The Economic and Social Review, Vol. 14, No. 4, July 1983, pp. 259-272

Comments on the Weighted Regression Approach to Missing Values

D. CONNIFFE
The Economic and Social Research Institute, Dublin

Abstract: These comments relate to those methods of dealing with missing values of explanatory variables in regression analysis that first "complete" the data by inserting estimates derived from regressions of explanatory variables on each other and then employ some form of weighted regression. It is argued that the choices of weights in the published methods are not optimal and that improvements are possible. This is verified for a simple case and the difficulties to extending the methodology to general cases are discussed.

I INTRODUCTION

There is an extensive literature on the topic of estimating parameters of equations when some observations are incomplete. Important papers in the statistical literature include Anderson (1957), Buck (1960), Hocking and Smith (1968), Hartley and Hocking (1971), Orchard and Woodbury (1972), Rubin (1974), Beale and Little (1975) and Dempster, Laird and Rubin (1977). The more specifically econometric literature includes Dagenais (1973), Kmenta (1978), Gourieroux and Monfort (1981), Dagenais and Dagenais (1982) and Conniffe (1983). However, this paper is concerned only with the sub-class of methods in which missing values of explanatory variables are first replaced by estimates before applying some form of weighted regression analysis to the "completed" data. Furthermore the estimates are presumed derived from regression equations of explanatory variables on each other. It is assumed that no values of the dependent variable are missing.

The idea of "filling in" missing values by this method and then employing standard least squares formulae is of long standing and is included in Afifi and Elashoff's (1966) review of methods. But the standard formulae treat all observations on the same basis whether or not they are initially incomplete. Intuitively, it seems more plausible to conduct a weighted regression where complete and incomplete observations need not be treated as of equal status. This approach was taken by Dagenais (1973) and Beale and Little (1975).

The Dagenais argument commenced by assuming the true model is

$$Y = XB + U \tag{1}$$

to which the standard assumptions of multiple regression would apply, but that some elements of some rows of X are unavailable. Let these be estimated using regression equations of explanatory variables on each other, as determined from the sub-set of complete data, and let Z denote the "completed" matrix of explanatory variables. Since

$$X = Z + (X - Z)$$

the model (1) becomes

$$Y = ZB + W \tag{2}$$

where $$W = (X - Z)B + U \tag{3}$$

For a complete observation the row of X is identical to that of Z and so the relevant component of W is identical to that of U. It is a random variable with mean zero and variance σ_u^2, say. When an observation is incomplete, the corresponding row of X - Z contains at least one non-zero element, which is the difference between the (unknown) true value and the regression estimate of that value. Provided assumptions of stochastic explanatory variables with linear regressions on each other are true, the estimate has the same expectation as that of the true value, where the expectation is taken over the distribution of the explanatory variable. So the component of W has again zero mean but the variance is a function of that of at least one explanatory variable and at least one component of B.

$$\begin{aligned} \text{Let } \Sigma &= E(WW^t) \\ &= I\sigma_u^2 + E[(X - Z)BB^t(X - Z)^t] \end{aligned} \tag{4}$$

Ignoring, for the present, the fact that the second term of (4) must be estimated and treating Σ as if it were a known matrix, analogy with generalised least squares suggests the estimator

$$(Z^t\Sigma^{-1}Z)^{-1}Z^t\Sigma^{-1}Y \tag{5}$$

This is essentially the estimator proposed by Dagenais (1973) although he neglected small terms in the second component of (4) including the off-diagonal terms. For example, for the case of two explanatory variables (apart from an intercept term) and two types of observations – either complete or with x_2 missing – his variances were σ_u^2 for complete observations and

$$\sigma_u^2 + b_2^2\sigma_2^2 \qquad (6)$$

for incomplete observations. Here b_2 is the regression coefficient of y on x_2 in model (1) and σ_2^2 the variance of x_2 about its regression on x_1. But the precise value of the ith diagonal term of (4) would be:

$$\sigma_u^2 + b_2^2\sigma_2^2 \left\{1 + \frac{1}{r} + (x_{1i} - \bar{x}_1)^2/Sx_1^2\right\} \qquad (7)$$

where r denotes the number of complete observations and the mean and sum of squared deviations in (7) are calculated over these complete observations. The three terms within the parentheses in (7) are familiar components of the variance of a prediction from a regression line and the second and third will be small if r is reasonably large. Similarly, from (4), the off-diagonal element of Σ corresponding to observations i and j is:

$$b_2^2\sigma_2^2 \quad \left\{\frac{1}{r} + (x_{1i} - \bar{x}_1)(x_{1j} - \bar{x}_1)/Sx_1^2\right\}$$

Even using (6) involves unknown parameters so Dagenais first estimated these from the complete observations before conducting the weighted regression. His simulation studies showed the procedure compared favourably with the alternative of just analysing complete observations only. A weighted regression allowing for the small terms in (7), etc., is also possible and Gourieroux and Monfort (1981) give an asymptotic result for one case.

