stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK (DEPARTMENT OF MATHEMATICAL STATISTICS) SW 61/78 NOVEMBER

R.D. GILL

REGRESSION ANALYSIS FOR MIXED CROSS-SECTION AND TIME-SERIES DATA WITH REFERENCE TO SOME "INCOMPLETE OBSERVATIONS" TECHNIQUES

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): 62P20, 62H99

Regression analysis for mixed cross-section and time-series data with reference to some "incomplete observations" techniques*)

by

R.D. Gill

ABSTRACT

An iterative method is proposed for estimating a certain regression model with mixed cross-section and time-series data, where each observational unit is not necessarily available at each time point of the time series. We prove consistency and asymptotic normality of the estimators of the regression coefficients as the size of the cross-section increases while the length of the time series remains bounded. We discuss both other models and the connection between our method and various "incomplete data" techniques from the literature.

KEY WORDS & PHRASES: *Mixed cross-section and time-series data, econometric models, incomplete observations, missing data, regression analysis, multivariate analysis.*

*) This report will be submitted for publication elsewhere.

1. INTRODUCTION

This report describes theoretical work*) carried out in connection with the consultation project "Basisonderzoek Kostenstructuur Ziekenhuizen" (Basis research on hospital cost-structure) commissioned by the Nationaal Ziekenhuis Instituut (Dutch national hospitals institute). The model described here arose from the needs of this particular project and its application is described in the N.Z.I. report VAN AERT and VAN MONTFORT [1979] and in VAN MONTFORT [1979]. Regression analyses had been carried out to investigate the relations between various hospital attributes such as size, degree of specialization, institutional form, case-mix, etc. as independent or *predictor* variables and deflated total costs or alternatively cost per case of a hospital as dependent or *criterion* variable, using data from a large number of Dutch hospitals (considered as independent observations) pertaining to each of the years 1968 to 1973; i.e. one regression analysis was done for the data of each of these six years. The need then arose to combine these analyses recognizing that (a) the hospitals taking part in one year's survey did not necessarily take part in another year's; (b) the effect of some predictor variables might vary over the years; in fact we wanted to discover such changes; (c) the disturbance term in each year's regression equation which, if the model is realistic, should reflect individual hospital effects such as special conditions, variation in efficiency, etc., will be highly correlated per hospital over the different years; (d) the combined model should not be in contradiction with the separate year models; and (e) none of the particular models which have been proposed in the literature for such data, such as the variance components or the random effects model, or the first order autocorrelation model, seemed a priori reasonable though it would be interesting to check how well they fit.

So much for the practical background. The plan of the rest of the paper is as follows: in the next section we describe our model which seems to be new in the econometric literature - perhaps because most time series are longer than ours! - and motivate and discuss the estimation and testing procedures used. Section 3 contains the proofs of asymptotic properties of our

*) A FORTRAN program implementing the method described here is also available

estimators (we only have heuristic justification for our testing procedures). In the fourth and last section we discuss some other possible approaches. In particular we note that our problem can be considered as one in the multivariate analysis of incomplete data - the data of some years being missing for some of the hospitals - and point out the connections with certain "missing data" methods. The various steps of our iterative estimation method are adaptations of some missing data techniques from the literature, so our theorems can be modified to give new results on consistency and asymptotic normality for these too, without assuming multivariate normality of the observations: an assumption we avoid making as far as possible.

2. DESCRIPTION OF THE MODEL AND ESTIMATORS

First we introduce some notation. Random variables are underlined, and E denotes the expectation operator. Indices $n = 1,2,\ldots,N$ refer to the N observations; j (or j') $= 1,2,\ldots,J$ refer to the J time points for which at least for some observations data is available; and k (or k') $= 1,2,\ldots,K$ refer to the K predictor variables. J and K are fixed, but the model is supposed to be specified for each $N = 1,2,\ldots$ as we will be interested in asymptotic properties of our estimators as N tends to infinity. We can now specify our

MODEL. *For* $n = 1,\ldots,N$ *let* $P^{(n)} = (P_j^{(n)})$ *be a* $j\times 1$*-vector with* $P_j^{(n)} = 1$ *or* 0 *for each* j, *and such that for each* n *a* j *exists with* $P_j^{(n)} = 1$, *and vice-versa. For* $n = 1,\ldots,N$, $j = 1,\ldots,J$ *such that* $P_j^{(n)} = 1$, *and for* $k = 1,\ldots,K$ *let* x_{njk} *be real numbers. For* $n = 1,\ldots,N$, $j = 1,\ldots,J$ *such that* $P_j^{(n)} = 1$ *suppose*

$$(1) \qquad \underline{y}_{nj} = \sum_k x_{njk}\beta_k + \underline{e}_{nj}$$

where

$$(2) \qquad E(\underline{e}_{nj}) = 0$$

and

$$(3) \qquad E(\underline{e}_{nj}\underline{e}_{n'j'}) = \begin{cases} 0 & n \neq n' \\ \sigma_{jj'} & n = n' \end{cases}$$

where $\beta = (\beta_k)$ *is a fixed* K×1*-vector and* $\Sigma = (\sigma_{jj'})$ *a fixed positive definite symmetric* K×K*-matrix.*

So $P^{(n)}$ denotes the *pattern* of available and missing time points for the n^{th} observation. "Fixed" means in the above specification "not depending on N". *All other quantities* may vary with N but we generally suppress this dependence in our notation. The symbols $\to_P$ and $\to_{\mathcal{D}}$ denote convergence in probability and convergence in distribution *as* N *tends to infinity*. For a sequence of random variables $\underline{X}_N$ we write $\underline{X}_N = O_P(1)$ if $\underline{X}_N$ *is bounded in probability as* $N \to \infty$: i.e. for all $\varepsilon > 0$ there exist finite C and N_0, such that $N \geq N_0$ implies $P(|\underline{X}_N| > C) < \varepsilon$. The x_{njk}'s are the fixed values taken by our K predictor variables, $\underline{y}_{nj}$ is our criterion variable; β and Σ are unknown parameters. We observe x_{njk} and $\underline{y}_{nj}$ for those n and j such that $P_j^{(n)} = 1$. For notational convenience define $x_{njk} = 0$ for n and j such that $P_j^{(n)} = 0$ and suppose that for *all* n and j random variables $\underline{e}_{nj}$ exist satisfying (1), (2) and (3).

We shall make more assumptions (mostly of a technical nature) as needed, in the following section. For the time being note that (i) the model is indeed not in contradiction with the "separate year" analyses; (ii) the model has as special cases (whose estimation is discussed in Section 4) (a) the "Random effects model" - $\underline{e}_{nj} = \underline{\varepsilon}_n + \underline{\delta}_{nj}$, where the $\underline{\varepsilon}_n$'s and $\underline{\delta}_{nj}$'s are all independent normally distributed with zero means and variances σ_ε^2 and σ_δ^2 respectively, so that consequently

$$\sigma_{jj'} = \begin{cases} \sigma_\varepsilon^2 & j \neq j' \\ \sigma_\varepsilon^2 + \sigma_\delta^2 & j = j' \end{cases}$$

- and (b) the "First order autocorrelation model" - $\underline{e}_{nj} = \rho \underline{e}_{n,j-1} + \underline{\varepsilon}_{nj}$, where the $\underline{\varepsilon}_{nj}$'s and the $\underline{e}_{n0}$'s are independent and normally distributed with zero means, and variances such that the $\underline{e}_{nj}$'s have constant variance σ_e^2, so that consequently

$$\sigma_{jj'} = \rho^{|j-j'|}\sigma_e^2$$

- while (iii) to let the effect of (some of) the predictor variables depend on time we might replace β_k in (1) with β_{jk} and rewrite (1) as

$$\underline{y}_{nj} = \sum_{j',k} (x_{njk}\chi_{jj'})\beta_{j'k} + \underline{e}_{nj}$$

where $\chi_{jj'} = 1$ if $j = j'$ and 0 otherwise; i.e. the same model with now JK predictor variables. Alternatively if we can assume that the effect of time is e.g. linear we can replace $x_{njk}\beta_k$ in (1) with $x_{njk}\beta_{0k} + (x_{njk}\cdot j)\beta_{1k}$ and again find ourselves back with the same model, but with more predictor variables.

