Sankhyā : The Indian Journal of Statistics
1986, Volume 48, Series B, Pt. 1, pp. 19-30.

A NOTE ON SOME METHODS FOR REGRESSION ANALYSIS WITH INCOMPLETE OBSERVATIONS

By R. D. GILL

Centre for Mathematics and Computer Science, Amsterdam, Netherlands

SUMMARY. Some recent related proposals for estimating regression coefficients with incomplete observations are discussed. The proposals included approximate standard errors for the estimators. It is shown that the estimators of the regression coefficients are consistent under fairly weak conditions, but that only under rather strong ones can the usual (asymptotic) tests of significance be validly based on the estimated coefficients and the computed standard deviations. The results depend heavily on whether a random or a fixed specification for the covariates can be assumed.

1. Introduction

Very many procedures, both specific and general, have been suggested in the literature for dealing with the problem of incomplete observations in regression analysis; see the papers of Afifi and Elashoff (1966, 1967, 1969a, b), Hartley and Hocking (1971) and Dempster, Laird and Rubin (1977). However few of the methods which are applicable in a general regression analysis situation give consistent estimators of the regression coefficients, and still fewer show how asymptotic standard deviations may be validly estimated (in order to carry out the usual t-tests, etc.). There are three very similar proposals which do at least give suggestions in this direction though little theoretical justification is given : these, the subject of this note, are Beale and Little's (1975) "method 5" and "method 6" and the method of Dagenais (1973).

Let us briefly sketch the kind of situation we are interested in. Each of N *observations* if complete would be a $(K+1)$-vector of values taken by K independent or *predictor* variables and 1 dependent or *criterion* variable. However for some observations, the values taken by some of the predictor variables are missing. We suppose that the mechanism causing this works independently of that generating the values of both predictor and criterion variables. This assumption is only implicitly made in the sequel, but it is an assumption of major importance (as is usual in the literature on this subject).

Mathematics subject classification (1980) : 62J05.

Key words and phrases : Incomplete observations, Missing data, Regression analysis.

We shall work conditionally on the realized patterns of missing and non-missing values. For the sake of simplicity we assume that none of the values taken by the criterion variable are missing.

We wish to make as few assumptions as possible about the predictor variables. In particular we certainly want to avoid the assumption that each complete $(K+1)$-vector observation is a drawing from a multivariate normal distribution. In practical applications of regression analysis it is usually possible to classify each predictor variable as being either a *fixed* "design variable" or a *random* "covariate". By design variables, we mean variables which are preset by the experimenter as part of the planned experimental design. Covariates on the other hand are variables which are generated by some stochastic mechanism jointly with the criterion variables, so that covariates and criterion variable together can be considered as drawn from some multivariate distribution (generally a different distribution for each set of values of the design variables).

Even if the natural specification of the predictor variables is random, the regression model may be specified in terms of the conditional distribution of the criterion variable given the predictor variables. So if one starts with a random specification of the predictor variables, one can step over to a fixed specification by conditioning. One also has the natural possibility, when some of the predictor variables are missing, of just conditioning on the non-missing predictor variables.

Our main aim is in fact to discover whether the methods for dealing with incomplete observations mentioned above are only appropriate with one of these possible specifications of the predictor variables. All three proposals work by filling in the missing values in each observation with least squares predictions based on the non-missing predictor variables in that observation; the coefficients needed for this are estimates based on all the present data. Then a standard weighted least squares regression analysis is carried out on the completed data set, supplying both estimates of the regression coefficients and standard errors for them. Weights are needed because the least squares prediction introduces an extra error of varying size in each incomplete observation. The proposals only differ in how the coefficients for the least squares predictions and how the weights for the final regression analysis are to be estimated (they all agree on what these coefficients and weights should be). Since it turns out that only consistency of the estimators of these quantities is needed, the differences between the proposals are not crucial.

We are interested in the problems of finding reasonable and non-technical conditions under which (i) the proposals yield consistent estimators of the regression coefficients, and (ii), suitably normed, these estimators are asymptotically normally distributed about the true regression coefficients with a covariance matrix which is *consistently estimated by that produced in the weighted least squares regression analysis.* Problem (i) turns out to have a satisfactory solution whatever the specification of the predictor variables. However in (ii), asymptotic normality is only easily proved with random predictor variables. Even when asymptotic normality can be proved, the asymptotic covariance matrix of the regression coefficient estimators only coincides with the limiting value of the covariance matrix estimator under conditions quite close to multivariate normality of the predictor variables.

