Sankhyā : The Indian Journal of Statistics
1988, Volume **50**, Series B, Pt. 1, pp. 95–102.

REGRESSION ANALYSIS FOR INCOMPLETE MIXED CROSS-SECTION AND TIME-SERIES DATA BY A MODIFIED EM ALGORITHM

By RICHARD D. GILL

Centre for Mathematics and Computer Science, Amsterdam

SUMMARY. 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 give theorems on consistency and asymptotic normality of estimators of the regression coefficients as the size of the cross-section increases while the length of the time series remains bounded. We discuss the connection between our method and the EM algorithm.

1. Introduction

In an econometric study of the cost-structure of Dutch hospitals (see Van Aert and Van Montfort (1979) the need arose to combine cross-sectional regression analyses over a (relatively short) time-series of years. Individual hospitals were to be treated as independent observations. It had to be taken into account that the hospitals taking part in each year's annual surveys varied over the years ; the effects of some explanatory variables might be allowed to vary over the years but those of others should be constant ; and the disturbance term for each observation could be highly correlated over the years. There is an extensive statistical and econometric literature on combining time-series and cross-sectional data, even with incomplete observations ; see the excellent survey by Dielman (1983). However the present paper addresses some novel issues and on the way provides some new methods and results on missing data in multivariate analysis. We do assume that observations are missing independently of the random components in our model (and condition on the observed patterns of missing data) but on the other hand make no special assumptions about the distributions of the disturbance term over the years. So the well known variance components (random effects) and first order auto-correlation models are statistically testable special cases of our model. In the next section we describe the model and our estimation and testing procedures. The following section

AMS (1980) *subject classification :* 62P20, 62H99.

Keywords and phrases : Mixed cross-section and time-series data, econometric models, incomplete observations, missing data, regression analysis, multivariate analysis, EM algorithm.

contains formal statements on asymptotic properties of the estimators under regularity assumptions; the (routine) proofs are contained in a technical report available from the author. In the final section we discuss some other possible approaches for estimation of our model. Also we show how our approach can be applied to the multivariate incomplete data problem.

2. Description of the model and estimators

First we introduce some notation. Random variables are set in bold. 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 explanatory 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 $P_n \subset \{1, \ldots, J\}$ be a non-empty set of indices j. Let x_{njk} ($n = 1, \ldots, N$, $j \in P_n$ and $k = 1, \ldots, K$) be real numbers. For $n = 1, \ldots, N$ and $j \in P_n$ suppose $\boldsymbol{y}_n = \Sigma_k x_{njk} \beta_k + \boldsymbol{e}_{nj}$ where the random varibles $\boldsymbol{e}_{nj}$ satisfy $\mathcal{E}(\boldsymbol{e}_{nj}) = 0$ and $\mathcal{E}(\boldsymbol{e}_{nj}\, \boldsymbol{e}_{n'j'}) = \delta_{nn'} \sigma_{jj'}$ (δ is the Kronecker symbol) where $\beta = (\beta_k)$ is a fixed $K \times 1$-vector and $\Sigma = (\sigma_{jj'})$ a fixed positive-definite symmetric $J \times J$-matrix.

So P_n denotes the *pattern* of non-missing time points for the n-th observations. "Fixed" means in the above specification "not depending on N_0" *All other quantities* may vary with N but we generally suppress this dependence in our notation. The symbol $\xrightarrow[\mathcal{L}]{}$ denotes convergence in distribution *as N tends to infinity.* The x_{njk}'s are the non-random values taken by our K explanatory variables, $\boldsymbol{y}_{nj}$ is the dependent variable; β and Σ are unknown parameters. We observe x_{njk} and $\boldsymbol{y}_{nj}$ for those n and j such that $j \in P_n$. Without loss of generality we suppose that $\boldsymbol{e}_{nj}$ is defined for *all* n and j (still satisfying the model assumptions).

The variance components and the first order autocorrelation models are obtained on placing appropriate restrictions on Σ. Some or all explanatory variables can be allowed to have arbitrarily varying effects over time by expanding the design matrix and vector of regression coefficients in the usual way.

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 $j \neq j'$; call this estimator $\boldsymbol{b}^{(0)}$.

