Estimation in a linear model with serially correlated errors when observations are missing

Wansbeek, T.J.; Kapteyn, A.J.

Published in:
International Economic Review

Publication date:
1985

INTERNATIONAL ECONOMIC REVIEW
Vol. 26, No. 2, June, 1985

ESTIMATION IN A LINEAR MODEL WITH SERIALLY CORRELATED ERRORS WHEN OBSERVATIONS ARE MISSING*

By Tom Wansbeek and Arie Kapteyn[1]

1. INTRODUCTION

There exists an extensive literature on estimation and testing in linear regression models with first-order serially correlated errors. For the case where a string of consecutive observations is missing there have appeared a number of recent articles dealing with various tests of autocorrelation (cf. Savin and White [1978], Richardson and White [1979], Honohan and McCarthy [1982]). Obviously, many time series suffer from missing observations, like long annual series from which observations on war years are missing, or daily series that are not observed during weekends.

The purpose of this paper is to develop the ML estimator for a linear regression model with serially correlated errors when observations are missing. The results derived are generalizations of those by Beach and MacKinnon [1978]. Using both actual and simulated data we compare computational and statistical aspects of the ML estimator to those of some 'intuitive' estimators based on adaptations of suggestions by Cochrane and Orcutt [1949], Prais and Winsten [1954] and Maeshiro [1976, 1979].

In section 2, we present the model. In section 3, we present some results on the structure of the error covariance matrix and develop a convenient matrix notation which facilitates the algebraic derivations. Section 4 presents the ML estimator and the information matrix. In section 5, some alternative two-stage estimators are defined. In section 6, we present results of experiments designed to compare the computational and statistical properties of the ML and two-stage estimators. Section 7 concludes.

2. THE MODEL

Consider the single-equation regression model

$$y = X\beta + \varepsilon, \tag{1}$$

* Manuscript received March, 1981; revised August, 1984.

[1] We are grateful to the editor and the referees for their constructive comments, to H. van Terheijden, D. Bijlsma and C. Mijderwijk for extensive programming support and to A. ten Cate for comments and discussion. A first version was written while Kapteyn was with the University of Southern California, Los Angeles. Throughout the writing of the various versions, Wansbeek was with the Netherlands Central Bureau of Statistics (CBS). The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the CBS.

where ε is an $n \times 1$-vector of disturbances ε_t $(t=1,\dots, n)$, X is an $n \times k$-matrix of explanatory variables, β is a $k \times 1$-vector of parameters to be estimated, and y is an $n \times 1$-vector of dependent variables. With respect to ε the following assumptions are made

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t, \qquad |\rho| < 1,\ t = 1,\dots, n \tag{2}$$

where the vector $u \equiv (u_1,\dots, u_n)'$ is distributed as

$$u_t \sim N(0, \sigma_u^2 I_n). \tag{3}$$

Moreover, we postulate

$$\varepsilon_0 \sim N\left(0, \frac{\sigma_u^2}{1-\rho^2}\right), \tag{4}$$

i.e., the process is stationary.

So far, the model is standard. In this paper, we consider the case where observations are missing. This may arise for instance when the data on y and X are gathered at irregular time-intervals. Let there be m actual observations out of the n possible observations $(m \leq n)$. So $(n-m)$ observations are missing. We identify the m actual observations in terms of the n possible observations as follows. Let the rank number of the i-th actual observation in the original set of observations be n_i. By assumption, $n_1 = 1$ and $n_m = n$. We then define the $m \times n$ *deletion matrix* D as the matrix that is obtained by deleting from the unit matrix of order n those rows that correspond to the missing observations. Hence, the (i, n_i) elements of D are unity, the remaining elements being zero.

The model with missing observations can be written in terms of the original model (1) as

$$Dy = DX\beta + D\varepsilon. \tag{5}$$

We call equation (5) the *missing observations model*. Model (1) will be referred to as the 'standard model'. In the sequel, we shall denote vectors and matrices that only refer to non-missing observations by a star subscript. For example, equation (5) can be rewritten as

$$y_* = X_*\beta + \varepsilon_*. \tag{6}$$

3. SOME PROPERTIES OF THE MISSING OBSERVATIONS MODEL

It is well-known that the disturbances in the model (1) follow a multivariate normal distribution

$$\varepsilon \sim N(0, \sigma_\varepsilon^2 V), \tag{7}$$

with

(8)
$$V \equiv \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & & & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}$$

(9)
$$\sigma_\varepsilon^2 \equiv \frac{\sigma_u^2}{1-\rho^2}.$$

See, e.g., Theil [1971, p. 252]. It follows immediately that

(10)
$$\varepsilon_* \equiv D\varepsilon \sim N(0, \sigma_\varepsilon^2 DVD').$$

The $m \times m$-matrix $V_* \equiv DVD'$ has the following structure:

(11)
$$V_* = \begin{bmatrix} 1 & \rho^{n_2-n_1} & \rho^{n_3-n_1} & \cdots & \rho^{n_m-n_1} \\ \rho^{n_2-n_1} & 1 & \rho^{n_3-n_2} & \cdots & \rho^{n_m-n_2} \\ \rho^{n_3-n_1} & \rho^{n_3-n_2} & 1 & \cdots & \rho^{n_m-n_3} \\ \vdots & & & & \\ \rho^{n_m-n_1} & \rho^{n_m-n_2} & \rho^{n_m-n_3} & \cdots & 1 \end{bmatrix}.$$

For what follows, it is useful to introduce some more notation. Let $t_i \equiv n_i - n_{i-1}$ $(i=2,\ldots, m)$, so when no observations are missing, all t_i are equal to one. Then we define

(12)
$$Q \equiv \begin{bmatrix} 1 & & & & & \\ -\rho^{t_2} & 1 & & & & \\ & -\rho^{t_3} & 1 & & & \\ & & & \ddots & & \\ & & & & 1 & \\ & & & & -\rho^{t_m} & 1 \end{bmatrix}$$

(13)
$$\Delta \equiv \operatorname{diag}(1, 1-\rho^{2t_2},\ldots, 1-\rho^{2t_m}).$$

Hence,

(14)
$$Q^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \rho^{n_2-n_1} & 1 & 0 & 0 \\ \rho^{n_3-n_1} & \rho^{n_3-n_2} & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n_m-n_1} & \rho^{n_m-n_2} & \cdots & 1 \end{bmatrix}$$

(15)
$$\Delta = Q' + Q - QQ'$$

(16)
$$V_* = Q^{-1} + (Q^{-1})' - I_m$$

as is easily verified. As a result of (15) and (16),

(17) $$V_* = Q^{-1} + (Q^{-1})' - I_m = Q^{-1}(Q'+Q-QQ')(Q')^{-1} = Q^{-1}\Delta(Q')^{-1}$$

So

(18) $$|V_*| = |Q^{-1}|\,|\Delta|\,|(Q')^{-1}| = \prod_{i=2}^{m}(1-\rho^{2t_i})$$

and

(19) $$V_*^{-1} = Q'\Delta^{-1}Q = (\Delta^{-\frac{1}{2}}Q)'(\Delta^{-\frac{1}{2}}Q),$$

where the matrix $\Delta^{-\frac{1}{2}}Q$ has the structure

(20) $$\Delta^{-\frac{1}{2}}Q = \begin{bmatrix} 1 & 0 & 0 & \cdots 0 & 0 \\ -\frac{\rho^{t_2}}{(1-\rho^{2t_2})^{\frac{1}{2}}} & \frac{1}{(1-\rho^{2t_2})^{\frac{1}{2}}} & 0 & \cdots 0 & 0 \\ 0 & -\frac{\rho^{t_3}}{(1-\rho^{2t_3})^{\frac{1}{2}}} & \frac{1}{(1-\rho^{2t_3})^{\frac{1}{2}}} & \cdots 0 & 0 \\ \vdots & \vdots & \vdots & \ddots \vdots & \vdots \\ 0 & 0 & 0 & \cdots \frac{1}{(1-\rho^{2t_{m-1}})^{\frac{1}{2}}} & 0 \\ 0 & 0 & 0 & \cdots -\frac{\rho^{t_m}}{(1-\rho^{2t_m})^{\frac{1}{2}}} & \frac{1}{(1-\rho^{2t_m})^{\frac{1}{2}}} \end{bmatrix}$$