Beale and Little (1975) arrived at a quite similar method of estimation. They argued that the "value" of an observation in which, say, k of the p explanatory variables were measured, could be taken as inversely proportional to the variance of the distribution of y conditionally on these k variables only. This conditional variance could be estimated by taking all observations for which the k variables were available and calculating the residual mean square of the regression of y on them. The weighted regression was based on these "values". However, the expectations of the residual mean squares are exactly the variances used by Dagenais (1973). For example, in the simple case already mentioned the expectation of the residual mean square in a regression of y on x_1 over all observations would be (6). So differences between the two methods are only a matter of computational detail. The Beale and Little paper examined other estimates also, including an iterative maximum likelihood method, and a simulation study showed the weighted regression approach to be usually inferior to this. Indeed, the weighted regression was sometimes inferior to analysis of complete observations only. The paper concluded by recommending the maximum likelihood method and only advised the use of weighted observations as a device to assign standard errors to coefficients.

There remains the possibility that a different choice of weights could lead to an estimator with better properties. The idea that incomplete observations should be assigned larger variances than complete ones in a weighted regression is intuitively appealing. The generalised least squares analogy is suggestive and the "value" arguments not implausible. However, Section II argues that these weights are sub-optimal and formulates another approach to optimal weighted regression. The approach is illustrated for a simple pattern of missing values in Section III and the results are contrasted with those obtained using the published methods. Unfortunately there are algebraic problems in extending the approach to complex patterns and some difficulties and possibilities are discussed in Section IV.

II WEIGHTING SCHEMES

It should not be assumed that (5) must be an optimal estimator (in the minimum variance unbiased sense) even if Σ were known. It is not as if the model (2) corresponds to a genuine generalisated least squares problem; that is, if Z may be treated as a matrix of constants with the first term on the right hand side of (2) the expectation of Y. Instead, some elements of Z are stochastic and the expectation of the first term is the same as the expectation of Y. Because one consequence is that some elements of Z and W are correlated, Kmenta (1978) stated that the weighted estimator and even the simpler one obtained by applying standard regression formulae to the "completed" data (which, for later reference, will be called the "unweighted" estimator) are biased. It will be shown in Section III that this is not necessarily true. On the other hand, minimum small-sample variance properties cannot be presumed and so it may be possible to improve on (5).

The argument for weights based on the concept of "values" of incomplete observations may also be interpreted somewhat differently. Consider the "marginal value" of an incomplete observation — the value of an extra observation of a particular type. The set of complete observations permit estimation of all the coefficients. The set of incomplete observations of a particular type — that is with the same variables missing — do not permit estimation of all the coefficients. What the sums of squares and cross-products calculable within the set do permit is the estimation of some functions of coefficients of the original equation and of the regressions of explanatory variables on each other. The complete observation estimates are required to transform this information on functions into improved estimates of all coefficients of the original equation. Suppose the number of incomplete observations of a particular type become very large. Now these functions are estimated very precisely but this is not necessarily true of the coefficients if the number of complete observations is small. So the marginal value of a complete observation relative to an incomplete one of this type is much greater when the frequency of the incomplete observations is large relative to the frequency of complete observations. Thus, the "values" and hence weights assigned to incomplete observations may need to reflect the frequencies of the types of observations as well as their composition in terms of missing variables. Kelejian (1969), although not dealing with weighted regression, uses a similar argument to this in assessing the usefulness of analysis of incomplete observations.

