stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK
(DEPARTMENT OF MATHEMATICAL STATISTICS)

SW 65/79 JANUARI

R.D. GILL

A NOTE ON SOME METHODS FOR REGRESSION ANALYSIS WITH INCOMPLETE OBSERVATIONS

Preprint

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O).

AMS(MOS) subject classification scheme 1970 62J05

A note on some methods for regression analysis with incomplete observations*)

by

R.D. Gill

ABSTRACT

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 consequences of assuming a random or a fixed specification for the covariates are also investigated.

KEYWORDS & PHRASES: *incomplete observations, missing data, regression analysis*

---

*) This report will be submitted for publication elsewhere.

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 & ELASHOFF (1966), (1967), (1969a) and (1969b), HARTLEY & HOCKING (1971) and DEMPSTER, LAIRD & 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 & 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). 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. As will become apparent later, the predictor variables are considered as "covariates" rather than as the "design variables" of a planned experiment. We consider both models with "random" and with "fixed" predictor variables. N is supposed large, making asymptotic results relevant; and missing values occur on a large scale, so that just dropping incomplete observations is an unacceptable throwing away of information. A final point is that we do not want to make strong distributional assumptions, such as that of multivariate normality, about the predictor variables, if indeed we assume them random at all.

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).

We are interested in the problems of finding reasonable and non-technical conditions under which (i) the proposals yield consistent estimators, and (ii), suitably normed, the 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. However in (ii), though asymptotic normality is easily proved under certain somewhat restrictive conditions, the asymptotic covariance matrix can often not be the one we want.

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^T denotes the transpose of the vector a. We specify a model for N observations for each N = 1,2,...; 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 $\to_p$ and $\to_{\mathcal{D}}$ 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 $N(\cdot,\cdot)$.

Let P and M (a *pattern* of observed predictor variables and its

complement of *missing* ones) denote sets of indices such that $P \cup M = \{1,\ldots K\}$ (where K is the number of predictor variables), $P \cap M = \phi$, $P \neq \phi$ 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×1 vector and Σ a K×K matrix then

$$\beta = \begin{pmatrix} \beta_P \\ \beta_M \end{pmatrix}, \tag{1}$$

$$\Sigma = \begin{pmatrix} \Sigma_{PP} & \Sigma_{PM} \\ \Sigma_{MP} & \Sigma_{MM} \end{pmatrix} = \begin{pmatrix} \Sigma_{\cdot P} & \Sigma_{\cdot M} \end{pmatrix}.$$

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

$$\underline{y}^n = \beta^T \underline{x}^n + \underline{e}^n \tag{2}$$

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

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

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.

$$\underline{x}_1^n = 1 \quad \text{almost surely for each } n. \tag{4}$$

We make the following 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×K symmetric positive definite matrix. Suppose

that for each P, the following convergences hold as $N \to \infty$:

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

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

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

A4 $\quad N^{-1} \sum_{n:P^n=P} \underline{x}_P^n e^n \to_P 0$

A5 $\quad A = \sum_P \rho_P \sigma_P^{-2} \Sigma_{\cdot P} \Sigma_{PP}^{-1} \Sigma_{P\cdot}$ is non-singular (σ_P^2 is defined in (8) below)

Interpreting Σ as a limiting average value of $E(\underline{x}^n \underline{x}^{nT})$, 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*. Suppose $(\underline{x}^n, \underline{e}^n)$, $n = 1,\ldots,N$ are independent over n and have the same distribution for all n and N. (This clearly doesn't exclude (4) from holding).
Then A2, A3 and A4 are consequences of A1, (3) and

(5) $\quad E(\underline{x}^n \underline{x}^{nT}) = \Sigma,$

which can be proved by applying the Weak Law of Large Numbers (with special attention for P such that $\rho_P = 0$ since we do not assume in such a case that $\#\{n;P^n=P\} \to \infty$ or is bounded. The proof on page 17 of GILL (1978) can be adapted for this situation).

EXAMPLE 2 *Fixed predictor variables*. For some vectors x^n,

(6) $\quad \underline{x}^n = x^n$ almost surely for each n.

A4 is now a consequence of A2, because

$$E\left(N^{-1}\sum_{n:P^n=P}\underline{x}_P^n\underline{e}^n\right)=0$$

$$E\left(N^{-1}\sum_{n:P^n=P}\underline{x}_P^n\underline{e}^n\right)^2=E\left(N^{-1}\sum_{n:P^n=P}x_P^n e^n\right)^2=$$

$$=N^{-1}\sigma^2N^{-1}\sum_{n:P^n=P}x_P^n\,x_P^{n\top}\to 0.$$

However A2 and A3 have to be explicitly assumed.