When ρ is known, applying OLS to the transformed model

(21) $$\Delta^{-\frac{1}{2}}Qy_* = \Delta^{-\frac{1}{2}}QX_*\beta + \Delta^{-\frac{1}{2}}Q\varepsilon_*$$

amounts to applying GLS to (6). The transformation leaves the first observation as it is. The other observations ($i=2,\ldots, m$) are transformed as follows:

(22) $$(1-\rho^{2t_i})^{-\frac{1}{2}}(y_{*i}-\rho^{t_i}y_{*i-1}) = (1-\rho^{2t_i})^{-\frac{1}{2}}\left\{\sum_{j=1}^{k}\beta_j(x_{*ij}-\rho^{t_i}x_{*i-1,j}) + \varepsilon_{*i} - \rho^{t_i}\varepsilon_{*i-1}\right\},$$

in obvious notation. For the case of a single gap in the data, this transformation (apart from a minor error) is also given by Dhrymes [1978]. If there are no missing observations (all t_i are equal to one), (21) and (22) reduce to the familiar transformation due to Prais and Winsten [1954] (see, e.g., Park and Mitchell [1980]).

There is a interesting interpretation of (22).[2] An error ε_{*i} in the set of actual observations satisfies

$$(23)\qquad \varepsilon_{*i} = \varepsilon_{n_i} = \rho^{t_i}\varepsilon_{n_{i-1}} + \rho^{t_i-1}u_{n_{i-1}+1} + \rho^{t_i-2}u_{n_{i-1}+2} + \cdots + \\ = \rho u_{n_{i-1}+(t_i-1)} + u_{n_{i-1}+t_i}\,.$$

Transformation (22) accomplishes two adjustments: autocorrelation adjustment and heteroskedasticity adjustment. The autocorrelation adjustment is

$$(24)\qquad \varepsilon_{*i} - \rho^{t_i}\varepsilon_{*i-1} = \varepsilon_{n_i} - \rho^{t_i}\varepsilon_{n_{i-1}} = \rho^{t_i-1}u_{n_{i-1}+1} + \cdots + u_{n_i}.$$

The heteroskedasticity adjustment stems from the fact that

$$(25)\qquad E(\varepsilon_{*i} - \rho^{t_i}{}_{*i-1})^2 = \sigma_u^2(1+\rho^2+\rho^4+\cdots+\rho^{2(t_i-1)}) = \frac{\sigma_u^2}{1-\rho^2}(1-\rho^{2t_i}).$$

So, dividing the i-th observation by $(1-\rho^{2t_i})^{\frac{1}{2}}$, for all $i \geq 2$, yields homoskedastic error terms with variance $\sigma_u^2/(1-\rho^2)$. This is also the variance of ε_1.

4. ML ESTIMATION

The log-likelihood corresponding to the model given in section 2 is given by

$$(26)\qquad \ln L = -\frac{1}{2}m\ln(2\pi\sigma_\varepsilon^2) - \frac{1}{2}\sum_{i=2}^{m}\ln(1-\rho^{2t_i}) \\ -\frac{1}{2\sigma_\varepsilon^2}\{\varepsilon_{*1}^2 + \sum_{i=2}^{m}(1-\rho^{2t_i})^{-1}(\varepsilon_{*i}-\rho^{t_i}\varepsilon_{*i-1})^2\},$$

with $\varepsilon_* \equiv y_* - X_*\beta$ (cf. (6)). Using results obtained by Magnus [1978], we show in appendix A that the first order conditions for a maximum of $\ln L$ with respect to β, σ_ε^2 and ρ are given by:

$$(27)\qquad \hat\beta = (X_*' V_*^{-1} X_*)^{-1} X_*' V_*^{-1} y$$

$$(28)\qquad \hat\sigma_\varepsilon^2 = \frac{1}{m}\{e_1^2 + \sum_{i=2}^{m}(1-\hat\rho^{2t_i})^{-1}(e_i - \hat\rho^{t_i}e_{i-1})^2\}$$

$$(29)\qquad \hat\sigma_\varepsilon^2 \sum_{i=2}^{m}(1-\hat\rho^{2t_i})^{-1}t_i\hat\rho^{2t_i-1} \\ = \sum_{i=2}^{m}(1-\hat\rho^{2t_i})^{-2}t_i\hat\rho^{t_i-1}(e_i-\hat\rho^{t_i}e_{i-1})(\hat\rho^{t_i}e_i - e_{i-1}),$$

where carets denote ML-estimates and $e \equiv y_* - X_*\hat\beta$. (Consistent notation would have e_* rather than e, but this would unnecessarily complicate the various expressions.) If (27)–(29) yield multiple roots, the roots that maximize $\ln L$ have to be chosen. For values of ρ, β and σ_ε^2 satisfying the first-order conditions, the last term of (26) becomes a constant.

[2] Due to a referee.

The information matrix I, of $\hat{\beta}$, $\hat{\rho}$ and $\hat{\sigma}_\varepsilon^2$ is derived in appendix B as

$$(30)\quad I=\begin{bmatrix} \frac{1}{\sigma_\varepsilon^2}X_*'V_*^{-1}X_* & 0 & 0 \\ 0 & \sum_{i=2}^{m}\frac{1+\rho^{2t_i}}{(1-\rho^{2t_i})^2}t_i^2\rho^{2t_i-2} & -\frac{1}{\sigma_\varepsilon^2}\sum_{i=2}^{m}\frac{t_i\rho^{2t_i-1}}{1-\rho^{2t_i}} \\ 0 & -\frac{1}{\sigma_\varepsilon^2}\sum_{i=2}^{m}\frac{t_i\rho^{2t_i-1}}{1-\rho^{2t_i}} & \frac{m}{2\sigma_\varepsilon^4} \end{bmatrix}$$

As usual, the inverse of this matrix can be taken as an approximation of the covariance matrix of the ML-estimators of the parameters β, ρ and σ_ε^2.

5. DISCUSSION

In this section, we make some general comments on the structure of the first-order conditions and their usefulness for computing a maximum of the likelihood. We also define some alternative 'intuitive' estimators. In section 6, we will compare the statistical and computational properties of these estimators.

To obtain some more insight into the structure of (29) we rewrite it somewhat. Define

$$(31)\qquad T \equiv \max_i \{t_i\}.$$

Denote the set $\{2 \leq i \leq m | t_i = j\}$ by I_j, and p_j, q_j and r_j by

$$(32)\qquad p_j \equiv n_j^{-1} \sum_{i\in I_j} e_i e_{i-1}$$

$$(33)\qquad q_j \equiv n_j^{-1} \sum_{i\in I_j} e_{i-1}^2$$

$$(34)\qquad r_j \equiv n_j^{-1} \sum_{i\in I_j} e_i^2, \qquad j = 1,\ldots, T$$

where n_j is the number of elements of I_j. Obviously, p_j, q_j and r_j are sample moments of residuals corresponding to equal values of t_i. Using the definitions, (29) can be written as

$$(35)\qquad \hat{\sigma}_\varepsilon^2 \sum_{j=1}^{T} n_j(1-\hat{\rho}^{2j})^{-1} j\hat{\rho}^{2j-1} = \sum_{j=1}^{T} n_j(1-\hat{\rho}^{2j})^{-2} j\hat{\rho}^{j-1}(\hat{\rho}^j r_j - p_j - \hat{\rho}^{2j} p_j + \hat{\rho}^j q_j),$$

or

$$(36)\qquad \sum_{j=1}^{T} n_j(1-\hat{\rho}^{2j})^{-2} j\hat{\rho}^{j-1}[-\hat{\sigma}_\varepsilon^2\hat{\rho}^{3j} + p_j\hat{\rho}^{2j} + (\hat{\sigma}_\varepsilon^2 - r_j - q_j)\hat{\rho}^j + p_j] = 0.$$

As an example, consider daily data that are collected on all days except Saturdays and Sundays. Let the first observation be made on a Monday. Then we have

$t_2=t_3=t_4=t_5=1$, $t_6=3$, $t_7=t_8=t_9=t_{10}=1$, $t_{11}=3$, etc. (It is implicitly assumed here that the data generation process does work on Saturdays and Sundays, but that the data are not observed.) If we collect data for 52 weeks, (36) becomes ($n_1=4\times 52=208$, $n_3=51$):

$$\frac{208}{(1-\hat{\rho}^2)^2}[-\hat{\sigma}_\varepsilon^2\hat{\rho}^3+p_1\hat{\rho}^2+(\hat{\sigma}_\varepsilon^2-r_1-q_1)\hat{\rho}+p_1] + \frac{153\hat{\rho}^2}{(1-\hat{\rho}^6)^2}[-\hat{\sigma}_\varepsilon^2\hat{\rho}^9+p_3\hat{\rho}^6+(\hat{\sigma}_\varepsilon^2-r_3-q_3)\hat{\rho}^3+p_3] = 0. \tag{37}$$

After multiplication by $(1-\hat{\rho}^2)^2(1-\hat{\rho}^6)^2$ this becomes a polynomial equation of degree 15. If, for instance, data are only collected on Mondays, Tuesdays, Thursdays and Fridays, the degree of the polynomial is 23.

