Estimation in a linear model with serially correlated errors when observations are missing
Wansbeek, T.J.; Kapteyn, A.J.

Publication date:
1984

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Wansbeek, T. J., & Kapteyn, A. J. (1984). Estimation in a linear model with serially correlated errors when observations are missing. (Research memorandum / Tilburg University, Department of Economics; Vol. FEW 139). Unknown Publisher.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
ESTIMATION IN A LINEAR MODEL WITH SERIALLY CORRELATED ERRORS WHEN OBSERVATIONS ARE MISSING

Tom Wansbeek *)
Arie Kapteyn *)

*) Netherlands Central Bureau of Statistics and Department of Econometrics, Tilburg University, respectively. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the Netherlands Central Bureau of Statistics. 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.

BPA no.: 13640-82-M1
January 10, 1984

Proj. M1-81-209
Second draft
ESTIMATION IN A LINEAR MODEL WITH SERIALLY CORRELATED ERRORS WHEN OBSERVATIONS ARE MISSING

Abstract

Time series data may be obtained at irregular intervals or observations may be missing. This paper deals with the estimation of the linear regression model with AR(1) errors on the basis of such data. The structure of the error covariance matrix is analyzed, and a simplifying transformation of the model is derived, which allows for OLS estimation. The MLE of the parameters is given and its asymptotic covariance matrix is established. The performance of ML vis-à-vis various two-stage estimation methods is assessed using both real and simulated data.
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 McKinnon (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:

\[ y = X\beta + \varepsilon, \quad (1) \]

where \( \varepsilon \) is an \( n \times 1 \)-vector of disturbances \( \varepsilon_t \) \( (t=1, \ldots, n) \), \( X \) is an \( n \times k \)-matrix of explanatory variables, \( \beta \) 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 \( \varepsilon \) the following assumptions are made:

\[
\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \quad |\rho| < 1, \ t=1, \ldots, n
\]  

(2)

where the vector \( u=(u_1, \ldots, u_n)' \) is distributed as

\[
u_t \sim N(0, \sigma^2_n)
\]

(3)

Moreover, we postulate

\[
\sigma^2_0 \sim N(0, \frac{\sigma^2_n}{1-\rho^2}),
\]

(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 \((m<n)\). 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:

\[
Dy = DXB + D\varepsilon.
\]

(5)

We call (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 (5) can be rewritten as
3. Some properties of the missing observations model

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

$$\varepsilon \sim N(0, \sigma^2_\varepsilon V),$$

with

$$V \equiv \begin{bmatrix}
1 & \rho & \rho^2 & \ldots & \rho^{n-1} \\
\rho & 1 & \rho & \ldots & \rho^{n-2} \\
\rho^2 & \rho & 1 & \ldots & \rho^{n-3} \\
& & & \ddots & \\
\rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \ldots & 1
\end{bmatrix}$$

$$\sigma^2_\varepsilon \equiv \frac{\sigma^2_\varepsilon}{1-\rho^2}.$$ 

See, e.g., Theil (1971, p. 252). It follows immediately that

$$\varepsilon_* \equiv \Delta \varepsilon \sim N(0, \sigma^2_\varepsilon V^D) .$$

The mxm-matrix $V_* \equiv V^D$ has the following structure:

$$V_* = \begin{bmatrix}
1 & n_2-n_1 & n_3-n_1 & \ldots & n_{m-1} \\
\rho & 1 & n_3-n_2 & \ldots & n_{m-2} \\
\rho & \rho & 1 & \ldots & n_{m-3} \\
& & & \ddots & \\
\rho & \rho & \rho & \rho & 1
\end{bmatrix}.$$
For what follows, it is useful to introduce some more notation. Let $t_i = n_i - n_{i-1}$ ($i=2, \ldots, m$), so when no observations are missing, all $t_i$ are equal to one. Then we define

$$ Q = \begin{bmatrix} 1 & t_2 & 0 & \cdots & 0 \\ -\rho & 1 & \ddots & \ddots & \vdots \\ 0 & -\rho & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & \cdots & -\rho & 1 \end{bmatrix} $$

(12)

$$ \Delta = \text{diag}(1, 1-\rho, \ldots, 1-\rho^m) $$

(13)