We propose an iterative method to estimate the regression coefficients β and the covariance matrix of the disturbances Σ, which we now present informally:

STEP 1. Estimate β by ordinary least squares (i.e. as if $\sigma_{jj} = \sigma^2$ for some $\sigma^2 > 0$ for each j, and $\sigma_{jj'} = 0$ for $j \neq j'$).

STEP 2. Estimate Σ from the residuals of step 1 by adding the product of the residuals for the time instants j and j' over n such that $P_j^{(n)}$ and $P_{j'}^{(n)} = 1$ and dividing by the number of such n to get an estimate of $\sigma_{jj'}$.

STEP 2r+1 (r = 1,2,...). With the estimate of Σ obtained from step 2r, re-estimate β by the method of generalized least squares (i.e. as if the estimate were the true value of Σ).

STEP 2r+2 (r = 1,2,...). With the estimate of β from step 2r+1 and the estimate of Σ from step 2r, construct a new estimate of Σ by (a) calculating the residuals - from now on we behave as if these residuals were the realized error terms and the estimate of Σ were its true value -, (b) using these to predict by least squares the error terms $\underline{e}_{nj}$ for those n and j such that $P_j^{(n)} = 0$, and (c) estimating Σ in the obvious way from the now "completed" set of error terms, except for a correction based on the old estimate of Σ which is added to a summand in the sums of squares or products or errors, c.q. predicted errors, whenever the product consists of two predicted errors. The correction term is the (estimate of) the partial covariance of the two errors which have been predicted given those on which the predictions are based.

To explain this let $\underline{e} = (\underline{e}_P', \underline{e}_M')'$ be a J×1 random vector (' denotes

transpose) partitioned into two according to a pattern of observed components P and its complement of missing ones M; so P + M is a J×1 vector of ones. Let us suppose

$$E(\underline{e}) = 0, \quad E(\underline{e}\underline{e}') = \Sigma = \begin{pmatrix} \Sigma_{PP} : \Sigma_{PM} \\ \Sigma_{MP} : \Sigma_{MM} \end{pmatrix}$$

where Σ is positive definite and partioned conform $\underline{e}$ itself. Then the linear least squares predictor of $\underline{e}_M$ given $\underline{e}_P$ is

$$\hat{E}(\underline{e}_M|\underline{e}_P) = \Sigma_{MP}\Sigma_{PP}^{-1}\underline{e}_P$$

which has the covariance matrix

$$E(\hat{E}(\underline{e}_M|\underline{e}_P)\hat{E}(\underline{e}_M|\underline{e}_P)') = \Sigma_{MP}\Sigma_{PP}^{-1}\Sigma_{PM}.$$

So

$$E(\underline{e}_M\underline{e}_M') = E(\hat{E}(\underline{e}_M|\underline{e}_P)\hat{E}(\underline{e}_M|\underline{e}_P)') + (\Sigma_{MM} - \Sigma_{MP}\Sigma_{PP}^{-1}\Sigma_{PM})$$

where the last term, also equal to the covariance matrix of $\underline{e}_M - \hat{E}(\underline{e}_M|\underline{e}_P)$, is conventionally called the partial covariance matrix of $\underline{e}_M$ given $\underline{e}_P$. On the other hand

$$E(\underline{e}_M\underline{e}_P') = E(\hat{E}(\underline{e}_M|\underline{e}_P)\underline{e}_P') = \Sigma_{MP}.$$

These estimators are defined formally in the corollary to theorems 1 to 4 (which deal in turn with steps 1, 2, 2r+1 and 2r+2 above). The estimator of step 2 is essentially GLASSER's [1964] method for estimating a covariance matrix with incomplete observations, while that of step 2r+2 is derived from the maximum likelihood method (assuming normality) developed by ORCHARD and WOODBURY [1972], BEALE and LITTLE [1975], and DEMPSTER *et al.* [1977].

It will be shown in section 3 that *under suitable conditions* these rules do (asymptotically) define estimators: i.e., as $N \to \infty$, the probability tends to one that the matrices which have to be inverted are nonsingular.

Also, denoting the estimator of β from step 2r+1 as $\underline{b}^{(r)}$ (r = 0,1,...) and that of Σ from step 2r+2 as $\underline{S}^{(r)}$ (r = 0,1,...), then $\underline{b}^{(r)}$ and $\underline{S}^{(r)}$ are consistent (as $N \to \infty$) for each $r \geq 0$: while for $r \geq 1$ $\underline{b}^{(r)}$ is asymptotically equivalent to the true generalized least squares estimator of β. We prove asymptotic normality of $N^{\frac{1}{2}}(\underline{b}^{(r)} - \beta)$ for each $r \geq 1$, with asymptotic mean zero and a (fixed) asymptotic covariance matrix which can be consistently estimated as $N \to \infty$. We also show that under multivariate normality of the $\underline{e}_{nj}$'s, $\underline{b}^{(r)}$ is an efficient estimator for β for each $r \geq 1$ and (this is the reason for iterating past r = 1) *if* $\underline{b}^{(r)}$ and $\underline{S}^{(r)}$ converge to say $\underline{b}$ and $\underline{S}$ as $r \to \infty$, then b and S are stationary points of the likelihood function $\underline{\ell}(\beta,\Sigma)$ for β and Σ given the data. In any case $\ell(\underline{b}^{(r)},\underline{S}^{(r)})$ is nondecreasing in r. So rough tests of hypotheses of interest may be carried out by assuming that the usual asymptotic maximum likelihood theory applies, and treating the likelihood at $\beta = \underline{b}^{(r)}$ and $\Sigma = \underline{S}^{(r)}$ for large r as the true maximum likelihood, if $\underline{b}^{(r)}$ and $\underline{S}^{(r)}$ appear to converge as $r \to \infty$.

We have not been able to prove anything on whether or not $\underline{b}^{(r)}$ and $\underline{S}^{(r)}$ converge, and have not tried to derive conditions under which the asymptotic maximum likelihood theory would hold. ORCHARD and WOODBURY [1972], BEALE and LITTLE [1975] and DEMPSTER *et al.* [1977] indicate that convergence is rather slow; our own practical experience with our method was that though the changes in $\underline{b}^{(r)}$, $\underline{S}^{(r)}$ and their likelihood under normality $\underline{\ell}(\underline{b}^{(r)},\underline{S}^{(r)})$ were not of practical significance for $r \geq 1$, convergence did not appear to be near even at the maximum number (r=30) of iterations we tried: this number being dictated by cost considerations as the number of independent variables K was rather large (25 or 30). In fact since our situation also had $N \approx 150$ and J = 6, and the assumption of normality could not be taken very seriously, such results would have been of doubtful relevance! However we still used likelihood theory to give rough tests using the likelihood evaluated at $\beta = \underline{b}^{(r)}$ and $\Sigma = \underline{S}^{(r)}$ for the last iteration.

In section 3 theorem 5 shows how the above method can be modified in a way which might be expected to speed up convergence, though the practical behaviour of the modified method was not much better.

A major assumption of the theorem is that for each pair of time instants j and j', the number of observations n for which $P_j^{(n)} = 1$ *and* $P_{j'}^{(n)} = 1$ tends to infinity as $N \to \infty$. This is obviously in general a necessary

condition for consistent estimation of Σ and hence for efficient estimation of β. If in a practical situation these numbers are not deemed large enough to justify the use of asymptotic theory, the only alternative would seem to be that of adopting a model such as one of the special cases mentioned above which involve less parameters.

The problem of how to look at residuals to check the assumption of multivariate normality or at least to look for serious outliers is a very difficult one in this situation and will also not be discussed here, though a number of procedures can be suggested and were applied.

3. ASYMPTOTIC RESULTS

The model and notation of the previous section is still supposed to hold throughout this one. In particular recall that dependence on N is generally suppressed, the only fixed quantities being J, K, β and Σ, and that we let $x_{njk} = 0$ when $P_j^{(n)} = 0$.