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 $j, j' \in P_n$ and dividing by the number of such n to get an estimate of $\sigma_{jj'}$. Call this estimator $\boldsymbol{S}^{(0)}$.

Step $2r+1$ ($r = 1, 2, \ldots$) : With the estimate of Σ obtained from step $2r$, reestimate β by the method of generalized least squares (i.e. as if the estimate were the value of Σ). This defines an estimator $\boldsymbol{b}^{(r)}$.

Step $2r+2$ ($r = 1, 2, \ldots$) : 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 $\boldsymbol{e}_{nj}$ for those n and j such that $j \notin P_n$, and (c) estimating Σ in the obvious way from the now "completed" set of error terms, except that a correction term based on the old estimate of Σ is added to a summand in the sums of squares or products of errors, e.q. predicted errors, whenever the product consists of two predicted errors. The correction term is (estimate of) the partial covariance of the two errors which have to be predicted given those on which the predictions are based. This defines an estimator $\boldsymbol{S}^{(r)}$.

To explain this last step and introduce some important notation let $\boldsymbol{e} = (\boldsymbol{e}_P^T, \boldsymbol{e}_M^T)^T$ be a $J \times 1$ random vector (T denotes transpose) partitioned according to a pattern of observed components P and it complement of missing ones $\boldsymbol{M}$; $\mathcal{E}(\boldsymbol{e}) = 0$, $\mathcal{E}(\boldsymbol{e}\boldsymbol{e}^T) = \Sigma = \begin{pmatrix} \Sigma_{PP} & \Sigma_{PM} \\ \Sigma_{MP} & \Sigma_{MM} \end{pmatrix} = (\Sigma_{\cdot P} \Sigma_{\cdot M})$ where Σ is positive definite and partitioned conform $\boldsymbol{e}$ itself. Then the linear least squares predictor of $\boldsymbol{e}_M$ given $\boldsymbol{e}_P$ is $\hat{\mathcal{E}}(\boldsymbol{e}_M \mid \boldsymbol{e}_P) = \Sigma_{MP} \Sigma_{PP}^{-1} \boldsymbol{e}_P$ which has the covariance matrix $\mathcal{E}(\hat{\mathcal{E}}(\boldsymbol{e}_M \mid \boldsymbol{e}_P) \hat{\mathcal{E}}(\boldsymbol{e}_M \mid \boldsymbol{e}_P)^T) = \Sigma_{MP} \Sigma_{PP}^{-1} \Sigma_{PM}$. So $\mathcal{E}(\boldsymbol{e}_M \boldsymbol{e}_M^T) = \mathcal{E}(\hat{\mathcal{E}}(\boldsymbol{e}_M \mid \boldsymbol{e}_P) \hat{\mathcal{E}}(\boldsymbol{e}_M \mid \boldsymbol{e}_P)^T) + (\Sigma_{MM} - \Sigma_{MP} \Sigma_{PP}^{-1} \Sigma_{PM})$ where the last term, also equal to the covariance matrix of $\boldsymbol{e}_M - \hat{\mathcal{E}}(\boldsymbol{e}_M \mid \boldsymbol{e}_P)$, is conventionally called the partial covariance matrix of $\boldsymbol{e}_M$ given $\boldsymbol{e}_P$. On the other hand $\mathcal{E}(\boldsymbol{e}_M \boldsymbol{e}_P^T) = \mathcal{E}(\hat{\mathcal{E}}(\boldsymbol{e}_M \mid \boldsymbol{e}_P \boldsymbol{e}_P^T) = \Sigma_{MP}$.