Hence:

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

(14)

$$ \Delta = Q' + Q - QQ' $$

(15)

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

(16)

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

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

(17)
So
\[ |v_\star| = |Q^{-1}||\Delta||Q'|^{-1}| = \prod_{i=2}^{m} (1 - \rho^{2t_i}) \]  \tag{18}
and
\[ v_\star^{-1} = Q'\Delta^{-1}Q = (\Delta^{-\frac{1}{2}}Q)'(\Delta^{-\frac{1}{2}}Q), \]  \tag{19}
where the matrix \( \Delta^{-\frac{1}{2}}Q \) has the structure

\[
\Delta^{-\frac{1}{2}}Q = \begin{bmatrix}
1 & 0 & 0 & \cdots & 0 & 0 \\
-\frac{\rho^{t_2}}{(1-\rho^{2t_2})^{\frac{1}{2}}} & 1 & 0 & \cdots & 0 & 0 \\
0 & -\frac{\rho^{t_3}}{(1-\rho^{3t_3})^{\frac{1}{2}}} & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & \frac{1}{(1-\rho^{m-1})^{\frac{1}{2}}} & 0 \\
0 & 0 & 0 & \cdots & -\frac{\rho^{t_m}}{(1-\rho^{mt_m})^{\frac{1}{2}}} & \frac{1}{(1-\rho^{mt_m})^{\frac{1}{2}}} \\
\end{bmatrix}
\]
\tag{20}

When \( \rho \) is known, applying OLS to the transformed model
\[ \Delta^{-\frac{1}{2}}Qy_\star = \Delta^{-\frac{1}{2}}QX_\star \beta + \Delta^{-\frac{1}{2}}Qe_\star \]  \tag{21}
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:
\[ (1 - \rho^t_{i-1})^{\frac{1}{2}} (y^*_{i-1} - \rho^t_i y^*_{i-1}) = \]

\[ = (1 - \rho^t_i)^{-\frac{1}{2}} \left\{ (1 - \rho^t_i)^{-\frac{1}{2}} \left( \sum_{j=1}^{k} \beta_j (x^*_{ij} - \rho^t_i x^*_{i-1,j}) + \varepsilon^*_{i-1} - \rho^t_i \varepsilon^*_{i-1} \right) \right\}, \quad (22) \]

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).\(^1\) An error \( \varepsilon^*_{ij} \) in the set of actual observations satisfies

\[ \varepsilon^*_{ij} = \rho^t_i \varepsilon^*_{ij-1} + \rho^t_i u_{i-1,1} + \rho^t_i u_{i-1,2} + \ldots + \rho^t_i u_{i-1,i} + u_{n_{i-1}}. \]

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

\[ \varepsilon^*_{ij} - \rho^t_i \varepsilon^*_{ij-1} = \varepsilon^*_{ij-1} - \rho^t_i \varepsilon^*_{ij-2} = \rho^t_i u_{i-1,1} + \ldots + u_{n_i}. \]

The heteroskedasticity adjustment stems from the fact that

\[ E(\varepsilon^*_{ij} - \rho^t_i \varepsilon^*_{ij-1})^2 = \sigma_u^2 (1 + \rho^2 + \rho^4 + \ldots + \rho^{2(t_i-1)}) = \frac{\sigma_u^2}{1 - \rho^2} (1 - \rho^{2t_i}). \]

So, dividing the \( i \)-th observation by \((1 - \rho^{2t_i})\), for all \( i \geq 2 \), yields homoskedastic error terms with variance \( \sigma_u^2/(1 - \rho^2) \). This is also the variance of \( \varepsilon^*_{1} \).
4. ML estimation