We also need the following notation. Define $r_n = \Sigma_j P_j^{(n)}$. Let X_n be the $r_n \times K$ matrix of elements x_{njk} such that $P_j^{(n)} = 1$, and similarly let $\underline{y}_n$ and $\underline{e}_n$ be the $r_n \times 1$ vectors of elements $\underline{y}_{nj}$ and $\underline{e}_{nj}$ respectively for which $P_j^{(n)} = 1$. Finally Σ_n is the $r_n \times r_n$ matrix of elements $\sigma_{jj'}$ of Σ for which $P_j^{(n)}$ and $P_{j'}^{(n)} = 1$.

Next we define

$$\tilde{X} = \begin{pmatrix} X_1 \\ \cdots \\ X_2 \\ \cdots \\ \vdots \\ \cdots \\ X_N \end{pmatrix} \qquad \tilde{\underline{y}} = \begin{pmatrix} \underline{y}_1 \\ \cdots \\ \underline{y}_2 \\ \cdots \\ \vdots \\ \cdots \\ \underline{y}_N \end{pmatrix} \qquad \tilde{\underline{e}} = \begin{pmatrix} \underline{e}_1 \\ \cdots \\ \underline{e}_2 \\ \cdots \\ \vdots \\ \cdots \\ \underline{e}_N \end{pmatrix} \qquad \text{and}$$

$$\tilde{\Sigma} = \begin{pmatrix} \Sigma_1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \Sigma_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \Sigma_N \end{pmatrix}$$

If $\underline{S}$ is some estimator of Σ, then $\underset{\sim}{\tilde{S}}$ is defined analogously. $\tilde{\Sigma}^{-1}$ and $\underset{\sim}{\tilde{S}}^{-1}$ denote $(\tilde{\Sigma})^{-1}$ and $(\underset{\sim}{\tilde{S}})^{-1}$ respectively. We always assume $\underline{S}$ is symmetric.

We can now write (1), (2) and (3) as

$$\underset{\sim}{\tilde{y}} = \tilde{X}\beta + \underset{\sim}{\tilde{e}} \tag{4}$$

$$E(\underset{\sim}{\tilde{e}}) = 0 \tag{5}$$

$$E(\underset{\sim}{\tilde{e}}\underset{\sim}{\tilde{e}}') = \tilde{\Sigma} \tag{6}$$

Let P denote an arbitrary pattern of missing and nonmissing time instants in one observation, i.e. a J×1 vector of zeros and ones. We shall need to refer to certain submatrices of Σ: for any P, $\Sigma_{PP} = (\sigma_{jj'})$ with j and j' restricted by $P_j = 1$ and $P_{j'} = 1$, $\Sigma_{\cdot P} = (\sigma_{jj'})$ with j' satisfying $P_{j'} = 1$ but j unrestricted, etc. If $\underline{S}$ is an estimator of Σ, $\underline{S}_{PP}$ and $\underline{S}_{\cdot P}$ are defined similarly. Σ_{PP}^{-1} means $(\Sigma_{PP})^{-1}$. So with this notation $\Sigma_n = \Sigma_{P(n)P(n)}$, n = 1,2,...,N.

Finally before stating our theorems we list the assumptions which will be made in some or all of them.

A1. For each P, j, j', k and k'

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n:P^{(n)}=P} x_{njk}x_{nj'k'} \qquad \text{exists (and is finite).}$$

A2. $$\lim_{N\to\infty} \frac{1}{N} \tilde{X}'\tilde{X}$$ (which exists if assumption A1 is made) is positive definite.

A3. $$\lim_{N\to\infty} \frac{1}{N} \tilde{X}'\tilde{\Sigma}^{-1}\tilde{X}$$ (which exists if assumption A1 is made) is positive definite.

A4. $(\underline{e}_{nj}; j = 1,...,J)$, n = 1,...,N are independent and, also over N = 1,2,..., identically distributed.

A5. For each j and j'

$$\lim_{N\to\infty} \sum_{n=1}^{N} P_j^{(n)} P_{j'}^{(n)} = \infty.$$

A6. for some constant $C < \infty$ not depending on N,

$$\sup_{n,j,k} |x_{njk}| \leq C.$$

A7. $\tilde{\underline{e}}$ is multivariate normally distributed.

To justify the assertions of existence in A2 and A3, note that for instance

$$(\tfrac{1}{N}\tilde{X}'\tilde{\Sigma}^{-1}\tilde{X})_{kk'} = \sum_{j,j',P} (\tfrac{1}{N} \sum_{n:P^{(n)}=P} x_{njk}x_{nj'k'})(\Sigma_{PP}^{-1})_{jj'}.$$

These assumptions are by no means the weakest under which our theorems remain valid (in particular A4 and A6 are stronger than necessary) but they are easy to state, lead to straightforward proofs, and are not prohibitively strong in a practical situation.

THEOREM 1. *Under* A1 *and* A2, $\tilde{X}'\tilde{X}$ *is for sufficiently large* N *nonsingular and defining*

$$\underline{b}^{(0)} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\underline{y}}, \tag{7}$$

$\underline{b}^{(0)}$ *is a consistent estimator of* β *and in fact* $N^{\frac{1}{2}}(\underline{b}^{(0)}-\beta)$ *is bounded in probability as* $N \to \infty$.

PROOF. Since $\frac{1}{N}(\tilde{X}'\tilde{X})$ converges to a nonsingular matrix, for N sufficiently large $\tilde{X}'\tilde{X}$ is nonsingular too. So by (4), for large enough N,

$$\underline{b}^{(0)} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\underline{y}} = \beta + (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\underline{e}}.$$

Therefore by (6)

$$E(N^{\frac{1}{2}}(\underline{b}^{(0)}-\beta)N^{\frac{1}{2}}(\underline{b}^{(0)}-\beta)') = (\tfrac{1}{N}\tilde{X}'\tilde{X})^{-1}(\tfrac{1}{N}\tilde{X}'\tilde{\tilde{\Sigma}}\tilde{X})(\tfrac{1}{N}\tilde{X}'\tilde{X})^{-1}$$

which converges to a finite matrix as $N \to \infty$, since $(\frac{1}{N}\tilde{X}'\tilde{X})^{-1}$ and $\frac{1}{N}\tilde{X}'\tilde{\tilde{\Sigma}}\tilde{X}$ converge to finite matrices. Applying Chebyshev's inequality we now easily see that $\underline{b}^{(0)} \to_p \beta$ as $N \to \infty$, i.e. $\underline{b}^{(0)}$ is a consistent estimator of β, and

in fact that $N^{\frac{1}{2}}(\underline{b}^{(0)}-\beta)$ is $O_P(1)$. □

THEOREM 2. *Suppose* A1, A4 *and* A5 *hold and that* $\underline{b}^{(0)}$ *is any consistent estimator of* β *such that* $N^{\frac{1}{2}}(\underline{b}^{(0)}-\beta) = O_P(1)$. *Then* $\underline{S}^{(0)}$ *defined by*

$$\underline{s}_{jj'}^{(0)} = \frac{1}{n_{jj'}} \sum_{n:P_j^{(n)}P_{j'}^{(n)}=1} (\underline{y}_{nj} - \sum_k x_{njk}\underline{b}_k^{(0)})(\underline{y}_{nj'} - \sum_{k'} x_{nj'k'}\underline{b}_{k'}^{(0)}) \tag{8}$$

where

$$n_{jj'} = \sum_n P_j^{(n)}P_{j'}^{(n)} \tag{9}$$

is a consistent estimator of Σ.