In the next section we specify our general model, define the estimators, and prove consistency. Section 3 looks at asymptotic normality while in the final section we briefly discuss some implications of our results.

2. Problem (i) : consistency

First some notation. Random variables will be underlined, so that the same symbol not underlined represents a possible value of the corresponding variable. a^τ denotes the transpose of the vector a. We specify a model for N observations for each $N = 1, 2, \ldots$; all quantities (including the underlying sample space) may depend on N unless explicitly stated otherwise, though this dependence is generally suppressed in the notation. We write $\xrightarrow[\mathcal{P}]{}$ and $\xrightarrow[\mathcal{D}]{}$ for convergence in probability and in distribution respectively (always as $N \to \infty$) and denote a multivariate normal distribution with given mean vector and covariance matrix by $\mathcal{N}(\cdot, \cdot)$.

Let P and M (a *pattern* of observed predictor variables and its complement of *missing* ones) denote sets of indices such that $P \bigcup M = \{1, \ldots, K\}$ (where K is the number of predictor variables), $P \bigcap M = \phi$, $P \neq \phi$ and if e.g. the *first* predictor variable is the constant 1, $1 \in P$. Vectors and matrices will often be partitioned according to P and M, e.g. if β is a $K \times 1$ vector and Σ a $K \times K$ matrix then

$$\beta = \begin{pmatrix} \beta_P \\ \beta_M \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} \Sigma_{PP} & \Sigma_{PM} \\ \Sigma_{MP} & \Sigma_{MM} \end{pmatrix} = (\Sigma_P \ \Sigma_M). \qquad \ldots \ (1)$$

Let $(\underline{x}^n, \underline{y}^n, \underline{e}^n)$, $n = 1, \ldots, N$, denote the complete $K \times 1$ vector of predictor variables, the criterion variable, and the disturbance variable for the n-th observation, related by

$$\underline{y}^n = \beta^T \underline{x}^n + \underline{e}^n \qquad \ldots \quad (2)$$

for some fixed $K \times 1$ vector β of regression coefficients which we want to estimate. Let P^n and M^n, $n = 1, \ldots, N$ be patterns of observed and missing predictor variables; the data consists of $(\underline{y}^n, \underline{x}_P^n, P^n)$, $n = 1, \ldots, N$, where we have written $\underline{x}_P^n$ for $\underline{x}_{Pn}^n$. Similarly we often write $\underline{x}_M^n$ for the unobserved $\underline{x}_{Mn}^n$. To (2) we add the usual assumptions

$$\begin{aligned} &\mathcal{E}(\underline{e}^n) = 0 && \text{for all } n \\ &\mathcal{E}(\underline{x}^n \underline{e}^{n'}) = 0 && \text{for all } n, n' \qquad \ldots \quad (3) \\ &\mathcal{E}(\underline{e}^n \underline{e}^{n'}) = \begin{cases} 0 & n \neq n' \\ \sigma^2 > 0 & n = n' \end{cases} \end{aligned}$$

where σ^2 like β does not depend on N. The second equality in (3) implies the first one if (2) includes a constant term, e.g.

$$\underline{x}_1^n = 1 \quad \text{almost surely for each } n. \qquad \ldots \quad (4)$$

Note that our model does not yet assume independence or identical distributions for the N observations, nor have we excluded the predictor variables from being fixed design variables. We make the following further assumptions, which we shall illustrate with some important examples in a moment. For each pattern P let ρ_P be a non-negative number, and let Σ be a fixed $K \times K$ symmetric positive definite matrix. The symbol $\#$ denotes the number of elements in a set. Suppose that for each P, the folloing convergences hold as $N \to \infty$:

A1 $$N^{-1} \# \{n : P^n = P\} \to \rho_P,$$

A2 $$N^{-1} \sum_{n\,:\,P^n = P} \underline{x}_P^n \underline{x}_P^{nT} \xrightarrow{\mathcal{P}} \rho_P \Sigma_{PP},$$

A3 $$N^{-1} \sum_{n\,:\,P^n = P} \underline{x}_P^n \underline{x}_M^{nT} \xrightarrow{\mathcal{P}} \rho_P \Sigma_{PM},$$

A4 $$N^{-1} \sum_{n\,:\,P^n = P} \underline{x}_P^n \underline{e}^n \xrightarrow{\mathcal{P}} 0.$$