EXAMPLE 3 *Conditional models*. If in Example 1, $(\underline{x}^n,\underline{e}^n,P^n)$, n = 1,...,N are the first N of a single infinite sequence, then the convergences in probability in A2, A3 and A4 are in fact almost sure convergences. So almost surely, after conditioning on $\underline{x}^n = x^n$, n = 1,2,..., A2, A3 and A4 remain valid and we have a special case of Example 2. Of course (3) is not necessarily valid if the expectations there are replaced by conditional expectations. In the next section we pay attention to the similar case where we only condition on $\underline{x}_P^n = x_P^n$, n = 1,2,..., which is interesting because it is in a way close to the spirit of the proposed estimators.

To define these estimators let us first work as if the parameters needed for the proposals (certain functions of σ^2, β and Σ) were known. Define

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

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

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

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

(11) $\hat{\Sigma}$ = the N×N diagonal matrix with diagonal elements $\sigma^2_{P^n}$, n=1,...,N

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

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

If α_{MP} and σ_P^2 were known, $\hat{\underline{\beta}}$ 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

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

we rewrite (2) as

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

where *in Example 1*

$$\begin{aligned} &E(\hat{\underline{e}}^n) = 0 \\ &E(\hat{\underline{x}}^n\hat{\underline{e}}^{n'}) = 0 \qquad\qquad \text{(c.f. (3))} \\ &E(\hat{\underline{e}}^n\hat{\underline{e}}^{n'}) = (\hat{\Sigma})_{nn'} \end{aligned} \tag{16}$$

After conditioning on $\underline{x}_P^n = x_P^n$, n = 1,...,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* $N(\hat{\underline{X}}^T\hat{\Sigma}^{-1}\hat{\underline{X}})^{-1}$ *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 with consistent estimators* $\hat{\underline{\alpha}}_{MP}$ *and* $\hat{\underline{\sigma}}_P^2$ *of the same quantities.*

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

$$N^{-1}\hat{\underline{X}}^T\hat{\Sigma}^{-1}\hat{\underline{X}} = \sum_P \sigma_P^{-2}\Sigma_{\cdot P}\Sigma_{PP}^{-1}\left(N^{-1}\sum_{n:P^n=P}\underline{x}_P^n\underline{x}_P^{nT}\right)\Sigma_{PP}^{-1}\Sigma_{P\cdot}.$$

$$\to_P \sum_P \sigma_P^{-2}\Sigma_{\cdot P}\Sigma_{PP}^{-1}\rho_P\Sigma_{PP}\Sigma_{PP}^{-1}\Sigma_{P\cdot} = A .$$

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

$$(17) \qquad N(\hat{\underline{X}}^T\hat{\Sigma}^{-1}\hat{\underline{X}})^{-1} \to_p A^{-1}.$$

Next defining

$$(18) \qquad \hat{\underline{E}} = (\hat{\underline{e}}^1,\ldots,\hat{\underline{e}}^N)^T$$

we can rewrite (15) as

$$(19) \qquad \underline{Y} = \hat{\underline{X}}\beta + \hat{\underline{E}}$$

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

$$(20) \qquad \hat{\underline{\beta}} = \beta + N(\hat{\underline{X}}^T\hat{\Sigma}^{-1}\hat{\underline{X}})^{-1}\, N^{-1}\hat{\underline{X}}^T\hat{\Sigma}^{-1}\hat{\underline{E}}\ .$$

So to prove

$$(21) \qquad \hat{\underline{\beta}} \to_p \beta$$

it suffices to establish

$$(22) \qquad N^{-1}\hat{\underline{X}}^T\hat{\Sigma}^{-1}\hat{\underline{E}} \to_p 0 \quad \text{as } N \to \infty.$$