One approach to optimal weighting is as follows:

$$(Z^t\Omega^{-1}Z)^{-1}Z^t\Omega^{-1}Y \qquad (8)$$

where Z and Y are as in (2). Taking $\Omega = I$ gives the "unweighted" estimator and $\Omega = \Sigma$ gives one weighted estimator. But the "best" Ω could be defined as that which makes (8) the unbiased estimator of minimum variance in the entire class that could be generated by varying the weightings of the incomplete observations. This is too broad a class to ensure estimability because since Ω is $(n \times n)$, where n is the number of observations, there could be more unknowns in an arbitrary Ω than there are observations. So assume that observations of the same type are assigned the same diagonal element, and off-diagonal elements corresponding to (i, j) and(i, k) are the same if j and k are the same type. The assumptions seem plausible because if i and j are of the same type, they are presumably of equal value, given assumptions of randomness of occurrence of missing values, and the weights assigned ought to be equal. It is also assumed that there are sufficient complete observations to permit estimation.

The process of obtaining an optimal Ω would require constraining the elements of Ω so that (8) is unbiased and then obtaining an expression for the variance matrix and seeking the choice of elements to minimise this (either in the sense that the difference between this variance matrix and any other is negative definite or, less strongly, in the sense of minimising the trace). It is evident that this is really a large-sample concept of optimality because the elements of Ω will undoubtedly be functions of the unknown parameters and will require estimation, so that small sample optimality (and even unbiasedness, perhaps) is not guaranteed. Even so, the proposed procedure is algebraically formidable except in simple cases.

III A SIMPLE CASE

Let there be two explanatory variables x_1, which is available for all n observations, and x_2, which is available only for the first r observations. For this case, the Zs, as defined in Section I, become

$$Z_{1i} = x_{1i}, \text{ for all i, and } Z_{2i} = x_{2i}, \text{ for } i \leqslant r,$$

and $$Z_{2i} = \bar{x}_2 + (x_{1i} - \bar{x}_1) Sx_1x_2 / Sx_1^2, \text{ for } i > r,$$

where $$Sx_1x_2 = \Sigma(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \text{ and } Sx_1^2 = \Sigma(x_{1i} - \bar{x}_1)^2$$

and the means and sums of squares and cross-products have been calculated over the first r observations.

Let Ω equal the partitioned matrix

$$\begin{bmatrix} I\sigma_u^2 & O \\ O^t & A \end{bmatrix}, \text{ where } A = \begin{bmatrix} D & C & \dots & C \\ C & D & \dots & C \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ C & C & \dots & D \end{bmatrix}$$

where I is an r X r unit matrix, O an r X (n - r) matrix of zeros and A an (n - r) X (n - r) matrix. The inverse of Ω is easily shown to be:

$$\begin{bmatrix} \frac{1}{I\sigma_u^2} & O \\ O^t & F \end{bmatrix}, \text{ where } F = \begin{bmatrix} G & H & \dots & H \\ H & G & \dots & H \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ H & H & \dots & G \end{bmatrix}$$

$$\text{where } G = \frac{D + (n - r - 2)C}{(D - C)\ \{D + (n - r - 1)C\}} \text{ and } H = \frac{-C}{(D - C)\ \{D + (n - r - 1)C\}}$$

Suppose the true model is $Y = b_0 + b_1x_1 + b_2x_2 + U$, then the estimators for b_0, b_1 and b_2 are:

$$(Z^t\Omega^{-1}Z)^{-1}Z^t\Omega^{-1}Y, \text{ where } Z = \begin{bmatrix} 1 & Z_{11} & Z_{21} \\ 1 & Z_{12} & Z_{22} \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & Z_{1n} & Z_{2n} \end{bmatrix}$$

Evaluating and simplifying, as outlined in the Appendix, gives the standard estimator, obtainable from the complete observations only, for b_2 (call it $\hat{b}_2$) and

$$b_1^* = \frac{Sx_1y + \ell S'x_1y + m(\bar{x}_1 - \bar{x}'_1)(\bar{y} - \bar{y}')}{Sx_1^2 + \ell S'x_1^2 + m(\bar{x}_1 - \bar{x}'_1)^2} - \hat{b}_2 \frac{Sx_1x_2}{Sx_1^2} \tag{9}$$

where primes indicate means or sums of squares and cross-products taken over the $n - r$ incomplete observations,

$$\ell = \frac{\sigma_u^2}{D - C} \qquad m = \frac{r(n - r)\sigma_u^2}{r(D - C) + (n - r)(\sigma_u^2 + rC)}$$