PROOF. Fixing j and j' and writing $\underline{y}_{nj} = \Sigma_k x_{njk}\beta_k + \underline{e}_{nj}$ and $\underline{b} = \underline{b}^{(0)}$, we see that

$$\begin{aligned}
s_{jj'}^{(0)} = {} & \frac{1}{n_{jj'}} \Big(\sum_{n:P_j^{(n)}P_{j'}^{(n)}=1} \underline{e}_{nj}\underline{e}_{nj'} \Big) \\
& - \sum_{k'} \frac{1}{n_{jj'}} \Big(\sum_{n:P_j^{(n)}P_{j'}^{(n)}=1} N^{-\frac{1}{2}}\underline{e}_{nj}x_{nj'k'} \Big) N^{\frac{1}{2}}(\underline{b}_{k'}-\beta_{k'}) \\
& - \sum_{k} \frac{1}{n_{jj'}} \Big(\sum_{n:P_j^{(n)}P_{j'}^{(n)}=1} N^{-\frac{1}{2}}\underline{e}_{nj'}x_{njk} \Big) N^{\frac{1}{2}}(\underline{b}_{k}-\beta_{k}) \\
& + \sum_k \sum_{k'} \frac{1}{n_{jj'}} \Big(\frac{1}{N} \sum_{n:P_j^{(n)}P_{j'}^{(n)}=1} x_{njk}x_{nj'k'} \Big) N^{\frac{1}{2}}(\underline{b}_k-\beta_k) N^{\frac{1}{2}}(\underline{b}_k-\beta_{k'}) \\
= {} & \underline{A}_N - \underline{B}_N - \underline{C}_N + \underline{D}_N \qquad \text{(say)}.
\end{aligned}$$

Now by A4 and A5 and the weak law of large numbers $\underline{A}_N \to_P \sigma_{jj'}$ as $N \to \infty$. Since $N^{\frac{1}{2}}(\underline{b}-\beta)$ is $O_P(1)$ as $N \to \infty$ and $n_{jj'} \to \infty$ as $N \to \infty$, to show that $\underline{B}_N \to_P 0$ it suffices to show that $\Sigma_{n:P_j^{(n)}P_{j'}^{(n)}=1} N^{-\frac{1}{2}} \underline{e}_{nj}x_{nj'k'}$ is $O_P(1)$. Now

$$E\Big(\Big(\sum_{n:P_j^{(n)}P_{j'}^{(n)}=1} N^{-\frac{1}{2}} e_{nj} x_{nj'k'}\Big)^2\Big) =$$

$$= \sum_{P:P_jP_{j'}=1} \frac{1}{N} \sum_{n:P^{(n)}=P} x_{nj'k'} x_{nj'k'} \sigma_{jj}$$

which converges to a finite limit as $N \to \infty$ by A1, so indeed $\underline{B}_N \to_P 0$ as $N \to \infty$. Similarly $\underline{C}_N \to_P 0$, and because

$$\frac{1}{N} \sum_{n:P_j^{(n)}P_{j'}^{(n)}=1} x_{njk} x_{nj'k'}$$

converges as $N \to \infty$, $\underline{D}_N \to_P 0$ too. □

THEOREM 3. *Suppose* A1 *and* A3 *hold and let* $\underline{S}^{(r)}$ *be any consistent estimator of* Σ. *Then with probability converging to* 1 *as* $N \to \infty$, $\tilde{\underline{S}}^{(r)}$ *and* $\tilde{X}'\tilde{\underline{S}}^{(r)-1}\tilde{X}$ *are nonsingular and defining*

$$\underline{b}^{(r+1)} = (\tilde{X}'\tilde{\underline{S}}^{(r)-1}\tilde{X})^{-1}\tilde{X}'\tilde{\underline{S}}^{(r)-1}\tilde{y}, \tag{10}$$

then

$$N^{\frac{1}{2}}(\underline{b}^{(r+1)}-\beta) = O_P(1) \tag{11}$$

(and so $\underline{b}^{(r+1)}$ *is a consistent estimator of* β*). If furthermore* A4 *and* A6 *hold, or* A7 *holds, then*

$$N^{\frac{1}{2}}(\underline{b}^{(r+1)}-\beta) \to_{\mathcal{D}} N(0,A) \tag{12}$$

where A *is defined by*

$$A^{-1} = \lim_{N\to\infty} \frac{1}{N} \tilde{X}'\tilde{\Sigma}^{-1}\tilde{X}; \tag{13}$$

in the latter case $\underline{b}^{(r+1)}$ *is an asymptotically efficient estimator of* β *(in the sense of e.g.* CRAMER [1946] §32.6). A *can consistently be estimated by* $(\frac{1}{N}\tilde{X}'\tilde{\underline{S}}^{-1}\tilde{X})^{-1}$.

PROOF. Write $\underset{\sim}{S} = \underset{\sim}{S}^{(r)}$ and $\underset{\sim}{b} = \underset{\sim}{b}^{(r+1)}$.

Since $\underset{\sim}{S} \to_P \Sigma$, and Σ_{PP} is non-singular for each pattern P, $P(\tilde{\underset{\sim}{S}}$ is non-singular) $\to 1$ as $N \to \infty$. Writing

$$(\frac{1}{N}\tilde{X}'\tilde{\underset{\sim}{S}}^{-1}\tilde{X})_{kk'} = \sum_{j,j';P:P_jP_{j'}=1} (\frac{1}{N} \sum_{n:P^{(n)}=P} x_{njk}x_{nj'k'})(\underset{\sim}{S}_{pp}^{-1})_{jj'}$$

we see that $\frac{1}{N}\tilde{X}'\tilde{\underset{\sim}{S}}^{-1}\tilde{X} \to_P A^{-1}$, and so with probability converging to 1 is non-singular. This also shows that A can be consistently estimated by $(\frac{1}{N}\tilde{X}'\tilde{\underset{\sim}{S}}^{-1}\tilde{X})^{-1}$. Now define

$$\underset{\sim}{b}^* = (\tilde{X}'\tilde{\Sigma}^{-1}\tilde{X})^{-1}\tilde{X}'\tilde{\Sigma}^{-1}\tilde{\underset{\sim}{y}} \quad . \tag{14}$$

We shall show that under A1 and A3

$$N^{\frac{1}{2}}(\underset{\sim}{b} - \underset{\sim}{b}^*) \to_P 0 \tag{15}$$

and that the theorem is also true if $\underset{\sim}{b}$ is replaced everywhere with $\underset{\sim}{b}^*$. These two facts establish the truth of the theorem itself.

Substituting (4) in (10) and (14) we see that with probability converging to 1 as $N \to \infty$

$$\underset{\sim}{b} = \beta + (\tilde{X}'\tilde{\underset{\sim}{S}}^{-1}\tilde{X})^{-1}\tilde{X}'\tilde{\underset{\sim}{S}}^{-1}\tilde{\underset{\sim}{e}} \tag{16}$$

and

$$\underset{\sim}{b}^* = \beta + (\tilde{X}'\tilde{\Sigma}^{-1}\tilde{X})^{-1}\tilde{X}'\tilde{\Sigma}^{-1}\tilde{\underset{\sim}{e}} \tag{17}$$

Thus again with probability converging to 1

$$\begin{aligned} N^{\frac{1}{2}}(\underset{\sim}{b}-\underset{\sim}{b}^*) &= (N^{-1}\tilde{X}'\tilde{\underset{\sim}{S}}^{-1}\tilde{X})^{-1}\{N^{-\frac{1}{2}}(\tilde{X}'\tilde{\underset{\sim}{S}}^{-1}\tilde{\underset{\sim}{e}} - \tilde{X}'\tilde{\Sigma}^{-1}\tilde{\underset{\sim}{e}})\} + \\ &\quad + \{(N^{-1}\tilde{X}'\tilde{\underset{\sim}{S}}^{-1}\tilde{X})^{-1} - (N^{-1}\tilde{X}'\tilde{\Sigma}^{-1}\tilde{X})^{-1}\}N^{-\frac{1}{2}}\tilde{X}'\tilde{\Sigma}^{-1}\tilde{\underset{\sim}{e}} \\ &= \underset{\sim}{A}_N\underset{\sim}{B}_N + \underset{\sim}{C}_N\underset{\sim}{D}_N \quad \text{(say)}. \end{aligned} \tag{18}$$

We have already seen that $\underline{A}_N \to_P A$ as $N \to \infty$ and that $\underline{C}_N \to_P 0$ as $N \to \infty$. So to establish (15) it remains to show that

$$\underline{B}_N \to_P 0 \tag{19}$$

and

$$\underline{D}_N = O_P(1). \tag{20}$$

Now

$$(\underline{B}_N)_k = \sum_{j,j',P} [N^{-\frac{1}{2}} \sum_{n:P^{(n)}=P} x_{njk}\underline{e}_{nj'}][(\underline{S}_{PP}^{-1})_{jj'} - (\Sigma_{PP}^{-1})_{jj'}]$$

where $[(\underline{S}_{pp}^{-1})_{jj'} - (\Sigma_{pp}^{-1})_{jj'}] \to_P 0$ as $N \to \infty$; while

$$E[N^{-\frac{1}{2}} \sum_{n:P^{(n)}=P} x_{njk}\underline{e}_{nj'}]^2 = \frac{1}{N} \sum_{n:P^{(n)}=P} x_{njk}x_{njk}\sigma_{j'j'}$$

which converges to a finite quantity as $N \to \infty$ by A1. So

$$N^{-\frac{1}{2}} \sum_{n:P^{(n)}=P} x_{njk}\underline{e}_{nj'} = O_P(1)$$

and hence (19) holds. Finally,

$$(\underline{D}_N)_k = \sum_{j,j',P} [N^{-\frac{1}{2}} \sum_{n:P^{(n)}=P} x_{njk}\underline{e}_{nj'}](\Sigma_{PP}^{-1})_{jj'} = O_P(1)$$

so (20) holds, and therefore (15) holds too.