Defining $\sigma_P^2 = \sigma^2 + \beta_M^T (\Sigma_{MM} - \Sigma_{MP}\, \Sigma_{PP}^{-1} \Sigma_{PM}) \beta_M$ suppose we also have

A5 $$A = \sum_P \rho_P \sigma_P^{-2} \Sigma_{\cdot P} \Sigma_{PP}^{-1} \Sigma_{P\cdot} \text{ is non-singular.}$$

Interpreting Σ as a limiting average value of $\underset{\sim}{x}^n\underset{\sim}{x}^{n\mathrm{T}}$, A2 and A3 together with A1 express the fact that the patterns of missing values are at least asymptotically not influenced by the predictor variables, while A4 expresses the same fact for the disturbance. The role of A5 will become clear later.

Example 1 : *Random predictor variables* : $(\underset{\sim}{x}^n, \underset{\sim}{e}^n)$, $n = 1, \ldots, N$, are independent over n and have the same distribution for all n and N, with

$$\mathcal{E}(\underset{\sim}{x}^n\underset{\sim}{x}^{n\mathrm{T}}) = \Sigma. \qquad \ldots \quad (5)$$

In this example A2, A3 and A4 follow from (3), (5) and A1 by the weak law of large numbers, taking special care for P such that $\rho_P = 0$. Here we use the fact that if $\underset{\sim}{u}_k \xrightarrow[\mathcal{P}]{} a$ as $k \to \infty$ for some sequence of random variables $\underset{\sim}{u}_k$, and if k_N satisfies $k_N/N \to \rho$ as $N \to \infty$, then $(k_N/N)\underset{\sim}{u}_{k_N} \xrightarrow[\mathcal{P}]{} \rho a$ as $N \to \infty$. This is trivial when $\rho > 0$; when $\rho = 0$ we still have that the sequence $\underset{\sim}{u}_{k_N}$ is bounded in probability and the result is then again trivial.

Example 2 : *Fixed predictor variables* : For some vector x^n (depending also on N), $\underset{\sim}{x}^n = x^n$ almost surely for each n and N; these vectors satisfy A2 and A3 with convergence in probability replaced by ordinary convergence.

We now have that A4 follows from (3) and A2, since $N^{-1} \sum_{n:P^n=P} x_P^n \underset{\sim}{e}^n$ has expectation zero and covariance matrix $(\sigma^2 N^{-1} \sum_{n:P^n=P} x_P^n x_P^{n\mathrm{T}})/N$. This example can be obtained from Example 1 by conditioning. Suppose we are in the situation of Example 1 with $(\underset{\sim}{x}^n, \underset{\sim}{e}^n, P^n)$, $n = 1, \ldots, N$ being the first N elements of a single infinite sequence. Then for almost all realizations $(\underset{\sim}{x}^1, \underset{\sim}{x}^2, \ldots) = (x^1, x^2, \ldots)$, conditions A2 and A3 continue to hold for the conditional distribution of the data given this realization of $(\underset{\sim}{x}^1, \underset{\sim}{x}^2, \ldots)$. We need to assume separately that (3) also holds with the expectations replaced by conditional expectations given $\underset{\sim}{x}^n = x^n$.

Example 3 : *Non-missing predictor variables fixed, missing predictor variables random* : for some vectors x_P^n (depending also on N), $\underset{\sim}{x}_P^n = x_P^n$ almost surely for each n and N; these vectors satisfy A_2 with convergence in probability replaced by ordinary convergence.

Again A4 follows from (3) and A2, but A3 needs to be assumed separately. This example will generally arise from Example 1 by conditioning. Consider the situation of Example 1 with $(\underset{\sim}{x}^n, \underset{\sim}{e}^n, P^n)$, $n = 1, \ldots, N$, being the first N elements of a single infinite sequence. Then for almost all realizations $(\underset{\sim}{x}_P^1, \underset{\sim}{x}_P^2, \ldots) = (x_P^1, x_P^2, \ldots)$, condition A2 continues to hold for the conditional distribution of the data given $(x_P^1, x_P^2, \ldots)$. We need to assume A3 and the conditional form of (3) separately.