Our 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 one step of the EM algorithm for determining the maximum likelihood estimator of Σ (assuming normality) ; see Dempster et all. (1977). Consistency asymptotic normality and efficiency properties of these estimators are given in the next section. In particular, under multivarite normality of the e_{nj}'s, $\boldsymbol{b}^{(r)}$ is an efficient estimator of β for each $r \geqslant 1$ and (this is the reason for iterating past $r=1$) *if* $\boldsymbol{b}^{(r)}$ and $\boldsymbol{S}^{(r)}$ converge to say $\boldsymbol{b}$ and $\boldsymbol{S}$ as $r \to \infty$, then $\boldsymbol{b}$ and $\boldsymbol{S}$ are stationary points of the likelihood function $\boldsymbol{l}(\beta, \Sigma)$ for β and Σ given the data. In any case $\boldsymbol{l}(\boldsymbol{b}^{(r)}, \boldsymbol{S}^{(r)})$ is nondecreasing in r. Rough tests of hypotheses of interest may be carried out by assuming the usual asymptotic maximum likelihood theory applies, provided convergence appears to have taken place. Our practical experience and the similarity to the EM algorithm suggest that it occurs, but rather slowly ; see Cziszar and Tusnady (1984) and Meilijson (1986).

A major assumption of the theorems is that for each pair of time instants j and j', the number of observations n for which $j, j' \in P_n$ tends to infinity as $N \to \infty$. This is obviously in general a necessary condition for consistent of Σ and hence for efficient estimation of β.

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.

We also need the following notation. Let r_n be the number of elements in P_n. Let X_n be the $r_n \times K$ matrix of elements x_{njk} such that $j \in P_n$, and similarly let $\boldsymbol{y}_n$ and $\boldsymbol{e}_n$ be the $r_n \times 1$ vectors of elements $\boldsymbol{y}_{nj}$ and $\boldsymbol{e}_{nj}$ respectively for which $j \in P_n$. Define the $r_n \times r_n$ matrix $\Sigma_n = \Sigma_{P_n P_n}$. Next we define $\tilde{X}$, $\tilde{\boldsymbol{y}}$, $\tilde{\boldsymbol{e}}$ and $\tilde{\Sigma}$ by $\tilde{X}^T = (X_1^T \ldots X_N^T)$, $\tilde{\boldsymbol{y}}^T = (\boldsymbol{y}_1^T \ldots \boldsymbol{y}_N^T)$, and $\tilde{\boldsymbol{e}}^T = (\boldsymbol{e}_1^T \ldots \boldsymbol{e}_N^T)$, while $\tilde{\Sigma}$ is the block-diagonal matrix with diagonal submatrices Σ_n. If $\boldsymbol{S}$ is some estimator of Σ, then $\boldsymbol{S}$ is defined analogously. We can now write the model assumptions as $\tilde{\boldsymbol{y}} = \tilde{X}\beta + \tilde{\boldsymbol{e}}$, $\mathcal{E}(\tilde{\boldsymbol{e}}) = 0$, and $\mathcal{E}(\tilde{\boldsymbol{e}}\tilde{\boldsymbol{e}}^T) = \tilde{\Sigma}$.

Finally before stating our theorems we list the assumptions which will be made in some or all of them. They can be much weakened, but lead to very easy proofs.

A1. For each P, j, j', k and k' $\lim_{N\to\infty} \frac{1}{N} \Sigma_{n:P_n=P} x_{njk} x_{nj'k'}$ exists (and is finite).

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

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

A4. $(e_{nj}; j = 1, \ldots, J)$, $n = 1, \ldots, N$ are independent and, also over $N = 1, 2, \ldots$, identically distributed random vectors.

A5. For each j and j' $\lim_{N\to\infty} \# \{n : j, j' \in P_n\} \to \infty$.

A6. For some constant $C < \infty$ not depending on N, $\sup_{n,j,k} |x_{njk}| \leqslant C$.

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

Theorem 1 : *Under* A1 *and* A2, $(\tilde{X}^T \tilde{X})$ *is for sufficiently large* N *nonsingular and defining* $b^{(0)} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{y}$, $b^{(0)}$ *is a* $\sqrt{N}$*-consistent estimator of* β ; *i.e.* $N^{\frac{1}{2}} (b^{(0)} - \beta)$ *is bounded in probability as* $N \to \infty$.

Theorem 2 : *Suppose* A1, A4 *and* A5 *hold and that* $b^{(0)}$ *is any consistent estimator of* β. *Then* $S^{(0)}$ *defined by* $s_{jj'}^{(0)} = n_{jj'}^{-1} \sum_{n:jj' \in P_n} (y_{nj} - \sum_k x_{njk} b_k^{(0)})(y_{nj'} - \sum_{k'} x_{nj'k'} b_{k'}^{(0)})$ *where* $n_{jj'} = \# \{n : j, j' \ P_n\}$ *is a consistent estimator of* Σ.