We now prove the theorem with $\underline{b}$ replaced by $\underline{b}^*$. For sufficiently large N $\tilde{X}'\tilde{\Sigma}^{-1}\tilde{X}$ is nonsingular, (17) holds and by (6) $E(N^{\frac{1}{2}}(\underline{b}^*-\beta)N^{\frac{1}{2}}(\underline{b}^*-\beta)') = (N^{-1}\tilde{X}'\tilde{\Sigma}^{-1}\tilde{X})^{-1} \to A$ as $N \to \infty$.

So $N^{\frac{1}{2}}(\underline{b}^*-\beta) = O_P(1)$ as $N \to \infty$. Next suppose A1, A3, A4 and A6 hold. By (17)

$$N^{\frac{1}{2}}(\underline{b}^*-\beta) \to_D N(0,A) \tag{21}$$

if

$$N^{\frac{1}{2}}(\tilde{X}'\tilde{\Sigma}^{-1}\tilde{X})^{-1}\tilde{X}'\tilde{\Sigma}^{-1}\tilde{\underline{e}} \to_D N(0,A) \tag{22}$$

Since $N^{-1}\tilde{X}'\tilde{\Sigma}^{-1}\tilde{X} \to A^{-1}$, it suffices to show that

$$N^{-\frac{1}{2}}\tilde{X}'\tilde{\Sigma}^{-1}\tilde{e} \to_{\mathcal{D}} N(0,A^{-1}) \tag{23}$$

Now $N^{-\frac{1}{2}}\tilde{X}'\tilde{\Sigma}^{-1}\tilde{e} = N^{-\frac{1}{2}}\sum_{n=1}^{N}(X_n'\Sigma_n^{-1}e_n)$ where the sum is a sum of N independent zero mean random vectors, 1/N times the sum of whose covariance matrices converges to A^{-1} as $N \to \infty$. So by the Lindebergh-Feller central limit theorem we must show that for each k,

$$N^{-1}\sum_{n=1}^{N} E[(X_n'\Sigma_n^{-1}e_n)^2 \chi_{\{(X_n'\Sigma_n^{-1}e_n)_k > N^{\frac{1}{2}}\varepsilon\}}] \to 0 \tag{24}$$

as $N \to \infty$ for each $\varepsilon > 0$, where $\chi_{\cdot}$ denotes the indicator random variables for the event denoted by $\cdot$. Since the elements of X_n and Σ_n^{-1} are bounded in $n = 1,\ldots,N$ and $N = 1,2,\ldots$ we need only show

$$N^{-1}\sum_{n=1}^{N}\sum_{j=1}^{J} E[e_{nj}^2 \chi_{\{e_{nj}>N^{\frac{1}{2}}\varepsilon\}}] \to 0. \tag{25}$$

But by A4, the left hand side of (25) equals

$$\sum_{j=1}^{J} E[e_{1j}^2 \chi_{\{e_{1j}>N^{\frac{1}{2}}\varepsilon\}}]$$

which converges to zero as N tends to infinity. We have now proved (21) under A1, A3, A4 and A6. Under A1, A3 and A7 $N^{\frac{1}{2}}(b^*-\beta)$ is distributed as $N(0,(N^{-1}\tilde{X}'\tilde{\Sigma}^{-1}\tilde{X})^{-1})$ and so converges in distribution to $N(0,A)$ as $N \to \infty$. Finally under A1, A3 and A7, by the usual theory of generalized least squares, b^* attains the Cramer-Rao lower bound for each N and is certainly asymptotically efficient (in the sense of asymptotically attaining the Cramer-Rao lower bound). □

THEOREM 4. *Suppose* A1 *and* A4, *hold, and suppose* $b^{(r+1)}$ *is any estimator of* β *such that* $N^{\frac{1}{2}}(b^{(r+1)}-\beta) = O_p(1)$ *and suppose* $S^{(r)}$ *is any consistent estimator of* Σ. *Then* $S^{(r+1)}$ *defined by*

$$\hat{e}_n = y_n - X_n b^{(r+1)} \tag{26}$$

$$\hat{\hat{\underline{e}}}_n = \underline{S}_P^{(r)} \underline{S}_{PP}^{(r)-1} \hat{\underline{e}}_n \tag{27}$$

where $P = P^{(n)}$ *and*

$$\underline{S}^{(r+1)} = \frac{1}{N} \sum_P \sum_{n:P^{(n)}=P} (\hat{\underline{e}}_n \hat{\underline{e}}_n' + \underline{S}^{(r)} - \underline{S}_{\cdot P}^{(r)} \underline{S}_{PP}^{(r)-1} \underline{S}_{P\cdot}^{(r)}). \tag{28}$$

is also a consistent estimator of Σ.

PROOF. Write $\underline{b} = \underline{b}^{(r+1)}$, $\underline{S} = \underline{S}^{(r)}$, $\underline{S}^* = \underline{S}^{(r+1)}$. Since $\underline{S} \to_p \Sigma$ and Σ is nonsingular, $P(\underline{S}$ is nonsingular$) \to 1$ as $N \to \infty$ and so $\underline{S}^*$ is well defined, at least with a probability converging to 1 as $N \to \infty$. Note that putting for any n $P^{(n)} = P$ and $r = \Sigma_j P_j$, $\hat{\underline{e}}_n$ is like $\underline{e}_n$ an $r \times 1$ random vector while $\hat{\hat{\underline{e}}}_n$ is a $J \times 1$ random vector. $\underline{S}_{\cdot p} = \underline{S}_p'$. is $J \times r$ and $\underline{S}_{pp}$ is $r \times r$. Also if $\underline{S}$ is nonsingular

$$\begin{cases} \hat{\hat{\underline{e}}}_{nj} = \hat{\underline{e}}_{nj} & \text{if } P_j = 1 \\ (\underline{S} - \underline{S}_{\cdot P} \underline{S}_{PP}^{-1} \underline{S}_{P\cdot})_{jj'} = 0 & \text{if } P_j = 1 \text{ or } P_{j'} = 1 \end{cases} \tag{29}$$

so $\underline{S}^*$ is indeed the estimator described in section 2, step 2r+2. To prove $\underline{S}^* \to_P \Sigma$ as $N \to \infty$ we combine (27) and (28) to give

$$\underline{S}^* = \frac{1}{N} \sum_P \sum_{n:P^{(n)}=P} (\underline{S}_{\cdot P} \underline{S}_{PP}^{-1} \hat{\underline{e}}_n \hat{\underline{e}}_n' \underline{S}_{PP}^{-1} \underline{S}_{P\cdot} + \underline{S} - \underline{S}_{\cdot P} \underline{S}_{PP}^{-1} \underline{S}_{P\cdot}) \tag{30}$$

$$= \underline{S} + \sum_P \underline{S}_{\cdot P} \underline{S}_{PP}^{-1} \{\frac{1}{N} \sum_{n:P^{(n)}=P} (\hat{\underline{e}}_n \hat{\underline{e}}_n' - \underline{S}_{PP})\} \underline{S}_{PP}^{-1} \underline{S}_{P\cdot} .$$

Since $\underline{S}_{\cdot P}$ and $\underline{S}_{PP}^{-1}$ converge in probability to $\Sigma_{\cdot p}$ and Σ_{PP}^{-1} respectively it suffices to show that for each pattern P

$$\frac{1}{N} \sum_{n:P^{(n)}=P} (\hat{\underline{e}}_n \hat{\underline{e}}_n' - \underline{S}_{PP}) \to_P 0. \tag{31}$$

