having assumed that

$$y_i = \alpha + \beta x_i + e_i,$$

where $E(e_i) = 0$, $V(e_i) = \sigma^2 (1-\rho^2)$ and X_i is $N(\mu_x, \sigma_x^2)$

Using their results, the mean square error of $\hat{\beta}$ is:

$$\frac{(r_2-1)\sigma^2(1-\rho^2)}{(r_1-3)(r_1-1)\sigma_x^2} + \frac{\beta^2(r_1-r_2)}{(r_1-1)^2}\left\{\frac{2(r_2-1)}{r_1+1} + (r_1-r_2)\right\} \qquad (2)$$

while that of β^* is

$$\frac{\sigma^2(1-\rho^2)}{(r_2-3)\sigma_x^2} \qquad (3)$$

The first term of (2) will clearly be much smaller than (3) provided r_1 is considerably greater than r_2. The second term of (2) depends on the true coefficient β and if this is small (2) can be less than (3). If β is not small enough (2) will exceed (3), so the use of $\hat{\beta}$ instead of β^* implies some degree of prior knowledge. But, if β is small, has the reduction in MSE resulted from information contained in the extra incomplete observations? This is a valid question because Hocking and Smith (1972) have shown that if y and x are bivariate Normal — or multivariate Normal if X is a vector of explanatory variables — and y values are missing from some observations, then the maximum likelihood estimate of β is just β^*. (This does not apply to the intercept term.)

In formula (1), Sx^2 is always less than $S'x^2$, so $\hat{\beta}$ is obtained from β^* by 'shrinking' it. Shrinkage estimators form a class of biased estimators obtained by multiplying least square estimators by quantities less than unity. Their properties have been discussed in the statistical literature; for example,

by Mayer and Wilke (1973). The optimal – in a minimum mean square error sense – shrinkage factor is

$$1/\left\{1+\frac{\sigma^2(1-\rho^2)}{\beta^2\ Sx^2}\right\} \tag{4}$$

so, provided β^2/σ^2 is bounded, an improvement over least squares is possible. The shrinkage estimator, defined by (4) is also superior to (1) although it uses only the complete observations. Since formula (4) involves unknown parameters that must be estimated from the data or from, probably vague, prior knowledge, the optimality may not be attainable in practice.

But how plausible is S^2x/S'^2x as a shrinkage factor? It does not involve σ^2, ρ or β and its magnitude depends on the number of incomplete observations. The degree of shrinkage increases with the number of extra incomplete observations. But from (4) the optimal shrinkage increases as β^2 decreases relative to $\sigma^2\ (1-\rho^2)$. Unless it could be postulated that missing values will increase in frequency in situations where the coefficient is small, $Sx^2/S'x^2$ is inappropriate as a shrinkage factor. So, even in cases where the choice of a 'good' shrinkage factor is far from clear, it can be stated, somewhat negatively, that the method of substituting the mean for missing Y values is unlikely to help.

III MISSING VALUES OF AN EXPLANATORY VARIABLE

The simplest case, the dependent variable measured on r_1 observations and one explanatory variable with missing values for r_1-r_2 of these, is a trivial one because substituting the sample mean just gives the ordinary estimator. Suppose then that there are two explanatory variables, x_1 measured on all r_1 observations and x_2, which is missing for r_1-r_2 of them. It is obvious that:

$$S'x_2^2 = Sx_2^2,\ S'x_1x_2 = Sx_1x_2 \text{ and } S'x_2y = Sx_2y$$

and then

$$\hat{\beta}_1 = \frac{Sx_2y\ S'x_1^2 - S'x_1y\ Sx_1x_2}{S'x_1^2\ Sx_2^2 - (Sx_1x_2)^2} \tag{5}$$

and

$$\hat{\beta}_2 = \frac{S'x_1y\ Sx_2^2 - Sx_2y\ Sx_1x_2}{S'x_1^2\ Sx_2^2 - (Sx_1x_2)^2}$$

Taking expectations, where the model $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$ is assumed, gives

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{Sx_2^2 (S^*x_1 x_2 - Sx_1 x_2)}{S'x_1^2 Sx_2^2 - (Sx_1 x_2)^2}$$

and (6)

$$E(\hat{\beta}_2) = \beta_2 \left\{ \frac{S'x_1^2 Sx_2^2 - Sx_1 x_2 S^*x_1 x_2}{S'x_1^2 Sx_2^2 - (Sx_1 x_2)^2} \right\}$$

$S^*x_1 x_2$ differs from $S'x_1 x_2$ in that it contains the true (but unknown) x_2 values instead of the sample means. Thus the expectations (6) are not actually estimable unless further assumptions are made about the x's. However, it is clear that the estimators (5) are biased; the extent of the bias depending on the magnitude of β_2. $\hat{\beta}_2$ is biased downwards and $\hat{\beta}_1$ biased upwards (assuming x_1 and x_2 positively related; if negatively related $\hat{\beta}_1$ is biased downwards). The variances of $\hat{\beta}_1$ and $\hat{\beta}_2$, conditional on the x's, are

$$\frac{\sigma_y^2 Sx_2^2}{S'x_1^2 Sx_2^2 - (Sx_1 x_2)^2} \quad \text{and} \quad \frac{\sigma_y^2 S'x_1^2}{S'x_1^2 Sx_2^2 - (Sx_1 x_2)^2} \tag{7}$$

respectively, and these are smaller than the variances of the ordinary least squares estimators, which are

$$\frac{\sigma_y^2 Sx_2^2}{Sx_1^2 Sx_2^2 - (Sx_1 x_2)^2} \quad \text{and} \quad \frac{\sigma_y^2 Sx_1^2}{Sx_1^2 Sx_2^2 - (Sx_1 x_2)^2} \tag{8}$$