Theorem 3 : *Suppose* A1 *and* A3 *hold and let* $S^{(r)}$ *be any consistent estimator of* Σ. *Then with probability converging to* 1 *as* $N \to \infty$, $\tilde{S}^{(r)}$ *and* $\tilde{X} \tilde{S}^{(r)-1} \tilde{X}$ *are nonsingular and defining* $b^{(r+1)} = (\tilde{X}^T \tilde{S}^{(r)-1} \tilde{X})^{-1} \tilde{X}^T \tilde{S}^{(r)-1} \tilde{y}$, *then* $b^{(r+1)}$ *is a* $\sqrt{N}$*-consistent estimator of* β. *If furthermore* A4 *and* A6 *hold or* A7 *holds, then* $N^{\frac{1}{2}}(b^{(r+1)} - \beta) \underset{\mathcal{A}}{\to} \mathcal{N}(0, A)$ *where* A *is defined by* $A^{-1} = \lim_{N\to\infty} N^{-1} \tilde{X}^T \tilde{\Sigma}^{-1} \tilde{X}$; *in the latter case* $b^{(r+1)}$ *is an asymptotically efficient estimator of* β. *The matrix* A *can be consistently estimated by* $(N^{-1} \tilde{X}^T \tilde{S}^{(r)-1} \tilde{X})^{-1}$.

Theorem 4 : *Suppose* A1 *and* A4 *hold, and suppose* $b^{(r+1)}$ *is any* $\sqrt{N}$*-consistent estimator of* β *and suppose* $S^{(r)}$ *is any consistent estimator of* Σ. *Then* $S^{(r+1)}$ *defined by* $\hat{e}_n = y_n - X_n b^{(r+1)}$, $\hat{\hat{e}}_n = S_{.P}^{(r)} S_{PP}^{(r)-1} \hat{e}_n$ *where* $P = P_n$, *and* $S^{(r+1)} = N^{-1} \Sigma_P \Sigma_{n:P_n=P} (\hat{\hat{e}}_n \hat{\hat{e}}_n^T + S^{(r)} - S_{.P}^{(r)} S_{PP}^{(r)-1} S_{P.}^{(r)})$, *is also a consistent estimator of* Σ.

Corollary to Theorems 1 to 4 : *Let* $b^{(0)}$ *and* $S^{(0)}$ *be defined as in Theorems* 1 *and* 2. *For* $r \geqslant 0$ *define* $b^{(r+1)}$ *and* $S^{(r+1)}$ *as in Theorems* 3 *and* 4. *Then under* A1 *to* A5, $b^{(r)}$ *and* $S^{(r)}$ *are consistent estimators of* β *and* Σ

respectively for each $r \geqslant 0$; *while for* $r \geqslant 1$, *under* A1 *to* A6, *or* A1 *to* A5 *and* A7, $N^{\frac{1}{2}}(\boldsymbol{b}^{(r)}-\beta) \underset{\mathcal{A}}{\rightarrow} (0, A)$ *where* A *can be consistently estimated by* $(N^{-1}\widetilde{X}^T\mathbf{S}^{(r)})^{-1}\widetilde{X})^{-1}$. *For each* $r \geqslant 1$, *under* A1 *to* A5 *and* A7, $\boldsymbol{b}^{(r)}$ *is an efficient estimator of* β.

When the data is complete and under multivariate normality $\mathbf{S}^{(r)}$ is actually efficient too, for each r. All these efficiency results are part of the general phenomenon observed and explained in Dzhaparidze (1983).

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

Other approaches : First of all we briefly discuss estimation of the random effects model and the first order autocorrelation model. These are both special cases of out model with restrictions on Σ. Wiorkowski (1975) describes a simple technique for obtaining maximum likelihood estimates even with incomplete data for the first of the two models ; the second can be treated in the same way. We can test such submodels by the usual asymptotic likelihood ratio test.