The expectation of (9), conditionally on x_2 is

$$b_1 + \left[\frac{Sx_1x_2 + \ell S'x_1x_2 + m(\bar{x}_1 - \bar{x}'_1)(\bar{x}_2 - \bar{x}'_2)}{Sx_1^2 + \ell S'x_1^2 + m(\bar{x}_1 - \bar{x}'_1)^2} - \frac{Sx_1x_2}{Sx_1^2}\right] b_2 \tag{10}$$

If we now take expectations over x_2, it is clear that (9) is unbiased since, given linear regression of x_2 on x_1 (and, of course, stationarity of the regression coefficient over all n observations)

$$E(Sx_1x_2) = kSx_1^2,\ E(S'x_1x_2) = kS'x_1^2,\ E(\bar{x}_2 - \bar{x}'_2) = k(\bar{x}_1 - \bar{x}'_1)$$

where k is the regression coefficient of x_2 on x_1. This, of course, assumes that ℓ and m are constants. In fact, we will find that the values of ℓ and m that minimise variance are functions of parameters and will have to be estimated in practice so that the foregoing argument will not suffice. But consider the case of "unweighted" regression: $D = \sigma_u^2$, $C = 0$. Then $\ell = 1$ and $m = r(n - r)/n$, both pure constants. So the "unweighted" estimator is unbiased. Note that this result contradicts the assertion by Kmenta (1978), mentioned in the previous section.

The variance of the estimator (9) may be calculated as the sum of two components. The first is the expectation, over x_2, of the variance of (9), conditionally on x_1. The second is the variance, over x_2, of the conditional expectation as given by (10). Note that in (9) the terms Sx_1y and $S'x_1y$, are independent because they are based on different y values; Sx_1y, $\bar{y}_1$ and $\hat{b}_2$ have zero covariances because they are orthogonal linear combinations of the ys; and $S'x_1y$, $\bar{y}'$ and $\hat{b}_2$ also have zero covariances for both of the previous reasons. So the conditional variance of (9) is

$$\sigma_u^2 \left\{ \frac{Sx_1^2 + \ell^2 S'x_1^2 + m^2(\bar{x}_1 - \bar{x}'_1)^2 n/r(n-r)}{[Sx_1^2 + \ell S'x_1^2 + m(\bar{x}_1 - \bar{x}'_1)^2]^2} \right\} + \frac{\sigma_u^2 (Sx_1 x_2)^2}{Sx_1^2 [Sx_1^2 Sx_2^2 - (Sx_1 x_2)]} \qquad (11)$$

So far, no probability distributions have been specified by y or x_2. In getting the expectation over x_2 of this conditional variance, no difficulty arises with the first term of (11) since it does not contain x_2. Some distributional assumption would be required to obtain the expectation of the second term, but since the term does not contain ℓ or m this cannot affect the choice of best estimator so we simply neglect the term. In calculating the variance, over x_2, of (10), note the independence of Sx_1x_2 and $S'x_1x_2$, the zero covariance of Sx_1x_2 and $\bar{x}_2$ and the zero covariance of $S'x_1x_2$ and $\bar{x}'_2$. We obtain

$$\frac{b_2^2\sigma_2^2 Q_2}{Q_1^2} - \frac{2b_2^2\sigma_2^2}{Q_1} + \frac{b_2^2\sigma_2^2}{Sx_1^2} \qquad (12)$$

where $Q_1 = Sx_1^2 + \ell S'x_1^2 + m(\bar{x}_1 - \bar{x}'_1)^2$

$Q_2 = Sx_1^2 + \ell^2 S'x_1^2 + m^2(\bar{x}_1 - \bar{x}'_1)^2 n/r(n-r)$

and σ_2^2 is the variance of x_2 about its regression on x_1. The third term of (12) is not a function of ℓ or m. Now we differentiate the sum of the first term of (11) and the first and second terms of (12) with respect to ℓ and m and equate to zero. Making use of the identity

$$S''x_1^2 = Sx_1^2 + S'x_1^2 + (\bar{x}_1 - \bar{x}'_1)^2 r(n-r)/n \qquad (13)$$

where the double prime denotes summation over all n observations, we find

$$m = \frac{\ell r(n-r)}{n} \quad \text{and} \quad \ell = \frac{\sigma_u^2}{\sigma_u^2 + b_2^2\sigma_2^2 S''x_1^2/Sx_1^2}$$

or,

$$C = \frac{b_2^2\sigma_2^2 S''x_1^2}{rSx_1^2} \quad \text{and} \quad D = \sigma_u^2 + \frac{b_2^2\sigma_2^2 S''x_1^2}{Sx_1^2} + C \qquad (14)$$