In general, the degree of (36) is at most equal to $2T(T+1)-1$. For given $\hat{\sigma}_\varepsilon^2$, $\hat{\beta}$, it is a polynomial equation in a single variable. If one has a computer program available which generates all roots in the $(-1, 1)$ interval, the following iterative procedure can be used to find a maximum of the likelihood. For given starting values of $\hat{\beta}$ and $\hat{\sigma}_\varepsilon^2$ calculate the roots of (36) in the $(-1, 1)$ interval. If there are multiple roots, pick the one that gives the highest value of the likelihood (cf. (26)). Use this value of ρ to calculate a new $\hat{\beta}$ and $\hat{\sigma}_\varepsilon^2$ from (27) and (28) and solve (36) again, and so forth until convergence. As in the standard model without missing observations, the value of the likelihood increases at each step, so eventually it will come arbitrarily close to a maximum (cf. Oberhofer and Kmenta [1974], Sargan [1964]. This maximum need not be a global maximum, however.

It appears that a computer program which generates all roots of a polynomial in a given interval is not generally available. Programs that calculate *all* roots of a polynomial are more widely available. This, of course, may lead to function evaluations outside the $(-1, 1)$ interval. If the degree of (36) is high, overflow in the computer may be the result.

Still another possibility is to use a general purpose computer program to find a maximum of a function in a given interval. This, of course, ignores the information contained in the first order conditions (27)–(29). As an alternative, one can do a grid search for ρ in the $(-1, 1)$ interval and compute $\hat{\beta}$, $\hat{\sigma}_\varepsilon^2$ and the value of the likelihood for each ρ value. If the grid is fine enough, one can be almost certain that a global maximum of the likelihood is obtained. Finally, one can solve (29) by using a general purpose computer program to find a root of nonlinear equations in a given interval. Computing this root for given values of $\hat{\beta}$ and $\hat{\sigma}_\varepsilon^2$ and next updating $\hat{\beta}$ and $\hat{\sigma}_\varepsilon^2$ gives an iterative procedure which, upon convergence, provides a solution of (27)–(29). In section 6, we report our computational experience with the various procedures described here, except the first one since we do not have an adequate computer program to find roots of a polynomial in a given interval.

Although the favorable asymptotic properties of ML are well enough known, it is important to compare its finite sample properties to those of other estimators.

To the extent that ML does better in finite samples than other estimators, it is important to know whether the difference is worth the extra computational complexity of ML. In section 6, we shall compare ML to seven two-step estimators. For each of the seven estimators, the first step consists of OLS in model (6). Next, an estimate of ρ is obtained from these residuals. Finally, this ρ is used to transform the model so that OLS is appropriate. Some more details follow:

1. ρ is estimated as the OLS-estimate of the coefficient of the regression of $\tilde{e}_i$ on $\tilde{e}_{i-1}$ for those $i \geq 2$ where $t_i = 1$ (i.e. there is no gap between observations i and $i-1$), and where $\tilde{e}_i$, $\tilde{e}_{i-1}$ are OLS-residuals. This is a straightforward generalization of the Cochrane-Orcutt procedure. Using the estimate of ρ, the data is transformed according to (22), but only those observations for which $t_i = 1$. The other observations, the first one and the first observation after each gap, are omitted. Then β and σ^2 are estimated by OLS on the transformed data. This is, once again, a straightforward generalization of the Cochrane-Orcutt procedure. We call this estimator COCO.
2. The second estimation method uses the same estimate of ρ, but transforms all data, except the first observation, according to (22). Then β and σ_ε^2 are estimated by OLS on the transformed data (including the first observation). Since the transformation (21)–(22) is a generalization of the Prais-Winsten procedure we denote this estimator as COPW.
3. ρ is estimated analogous to the procedure in 1 but in the denominator of the least squares formula we omit the first term. This estimation method generalizes Prais-Winsten (cf. Park and Mitchell [1980, eq. (9b)]). This estimate of ρ is used to transform the data as with the first estimator. We call the estimator PWCO.
4. ρ is estimated as under 3 and the data is transformed as with the second estimator. This estimator is denoted by PWPW.
5. A two-step ML method: First, ρ is set at zero and β is estimated by OLS. Next, σ_ε^2 is estimated from (28) with $\rho = 0$ and (29) is used to estimate ρ. With this estimate of ρ, β in (27) and σ^2 in (28) are reestimated. This method, which produces asymptotically efficient estimators for β and σ^2, is denoted as ML2.
6. As the COCO-method, but the first observation is retained when estimating β and σ^2. So only the first observation after a gap is omitted. Since this method focuses on the importance of retaining the first observation, a point made repeatedly by Maeshiro [1976, 1979], we denote this method by COMA.
7. Analogously we also employed the PWMA method, whose description is clear from its name.

6. THE EXPERIMENTS AND THE RESULTS

Three sets of experiments have been performed. Within each set, experiments have been performed 27 times: both on a 'complete' data set (i.e. with no missing

Rank number of deleted observations

Pattern	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Number of retained observations
A	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	19
B	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	19
C	○	○	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	19
D	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	●	○	19
E	○	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	18
F	○	○	○	○	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	18
G	○	○	○	○	○	○	○	○	○	●	●	○	○	○	○	○	○	○	○	○	18
H	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	●	●	○	18
I	○	●	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	18
J	○	●	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	18
K	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	●	○	18
L	○	●	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	17
M	○	○	○	○	○	○	○	○	○	●	●	●	○	○	○	○	○	○	○	○	17
N	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	●	●	●	○	17
O	○	●	○	○	○	○	○	○	○	●	○	○	○	○	●	○	○	○	○	○	17
P	○	●	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	●	○	17
Q	○	●	●	○	○	○	○	○	○	○	○	●	●	○	○	○	○	○	○	○	16
R	○	○	○	○	○	○	○	○	○	●	●	○	○	○	○	○	○	●	●	○	16
S	○	●	●	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	16
T	○	○	○	○	○	○	○	○	○	●	●	●	●	○	○	○	○	○	○	○	16
U	○	●	●	●	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	15
V	○	○	○	○	○	○	○	○	○	●	●	●	●	●	○	○	○	○	○	○	15
W	○	●	●	○	○	○	○	○	○	○	○	●	●	●	○	○	○	○	○	○	15
X	○	○	○	○	○	○	○	●	●	●	●	●	●	○	○	○	○	○	○	○	14
Y	○	●	○	●	○	●	○	●	○	●	○	●	○	●	○	●	○	●	○	○	11
Z	○	●	○	●	○	●	○	●	○	●	○	●	○	●	○	●	○	●	○	●	10

FIGURE 1

PATTERNS OF DELETED OBSERVATIONS

observations), and on data that are obtained from the complete set by deleting observations according to 26 different patterns. These patterns are defined in Figure 1.

The first set of experiments deals with real-life data, consisting of ten sets of time-series for twenty years. This set has been mainly used to assess the computational burden of the various methods for the different patterns of missing observations. The second set deals with simulated data. Here we pay explicit attention to the differences in results between trended and non-trended data. The third set further explores properties of estimators in the context of trended data, employing a real-life trended independent variable and a simulated dependent variable.