The difference will not be appreciable in the case of $\hat{\beta}_2$ if $Sx_1 x_2$ is small. The differences could be very large, however, if the denominators in (8) were close to zero. Excluding this case for the present and remembering that mean square error is the sum of variance and squared bias, it is evident that the estimators (5) will only have better mean square errors than the least squares estimators if β_2 is small. As in Section II it appears that prior knowledge is needed to justify the estimators.

There are other estimators for this situation, of course. Another approach is to replace the missing values by

$$\hat{x}_{2i} = \bar{x}_2 + (x_{1i} - \bar{x}_1) Sx_1 x_2 / Sx_1^2, \tag{9}$$

where x_{1i} is the value of x_1 corresponding to the missing x_2 value and $\bar{x}_2$ is the sample mean over the r_2 complete observations. Now the standard analysis of the 'completed' data gives the usual least squares estimator $\beta_2^{\ddagger}$ say, for β_2 and

$$\frac{S'x_1 y}{S'x_1^2} - \frac{Sx_1 x_2}{Sx_1^2}\beta_2^{\dagger} \qquad (10)$$

for β_1. Conditionally on the x's, this has expectation

$$\beta_1 + \beta_2 \left\{ \frac{S'x_1 x_2}{S'x_1^2} - \frac{Sx_1 x_2}{Sx_1^2} \right\},$$

and if x_2 is itself assumed to have a linear regression on x_1, the further expectation over x_2 is β_1. So with this frequently made assumption, (10) is unbiased, unlike the estimator $\hat{\beta}_1$ given by (5). The mean square error of (10) is easily shown to be smaller than that of $\hat{\beta}_1$ except for small β_2 and for the circumstances to be discussed next. The estimator (10) can be improved on by various modifications; for example, starting with (9) and conducting a weighted instead of unweighted analysis. The approach of Dagenais (1973) and Methods 4 and 5 of Beale and Little (1975) are of this type.

But there is a situation where the mean square error properties of the simple method of substituting means are clearly superior to ordinary least squares on the complete observations and even superior to the more sophisticated methods mentioned. If

$$Sx_1^2 \, Sx_2^2 - (Sx_1 x_2)^2 \qquad (11)$$

is close to zero, that is, if there is a severe multicolliniarity problem, the variances given by (8) will be very large. Those given by (7) will be much smaller and this reduction will more than offset the squared bias, provided β_2 is bounded. The simple method is also superior to (10) because the variance of that estimator may be shown to contain (11) as a denominator. It is easy to see why any approach, based on predicting missing values from values of another explanatory variable, will fail in an extreme multicolliniarity situation. Formula (9) sets up an exact linear relationship between the explanatory variables for the incomplete observations and so makes the original multicolliniarity problem even more severe.

This superiority, in a mean square error sense, of the method of substituting means, is another example of the fairly general finding that, in regression situations, biased estimators are most effective when extreme multicolliniarity is present in the data. In summary then, although the method of substituting means for missing explanatory variables is not generally sound, it could outperform other approaches, at least in terms of mean square error, in studies where the data-set is close to multicolliniar.

REFERENCES

AFIFI, A. A. and R. M. ELASHOFF, 1966. "Missing Observations in Multivariate Statistics I. Review of the Literature." *Journal of American Statistical Association, 61,* 595–604.

AFIFI, A. A. and R. M. ELASHOFF, 1967. "Missing Observations in Multivariate Statistics II. Point Estimation in Simple Linear Regression." *Journal of American Statistical Association, 62,* 10–29.

BEALE, E. M. L. and R. J. A. LITTLE, 1975. "Missing Observations in Multivariate Analysis." *Journal Royal Statistical Society B, 37,* 129–145.

BUCK, S. F. 1960. "A Method of estimation of Missing Values in Multivariate data suitable for use with an electronic computer." *Journal Royal Statistical Society B, 22,* 302–306.

DAGENAIS, M. G., 1973. "The use of Incomplete observations in Multiple Regression Analysis." *Journal of Econometrics,* 1, 317–328.

HAITOVSKY, Y., 1968. "Missing Data in Regression Analysis." *Journal Royal Statistical Society B, 30,* 67–82.

HOCKING, R. R. and W. B. SMITH, 1968. "Estimation of Parameters in the Multivariate Normal Distribution with missing observations." *Journal of American Statistical Association* 63, 159–173.

HOCKING, R. R. and W. B. SMITH, 1972. "Optimum Incomplete Multinormal Samples" *Technometrics 14,* 299–307.

KMENTA, J., 1971. *Elements of Econometrics,* Chapter 9, New York: Macmillan.

MAYER, L. S. and T. A. WILKE, 1973. "On Biased Estimation in Linear Models." *Technometrics,* 15, 497–508.