The Dagenais (1973) or Beale and Little (1975) estimates, for this case, would be given by

$$C = 0 \qquad \text{and} \qquad D = \sigma_u^2 + b_2^2\sigma_2^2 \qquad (15)$$

The "optimal" values of C and D as given by (14) supports the intuitive argument given in the previous section. D increases with $S''x_1^2/Sx_1^2$; that is, the greater n relative to r. So the greater the relative frequency of incomplete observations the more they are weighted against in the regression. The estimates (14) and (15) are not asymptotically equivalent because although C in (14) will tend to zero, D will tend to $\sigma_u^2 + \delta b_2^2 \sigma_2^2$ where

$$\delta = \underset{\substack{r \to \infty \\ n \to \infty}}{\mathrm{Lt}} \quad S''x_1^2/Sx_1^2$$

Substituting (14) back into the variance formula would give a lower bound to the small sample variance. It is not the exact variance because C and D require estimates of parameters. Similarly, unbiasedness is not immediately evident. However, if (14) are substituted back into (9) via ℓ and m, the expression for the estimator can be rewritten in the form:

$$\lambda \frac{S''x_1 y}{S''x_1^2} + (1 - \lambda) \frac{Sx_1 y}{Sx_1^2} - \hat{b}_2 \frac{Sx_1 x_2}{Sx_1^2} \tag{16}$$

where $\lambda = \sigma_u^2/(\sigma_u^2 + b_2^2\sigma_2^2)$. Now suppose λ is estimated by the ratio of residual mean squares of y on x_1 and x_2, and y on x_1, using the complete observations as data. Then (16) is a special case of the estimators discussed by Conniffe (1983) which were shown to be unbiased. (This coincidence of estimators seems to be a special case. If the argument of this section is extended to several explanatory variables with missing values and several without, the resulting "optimal" weighted regression estimators do not seem to coincide with the corresponding generalisation of (16).)

It is interesting to quantify the differences in variance between the estimators given by (14) and (15). Although the exact small sample variance formula given by Conniffe (1983) can be applied to (16) there is no corresponding algebraic formula for the estimator corresponding to (15) so a simulation study was undertaken, with D in (15) also estimated by the residual mean square. For the study, n = 18, r = 9 and the x_1 values were the integers 1 and 18. This choice, by including a trend in x_1, deliberately makes $S''x_1^2/Sx_1^2$ large so as to make the Ds in (14) and (15) appreciably different. Conditional normality of y given x_1 and x_2 and of x_2 given x_1 was assumed. Obviously, if x_2 was too closely related to x_1 a multicollinearity problem would arise and both (14) and (15) would give estimators with large variances (as indeed would have occurred with the usual estimator even without missing values). So a squared "correlation" between x_2 and x_1 of 0.5 was chosen. For completeness, the variance of the "unweighted" estimator was also computed. In Table 1 the variances of all these estimators are expressed as ratios of the variance of the standard estimator based on com-

Table 1: *Ratios of estimator variances to variance of complete observation estimator*

$b_2\sigma_2/\sigma_u$	*Unweighted*	*"Published" Weights*	*"Optimal" Weights*	M_1	M_2
4.0	7.90	2.68	0.96	1.11	1.86
2.0	2.38	1.67	0.90	1.01	1.42
1.0	1.01	0.93	0.77	0.80	0.89
0.5	0.67	0.66	0.64	0.65	0.65

plete observations only.

The weights (14) and (15) are referred to as "Optimal" and "Published" respectively. The columns headed M_1 and M_2 will be explained in the next section. The simulation was conducted for a range of values of $b_2\sigma_2/\sigma_u$ because the consequences of missing values of x_2 clearly depend on the magnitude of the standardised regression coefficient. If it is zero or small, little information is lost by missing values of x_2 and, hence, an analysis including the incomplete observations will be considerably more efficient than one ignoring them, while if it is large the converse is true. Each tabular value is based on 1,000 replications. The "Optimal" is truly best except when $b_2\sigma_2/\sigma_u$ is small when there is no difference between estimators, all three being considerably superior to analysis of complete observations only. But when $b_2\sigma_2/\sigma_u$ becomes larger only the "Optimal" retains its superiority and the others can become alarmingly inferior. So using the "Published" rather than "Optimal" weights in this case may not mean just a reduction in efficiency but will determine whether or not the incomplete observations were worth including in the analysis at all.