6.1. *Computational burden.* The first set of experiments concerns the so-called Grunfeld data (Maddala [1977, table 10–4]). These data consist of annual observations from 1935 through 1954 for 10 large U. S. companies of the following variables: Gross investment (I_t), Value of the firm (F_t) and Stock of plant and equipment (C_t). Annual investment of a firm is explained by the following model:

$$I_t = \beta_0 + \beta_1 F_{t-1} + \beta_2 C_{t-1} + \varepsilon_t . \tag{38}$$

We allow for serial correlation in the ε_t according to equation (2). Model (38) is estimated for each of the ten companies by means of ML and the two-step estimation methods defined at the end of section 5. The estimations were repeated for 26 different patterns of missing observations, apart from COMA and PWMA.[3]

Table 1 gives an overview of the computational burden of the various methods for the different patterns. Comparing the methods, the five two-step methods are about four times faster than the cheapest ML-method, optimization using the first-order conditions. As to ML, using the first-order conditions saves roughly a third in computer time compared to direct optimization. Grid search is many times more expensive, although it can be sped up by requiring less than the four-decimal accuracy used here.

Over the patterns, the two-step estimators become gradually somewhat cheaper as the number of 'holes' increases, i.e. as the amount of data to be processed decreases. The same holds for grid search ML. ML2 tends to become somewhat more expensive as the polynomial equation becomes more complicated. The cost of the remaining two ML approaches does not show a clear relation with the patterns.

6.2. *Simulated data.* To provide more insight into the finite sample statistical properties of the estimators, we present results of simulations, which are variations on the simulations carried out by Beach and MacKinnon [1978]. The model considered is the following

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t,\ \varepsilon_t = \rho\varepsilon_{t-1} + u_t,\ u_t \sim \mathrm{NID}(0,\ 0.0036). \tag{39}$$

Two kinds of x_t-series are generated. One is a trending series generated according to

$$x_t = \exp(0.04t) + w_t,\ w_t \sim \mathrm{NID}(0,\ 0.0009). \tag{40}$$

The second one is a non-trending series generated according to

$$x_t \sim \mathrm{NID}(0,\ 0.0625). \tag{41}$$

We consider three values of ρ: 0.8, 0.6 and -0.8, and two sample sizes: 20 and 60. For sample size 20, we delete observations according to the patterns defined in

[3] These estimators were added later on suggestion of a referee. To save computer costs, we did not repeat all simulations with these estimators. For the present experiment, for example, it is clear that the computational burdens of COMA and PWMA will be similar to these of COCO and PWPW.

TABLE 1

COMPARISON OF METHODS

Pattern	Computational burden[a)]							
	ML[b)]	ML[c)]	ML[d)]	COCO	COPW	PWCO	PWPW	ML2
complete	690	74	49	15	15	15	15	15
A	653	63	45	14	14	14	15	16
B	655	64	45	14	15	14	15	15
C	658	65	47	14	14	14	15	16
D	656	70	49	14	15	14	14	15
E	625	62	45	13	14	13	14	17
F	629	61	47	13	14	13	14	16
G	621	70	49	13	14	14	14	17
H	621	68	49	14	14	13	14	16
I	634	54	39	13	14	13	14	15
J	634	58	42	13	14	13	14	16
K	631	67	45	13	14	13	14	15
L	600	62	45	13	13	13	13	17
M	600	69	49	13	14	13	14	17
N	603	62	45	13	14	13	13	17
O	608	56	40	12	14	12	13	15
P	604	63	45	12	13	12	13	15
Q	581	59	43	12	13	12	13	16
R	577	69	48	12	13	12	13	16
S	576	59	46	12	13	13	13	18
T	577	91	64	12	13	12	13	18
U	550	66	49	12	13	12	13	20
V	550	71	49	12	12	12	13	20
W	538	54	42	11	12	11	12	17
X	527	147	101	11	12	11	12	20
Y	442	43	31	e)	e)	e)	e)	13
Z	415	—	—	e)	e)	e)	e)	—

a) Measured in tens of milli-seconds on an ICL 2966. All programs are written in ALGOL 68. The entries are averages over the 10 companies.

b) Grid search method; ρ is increased in steps of 0.1 from -0.95 to 0.95 and for each value of ρ the value of the likelihood is computed. Let r be the value which gives the highest likelihood, a new search is then started in the interval $[r-0.10, r+0.10]$ etc. until an accuracy of 4 decimal places is obtained.

e) Direct maximization of the likelihood. We used the E04 ABF routine from NAG, adapted for use in ALGOL 68, which employs the 'safeguarded quadratic-interpolation method' of Gill and Murray [1973].

d) Maximization of the likelihood using first-order conditions. The C05 ADF routine from NAG (adapted for use in ALGOL 68) was used to find a solution for (36) in the interval $(-1, 1)$. This routine is based on a procedure due to Bus and Dekker [1975].

e) These estimation methods are not defined for pattern Z. PWCO and PWPW are not defined for pattern Y either, whereas COCO and COPW would estimate ρ on the basis of one observation. Thus, we do not report results for any of these four methods for patterns Y and Z.

TABLE 2

RMSE'S AND MEANS OF ESTIMATORS

Pattern		Trending							
		ML		COCO COPW COMA		PWCO PWPW PWMA		ML2	
		mean	RMSE	mean	RMSE	mean	RMSE	mean	RMSE
	complete	49	38	46	40	51	38	47	39
	C	48	40	45	42	49	39	45	41
	G	47	41	43	44	48	41	45	42
	J	46	43	43	44	46	42	44	43
$\rho=0.8$	*M*	45	43	42	45	47	43	43	45
	P	45	46	41	46	45	44	43	45
	Q	44	45	39	48	42	47	40	47
	T	42	47	38	49	43	47	39	49
	V	40	50	36	52	42	50	37	51
	W	43	47	38	50	41	49	39	49
	complete	36	33	34	33	37	32	35	33
	C	34	34	32	35	35	34	33	35
	G	33	37	30	37	33	37	31	37
	J	32	37	31	37	33	37	31	38
$\rho=0.6$	*M*	32	38	30	39	33	38	30	39
	P	30	42	30	39	32	39	29	40
	Q	29	43	27	42	29	42	28	42
	T	28	42	26	43	29	42	27	42
	V	26	43	25	44	28	43	25	43
	W	29	43	27	43	28	43	27	43
	complete	−76	12	−75	13	−79	12	−75	13
	C	−76	12	−75	14	−80	13	−75	13
	G	−77	13	−75	14	−81	13	−75	13
	J	−76	13	−72	19	−75	17	−71	17
$\rho=-0.8$	*M*	−77	12	−74	15	−79	14	−75	13
	P	−75	17	−65	26	−69	24	−64	24
	Q	−76	14	−72	20	−77	19	−72	16
	T	−76	13	−75	15	−80	14	−74	14
	V	−76	14	−74	16	−80	16	−74	15
	W	−76	15	−71	22	−77	21	−72	18

table 1. For sample size 60, we consider two cases. In the first case, the patterns defined in table 1 are repeated three times. In the second case, the patterns of table 1 are 'stretched' by a factor of 3. So a gap of two becomes a gap of six, a string of 5 consecutive observations becomes a string of 15 consecutive observations, etc.