Now

$$N^{-1}\hat{\underline{X}}^T\hat{\Sigma}^{-1}\hat{\underline{E}} = \sum_P \sigma_P^{-2}\Sigma_{\cdot P}\Sigma_{PP}^{-1}\left(N^{-1}\sum_{n:P^n=P}\underline{x}_P^n\hat{\underline{e}}^n\right) =$$

$$(23) \qquad = \sum_P \sigma_P^{-2}\Sigma_{\cdot P}\Sigma_{PP}^{-1}\left(N^{-1}\sum_{n:P^n=P}(\underline{x}_P^n\underline{e}^n + (\underline{x}_P^n\underline{x}_M^{nT} - \underline{x}_P^n\underline{x}_P^{nT}\alpha_{MP}^T)\beta_M)\right)$$

$$\to_p 0$$

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 & 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 normality 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 $$N^{-1} \sum_{n:P^n=P} (\underline{e}^n)^2 \to_p \rho_P \sigma^2 \quad \text{for all P.}$$

We have not yet investigated the other methods in this respect though a similar approach should be applicable. The proof of Theorem 1 actually also shows consistency of BEALE & 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* $\underline{y}^n$ 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

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

then we find that in general $E\underline{x}_P^n \hat{\underline{e}}^n \neq 0$ and so it does not hold that $N^{-1} \Sigma_{n:P^n=P} \underline{x}_P^n \hat{\underline{e}}^n \to_p 0$ if $\rho_P > 0$. This fact makes another of BEALE & LITTLE's (1975) proposals (see their section 6) rather difficult to motivate, though this proposal is made in a different context to ours.

3. PROBLEM (ii): ASYMPTOTIC NORMALITY WITH CORRECT COVARIANCE MATRIX

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

(24) $$N^{\frac{1}{2}}(\hat{\underline{\beta}}-\beta) \to_D N(0,A^{-1})$$

if and only if

(25) $$N^{-\frac{1}{2}} \hat{\underline{X}}^T \hat{\Sigma}^{-1} \hat{\underline{E}} \to_D N(0,A).$$

If (17) holds too we can indeed validly use $(\hat{\underline{X}}^T \hat{\Sigma}^{-1} \hat{\underline{X}})^{-1}$ as an asymptotic

covariance matrix for $\hat{\underline{\beta}}$ 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. *Suppose* $(\underline{x}^n, \underline{e}^n)$, n = 1,...,N *are independent over* n *and have the same distribution for all* n *and* N. *Suppose* A1 *to* A5 *hold and that furthermore*

$$
\begin{aligned}
&E(\underline{x}^n \underline{x}^{nT}) = \Sigma \\
&E(\underline{x}^n \underline{e}^n) = 0 \\
&E((\underline{e}^n)^2) = \sigma^2 \\
&E((\hat{\underline{e}}^n)^2 \underline{x}_P^n \underline{x}_P^{nT}) = \Psi_P
\end{aligned}
\tag{26}
$$

for some finite matrices Ψ_P. *Then*

$$
N^{-\frac{1}{2}} \hat{\underline{X}}^T \hat{\Sigma}^{-1} \hat{\underline{E}} \to_D N(0,B) \tag{27}
$$

where

$$
B = \sum_P \rho_P \, \sigma_P^{-4} \Sigma_{\cdot P} \Sigma_{PP}^{-1} \Psi_P \Sigma_{PP}^{-1} \Sigma_{P\cdot}. \tag{28}
$$

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

$$
\Psi_P = \sigma_P^2 \Sigma_{PP} \quad \text{for all P.} \tag{29}
$$

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 $\rho_P = 0$),

$$
N^{-\frac{1}{2}} \hat{\underline{X}} \, \hat{\Sigma}^{-1} \hat{\underline{E}} \to_D N(0,B) \tag{30}
$$

Obviously if (29) holds, then A = B. □

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 $(\underline{\hat{e}}^n)^2$ uncorrelated with $\underline{x}_P^n \underline{x}_P^{nT}$, at least, provided α_{MP} and σ_P^2 may be replaced with consistent estimators $\underline{\hat{\alpha}}_{MP}$ and $\underline{\hat{\sigma}}_P^2$. This turns out to be possible if one adds the assumption that $N^{\frac{1}{2}}(\underline{\hat{\alpha}}_{MP} - \alpha_{MP})$ is bounded in probability as $N \to \infty$.

Of course one could often be reluctant to assume that the $\underline{x}^n$'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) $\underline{\hat{X}} = \hat{X}$ is now non-random, and $\underline{\hat{E}}$ is the sum of a random and a non-random component. There is no reason why the non-random contribution to the left hand side of (27) should converge at all. For instance suppose (as in the first part of Example 3) that we wish to work conditionally on $\underline{x}^n = x^n$, $n = 1,2,\ldots$, arising from an infinite sequence of independent and identically distributed $(\underline{x}^n, \underline{e}^n)$'s, where furthermore $\underline{x}^n$ is independent of $\underline{e}^n$. 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 $\underline{x}^n = x^n$, $n = 1,2,\ldots$ 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.