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

\[
\ln L = -\frac{1}{2}m \ln (2\pi \sigma^2_e) - \frac{1}{2} \sum_{i=2}^{m} \ln (1 - \rho^{-i}) - \frac{1}{2\sigma^2_e} \left( \epsilon^2_{x1} + \sum_{i=2}^{m} (1 - \rho^{-i})^{-1}(\epsilon_{x1} - \rho \epsilon_{x1-1})^2 \right),
\]

with \( \epsilon_{x1} \equiv y_{\ast} - X_{\ast} \beta \) (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 \( \beta, \sigma^2_e \) and \( \rho \) are given by:

\[
\hat{\beta} = (X'V_{\ast}^{-1}X_{\ast})^{-1}X'V_{\ast}^{-1}y
\]

\[
\hat{\sigma}^2_e = \frac{1}{m} \left( e_{1}^2 + \sum_{i=2}^{m} (1 - \rho^{-i})^{-1}(e_{1} - \rho e_{i-1})^2 \right)
\]

\[
\hat{\sigma}^2 = \sum_{i=2}^{m} (1 - \rho^{-i})^{-1}(e_{1} - \rho e_{i-1})^2 / \sum_{i=2}^{m} (1 - \rho^{-i})^{-1}
\]

where carets denote ML-estimates and \( \epsilon_{\ast} \equiv y_{\ast} - X_{\ast} \beta \). (Consistent notation would have \( \epsilon_{\ast} \) rather than \( \epsilon \), 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 \( \rho, \beta \) and \( \sigma^2_e \) satisfying the first order conditions, the last term of (26) becomes a constant.

The information matrix \( I \), of \( \hat{\beta}, \hat{\sigma}^2_e \) and \( \hat{\sigma}^2_e \) is derived in appendix B as

\[
I = \begin{bmatrix}
\frac{1}{\sigma^2_e} X'V_{\ast}^{-1}X_{\ast} & 0 \\
0 & \frac{1}{\sigma^2_e} X'V_{\ast}^{-1}X_{\ast}
\end{bmatrix}
\]

\[
I = \begin{bmatrix}
\frac{1}{\sigma^2_e} X'V_{\ast}^{-1}X_{\ast} & 0 \\
0 & \frac{1}{\sigma^2_e} X'V_{\ast}^{-1}X_{\ast}
\end{bmatrix}
\]

\[
I = \begin{bmatrix}
\frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho & \frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho \\
\frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho & \frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho
\end{bmatrix}
\]

\[
I = \begin{bmatrix}
\frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho & \frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho \\
\frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho & \frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho
\end{bmatrix}
\]

\[
I = \begin{bmatrix}
\frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho & \frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho \\
\frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho & \frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho
\end{bmatrix}
\]

\[
I = \begin{bmatrix}
\frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho & \frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho \\
\frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho & \frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho
\end{bmatrix}
\]

\[
I = \begin{bmatrix}
\frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho & \frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho \\
\frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho & \frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho
\end{bmatrix}
\]

\[
I = \begin{bmatrix}
\frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho & \frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho \\
\frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho & \frac{m}{\sigma^2_e} (1 + \rho) t_1^\rho
\end{bmatrix}
\]
As usual, the inverse of this matrix can be taken as an approximation of the covariance matrix of the ML-estimators of the parameters \( \beta \), \( \rho \) and \( \sigma^2 \).

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

\[ T = \max \{ t_i \} . \]  

(31)

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

\begin{align*}
  p_j & = n_j^{-1} \sum_{i \in I_j} e_i e_i^\prime, \\
  q_j & = n_j^{-1} \sum_{i \in I_j} e_i^2, \\
  r_j & = n_j^{-1} \sum_{i \in I_j} e_i^2, \\
\end{align*}

\[ j = 1, \ldots, T \]  

(32) \hspace{1cm} (33) \hspace{1cm} (34)

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

\begin{equation}
\begin{aligned}
  \hat{\sigma}^2 & = \sum_{j=1}^{T} n_j (1 - \hat{\rho}^{2j})^{-1} \hat{\rho}_{j}^{2j-1} = \\
  & = \sum_{j=1}^{T} n_j (1 - \hat{\rho}^{2j})^{-2} \hat{\rho}_{j}^{3j-1} (\hat{\rho}^{j} r_{j} - \hat{\rho}^{2j} p_{j} + \hat{\rho}^{j} q_{j}) .
\end{aligned}
\end{equation}