FOR ρ ($\times 100$), $N=20$

Non-trending								
ML		COCO COPW COMA		PWCO PWPW PWMA		ML2		Number of observations missing
mean	RMSE	mean	RMSE	mean	RMSE	mean	RMSE	
64	25	60	28	64	26	60	27	0
63	26	60	28	65	26	59	29	1
63	27	59	30	64	28	60	29	2
62	27	58	30	63	28	57	30	2
62	28	59	31	65	29	58	31	3
59	35	54	36	59	34	54	36	3
60	31	55	35	60	33	55	34	4
61	29	57	32	63	30	57	31	4
60	30	57	33	63	31	57	32	5
60	33	54	37	60	35	55	35	5
48	23	45	24	48	24	45	24	0
48	23	45	25	48	24	45	25	1
48	24	44	27	48	26	45	26	2
47	25	43	27	47	26	43	26	2
47	26	45	27	48	27	44	27	3
43	34	41	32	44	31	40	33	3
44	31	39	34	43	33	40	32	4
45	27	42	29	46	29	42	29	4
45	28	42	30	47	29	42	29	5
44	32	40	34	44	34	40	32	5
−75	13	−66	22	−70	20	−66	21	0
−74	15	−67	22	−71	19	−66	21	1
−75	15	−67	21	−72	19	−67	21	2
−75	15	−66	24	−69	22	−65	22	2
−74	16	−68	22	−72	19	−67	21	3
−74	17	−61	29	−64	28	−61	26	3
−73	21	−65	26	−70	24	−65	24	4
−73	20	−67	24	−72	21	−66	23	4
−73	21	−67	26	−72	24	−66	24	5
−74	21	−64	27	−70	26	−64	25	5

Some results for $N=20$ are given in table 2 for ρ, and in table 3 for β_2. To save space, we present only some selected patterns, and only means and RMSE's[4]

[4] A full set of tables with simulation results is available on request.

TABLE 3
RMSE'S OF ESTIMATORS

Pattern		Trending							
		ML	COCO	COPW	COMA	PWCO	PWPW	PWMA	ML2
complete		108	136	109	109	147	109	109	109
	C	108	133	109	110	142	109	111	109
$\rho=0.8$	M	109	134	110	111	146	110	112	110
	W	110	187	114	113	196	115	112	112
complete		77	93	77	77	97	77	77	77
	C	77	91	77	77	94	77	77	77
$\rho=0.6$	M	77	91	77	78	95	77	78	77
	W	78	116	79	80	118	79	81	78
complete		25	25	25	25	25	25	25	25
	C	25	25	25	25	25	25	25	25
$\rho=-0.8$	M	25	25	25	25	25	25	25	25
	W	29	29	28	29	29	29	30	28

a) Since all estimators are unbiased, the RMSE's are also standard errors. Given that the errors by the square root of $2\times(\text{s.e.})^4/100$, if s.e. is the entry we are concerned with. 108, the associated standard error $(\times 1000)=1000\times 0.14\times(0.108)^2=1.6$; for the south–

Each number presented is based on 100 replications.[5] The main impression from table 2 is that the different estimators for ρ have very similar small sample properties. Generally, ML exhibits the smallest RMSE very closely followed by the PW-estimator. Next comes ML2 and finally CO. All estimators are biased towards zero, especially for positive ρ and trending x_t, with the PW-estimator usually showing the smallest bias and CO the largest one. The smaller bias of PW is due to the omission of the first term in the denominator of the least squares formula (see the description of the PW-estimator in the preceding section), which increases its magnitude in absolute value. At the same time, this also increases its variance. As a result, ML tends to have a slightly smaller RMSE.

Bias and RMSE are largest for positive ρ and trending x_t. The case of a negative ρ and a non-trending x_t is the only instance where ML is markedly better than the other estimators. There is no discernable relation between the relative performance of the estimators and the pattern of missing observations. Of course, both bias and RMSE tend to increase when the number of observations left decreases.

The results for $N=60$ are very similar to the ones reported here and will there-

[5] Since the means reported here are based on the rather small number of 100 replications, the reported means are subject to some sample variability. The standard error associated with the means in table 2 are 0.02 or less. For the standard errors associated with the entries in table 2, see the footnote of that table.

FOR β_2 ($\times 1000$), $N=20$[a)]

Non-trending								Number of observations missing
ML	COCO	COPW	COMA	PWCO	PWPW	PWMA	ML2	
37	38	37	37	38	37	37	38	0
39	40	40	39	39	40	39	40	1
40	41	41	40	40	41	40	41	3
61	72	62	72	69	61	69	62	5
42	43	43	43	43	43	43	43	0
44	45	44	44	44	44	44	45	1
44	45	45	45	45	45	45	45	3
66	78	66	78	76	66	76	68	5
67	74	76	76	73	74	74	75	0
75	79	81	81	78	79	80	81	1
77	78	84	81	77	82	79	83	3
76	84	76	82	81	76	84	75	5

entries of the table are based on 100 replications, we can approximate their standard
This equals $\sqrt{2}/10$ times $(\text{s.e.})^2 = 0.14 \times (\text{s.e.})^2$. For example, for the north-west entry
west entry 29 it is 0.1.

fore not be presented. Naturally, for $N=60$ RMSE and bias are substantially smaller. For example, for the patterns considered in table 2, the bias in ρ is now generally 0.10 or less.

Table 3 makes it clear that for trended data and positive ρ it is very important to exploit the first observation, confirming Maeshiro's findings. The reason is that the first observation is treated differently from the other observations, which stretches the scatter of points through which the regression line is fitted. This is especially important when x_t is trending because the autoregressive transformation tends to reduce the variability of the other x_t (cf. Maeshiro [1980]). Maintaining the first observation is more important in this case than maintaining the first observation after each gap, so that for trending x_t and $\rho=0.8$ or $\rho=0.6$, ML, COPW, PWPW, PWMA, ML2 have a similar performance. For $\rho=-0.8$ and trending data, the data are stretched very thinly after the autoregressive transformation so that all estimators of β_2 are quite accurate (cf. Maeshiro [1976]).

For non-trending data, it is not the transformation that is very important but rather the number of observations used. The performance of COCO, PWCO, COMA and PWMA relative to the other estimators gets worse with an increase in the number of gaps, because these estimators neglect the information contained in the observation after each gap.

Some further insight can be gained by considering table 4, where RMSE's of estimators of β_2 are given of four selected estimators for $\rho=0.8$ and all patterns of missing observations. Notice that PWCO, PWPW and PWMA all use the same

TABLE 4
RMSE'S OF SELECTED ESTIMATORS OF β_2 ($\times 1000$) $\rho=0.8$, $N=20$

Pattern	Trending				Non-trending				Number of observ. missing	Number of gaps
	ML	PWCO	PWPW	PWMA	ML	PWCO	PWPW	PWMA		
complete	108	147	109	109	37	38	37	37	0	0
A	110	154	111	109	43	44	43	44	1	1
B	109	151	109	110	37	39	38	39	1	1
C	108	142	109	111	39	39	40	39	1	1
D	108	148	110	116	37	37	37	37	1	1
E	109	176	110	110	51	55	51	55	2	1
F	108	151	108	110	39	39	39	40	2	1
G	109	143	109	113	39	40	39	40	2	1
H	108	150	108	122	37	38	38	38	2	1
I	110	150	109	110	43	49	45	49	2	2
J	110	150	111	110	46	47	47	46	2	2
K	111	160	112	116	44	45	44	45	2	2
L	108	184	110	107	62	62	65	63	3	1
M	109	146	110	112	40	40	41	40	3	1
N	108	147	109	129	38	39	39	39	3	1
O	111	150	115	114	47	46	48	46	3	3
P	111	157	111	118	46	48	47	48	3	3
Q	110	196	114	112	59	67	60	68	4	2
R	109	150	108	129	40	41	40	41	4	2
S	107	180	108	109	61	60	64	62	4	1
T	109	147	109	112	41	42	41	42	4	1
U	106	200	108	110	61	62	64	64	5	1
V	109	148	111	114	42	42	42	42	5	1
W	110	196	115	112	61	69	61	69	5	2
X	110	153	110	113	45	46	45	45	6	1

estimator for ρ; ML has been added as a bench-mark.