First we show that

$$\frac{1}{N} \sum_{n:P^{(n)}=P} (\underline{e}_n \underline{e}_n' - \hat{\underline{e}}_n \hat{\underline{e}}_n') \to_P 0. \tag{32}$$

Now by (26) and the fact that $\underline{e}_n = \underline{y}_n - X_n\beta$,

$$\hat{\underline{e}}_n = \underline{e}_n - X_n(\underline{b}-\beta)$$

and so

$$(33) \qquad \frac{1}{N} \sum_{n:P^{(n)}=P} (\underline{e}_n\underline{e}_n' - \hat{\underline{e}}_n\hat{\underline{e}}_n')$$

$$= \frac{1}{N} \sum_{n:P^{(n)}=P} [X_n(\underline{b}-\beta)\underline{e}_n' + \underline{e}_n(\underline{b}-\beta)'X_n' - X_n(\underline{b}-\beta)(\underline{b}-\beta)'X_n']$$

Looking at the three parts of this in turn, we see first that for j and j' such that P_j and $P_{j'} = 1$,

$$(\frac{1}{N} \sum_{n:P^{(n)}=P} X_n(\underline{b}-\beta)\underline{e}'_n)_{jj'} = \sum_k (\frac{1}{N} \sum_{n:P^{(n)}=P} x_{njk}\underline{e}_{nj'})(\underline{b}_k-\beta_k)$$

where

$$E(\frac{1}{N} \sum_{n:P^{(n)}=P} x_{njk}\underline{e}_{nj'})^2 = \frac{1}{N} \sum_{n:P^{(n)}=P} x_{njk}x_{njk}\sigma_{j'j'}.$$

So by A1

$$\frac{1}{N} \sum_{n:P^{(n)}=P} x_{njk}\underline{e}_{nj'} = O_P(1)$$

while $(\underline{b}_k-\beta_k) \to_P 0$. This shows that the first two parts of the right hand side of (33) converge in probability to zero as $N \to \infty$. For the third part,

$$(\frac{1}{N} \sum_{n:P^{(n)}=P} X_n(\underline{b}-\beta)(\underline{b}-\beta)'X_n')_{jj'} =$$

$$= \sum_k \sum_{k'} (\frac{1}{N} \sum_{n:P^{(n)}=P} x_{njk}x_{nj'k'})(\underline{b}_k-\beta_k)(\underline{b}_{k'}-\beta_{k'})$$

so this too converges in probability to zero as $N \to \infty$.

This establishes (32). Now obviously

$$\frac{1}{N} \sum_{n:P^{(n)}=P} (\underline{S}_{PP} - \Sigma_{PP}) \to_P 0 \qquad \text{as } N \to \infty,$$

so we finally prove

$$(34) \qquad \frac{1}{N} \sum_{n:P^{(n)}=P} (\underline{e}_n \underline{e}_n' - \Sigma_{PP}) \to_P 0 \qquad \text{as } N \to \infty.$$

Let $r_P = \#\{n:P^{(n)}=P\}$, and fix j and j' such that P_j and $P_{j'} = 1$. Since by A4 the $\underline{e}_n$'s with $P^{(n)} = P$ are independent and identically distributed and $E(\underline{e}_n \underline{e}_n') = \Sigma_{PP}$ if $P^{(n)} = P$, given ε and δ there exists an r such that

$$r_P(= r_P(N)) > r \Rightarrow P\left(\left|\frac{1}{r_P} \sum_{n:P^{(n)}=P} (e_{nj} e_{nj'} - \sigma_{jj'})\right| > \varepsilon\right) < \delta.$$

So since $r_P \le N$, for all N such that $r_P > r$,

$$P\left(\left|\frac{1}{N} \sum_{n:P^{(n)}=P} (e_{nj} e_{nj'} - \sigma_{jj'})\right| > \varepsilon\right) < \delta.$$

On the other hand, for all N such that $r_P \le r$,

$$\sum_{n:P^{(n)}=P} (e_{nj} e_{nj'} - \sigma_{jj'})$$

is bounded in probability. So there exists an N_0 such that if $N \ge N_0$,

$$r_P(N) \le r \Rightarrow P\left(\left|\frac{1}{N} \sum_{n:P^{(n)}=P} (e_{nj} e_{nj'} - \sigma_{jj'})\right| > \varepsilon\right) < \delta.$$

Thus for sufficiently large N, whether $r_P(N) \le r$ or not,

$$P\left(\left|\frac{1}{N} \sum_{n:P^{(n)}=P} (e_{nj} e_{nj'} - \sigma_{jj'})\right| > \varepsilon\right) < \delta$$

and we have proved (34). □

COROLLARY TO THEOREMS 1 TO 4. *Let* $\underline{b}^{(0)}$ *be defined by* (7) *and then* $\underline{S}^{(0)}$ *by* (8) *and* (9). *For* $r \ge 0$ *define* $\underline{b}^{(r+1)}$ *by* (10) *and then* $\underline{S}^{(r+1)}$ *by* (26), (27) *and* (28). *Then under* A1 *to* A5, $\underline{b}^{(r)}$ *and* $\underline{S}^{(r)}$ *are consistent estimators of* β *and* Σ *respectively for each* $r \ge 0$; *while for* $r \ge 1$, *under* A1 *to* A6, *or* A1 *to* A5 *and* A7, $N^{\frac{1}{2}}(\underline{b}^{(r)} - \beta) \to_D N(0,A)$ *where* A *is defined by* (13) *and can be consistently estimated by* $(N^{-1}\tilde{X}'(\tilde{\underline{S}}^{(r)})^{-1}\tilde{X})^{-1}$. *For each* $r \ge 1$, *under* A1 *to* A5 *and* A7, $\underline{b}^{(r)}$ *is an efficient estimator of* β.

PROOF. The proof is immediate. □

THEOREM 5. *Define* $(\underline{b},\underline{S})$ *as the limit as* $r \to \infty$ *of* $(\underline{b}^{(r)},\underline{S}^{(r)})$ *if this limit exists. Then under* A7, $(\underline{b},\underline{S})$ *is a stationary point of the likelihood function for* (β,Σ) *given the data. The likelihood function,* $\underline{\ell}(\beta,\Sigma)$, *evaluated at* $(\beta,\Sigma) = (\underline{b}^{(r)},\underline{S}^{(r)})$ *is non-decreasing in* r.

PROOF. If $(\underline{b}^{(r)},\underline{S}^{(r)})$ converges to a limit $(\underline{b},\underline{S})$ as $r \to \infty$, then $(\underline{b},\underline{S})$ satisfies the fixed point equations

$$\underline{b} = (\tilde{X}'\tilde{\underline{S}}^{-1}\tilde{X})^{-1}\tilde{X}'\tilde{\underline{S}}^{-1}\tilde{\underline{y}} \tag{35}$$

$$\hat{\underline{e}}_n = \underline{y}_n - X_n\underline{b}, \qquad n = 1,\ldots,N, \tag{36}$$

$$\hat{\underline{e}}_n = \underline{S}_{.p}\underline{S}_{pp}^{-1}\hat{\underline{e}}_n \text{ where } P = P^{(n)}, \qquad n = 1,\ldots,N, \tag{37}$$

and

$$\underline{S} = \frac{1}{N}\sum_{P}\sum_{n:P^{(n)}=P}(\hat{\underline{e}}_n\hat{\underline{e}}_n' + \underline{S} - \underline{S}_{.p}\underline{S}_{pp}^{-1}\underline{S}_{p.}) \tag{38}$$

Now denoting the likelihood function as $\underline{\ell}(\beta,\Sigma)$ (where β and Σ are variables and not the true values of the unknown paramaters) by (35),

$$\frac{\partial\underline{\ell}}{\partial\beta}(\beta,\Sigma)\Big|_{\beta=\underline{b},\Sigma=\underline{S}} = 0$$

while by (36) to (38),

$$\frac{\partial\underline{\ell}}{\partial\Sigma}(\beta,\Sigma)\Big|_{\beta=\underline{b},\Sigma=\underline{S}} = 0$$