We shall not give explicit definitions of the estimators proposed in Dagenais (1973) and in Beale and Little's (1975) "method 5" and "method 6" but first work as if the parameters needed for the proposals (certain functions of σ^2, β and Σ) were known. Define

$$\alpha_{MP} = \Sigma_{MP}\ \Sigma_{PP}^{-1} \quad (\text{where } \Sigma_{PP}^{-1} = (\Sigma_{PP})^{-1}.) \qquad \ldots \quad (7)$$

$$\sigma_P^2 = \sigma^2 + \beta_M^T(\Sigma_{MM} - \Sigma_{MP}\ \Sigma_{PP}^{-1}\Sigma_{PM})\beta_M \qquad \ldots \quad (8)$$

$$\hat{\underline{x}}^n = \Sigma_{\cdot P}\Sigma_{PP}^{-1}\underline{x}_P^n = \begin{pmatrix} \underline{x}_P^n \\ \alpha_{MP}\underline{x}_P^n \end{pmatrix} \text{ where } P = P_n \qquad \ldots \quad (9)$$

$$\hat{\underline{x}} = (\underline{x}^1, \ldots, \hat{\underline{x}}^N)^T \qquad \ldots \quad (10)$$

$\hat{\Sigma} =$ the $N \times N$ diagonal matrix with diagonal elements σ_{Pn}^2, $n = 1, \ldots, N$ (11)

$$\underline{Y} = (\underline{y}^1, \ldots, \underline{y}^N)^T \qquad \ldots \quad (12)$$

$$\hat{\beta} = (\hat{\underline{X}}^T\hat{\Sigma}^{-1}\underline{X})^{-1}(\hat{\underline{X}}^T\hat{\Sigma}^{-1}\underline{Y}) \text{ if } \hat{\underline{X}}^T\hat{\Sigma}^{-1}\hat{\underline{X}} \text{ is non-singular} \qquad \ldots \quad (13)$$

If α_{MP} and σ_P^2 were known, $\hat{\beta}$ would be the proposed estimator of β and $(\hat{\underline{X}}^T\hat{\Sigma}^{-1}\hat{\underline{X}})^{-1}$ the proposed approximate covariance matrix for it. In fact in Example 1 $\hat{\underline{x}}_M^n$ is the best linear predictor of $\underline{x}_M^n$ based on $\underline{x}_P^n$ while σ_{Pn}^2 is the expanded variance of the error term in (2) if $\underline{x}_M^n$ is replaced there by $\hat{\underline{x}}_M^n$. For defining

$$\hat{\underline{e}}^n = \underline{e}^n + \beta_M^T(\underline{x}_M^n - \alpha_{MP}\underline{x}_P^n) \text{ (where } P = P^n \text{ and } M = M^n \qquad \ldots \quad (14)$$

we rewrite (2) as

$$\underline{y}^n = \beta^T\hat{\underline{x}}^n + \hat{\underline{e}}^n \quad n = 1, \ldots, N \qquad \ldots \quad (15)$$

where *in Example* 1

$$\mathcal{E}(\hat{\underline{e}}^n) = 0$$

$$\mathcal{E}(\hat{\underline{x}}^n\hat{\underline{e}}^{n'}) = 0 \qquad \text{(c.f. (3))} \qquad \ldots \quad (16)$$

$$\mathcal{E}(\hat{\underline{e}}^n\hat{\underline{e}}^{n'}) = (\hat{\Sigma})_{nn'}$$

After conditioning on $\underline{x}_P^n = x_P^n$, $n = 1, \ldots, N$ (Example 3), (16) no longer necessarily holds, while in Example 2 it is generally false.

Theorem 1 : *Under* A1 *to* A5, $\hat{\underline{\beta}}$ *defined by* (13) *is a consistent estimator of* β *and* $N(\hat{\underline{X}}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\underline{X}})^{-1}$ *is a consistent estimator of* A^{-1}. *These statements are also true if in the definitions* (7) *to* (13), α_{MP} *and* σ_P^2 *are replaced by consistent estimators* α_{MP} *and* σ_P^2 *of the same quantities.*

Proof: We first look at the estimation of A^{-1}.

$$N^{-1}\underset{\sim}{\hat{X}}^{\intercal}\hat{\mathbf{\Sigma}}^{-1}\underset{\sim}{\hat{X}} = \sum_{P} \sigma_P^{-2}\, \Sigma_{\cdot P}\Sigma_{PP}^{-1}\left(N^{-1}\sum_{n:\, P^n=P} \underset{\sim}{x}_P^n \underset{\sim}{x}_P^{n\intercal}\right)\Sigma_{PP}^{-1}\Sigma_{P\cdot}$$

$$\underset{\mathcal{P}}{\to} \sum_{P} \sigma_P^{-2}\Sigma_{\cdot P}\,\Sigma_{PP}^{-1}\rho_P\,\Sigma_{PP}\,\Sigma_{PP}^{-1}\,\Sigma_{P\cdot} = A.$$