Let us first consider the case of trending x_t. Obviously, PWCO is inferior to the other estimators, but its efficiency loss varies over patterns. For patterns *A*, *B*, *C*, *D* (one observation missing) the loss is smallest for *C*, where the tenth observation is missing. The reason why gaps at the end of the data series cause a greater efficiency loss for PWCO than gaps in the middle can be seen as follows. Let the data be trended according to $x_t = \exp(\alpha t)$. Then transformation (22) carries x_t over into $x_t(1-\rho/\exp(\alpha))/(1-\rho^2)^{\frac{1}{2}}$ if there is no gap between x_t and x_{t-1}; if there is a gap of one, then x_t becomes $x_t(1-\rho^2/\exp(2\alpha))/(1-\rho^4)^{\frac{1}{2}}$. The ratio of these two expressions equals $(1+\rho/\exp(\alpha))/(1+\rho^2)^{\frac{1}{2}}$ or, for ρ close to 1 and small α, roughly $\sqrt{2}$. When, for instance for pattern *D* (a gap at $t=19$), PWCO and PWPW are compared, an observation is neglected that is — after transformation — sizeably larger than the neighboring ones. This leads to a loss in efficiency. Of course, the same reasoning applies to a gap at $t=2$, but then we have in addition that ML, PWPW and PWMA treat the first observation differently, so that it moves even further away from the other observations.

This intuitive argument makes it also easier to understand why, of the patterns E, F, G and H, the efficiency loss of PWCO is large for E and H and smaller for G; why of I, J, K, the loss is largest for K; of L, M, N the smallest loss is for M; of O, P the largest loss is for P; of Q, R, S, T the smallest loss is for T; of U, V, W the smallest loss is for V.

Regarding PWMA, the preceding argument makes it clear that it will perform relatively bad if there are gaps at the end, i.e. for patterns D, H, K, N, P, R.

The case of non-trending x_t does not show much variation across estimators although the estimators that use all observations (ML, PWPW) have a slight edge over the estimators that ignore one or more observations. For the case of trending x_t, it is noteworthy that the RMSE's of the efficient methods (ML, PWPW) do not vary appreciably with the number of observations that remain. Evidently, it is not the number of data point that matters most, but rather their dispersion.

From the results so far it appears that ML and PWPW are performing very well in all cases, with ML2 and COPW following closely behind. For all other estimators (COCO, PWCO, COMA, PWMA) there are certain cases in which they are doing rather badly (COCO, PWCO) or not so good (COMA, PWMA). The COCO and PWCO estimators suffer from an extra problem. Sometimes the estimate of ρ does not lie in the interval $(-1, 1)$. The standard approach taken for that event is to set ρ equal to -0.99999 or 0.99999. In the case where ρ is equal to 0.99999, the Cochrane-Orcutt transformation turns the ones corresponding to the constant term practically into zeros. Consequently, β_1 is (almost) unidentified and its estimate may be (almost) any real number. As a result, the RMSE's of the COCO and PWCO estimates of β_1 are very large (between 10^2 and 10^3) for some patterns. In practice, this does not have to be too serious a problem as long as one is not interested in β_1, since one can simply apply the first difference transformation.

Finally, it is of importance to know whether the information matrix provides a useful approximation of the true standard errors of the estimates. It turns out that the approximation of the standard error of the estimates of ρ is generally very good: the means (over 100 replications) of the standard errors computed from the information matrix usually differ no more than 10% from the true standard errors. Of course, this is not too helpful, because the estimators of ρ are heavily biased. The approximations of the standard errors of β_2 are substantially worse: computed and true standard error may differ as much as 100%. Of course, this is a consequence of the often poor estimates of ρ, which are used to compute Ω.

6.3. *Combined real life-simulated data.* Given the importance of trending variables, a third set of experiments has been performed focusing on this type of data. The model is

(42) $$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \quad \varepsilon_t = \rho\varepsilon_{t-1} + u_t, \quad u_t \sim \text{NID}(0, 0.0036).$$

For x_t, the U. S. GNP data are taken, as in Maeshiro [1976, 1979] and in Park and Mitchell [1980] ($t = 1950, \ldots, 1969$). Again, 100 experiments were performed for all patterns and $\rho = 0.8$, 0.6 and -0.8. The results turn out to be very similar to those obtained with the simulated x_t where x_t is trending.

7. CONCLUSIONS

Of the eight estimators considered here (ML and the seven two-step estimators defined at the end of section 5), ML is the most complicated one, but also the most efficient one. However, the performance of PWPW is so close to that of ML that this simple two-step estimator will presumably be the preferred estimator for practical work.

As is shown most clearly in table 4, in the common situation where exogenous variables are trending and errors are positively correlated, missing data generally have a very minor effect on the efficiency of estimators. The information matrix appears to give a good approximation of the standard error of the estimate of ρ (but not of its RMSE) and a rather poor one of the standard error of the slope coefficient. These findings apply equally well to complete data as to data with some observations missing.

In conclusion, missing observations in a linear model with serially correlated errors do not create any great difficulties in addition to those already present in models for a complete set of observations.

Groningen University, The Netherlands
Tilburg University, The Netherlands

APPENDIX A

First Order Conditions for ML

We derive (27)–(29). A general treatment of ML estimation of the GLS model was given by Magnus [1978]. From his results, it follows that the first-order conditions for ML are:

(A.1) $$\hat{\beta} = (X'_*\hat{\Omega}^{-1}X_*)^{-1}X'_*\hat{\Omega}^{-1}y$$

(A.2) $$\operatorname{tr}\left(\frac{\partial\Omega^{-1}}{\partial\sigma_\varepsilon^2}\Omega\right)_{\sigma_\varepsilon^2=\hat{\sigma}_\varepsilon^2} = e'\left(\frac{\partial\Omega^{-1}}{\partial\sigma_\varepsilon^2}\right)_{\sigma_\varepsilon^2=\hat{\sigma}_\varepsilon^2} e$$

(A.3) $$\operatorname{tr}\left(\frac{\partial\Omega^{-1}}{\partial\rho}\Omega\right)_{\rho=\hat{\rho}} = e'\left(\frac{\partial\Omega^{-1}}{\partial\rho}\right)_{\rho=\hat{\rho}} e,$$

with $e \equiv y_* - X_*\hat{\beta}$ and $\Omega \equiv \sigma_\varepsilon^2 V_*$. Of course, (27) follows immediately from (A.1).

First consider (A.2). Since

$$\frac{\partial \Omega^{-1}}{\partial \sigma_\varepsilon^2} = -\sigma_\varepsilon^4 V_*^{-1}, \tag{A.4}$$

(A.2) reduces to

$$\operatorname{tr}(-\hat\sigma_\varepsilon^{-4}\hat V_*^{-1}\hat\sigma_\varepsilon^2 \hat V_*) = e'(-\hat\sigma_\varepsilon^{-4}\hat V_*^{-1})e. \tag{A.5}$$

Using

$$Qe = \begin{bmatrix} e_1 \\ e_2 - \rho^{t_2}e_1 \\ \vdots \\ e_m - \rho^{t_m}e_{m-1} \end{bmatrix}, \tag{A.6}$$

we can rewrite (A.5) as

$$\begin{aligned}\hat\sigma_\varepsilon^2 &= \frac{1}{m}e'\hat V_*^{-1}e = \frac{1}{m}e'\hat Q'\hat\Delta^{-1}\hat Q e \\ &= \frac{1}{m}\{e_1^2 + \sum_{i=2}^{m}(1-\hat\rho^{2t_i})^{-1}(e_i - \hat\rho^{t_i}e_{i-1})^2\},\end{aligned} \tag{A.7}$$

which is (28).