(using BEALE and LITTLE [1975] Theorems 1 and 2 and formulas (2.5), (2.6), (2.8) and (2.9) with $\mu \equiv 0$). To prove the final assertion we observe that for $r \geq 1$,

$$\underline{\ell}(\underline{b}^{(r)},\underline{S}^{(r-1)}) = \sup_{\beta}\underline{\ell}(\beta,\underline{S}^{(r-1)})$$

because $\underline{b}^{(r)}$ is the generalized least squares estimator of β assuming that $\Sigma = \underline{S}^{(r-1)}$. So

$$\underline{\ell}(\underline{b}^{(r)},\underline{S}^{(r-1)}) \geq \underline{\ell}(\underline{b}^{(r-1)},\underline{S}^{(r-1)}).$$

Also by DEMPSTER *et al.* [1977] Theorem 1 and section 4.1.3,

$$\underline{\ell}(\underline{b}^{(r)},\underline{S}^{(r)}) \geq \underline{\ell}(\underline{b}^{(r)},\underline{S}^{(r-1)})$$

because $\underline{S}^{(r)}$ is computed from $\underline{b}^{(r)}$ and $\underline{S}^{(r-1)}$ by one step of the EM algorithm assuming that $\beta = \underline{b}^{(r)}$. □

THEOREM 6. *For some estimator* $\underline{S}$ *of* Σ, *define*

$$\underline{b} = (\tilde{X}'\tilde{\underline{S}}^{-1}\tilde{X})^{-1}\tilde{X}'\tilde{\underline{S}}^{-1}\tilde{y}$$

Then under A7 *the likelihood function for* β *and* Σ *given the data evaluated at* $\underline{b}$ *and* $\alpha\underline{S}$ *is maximized over* $\alpha > 0$ *by choosing*

$$\underline{\alpha} = \Big(\sum_{n,j} P_j^{(n)}\Big)^{-1} \cdot \sum_n (\underline{y}_n - X_n\underline{b})'\underline{S}_n^{-1}(\underline{y}_n - X_n\underline{b}). \tag{39}$$

If $N^{\frac{1}{2}}(\underline{b}-\beta) = O_P(1)$ *as* $N \to \infty$ *and* $\underline{S} \to_P \Sigma$ *as* $N \to \infty$, *and* A1 *and* A4 *hold, then* $\underline{\alpha} \to_P 1$ *as* $N \to \infty$. *Finally, both the corollary to theorems* 1 *to* 4, *and theorem* 5, *remain true if for* $r \geq 0$, $\underline{S}^{(r+1)}$ *is now defined by* (26), (27) *and* (28) *with* $\underline{S}^{(r)}$ *replaced by* $\underline{\alpha}^{(r)}\underline{S}^{(r)}$ *where* $\underline{\alpha}^{(r)}$ *is defined by* (39) *with* $\underline{b} = \underline{b}^{(r+1)}$ *and* $\underline{S} = \underline{S}^{(r)}$.

PROOF. Under A7 the log of the likelihood of the data evaluated at $\beta = \underline{b}$, $\Sigma = \alpha\underline{S}$ is equal to a constant plus

$$-\tfrac{1}{2}\sum_n [\log\det(\alpha\underline{S}_n) + \underline{\hat{e}}_n'(\alpha\underline{S}_n)^{-1}\underline{\hat{e}}_n]$$

$$= -\tfrac{1}{2}[\Big(\sum_{n,j} P_j^{(n)}\Big)\log\alpha + \alpha^{-1}\sum_n \underline{\hat{e}}_n'\underline{S}_n^{-1}\underline{\hat{e}}_n] - \tfrac{1}{2}\sum_n \log\det\underline{S}_n$$

where $\underline{\hat{e}}_n = \underline{y}_n - X_n\underline{b}$. By differentiating, we see that this as a function of α is maximized by

$$\underline{\alpha} = \Big(\sum_{n,j} P_j^{(n)}\Big)^{-1}\sum_n \underline{\hat{e}}_n'\underline{S}_n^{-1}\underline{\hat{e}}_n$$

To prove that under the stated conditions $\underline{\alpha} \to_P 1$, arguments of exactly the same nature as those used in Theorem 4 can be employed. Note that $\hat{e}_n' S_n^{-1} \hat{e}_n = \text{trace}(\hat{e}_n \hat{e}_n' S_n^{-1})$. The final assertions on the validity of the corollary to theorems 1 to 4 and of theorem 5 with redefined $\underline{S}^{(r+1)}$ for $r \geq 0$ are straightforward to check. Note that we now have for $r \geq 1$ the inequalities

$$\underline{\ell}(\underline{b}^{(r-1)}, \underline{S}^{(r-1)}) \leq \underline{\ell}(\underline{b}^{(r)}, \underline{S}^{(r-1)}) \leq \underline{\ell}(\underline{b}^{(r)}, \underline{\alpha}^{(r-1)} S^{(r-1)}) \leq$$

$$\leq \underline{\ell}(\underline{b}^{(r)}, \underline{S}^{(r)}). \qquad \square$$

4. OTHER APPROACHES

First of all we briefly discuss estimation of the "random effects model" (a) and the "first order autocorrelation model" (b) described in section 2. These are both special cases of our model with restrictions on Σ. Econometric applications of the random effects model, or very similar models, are described by KMENTA [1971] §12.2 and WALLACE and HUSSAIN [1969]. Both articles assume that the data is complete - i.e. $P_j^{(n)} = 1$ for all n and j - and neither describe maximum likelihood methods under multivariate normality. HAN [1969] shows how to obtain maximum likelihood estimates, again supposing the data is complete, while WIORKOWSKI [1975] describes a technique for obtaining maximum likelihood estimates even with incomplete data. His method works as follows. Defining $\rho = \sigma_\varepsilon^2/(\sigma_\varepsilon^2 + \sigma_\delta^2)$ (see section 2) we see that in the model specified by (4), (5) and (6) together with assumption A7,

$$\tilde{\Sigma} = (\sigma_\varepsilon^2 + \sigma_\delta^2) \cdot H(\rho) = \sigma^2 H(\rho) \tag{40}$$

where $H(\rho)$ is a $\Sigma_{n,j} P_j^{(n)} \times \Sigma_{n,j} P_j^{(n)}$ matrix whose elements are either 0, ρ or 1. For a range of values of ρ it is now easy to estimate β and σ^2 by maximum likelihood given ρ (i.e. by generalized least squares), and for each value of ρ the maximum likelihood can be calculated. This function of ρ can now be maximized, giving the maximum likelihood estimator of ρ and corresponding estimators of β and σ. The usefulness of this method is that under A7, by computing the maximum likelihood for our model and for the more

restricted model of random effects, we can test the latter by using the usual asymptotic likelihood ratio test.

The first order autocorrelation model (b) can be estimated, under A7, in exactly the same way as in WIORKOWSKI's [1975] method for the random effects model because again in the model (4), (5) and (6) together with A7, (40) holds, where $H(\rho)$ is now another matrix function of the parameter ρ.

Next we look at other ways of estimating our own model. If the data is complete it is very easy under multivariate normality of $\tilde{\underline{e}}$ to write down the maximum over β of the likelihood function for β and Σ. This gives a (random) function of Σ which can itself be maximized over Σ by (iterative) numerical optimization techniques. When there are no missing observations the model can be written as

$$\underline{y}_n = \tilde{\tilde{\beta}}\tilde{\tilde{X}}_n + \underline{e}_n, \qquad n = 1,\ldots,N \tag{41}$$

where $\underline{y}_n$ and $\underline{e}_n$ are $J \times 1$ random vectors, $\tilde{\tilde{\beta}}$ is $J \times JK$ and $\tilde{\tilde{X}}_n$ is $JK \times 1$,

$$\tilde{\tilde{\beta}} = \begin{pmatrix} \beta' & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \beta' & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ \vdots & \vdots & \vdots & \vdots \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \beta' \end{pmatrix}, \tag{42}$$

$$\tilde{\tilde{X}}_n = (x_{n11},\ldots,x_{n1K},x_{n21},\ldots,x_{n2K},\ldots,x_{nJK})', \tag{43}$$

and the $\underline{e}_n$'s are uncorrelated with zero means and fixed $J \times J$ covariance matrix. All the same, the usual theory of estimating the general linear model (41) does not tell us how to estimate $\tilde{\tilde{\beta}}$ when $\tilde{\tilde{\beta}}$ is of the form given by (42). Such a model is a special case of the "ACOVSM" model of JÖRESKOG [1970], though using this fact would probably lead to computationally inefficient ways of estimating it.