However the other possibility of conditioning only on the observed values of the predictor variables $\underline{x}_P^n = x_P^n$, which was touched on in Example 3, fits rather nicely with the estimator $\underline{\hat{\beta}}$, at least if strong enough conditions are made. If A1 holds, then with probability 1 after conditioning A2 holds too, and this implies that

$$N^{-1}\hat{X}^T\hat{\Sigma}^{-1}\hat{X} \to A \tag{31}$$

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

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

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

$$
\begin{aligned}
&E(\hat{\underline{\beta}} \mid \hat{\underline{X}} = \hat{X}) = \beta \\
&E((\hat{\underline{\beta}}-\beta)(\hat{\underline{\beta}}-\beta)^T \mid \hat{\underline{X}} = \hat{X}) = (\hat{X}^T \hat{\Sigma}^{-1} \hat{X})^{-1}
\end{aligned}
\tag{33}
$$

By (31) and (33), $\hat{\underline{\beta}}$ 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 $\hat{\underline{\beta}}$. Does such a simple argument also give asymptotic normality of $N^{\frac{1}{2}}(\hat{\underline{\beta}}-\beta)$?

In the first place, proving a central limit theorem for $N^{-\frac{1}{2}}\hat{X}^T\hat{\Sigma}^{-1}\hat{\underline{E}}$ (which by (25) is what is needed) 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 $\underline{e}^n + \beta_M^T(\underline{x}_M^n - \alpha_{MP}\underline{x}_P^n)$ are independent for each P. Taking P = {1,2} and M = ϕ, $\underline{e}^n$ 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 & RAO (1973) Theorem 3.1.1) it now follows that if all the coefficients involved above 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 $\hat{\underline{X}} = \hat{X}$, $\hat{\underline{\beta}}$ has the $N(\beta, (\hat{X}^T\hat{\Sigma}^{-1}\hat{X})^{-1})$ distribution and there is little need for asymptotic results.

4. 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 $(\hat{\underline{X}}^T\hat{\Sigma}^{-1}\hat{\underline{X}})^{-1}$ as an approximate covariance matrix for $\hat{\underline{\beta}}$: namely randomness of the covariates, indepence between the N observations, and uncorrelatedness of $(\hat{\underline{e}}^n)^2$ and $\underline{x}_p^n\underline{x}_p^{nT}$. It is worth pointing out that small sample simulation results in the literature are nearly always based on a multivariate normal distribution for $(\underline{x}^n,\underline{e}^n)$.

REFERENCES

[1] AFIFI, A.A. & R.M. ELASHOFF (1966), *Missing observations in multivariate statistics - I. Review of the literature*, J. Amer. Statist. Ass. 61, 595-604.

[2] AFIFI, A.A. & R.M. ELASHOFF, (1967), *Missing observations in multivariate statistics - II. Point estimation in simple linear regression*, J. Amer. Statist. Ass. 62, 10-29.

[3] AFIFI, A.A. & R.M. ELASHOFF (1969a), *Missing observations in multivariate statistics - III. Large sample analysis of simple linear regression*, J. Amer. Statist. Ass. 64, 337-358.

[4] AFIFI, A.A. & R.M. ELASHOFF (1969b), *Missing observations in multivariate statistics - IV. A note on simple linear regression*, J. Amer. Statist. Ass. 64, 359-365.

[5] BEALE, E.M.L. & R.J. LITTLE (1975), *Missing values in multivariate analysis*, J.R. Statist. Soc. (B) 37, 129-145.

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

[7] DEMPSTER, A.P., N.M. LAIRD & D.B. RUBIN (1977), *Maximum likelihood from incomplete data via the EM algorithm*, J.R. Statist. Soc. (B) 39, 1-38.

[8] 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 Statictics. J.R. Barra et al. (eds.), North-Holland, Amsterdam.

[9] GILL, R.D. (1978), *Regression analysis for mixed cross-section and time-series data with reference to some "incomplete observations" techniques*, Report SW61, Mathematisch Centrum, Amsterdam.

[10] HARTLEY, H.O. & R.R. HOCKING (1971), *The analysis of incomplete data*, Biometrics 27, 783-823.

[11] KAGAN, A.M., Y.V. LINNIK & C.R. RAO (1973), *Characterization problems in mathematical statistics*, Wiley, New York.