IV PROBLEMS OF GENERALISATION

The previous section has illustrated the importance of "Optimal" weights in a simple case and there seems no reason why choice of weights should be less important in more complex cases. Unfortunately, the algebra of deriving weights was not trivial even with the simple patterns. For more complicated cases the expectations and variances of estimators will be more difficult to obtain. For example, the fact that no x_1 values were missing in the example permitted working with expectation and variances conditional on x_1. In more general cases, it may be necessary to take expectations over all variables and this could require specification of the multivariate distribution of the variables. Also in complex cases, and even if working with the average variance of coefficients, the optimisation stage would involve the solution of an apparently complicated set of non-linear equations.

At this point it is as well to review why one might prefer a weighted regression procedure to an iterative maximum likelihood method. While it is conceivable that with appropriate weights the estimators may not be less efficient than maximum likelihood in small samples (the simulation in Conniffe, 1983, would justify this for the case of Section III) it seems unlikely they could ever be much superior. So efficiency is not a reason. There are two other possible reasons. The first is that iterative maximum likelihood is computationally demanding and estimates the regression equation only after estimating all the parameters of the multivariate distribution. If only a single regression equation is of interest one feels there should be a short-cut. But this argument loses whatever virtue it possesses unless the weights can be fairly easily determined. The second reason is the distributional assumptions required for likelihood methods. But, as already remarked, the derivation of "Optimal" weights would seem to require these too.

Another possibility is to seek weights that if not "Optimal" are closer than the "Published" ones. The columns M_1 and M_2 in Table 1 correspond to the estimators given by the weights:

$$D = \sigma_u^2 + b_2^2\sigma_2^2 S''x_1^2/Sx_1^2 \qquad C = 0,$$

and $$D = \sigma_u^2 + b_2^2\sigma_2^2 n/r \qquad C = 0.$$

The table shows that setting off-diagonal terms to zero does not greatly impair efficiency. The further modification to a weight involving a simple relative frequency does have a substantial detrimental effect but is still superior to the "Published" weights. Furthermore, the choice of the positive integers for x_1 was extreme. If x_1 was normally distributed the expectation of $S''x_1^2/Sx_1^2$ would be $(n - 3)/(r - 3)$. So a simple modification of "Published" weights by the relative frequency might produce a good improvement.

Unfortunately the appropriate extension to general cases is unclear, at least to the author. It is easy to modify each weight by the relative frequency of that type of observation but, with more than two types of observation, this is ignoring a lot about pattern. For example, in a case of four explanatory variables with three types of observation, including complete, observations with x_3 and x_4 missing are of greater value if the other incomplete observations have x_1 and x_2 missing than if they have x_2 and x_3 missing. Although simulation can be used to assess any particular weighting scheme, it would be an immense task to evaluate the possible range of schemes. An insight into the relationship between pattern and appropriate weighting scheme is required to make progress. So far, the author has failed to find it. Perhaps some reader might.

REFERENCES

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

ANDERSON, T.W., 1957. "Maximum Likelihood Estimates for a Multivariate Normal Distribution when some Observations are Missing", *Journal of the American Statistical Association*, 52, pp. 200-203.

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

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, SeriesB*, 22, pp. 302-306.

CONNIFFE, D., 1983. "Small-Sample Properties of Estimators of Regression Coefficients Given a Common Pattern of Missing Data", *Review of Economic Studies*, 50, pp. 111-120.

DAGENAIS, M.G., 1973. "The Use of Incomplete Observations in Multiple Regression Analysis", *Journal of Econometrics*, 1, pp. 317-328.

DAGENAIS, M.G. and L.D. DAGENAIS, 1982. "A General Approach for Estimating Econometric Models with Incomplete Observations", In J. Paelinck (ed.,) *Qualitative and Quantitative Economics*, The Hague: Martinus Nijhoff, pp. 89-113.

DEMPSTER, A.P., N.M. LAIRD and D.B. RUBIN, 1977. "Maximum Likelihood from Incomplete Data via the E.M. Algorithm", (with discussion), *Journal of the Royal Statistical Society, Series B*, 39, pp. 1-38.

GOURIEROUX, C. and A. MONFORT, 1981. "On the Problem of Missing Data in Linear Models", *Review of Economic Studies*, 48, pp. 579-586.