Next we look at other ways of estimating our own model. If the data is complete it is very easy under multivariate normality of $\tilde{\boldsymbol{e}}$ 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 (heavy) iterative numerical optimization techniques. One can actually write the model as a special case of the "ACOVSM" model of Jöreskog (1970), but this does not help matters.

We next consider the question of whether our model could have been estimated by the EM algorithm of Dempster *et al.* (1977). 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 ; using maximum likelihood at each step based on predicted complete data sufficient statistics (we have a curved exponential family). However, even with complete data a numerically intensive method has to be used to get maximum likelihood estimates so this is not very feasible.

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); Harville (1977)). In these models $\tilde{\Sigma}$ 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 $\tilde{\boldsymbol{y}}$ into two parts by means of two linear transformations of $\tilde{\boldsymbol{y}}$, such that the disribution of one of these parts depends only on $\sigma^2 H(\theta)$ and not on β (assuming multivariate normality of $\tilde{e}$). θ and σ^2 are estimated by maximum likelihood applied to this part of the data. Then with the estimate of θ so obtained, β is estimated by the obvious generalized least squares formula.

Finally, it is sometimes reasonable to consider the X_n's as being the realized values of stochastic variables $\boldsymbol{X}_n$; e.g. suppose that $(\boldsymbol{y}_n, \boldsymbol{X}_n)$, $n = 1, \ldots, N$, are independent observations each with $(J-r_n).(K+1)$ 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; β and Σ are functions of these parameters. The fact that $\boldsymbol{y}_{nj}$ has the same regression on x_{njk}, $k=1, \ldots, K$ for each j means that some constraints should be introduced.

The method described in this paper itself supplies a "modified EM algorithm" for 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. Even when the observations are not multivariate normally distributed, "maximum likelihood estimation under multivariate normality" can still give consistent and even asymptotically normally distributed estimators of mean vector and covariance matrix; see Gill (1977, 1986) and van Praag, De Leeuw and Kloek (1986). The advantage of our modification is that the estimators have good statistical properties right from the first iteration step.

References

Van Aert, J. H. and Van Montfort, A. P. W. P. (1979): *Basic Research on the Cost-Structure of Hospitals* (in Dutch), National Hospitals Institute (NZI), Utrecht.

Corbeil, R. R. and Searle J. R. (1976): Restricted maximum likelihood estimation of variance components in the mixed model. *Technometrics*, **16**, 833-834.

Cziszar, I. and Tusnady, G. (1984): Information geometry and alternating minimization procedures. *Statistics and Decisions*, Supplement Issue No 1, 205-237.

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., LAIRD, N. M. and RUBIN, D. B. (1977) : Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.* (B), **39**, 1-38.

DIELMAN, T. E. (1983) : Pooled cross-sectional and time-series data : A survey of current statistical methodology. *The American Statistician*, **37**, 111-122.

DZHAPARIDZE, K. (1983) : On iterative procedures of statistical inference. *Statistical Neerlandica*, **37**, 181-189.

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

——— (1986) : A note on some methods for regression analysis with incomplete data. *Sankhyā* (B), **48**, 19-30.

GLASSER, M. (1964) : Linear regression analysis with missing observations among the independent variables. *J. Amer. Statist. Ass.*, **59**, 834-844.

HARVILLE, D. A. (1977) : Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.*, **72**, 320-340.

JÖRESKOG, K. G. (1970) : A general method for analysis of covariance structures. *Biometrika*, **57**, 239-251.

MEILIJSON, I. (1986) : *A Fast Improvement of the EM Algorithm on its Own Terms*, Preprint, School of Mathematical Sciences, Tel-Aviv University. (To appear in *J. R. Statist. Soc.* (B)).

VAN PRAAG, B. M. S., DE LEEUW J. and KLOEK, T. (1986) : The population-sample decomposition approach to multivariate estimation methods. *Appl. Stoch. Models and Data Anal.*, **2**, 99-119.

WIORKOWSKI, J. J. (1975) : Unbalanced regression analysis with residuals having a covariance structure of intra-class form. *Biometrics*, **31**, 611-618.

Paper received : March, 1982.

Revised : June, 1987.