Because A is non-singular, the probability that $N^{-1}\underset{\sim}{\hat{X}}\,\hat{\mathbf{\Sigma}}^{-1}\underset{\sim}{\hat{X}}$ is non-singular converges to 1 as $N \to \infty$ and hence

$$N(\underset{\sim}{\hat{X}}\,\hat{\mathbf{\Sigma}}^{-1}\underset{\sim}{\hat{X}})^{-1} \underset{\mathcal{P}}{\to} A^{-1}. \qquad \ldots \quad (17)$$

Next defining

$$\underset{\sim}{\hat{E}} = (\underset{\sim}{\hat{e}}^1, \ldots, \underset{\sim}{\hat{e}}^N)^{\intercal} \qquad \ldots \quad (18)$$

we can rewrite (15) as

$$\underset{\sim}{Y} = \underset{\sim}{\hat{X}}\beta + \underset{\sim}{\hat{E}} \qquad \ldots \quad (19)$$

and so by (13) and (17), with probability converging to 1,

$$\underset{\sim}{\hat{\beta}} = \beta + N(\underset{\sim}{\hat{X}}^{\intercal}\hat{\mathbf{\Sigma}}^{-1}\underset{\sim}{\hat{X}})^{-1}N^{-1}\underset{\sim}{\hat{X}}^{\intercal}\hat{\mathbf{\Sigma}}^{-1}\underset{\sim}{\hat{E}}. \qquad \ldots \quad (20)$$

So to prove

$$\underset{\sim}{\hat{\beta}} \underset{\mathcal{P}}{\to} \beta \qquad \ldots \quad (21)$$

it suffices to establish

$$N^{-1}\underset{\sim}{\hat{X}}^{\intercal}\hat{\mathbf{\Sigma}}^{-1}\underset{\sim}{\hat{E}} \underset{\mathcal{P}}{\to} 0 \text{ as } N < \infty. \qquad \ldots \quad (22)$$

Now,

$$N^{-1}\underset{\sim}{\hat{X}}^{\intercal}\hat{\mathbf{\Sigma}}^{-1}\underset{\sim}{\hat{E}} = \sum_{P} \sigma_P^{-2}\,\Sigma_{\cdot P}\Sigma_{PP}^{-1}\left(N^{-1}\sum_{n:\, P^n=P} \underset{\sim}{x}_P^n\, \underset{\sim}{\hat{e}}^n\right)$$

$$= \sum_{P} \sigma_P^{-2}\,\Sigma_{\cdot P}\,\Sigma_{PP}^{-1}\left(N^{-1}\sum_{n:\, P^n=P} (\boldsymbol{x}_P^n\, \underset{\sim}{e}^n + (\underset{\sim}{x}_P^n\, \underset{\sim}{x}_M^{n\intercal} - \underset{\sim}{x}_P^n\, \boldsymbol{x}_P^{n\intercal}\, \alpha_{MP}^{\intercal})\beta_M)\right)$$

$$\underset{\mathcal{P}}{\to} 0 \qquad \ldots \quad (23)$$

by A2, A3, A4 and (7).

Finally even if α_{MP} and σ_P^2 are everywhere replaced by consistent estimators of the same quantities, all the above arguments remain valid. □

Remark 1: The consistency of the estimators of α_{MP} and σ_P^2 in Beale and Little's (1975) "method 6" can be established by the same type of arguments as in Gill (1977) even though they derive their estimators from

considerations of maximum likelihood under multivariate normalyty of $(\underline{x}^n, \underline{e}^n)$, $n = 1, \ldots, N$. Suitable conditions for consistency are A1, A2 and A4 supplemented with

A6 For P such that $\rho_P = 0, \#\{n: P^n = P\} = 0$ for sufficiently large N, and

A7 $$N^{-1} \sum_{n\,:\,P^n=P} (e^n)^2 \underset{\mathcal{P}}{\to} \rho_P \sigma^2 \text{ for all } P.$$

The (more complicated) other two proposals are harder to analyse though a similar approach is applicable; we have not worked out the full details. Since consistency of the estimators of α_{MP} and σ_P^2 is all that is needed ($\sqrt{N}$-consistency of the estimator of α_{MP} for asymptotic normality; see next section), less complicated proposals can easily be made which still have the required properties. The proof of Theorem 1 actually also shows consistency of Beale and Little's (1975) "method 4", where weights are not introduced.