HARTLEY, H.O. and R.R. HOCKING, 1971. "The Analysis of Incomplete Data", *Biometrics*, 27, pp. 783-823.

HOCKING, R.R. and W.B. SMITH, 1968. "Estimation of Parameters in the Multivariate Normal Distribution with Missing Observations", *Journal of the American Statistical Association*, 63, pp. 159-173.

KELEJIAN, H.H., 1969. "Missing Observations in Multivariate Analysis: Efficiency of a First-Order Method", *Journal of the American Statistical Association*, 64, pp. 1609-1616.

KMENTA, J., 1978. "On the Problem of Missing Measurements in the Estimation of Economic Relationships", Report 102, Department of Economics, University of Michigan.

ORCHARD, T. and M.A. WOODBURY, 1972. "A Missing Information Principle: Theory and Application", Proceedings of 6th Berkeley Symposium on Mathematical Statistics and Probability, 1, pp. 697-715.

RUBIN, D.B., 1974. "Characterising the Estimation of Parameters in Incomplete Data Problems", *Journal of the American Statistical Association*, 69, pp. 467-474.

APPENDIX

Noting that $G - (n - r - 1)H = \dfrac{1}{D + (n - r - 1)C}$ and $G - H = \dfrac{1}{D - C}$,

$$Z^t\Omega^{-1} = \begin{bmatrix} \dfrac{1}{\sigma_u^2} & \cdots & \dfrac{1}{\sigma_u^2} & \dfrac{1}{D + (n - r - 1)C} & \cdots & \dfrac{1}{D + (n - r - 1)C} \\ \dfrac{Z_{11}}{\sigma_u^2} & \cdots & \dfrac{Z_{1r}}{\sigma_u^2} & \dfrac{Z_{1r+1}}{D - C} + H\Sigma' Z_1 & \cdots & \dfrac{Z_{1n}}{D - C} + H\Sigma' Z_1 \\ \dfrac{Z_{21}}{\sigma_u^2} & \cdots & \dfrac{Z_{2r}}{\sigma_u^2} & \dfrac{Z_{2r+1}}{D - C} + H\Sigma' Z_2 & \cdots & \dfrac{Z_{2n}}{D - C} + H\Sigma' Z_2 \end{bmatrix}$$

where Σ' denotes summation over the $n - r$ incomplete observations. The terms

$$Z^t\Omega^{-1}Z \qquad \text{and} \qquad Z^t\Omega^{-1}Y$$

follow easily. Then eliminating the constant b_0 from the equations

$$Z^t\Omega^{-1}Z \quad \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = Z^t\Omega^{-1}Y$$

by subtracting multiples of the first equation from the other two, and replacing the Zs by Xs as specified in Section III of the paper, gives

$$\left\{ \frac{1}{\sigma_u^2}\left(\frac{r}{\sigma_u^2} + \theta\right) \begin{bmatrix} Sx_1^2 & Sx_1x_2 \\ Sx_1x_2 & Sx_1^2 \end{bmatrix} + \left\{ \left(\frac{r}{\sigma_u^2} + \theta\right) \frac{S'x_1^2}{D - C} + \frac{r\theta}{\sigma_u^2}(\bar{x}_1 - \bar{x}_1')^2 \right\} \begin{bmatrix} 1 & \dfrac{Sx_1x_2}{Sx_1^2} \\ \dfrac{Sx_1x_2}{Sx_1^2} & \left(\dfrac{Sx_1x_2}{Sx_1^2}\right)^2 \end{bmatrix} \right\} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

$$= \frac{1}{\sigma_u^2} \left(\frac{r}{\sigma_u^2} + \theta\right) \begin{bmatrix} Sx_1y \\ Sx_2y \end{bmatrix} + \left\{ \frac{1}{D - C} \left(\frac{r}{\sigma_u^2} + \theta\right) S'x_1y + \frac{r\theta}{\sigma_u^2} (\bar{x}_1 - \bar{x}_1')(\bar{y} - \bar{y}') \right\} \begin{bmatrix} 1 \\ Sx_1x_2/Sx_1^2 \end{bmatrix}$$

where $\theta = \dfrac{n - r}{D + (n - r - 1)C}$

and primes denote means or sums of squares and cross-products taken over the incomplete observations. Solving these equations for b_1 and b_2 gives equation (9) of Section III.