Now consider (A.3). As

$$\frac{\partial \Omega^{-1}}{\partial \rho}\Omega = \sigma_\varepsilon^{-2}\frac{\partial V_*^{-1}}{\partial \rho}\sigma_\varepsilon^2 V_*, \tag{A.8}$$

(A.3) reduces to

$$\operatorname{tr}\left(\frac{\partial V_*^{-1}}{\partial \rho}V_*\right)_{\rho=\hat\rho} = \hat\sigma_\varepsilon^{-2}e'\left(\frac{\partial V_*^{-1}}{\partial \rho}\right)_{\rho=\hat\rho}e. \tag{A.9}$$

In view of (19), there holds

$$\frac{\partial V_*^{-1}}{\partial \rho} = \frac{\partial Q'}{\partial \rho}\Delta^{-1}Q + Q'\frac{\partial \Delta^{-1}}{\partial \rho}Q + Q'\Delta^{-1}\frac{\partial \Omega}{\partial \rho} \tag{A.10}$$

$$\begin{aligned}\operatorname{tr}\left(\frac{\partial V_*^{-1}}{\partial \rho}V_*\right) &= \operatorname{tr}\left(\frac{\partial V_*^{-1}}{\partial \rho}Q^{-1}\Delta(Q')^{-1}\right) \\ &= \operatorname{tr}\left(\frac{\partial Q'}{\partial \rho}(Q')^{-1} + \frac{\partial \Delta^{-1}}{\partial \rho}\Delta + \frac{\partial Q}{\partial \rho}Q^{-1}\right) \\ &= 2\operatorname{tr}\left(\frac{\partial \Omega}{\partial \rho}Q^{-1}\right) + \operatorname{tr}\left(\frac{\partial \Delta^{-1}}{\partial \rho}\Delta\right).\end{aligned} \tag{A.11}$$

The first of these two terms equals zero, because Q^{-1} is lower triangular, and $\partial Q/\partial \rho$ has a zero diagonal and a zero upper triangle. The second term is

$$\begin{aligned}\operatorname{tr}\left(\frac{\partial \Delta^{-1}}{\partial \rho}\Delta\right) &= -\operatorname{tr}\left(\Delta^{-1}\frac{\partial \Delta}{\partial \rho}\Delta^{-1}\Delta\right) = -\operatorname{tr}\left(\Delta^{-1}\frac{\partial \Delta}{\partial \rho}\right) \\ &= 2\sum_{i=2}^{m}(1-\rho^{2t_i})^{-1}t_i\rho^{2t_i-1}.\end{aligned} \tag{A.12}$$

Putting $\rho = \hat{\rho}$ in this expression gives the LHS of (A.9) and hence of (A.3).

We next evaluate the RHS of (A.3) and (A.9). There holds, in view of (A.10):

$$\text{(A.13)} \qquad e' \frac{\partial V_*^{-1}}{\partial \rho} e = 2e'Q'\Delta^{-1} \frac{\partial Q}{\partial \rho} e + e'Q' \frac{\partial \Delta^{-1}}{\partial \rho} Qe = 2e'Q'\Delta^{-1} \frac{\partial Q}{\partial \rho} e - e'Q'\Delta^{-1} \frac{\partial \Delta}{\partial \rho} \Delta^{-1} Qe.$$

Since

$$\text{(A.14)} \qquad \frac{\partial Q}{\partial \rho} e = - \begin{bmatrix} 0 \\ t_2 \rho^{t_2 - 1} e_1 \\ \vdots \\ t_m \rho^{t_m - 1} e_{m-1} \end{bmatrix}$$

$$\text{(A.15)} \qquad \frac{\partial \Delta}{\partial \rho} = -2 \begin{bmatrix} 0 & & & \\ & t_2 \rho^{2t_2 - 1} & & \\ & & \ddots & \\ & & & t_m \rho^{2t_m - 1} \end{bmatrix},$$

(A.13) can be further written as

$$\text{(A.16)} \qquad e' \frac{\partial V_*^{-1}}{\partial \rho} e = -2 \sum_{i=2}^{m} (1-\rho^{2t_i})^{-1} t_i \rho^{t_i - 1} e_{i-1}(e_i - \rho^{t_i} e_{i-1}) + 2 \sum_{i=2}^{m} (1-\rho^{2t_i})^{-2} t_i \rho^{2t_i - 1} (e_i - \rho^{t_i} e_{i-1})^2 = 2 \sum_{i=2}^{m} (1-\rho^{2t_i})^{-2} t_i \rho^{t_i - 1} (e_i - \rho^{t_i} e_{i-1})(\rho^{t_i} e_i - e_{i-1}).$$

Putting $\rho = \hat{\rho}$ in this expression gives the RHS of (A.9) and hence of (A.3), apart from the factor $\hat{\sigma}_\varepsilon^2$. Combining (A.3), (A.9), (A.12) and (A.16) gives (29).

APPENDIX B

The Information Matrix

Let $\phi_1 \equiv \rho$ and $\phi_2 \equiv \sigma_\varepsilon^2$, and let $\Psi(2 \times 2)$ be a matrix with typical element

$$\text{(B.1)} \qquad \psi_{ij} = \operatorname{tr}\left(\frac{\partial \Omega^{-1}}{\partial \phi_i} \Omega \frac{\partial \Omega^{-1}}{\partial \phi_j} \Omega \right).$$

The information matrix I corresponding with the likelihood function is

$$\text{(B.2)} \qquad I = \begin{bmatrix} X_*' \Omega^{-1} X_* & 0 \\ 0 & \frac{1}{2} \Psi \end{bmatrix}$$

(Magnus [1978, p. 288]). It remains to evaluate Ψ. First, let $i = j = 2$. Then

(B.3)
$$\psi_{22} = \operatorname{tr}\Big(\frac{\partial \Omega^{-1}}{\partial \sigma_\varepsilon^2}\,\Omega\,\frac{\partial \Omega^{-1}}{\partial \sigma_\varepsilon^2}\,\Omega\Big)$$
$$= \operatorname{tr}\{\sigma_\varepsilon^{-4} V_*^{-1} \times \sigma_\varepsilon^2 V_* \times \sigma_\varepsilon^{-4} V_*^{-1} \times \sigma_\varepsilon^2 V_*\} = m\sigma_\varepsilon^{-4}.$$

Next, let $i=2, j=1$. Then

(B.4)
$$\psi_{21} = \operatorname{tr}\Big(\frac{\partial \Omega^{-1}}{\partial \sigma_\varepsilon^2}\,\Omega\,\frac{\partial \Omega^{-1}}{\partial \rho}\,\Omega\Big) = -\operatorname{tr}\Big(\sigma_\varepsilon^{-4} V_*^{-1} \times \sigma_\varepsilon^2 V_* \times \frac{\partial V_*^{-1}}{\partial \rho} \times V_*\Big)$$
$$= -\sigma_\varepsilon^{-2}\operatorname{tr}\Big(\frac{\partial V_*^{-1}}{\partial \rho} V_*\Big) = -2\sigma_\varepsilon^{-2}\sum_{i=2}^{m}(1-\rho^{2t_i})^{-1} t_i \rho^{2t_i-1},$$

using (A.11) and (A.12). Finally, consider the case $i=j=1$:

(B.5)
$$\psi_{11} = \operatorname{tr}\Big(\frac{\partial \Omega^{-1}}{\partial \rho}\,\Omega\,\frac{\partial \Omega^{-1}}{\partial \rho}\,\Omega\Big) = \operatorname{tr}\Big(\frac{\partial V_*^{-1}}{\partial \rho} V_* \frac{\partial V_*^{-1}}{\partial \rho} V_*\Big).$$

Insertion of $\partial V_*^{-1}/\partial \rho$ as given in (A.10) into (B.5) yields an expression which is the trace of a sum of nine matrices. Using the well-known properties $\operatorname{tr}(P) = \operatorname{tr}(P')$ and $\operatorname{tr}(AB) = \operatorname{tr}(BA)$, one easily obtains

(B.6)
$$\psi_{11} = 2\operatorname{tr}\Big(\frac{\partial Q}{\partial \rho} Q^{-1} \frac{\partial Q}{\partial \rho} Q^{-1}\Big) - 4\operatorname{tr}\Big(\frac{\partial Q}{\partial \rho} Q^{-1} \frac{\partial \Delta}{\partial \rho} \Delta^{-1}\Big)$$
$$+ 2\operatorname{tr}\Big(\frac{\partial Q}{\partial \rho} V_* \frac{\partial Q'}{\partial \rho} \Delta^{-1}\Big) + \operatorname{tr}\Big(\frac{\partial \Delta}{\partial \rho} \Delta^{-1} \frac{\partial \Delta}{\partial \rho} \Delta^{-1}\Big).$$

Of these four terms, the first two vanish since all elements of $\partial Q/\partial \rho$ are zero apart from those directly below its main diagonal, and since Q^{-1} is lower-triangular; hence their product is lower-triangular with zero elements on the main diagonal.