Remark 2 : If $\underline{x}_M^n$ is predicted by regression on $\underline{x}_P^n$ *and* y^n for each n, the resulting estimator of β is generally inconsistent. For instance in Example 1, if we let $\hat{\underline{x}}_M^n$ be the best linear predictor of $\underline{x}_M^n$ based on $\underline{y}^n$ and $\underline{x}_P^n$, and write

$$\underline{y}^n = \beta_P^\intercal \underline{x}_P^n + \beta_M^\intercal \hat{\underline{x}}_M^n + \hat{\underline{e}}^n,$$

then we find that in general $\mathcal{E}\underline{x}_P^n \hat{\underline{e}}^n \neq 0$ and so it does not hold that $N^{-1} \Sigma_{n\,:\,P^n=P} \underline{x}_P^n \hat{\underline{e}}^n \underset{\mathcal{P}}{\to} 0$ if $\rho_P > 0$. This fact makes another of Beale and Little's (1975) proposals (middle of their section 5) rather difficult to motivate.

3. Problem (ii) : asymptotic normality with corbect covariance matrix

Reviewing the proof of Theorem 1, we see that under the conditions of that theorem,

$$N^{\frac{1}{2}}(\underline{\hat{\beta}} - \beta) \underset{\mathcal{D}}{\to} \mathcal{N}(0, A^{-1}) \qquad \ldots \quad (24)$$

if and only if

$$N^{-\frac{1}{2}} \hat{\underline{X}}^\intercal \hat{\Sigma}^{-1} \hat{\underline{E}} \underset{\mathcal{D}}{\to} \mathcal{N}(0, A). \qquad \ldots \quad (25)$$

Since then (17) holds too we can indeed validly use $(\hat{\underline{X}}^\intercal \hat{\Sigma}^{-1} \hat{\underline{X}})^{-1}$ as an asymptotic covariance matrix for $\underline{\hat{\beta}}$ and carry out the usual tests of significance on regression coefficients. We shall prove a theorem giving conditions for (24) to hold in the special case of Example 1, but shall give some heuristic arguments that it cannot hold in Example 2, and only holds under rather special conditions in Example 3.

Theorem 2 : *Consider the situation of Example* 1 *and suppose furthermore*

$$\mathcal{E}((\hat{\underset{\sim}{e}}^n)^2 \underset{\sim}{x}_P^n \underset{\sim}{x}_P^{n\tau}) = \Psi_P \qquad \ldots \quad (26)$$

for some finite matrices Ψ_P. *Then*

$$N^{-\frac{1}{2}} \hat{\underset{\sim}{X}}^\tau \hat{\Sigma}^{-1} \hat{\underset{\sim}{E}} \underset{\mathcal{L}}{\to} \mathcal{N}(0, B) \qquad \ldots \quad (27)$$

where

$$B = \sum_P \rho_P \sigma_P^{-4} \Sigma_P . \Sigma_{PP}^{-1} \Psi_P \Sigma_{PP}^{-1} \Sigma_P . \qquad \ldots \quad (28)$$

A sufficient condition for (24) *to hold* (*i.e. for equality of A and B*) *is*

$$\Psi_P = \sigma_P^2 \Sigma_{PP} \text{ for all } P. \qquad \ldots \quad (29)$$

These results also hold with α_{MP} *and* σ_P^2 *replaced by consistent estimators provided the estimator of* α_{MP} *is actually* $\sqrt{N}$*-consistent, i.e.* $N^{\frac{1}{2}}(\hat{\underset{\sim}{\alpha}}_{MP} - \alpha_{MP})$ *is bounded in probability as* $N \to \infty$.

Proof : Multiplying (23) by $N^{\frac{1}{2}}$ and recalling (16), we see that by the Central Limit Theorem, (again with special care for P such that $\rho_P = 0$),

$$N^{-\frac{1}{2}} \hat{X}^\tau \hat{\Sigma}^{-1} \hat{\underset{\sim}{E}} \underset{\mathcal{L}}{\to} \mathcal{N}(0, B). \qquad \ldots \quad (30)$$

Obviously if (29) holds, then $A = B$. The proof of the last part of the theorem is straightforward. □

Remark 3 : Theorem 2 is a satisfactory solution to problem (ii) if we can consider the predictor variables as random and can assume that the complete observations would have been independent and identically distributed, with $(\hat{\underset{\sim}{e}})^2$ uncorrelated with $\underset{\sim}{x}_P^n \underset{\sim}{x}_P^{n\tau}$. This does not seem a very heavy assumption; we already have that $\hat{\underset{\sim}{e}}_n$ is uncorrelated with $\underset{\sim}{x}_P^n$.