We next consider the question of whether our model could have been estimated by the EM algorithm of DEMPSTER *et al.* [1977] in its proper form,

instead of the adaptation we have chosen. This algorithm is a general one for maximum likelihood estimation with incomplete data; here we would assume multivariate normality of the disturbances $\underline{e}_{nj}$. Our method works by switching between estimating β and Σ: we estimate β by maximum likelihood as if the current estimate of Σ were the true value of Σ, and then improve our estimate of Σ by carrying out one iteration of the EM algorithm, as if the current estimate of β were *its* true value. However the EM algorithm could be applied to improve the current estimates of β and Σ simultaneously. Considering our data as coming from the model specified by (41), (42) and (43) together with the usual assumptions, including that of multivariate normality, about the $\underline{e}_n$'s, but with some components of the vectors $\underline{y}_n$ missing, the EM algorithm could be applied at the so called second level of generality (DEMPSTER *et al.* [1977], bottom of page 5) appropriate when the distribution of the complete data comes from an exponential family of distributions whose natural parameters $(\Sigma^{-1})_{jj'}$ $(1 \leq j \leq j' \leq J)$ and $(\Sigma^{-1}\tilde{\beta})_{j\ell}$ $(1 \leq j \leq J,\ 1 \leq \ell \leq JK)$, which vary in the convex region of $\mathbb{R}^{\frac{1}{2}j(j+1)+j.jk}$ where Σ^{-1} is positive definite, are restricted to lying in the curved submanifold specified by (42). So this model is certainly one to which their method is applicable; however the difficulty is that at repeated steps of the algorithm the parameters have to be reestimated by maximum likelihood based on current predictions of the sufficient statistics for the complete data (i.e. as if the predictions were the actual values taken by these sufficient statistics). Now as we have seen, even with complete data an iterative method has to be used to get maximum likelihood estimates, which makes this approach computationally unattractive.

Another way of estimating our usual model as well as the models of random effects and first order autocorrelation under multivariate normality is to make use of the method of restricted maximum likelihood (CORBEIL and SEARLE [1976]). In the model specified by (4), (5) and (6), $\tilde{\Sigma}$ is in each case equal to $\sigma^2 H(\theta)$ for some $\sigma^2 > 0$ and a vector θ of (a fairly small number of) parameters. CORBEIL and SEARLE [1976] suggest transforming $\tilde{\underline{y}}$ into two parts by means of two linear transformations of $\tilde{\underline{y}}$, such that the distribution of one of these parts depends only on $\sigma^2 H(\theta)$ and not on β (assuming multivariate normality of $\tilde{\underline{e}}$). θ and σ^2 are estimated by maximum likelihood applied to this part of the data (using direct numerical optimization

techniques, though perhaps special methods related to the ones we have described here would work faster). Then with the estimate of θ so obtained, β is estimated by the obvious generalized least squares formula

$$\hat{\underline{\beta}} = (\tilde{X}'H(\hat{\underline{\theta}})^{-1}\tilde{X})^{-1}\tilde{X}'H(\hat{\underline{\theta}})^{-1}\tilde{\underline{y}}.$$

This method is discussed and compared with others in HARVILLE [1977] and seems rather promising.

Finally, it is sometimes reasonable to consider the X_n's as being the realized values of stochastic variables $\underline{X}_n$; e.g. suppose that $(\underline{y}_n, \underline{X}_n)$, n = 1,,..,N, are independent observations each with $(J - \Sigma_j P_j^{(n)})$. (K + 1) missing components from some J.(K + 1)-variate distribution, the observations being independent of one another. We could now estimate the mean vector and covariance matrix of the underlying joint distribution by some incomplete data technique; for β and Σ are functions of these parameters. Here again there is a problem because the fact that $\underline{y}_{nj}$ has the same regression on $\underline{x}_{njk}$, k = 1,...,K for each j means that some constraints should be introduced.

The method described in this report itself supplies a "modified EM algorithm" for estimating mean vector and covariance matrix from observations from a multivariate distribution with components missing according to some fixed patterns. For setting K = J and x_{njk} = 1 if j = k and 0 otherwise gives us exactly this model. We have not studied the differences between this modification and the true EM algorithm. Of course, even when the observations are not multivariate normally distributed, "maximum likelihood estimation under multivariate normality" can still give consistent estimators of mean vector and covariance matrix; this can for instance be shown by adapting the technique used by GILL [1977].

It is not our purpose here to discuss the general problem of missing observations in multivariate analysis: surveys including large reference lists are given by AFIFI and ELASHOF [1966] and HARTLEY and HOCKING [1971]. BEALE and LITTLE [1975] give some useful regression analysis methods, and DEMPSTER *et al.* also contains very many references. Another interesting article (with econometric applications) is by DAGENAIS [1973], who proposes a method very similar to "methods (5) and (6)" of BEALE and LITTLE [1975]. Unfortunately,

DAGENAIS' somewhat sketchy proof of asymptotic normality of regression coefficient estimators seems to require rather stronger conditions than he suggests. We study this aspect of these methods further in GILL [1979].

REFERENCES

[1] AERT, J.H. VAN & A.P.W.P. VAN MONTFORT [1979], *Basisonderzoek kostenstructuur ziekenhuizen deel 7* (forthcoming), Nationaal Ziekenhuisinstituut, Utrecht.

[2] AFIFI, A.A. & R.M. ELASHOFF [1966], *Missing Observations in Multivariate Statistics: I, Review of the Literature*, J. Amer. Stat. Ass. 61, 595-604.

[3] BEALE, E.M.L. & R.J.A. LITTLE [1975], *Missing Values in Multivariate Analysis*, J.R. Statist. Soc. (B) 37, 129-145.

[4] CRAMÉR, H. [1946], *Mathematical Methods of Statistics*, Princeton University Press, Princeton.

[5] CORBEIL, R.R. & J.R. SEARLE [1976], *Restricted Maximum Likelihood Estimation of Variance Components in the Mixed Model*, Technometrics 16, 833-834.

[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 Statistics, J.R. Barra *et al.*(eds.), North-Holland, Amsterdam.

[9] GILL, R.D. [1979], *A Note on some Methods for Regression Analysis with Incomplete Data*, Mathematisch Centrum, Amsterdam.

[10] GLASSER, M. [1964], *Linear Regression Analysis with Missing Observations among the Independent Variables*, J. Amer. Statist. Ass. 59, 834-844.

[11] HAN, C.P. [1969], *Maximum Likelihood Estimate in Intraclass Correlation Model*, Technometrics 11, 833-834.

[12] HARTLEY, H.O. & R.R. HOCKING [1971], *The Analysis of Incomplete Data*, Biometrics 27, 783-823.

[13] HARVILLE, D.A. [1977], *Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems*, J. Amer. Statist. Ass. 72, 320-340.

[14] JÖRESKOG, K.G. [1970], *A General Method for Analysis of Covariance Structures*, Biometrika 57, 239-251.

[15] KMENTA, J. [1971], *Elements of Econometrics*, Macmillan, New-York.

[16] MONTFORT, A.P.W.P. VAN [1979], Ph.D. Thesis (forthcoming), Nationaal Ziekenhuisinstituut, Utrecht.

[17] ORCHARD, T. & M.A. WOODBURY [1972], *A Missing Information Principle: Theory and Applications*, Proc. 6th Berkeley Symp. Math. Statist. Prob. 1, 697-715.

[18] WALLACE, T.D. & A. HUSSAIN [1969], *The Use of Error Components in Combining Cross Section with Time Series Data*, Econometric 37, 55-72.

[19] WIORKOWSKI, J.J. [1975], *Unbalanced Regression Analysis with Residuals having a Covariance Structure of Intra-Class Form*, Biometrics 31, 611-618.

ACKNOWLEDGEMENTS

I wish to thank Drs. A.P.W.P. van Montfort and Dr. J.H. van Aert for supplying the problem and for many useful discussions, and Mr. F.J. Burger (of the Mathematical Centre, Amsterdam) for implementing the method on the computer.