It remains to evaluate the third and fourth term. Let e_i denote an $m \times 1$-vector with a unit element in position i, the other elements being zero. Denote an $m \times 1$-vector of zero elements by 0_m. Then

(B.7)
$$\frac{\partial Q'}{\partial \rho} = -(0_m, t_2\rho^{t_2-1} \times e_1, \ldots, t_m\rho^{t_m-1} \times e_{m-1}),$$

and so

(B.8)
$$\Big[\frac{\partial Q}{\partial \rho} V_* \frac{\partial Q'}{\partial \rho}\Big]_{ii} = t_i^2 \rho^{2t_i-2} \times e_i' V_* e_i = t_i^2 \rho^{2t_i-2},$$

for $i=2,\ldots, m$; for $i=1$ the expression evidently vanishes. So the third term on the RHS of (B.6) equals:

(B.9)
$$2\sum_{i=2}^{m}(1-\rho^{2t_i})^{-1} t_i^2 \rho^{2t_i-2}.$$

The fourth term equals

(B.10)
$$4\sum_{i=2}^{m}(1-\rho^{2t_i})^{-2} t_i^2 \rho^{4t_i-2},$$

because $\partial \Delta / \partial \rho$ is a diagonal matrix with i-th diagonal element equal to $-2t_i\rho^{2t_i-1}$ for $i=2,\ldots, m$ (and equal to zero for $i=1$). Collecting (B.9) and (B.10) one gets:

$$\text{(B.11)} \qquad \psi_{11} = 2 \sum_{i=2}^{m} (1-\rho^{2t_i})^{-2} t_i^2 \rho^{2t_i-2} \{(1-\rho^{2t_i}) + 2\rho^{2t_i}\}$$

$$= 2 \sum_{i=2}^{m} (1-\rho^{2t_i})^{-2} (1+\rho^{2t_i}) t_i^2 \rho^{2t_i-2}.$$

Together, (B.3), (B.4) and (B.11) give the elements of Ψ, the lower right part of the information matrix.

REFERENCES

BEACH, C. M. AND J. G. MACKINNON, "A Maximum Likelihood Procedure for Regression with Autocorrelated Errors," *Econometrica*, 46 (January, 1978), 51–58.

BUS, J. C. P. AND T. J. DEKKER, "Two Efficient Algorithms with Guaranteed Convergence for Finding a Zero of a Function," *ACM Transactions on Mathematical Software*, 1 (December, 1975), 330–345.

COCHRANE, D. AND G. H. ORCUTT, "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms," *Journal of the American Statistical Assocation*, 44 (March, 1949), 32–61.

DHRYMES, P. J., *Introductory Econometrics* (New York: Springer, 1978).

GILL, P. E. AND W. MURRAY, "Safeguarded Steplength Algorithms for Optimization using Descent Methods," National Physical Laboratory Report, No. NAC 37, National Physical Laboratory (1973).

HONOHAN, P. AND C. MCCARTHY, "On the Use of Durbin-Watson Type Statistics where there are Missing Observations," *The Statistician*, 31 (1982), 149–152.

MADDALA, G. S., *Econometrics* (New York: McGraw-Hill, 1977).

MAESHIRO, A., "Autoregressive Transformation, Trended Independent Variables and Autocorrelated Disturbance Terms," *The Review of Economics and Statistics*, 58 (November, 1976), 497–500.

———, "On the Retention of the First Observations in Serial Correlation Adjustment of Regression Models," *International Economic Review*, 20 (February, 1979), 259–265.

———, "Autocorrelation and Trended Explanatory Variables: A Reply," *The Review of Economics and Statistics*, 62 (August, 1980), 487–489.

MAGNUS, J. R., "Maximum Likelihood Estimation of the GLS Model with Unknown Parameters in the Disturbance Covariance Matrix," *Journal of Econometrics*, 7 (June, 1978), 281–312.

OBERHOFER, W. AND J. KMENTA, "A General Procedure for Obtaining Maximum Likelihood Estimates in Generalized Regression Models," *Econometrica*, 42 (May, 1974), 579–590.

PARK, R. E. AND B. M. MITCHELL, "Estimating the Autocorrelated Error Model with Trended Data," *Journal of Econometrics*, 13 (June, 1980), 185–201.

PRAIS, G. J. AND C. B. WINSTEN, "Trend Estimators and Serial Correlation," Cowles Commission Discussion Paper No. 383, (1954).

RICHARDSON, S. M. AND K. J. WHITE, "The Power of Tests for Autocorrelation with Missing Observations," *Econometrica*, 47 (May, 1979), 785–788.

SARGAN, J. D., "Wages and Prices in the United Kingdom: A Study in Econometric Methodology," in, P. E. Hart, G. Mills and J. K. Whitaker, eds., *Econometric Analysis for National Economic Planning* (London: Butterworths, 1964), 25–63.

SAVIN, N. E. AND K. J. WHITE, "Testing for Autocorrelation with Missing Observations," *Econometrica*, 46 (January, 1978), 59–67.

THEIL, H., *Principles of Econometrics* (New York: Wiley, 1971).

Estimation in a Linear Model with Serially Correlated Errors When Observations are Missing

Tom Wansbeek; Arie Kapteyn

International Economic Review, Vol. 26, No. 2. (Jun., 1985), pp. 469-490.

Stable URL:

http://links.jstor.org/sici?sici=0020-6598%28198506%2926%3A2%3C469%3AEIALMW%3E2.0.CO%3B2-3

References

A Maximum Likelihood Procedure for Regression with Autocorrelated Errors

Charles M. Beach; James G. MacKinnon

Econometrica, Vol. 46, No. 1. (Jan., 1978), pp. 51-58.

Stable URL:

http://links.jstor.org/sici?sici=0012-9682%28197801%2946%3A1%3C51%3AAMLPFR%3E2.0.CO%3B2-0

On the Use of Durbin-Watson Type Statistics Where There are Missing Observations

Patrick Honohan; Colm McCarthy

The Statistician, Vol. 31, No. 2. (Jun., 1982), pp. 149-152.

Stable URL:

http://links.jstor.org/sici?sici=0039-0526%28198206%2931%3A2%3C149%3AOTUODT%3E2.0.CO%3B2-J

Autoregressive Transformation, Trended Independent Variables and Autocorrelated Disturbance Terms

Asatoshi Maeshiro

The Review of Economics and Statistics, Vol. 58, No. 4. (Nov., 1976), pp. 497-500.

Stable URL:

http://links.jstor.org/sici?sici=0034-6535%28197611%2958%3A4%3C497%3AATTIVA%3E2.0.CO%3B2-L

LINKED CITATIONS

On the Retention of the First Observations in Serial Correlation Adjustment of Regression Models

Asatoshi Maeshiro

International Economic Review, Vol. 20, No. 1. (Feb., 1979), pp. 259-265.

Stable URL:

http://links.jstor.org/sici?sici=0020-6598%28197902%2920%3A1%3C259%3AOTROTF%3E2.0.CO%3B2-S

Autocorrelation and Trended Explanatory Variables: A Reply

Asatoshi Maeshiro

The Review of Economics and Statistics, Vol. 62, No. 3. (Aug., 1980), pp. 487-489.

Stable URL:

http://links.jstor.org/sici?sici=0034-6535%28198008%2962%3A3%3C487%3AAATEVA%3E2.0.CO%3B2-K

A General Procedure for Obtaining Maximum Likelihood Estimates in Generalized Regression Models

W. Oberhofer; J. Kmenta

Econometrica, Vol. 42, No. 3. (May, 1974), pp. 579-590.

Stable URL:

http://links.jstor.org/sici?sici=0012-9682%28197405%2942%3A3%3C579%3AAGPFOM%3E2.0.CO%3B2-1

The Power of Tests for Autocorrelation with Missing Observations

Sally M. Richardson; Kenneth J. White

Econometrica, Vol. 47, No. 3. (May, 1979), pp. 785-788.

Stable URL:

http://links.jstor.org/sici?sici=0012-9682%28197905%2947%3A3%3C785%3ATPOTFA%3E2.0.CO%3B2-C

Testing for Autocorrelation with Missing Observations

N. E. Savin; Kenneth J. White

Econometrica, Vol. 46, No. 1. (Jan., 1978), pp. 59-67.

Stable URL:

http://links.jstor.org/sici?sici=0012-9682%28197801%2946%3A1%3C59%3ATFAWMO%3E2.0.CO%3B2-N