Of course one could often be reluctant to assume that the $\underset{\sim}{x}^n$'s are random variables at all. However, it is easy to see that (27) cannot hold under reasonable conditions in the case of Example 2, even with a different definition of the matrix B. In (27) $\hat{\underset{\sim}{X}} = \hat{X}$ is now non-random, and $\hat{\underset{\sim}{E}}$ is the sum of a random and a non-random component, so the left hand side of (27) itself splits into a random and a non-random part. There is no reason why the non-random part should converge at all. For instance suppose that the situation of Example 2 has arisen by starting in Example 1 and conditioning on $\underset{\sim}{x}^n = x^n$, $n = 1, 2, \ldots$, as in the discussion after Example 2. Suppose the conditional distribution of $\underset{\sim}{e}^n$ does not depend on x^n.

Looking at (27), (18) and (14) we see that under the assumptions of Theorem 2, *unconditionally*, both parts of the left hand side of (27) converge in distribution to in general non-degenerate normal distributions. *Conditional* on $\underset{\sim}{x}^n = x^n$, $n = 1, 2, \ldots$ the random part still converges in distribution to a normal distribution with mean vector zero. The other part, now non-random, would have to converge to zero for (27) to be valid. But the probability must be zero that such x^n's have been realized, in view of the unconditional non-degenerate limiting normal distribution.

We finally turn to Example 3, supposing it has arisen by starting from Example 1 and conditioning on the non-missing predictor variables as described previously. If A1 holds, then with probability 1 after conditioning A2 holds too, and this implies that

$$N^{-1}\hat{X}^T\hat{\Sigma}^{-1}\hat{X} \to A \qquad \ldots (31)$$

($\hat{X}$ is now non-random). However the validity of A3 and A4 depends on the conditional distributions of $\underset{\sim}{x}_M^n$ and $\underset{\sim}{e}^n$ given $\underset{\sim}{x}_P^n = x_P^n$. Let us make the rather strong assumption (it implies for instance (29), and is itself implied by multivariate normality of $(\underset{\sim}{x}^n, \underset{\sim}{e}^n)$) that these are such that for *all* P and x_P^n

$$\begin{aligned} &\mathcal{E}(\hat{\underset{\sim}{e}}^n \mid \underset{\sim}{x}_P^n = x_P^n) = 0 \\ &\qquad\qquad\qquad\qquad \text{(c.f. (16))} \qquad\qquad \ldots (32) \\ &\mathcal{E}((\hat{\underset{\sim}{e}}^n)^2 \mid \underset{\sim}{x}_P^n = x_P^n) = \sigma_P^2 \end{aligned}$$

or in words, every regression on $\underset{\sim}{y}^n$ on a group of variables from $\underset{\sim}{x}^n$ is linear and homoscedastic. Looking at (20), this now implies that

$$\begin{aligned} &\mathcal{E}(\hat{\underset{\sim}{\beta}} \mid \hat{\underset{\sim}{X}} = \hat{X}) = \beta \\ &\mathcal{E}((\hat{\underset{\sim}{\beta}}-\beta)(\hat{\underset{\sim}{\beta}}-\beta)^T \mid \hat{\underset{\sim}{X}} = \hat{X}) = (\hat{X}^T\hat{\Sigma}^{-1}\hat{X})^{-1}. \qquad \ldots (33) \end{aligned}$$

By (31) and (33), $\hat{\underset{\sim}{\beta}}$ is consistent; but more importantly, (33) gives new motivation for using $(\hat{X}^T\hat{\Sigma}^{-1}\hat{X})^{-1}$, with α_{MP} and σ_P^2 replaced by estimates, as an approximate covariance matrix for $\hat{\underset{\sim}{\beta}}$. Does such a simple argument also give asymptotic normality of $N^{\frac{1}{2}}(\hat{\underset{\sim}{\beta}}-\beta)$?

By (25), what is needed is a central limit theorem for $N^{-\frac{1}{2}}\hat{X}^T\hat{\Sigma}^{-1}\hat{E}$ which by (32) can be written as a sum of N independent zero mean random vectors. Moreover, by (31) the covariance matrix of this sum converges to A. So we need only add a Lindebergh-type condition ensuring that each term in the sum is asymptotically negligeable in order to guarantee asymptotic normality of $\hat{\underset{\sim}{\beta}}$. This condition is going to involve the conditional distributions of

$\hat{\underline{e}}^n$ given $\underline{x}_P^n = x_P^n$, which could depend on P^n and x_P^n in a very complicated way. For simplicity we might assume them only to depend on P^n; we have already assumed this for the conditional expectations and variances. However this is rather close to assuming multivariate normality of $\underline{x}^n$ as the following special case, $K = 2$, shows. The new assumption is equivalent to assuming that $\underline{x}_P^n$ and $\underline{e}^n + \beta_M^{\mathrm{T}}(\underline{x}_M^n - \alpha_{MP}\underline{x}_P^n)$ are independent for each P. Taking $P = \{1, 2\}$ and $M = \phi$, $\underline{e}^n$ and $\underline{x}^n$ are independent; taking $P = \{1\}$ and then $P = \{2\}$ we find that $\underline{x}_1^n$ is independent of $\beta_2(\underline{x}_2^n - \alpha_{21}\underline{x}_1^n)$ and $\underline{x}_2$ of $\beta_1(\underline{x}_1^n - \alpha_{12}\underline{x}_2^n)$. By the theorem of Skitovich (see Kagan, Linnik and Rao (1973) Theorem 3.1.1) it now follows that if all the coefficients involved here are non-zero then $\underline{x}^n$ is bivariate normally distributed.

Of course, if $(\underline{x}^n, \underline{e}^n)$ is multivariate normally distributed, then conditional on $\hat{\underline{X}} = \hat{X}, \hat{\beta}$ has the $\mathcal{N}(\beta, (\hat{X}^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{X})^{-1})$ distribution and the required results can be obtained very easily.

Conclusion

Though under reasonable conditions the estimators considered are consistent—it is not even necessary to assume the covariates are random—fairly strong conditions are needed to justify the use of $(\hat{\underline{X}}^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{\underline{X}})^{-1}$ as an approximate covariance matrix for $\hat{\underline{\beta}}$: namely randomness of the covariates, independence between the N observations, and uncorrelatedness of $(\hat{\underline{e}}^n)^2$ and $\underline{x}_P^n \underline{x}_P^{n\mathrm{T}}$. It is worth pointing out that small sample simulation results in the literature are nearly always based on a multivariate normal distribution for $(\underline{x}^n, \underline{e}^n)$; see for instance Little (1979).

References

Afifi, A. A. and Elashoff R. M. (1966) : Missing observations in multivariate statistics—I. Review of the literature. *J. Amer. Statist. Ass.*, **61**, 595-604.

———— (1967) : Missing observations in multivariate statistics—II. Point estimation in simple linear regression. *J. Amer. Statist. Ass.*, **62**, 10-29.

———— (1969a) : Missing observations in multivariate statistics—III. Large sample analysis of simple linear regression. *J. Amer. Statist. Ass.*, **64**, 337-358.

———— (1969b) : Missing observations in multivariate statistics—IV. A note on simple linear regression. *J. Amer. Statist. Ass.*, **64**, 359-365.

Beale, E. M. L. and Little R. J. (1975) : Missing values in multivariate analysis. *J. R. Statist. Soc.*, (B) **37**, 129-145.

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

Dempster, A. P., N. M. Laird and D. B. Rubin (1977) : Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc.* (B), **39**, 1-38.

GILL, R. D. (1977): Consistency of maximum likelihood estimators of the factor analysis model, when the observations are not multivariate normally distributed. *Recent Developments in Statistics*, J. R. Barra et al. (eds.), North-Holland, Amsterdam.

HARTLEY, H. O. and R. R. HOCKING (1971): The analysis of incomplete data. *Biometrics*, **27**, 783-823.

KAGAN, A. M., Y. V. LINNIK and C. R. RAO (1973): *Characterization Problems in Mathematical Statistics*, Wiley, New York.

LITTLE, R. J. A. (1979): Maximum likelihood inference for multiple regression with missing values: a simulation study. *J. Roy Statist. Soc.*, (B) **41**, 76-87.

Paper received : March, 1982.

Revised : December, 1983.