(35)
or
\[ T \sum_{j=1}^{T} n_j (1-\hat{\rho}^2)^{-2} \hat{\rho}^{-1} \left[ -\hat{\sigma}_e^2 \hat{\rho}^3 + p_j \hat{\rho}^2 \right] + (\hat{\sigma}_e^2 - r_j - q_j) \hat{\rho}^3 + p_j = 0. \] (36)

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 \( (n_1 = 4 \times 52 = 208, n_3 = 51) \):

\[ \frac{208}{(1-\hat{\rho}^2)^2} \left[ -\hat{\sigma}_e^2 \hat{\rho}^3 + p_1 \hat{\rho}^2 + (\hat{\sigma}_e^2 - r_1 - q_1) \hat{\rho} + p_1 \right] + \]
\[ + \frac{153 \hat{\rho}^2}{(1-\hat{\rho}^6)^2} \left[ -\hat{\sigma}_e^2 \hat{\rho}^9 + p_3 \hat{\rho}^6 + (\hat{\sigma}_e^2 - r_3 - q_3) \hat{\rho}^3 + p_3 \right] = 0. \] (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 \( \hat{\sigma}_e^2, \hat{\rho}, \) 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 \( \hat{\rho} \) and \( \hat{\sigma}_e^2 \), 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 \( \hat{\rho} \) to calculate a new \( \hat{\sigma}_e \) and \( \hat{\sigma}_e^2 \) 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 \( \rho \) in the (-1,1) interval and compute \( \hat{\beta}, \hat{\sigma}_\varepsilon^2 \) and the value of the likelihood for each \( \rho \) 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 \( \rho \) is obtained from these residuals. Finally, this \( \rho \) is used to transform the model so that OLS is appropriate. Some more details follow:

1. \( \rho \) is estimated as the OLS-estimate of the coefficient of the regression of \( \tilde{\varepsilon}_i \) on \( \tilde{\varepsilon}_{i-1} \) for those \( i \) where \( t_{i-1} = 1 \) (i.e. there is no gap between observations \( i \) and \( i-1 \)), and where \( \tilde{\varepsilon}_i, \tilde{\varepsilon}_{i-1} \) are OLS-residuals. This is a straightforward generalization of the Cochrane-Orcutt procedure. Using the estimate of \( \rho \), the data is transformed according to (22), but only those observations for which \( t_{i-1} = 1 \). The other observations, the first one and the first observation after each gap, are omitted. Then \( \hat{\beta} \) and \( \hat{\sigma}_\varepsilon^2 \) 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 \( \rho \), but transforms all data, except the first observation, according to (22). Then \( \beta \) and \( \sigma^2 \) 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. \( \rho \) 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 \( \rho \) is used to transform the data as with the first estimator. We call the estimator PWCO.

4. \( \rho \) 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, \( \rho \) is set at zero and \( \beta \) is estimated by OLS. Next \( \sigma^2 \) is estimated from (28) with \( \rho=0 \) and (29) is used to estimate \( \rho \). With this estimate of \( \rho \), \( \beta \) in (27) and \( \sigma^2 \) in (28) are reestimated. This method, which produces asymptotically efficient estimators for \( \beta \) en \( \sigma^2 \), is denoted as ML2.

6. As the COCO-method, but the first observation is retained when estimating \( \beta \) and \( \sigma^2 \). 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 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
Figure 1. Patterns of deleted observations

<table>
<thead>
<tr>
<th>Rank number of deleted observations</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>H</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>J</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>U</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>W</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Z</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
patterns of missing observations. The second set deals with simulated data. Here we pay explicit attention to the differences in result 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:

\[
I_t = \beta_0 + \beta_1 F_{t-1} + \beta_2 C_{t-1} + \epsilon_t .
\]  

(38)

We allow for serial correlation in the \(\epsilon_t\) 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.2)

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 of course 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.
Table 1. Comparison of methods

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Computational burden$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ML $^b$</td>
</tr>
<tr>
<td>complete</td>
<td>690</td>
</tr>
<tr>
<td>A</td>
<td>653</td>
</tr>
<tr>
<td>B</td>
<td>655</td>
</tr>
<tr>
<td>C</td>
<td>658</td>
</tr>
<tr>
<td>D</td>
<td>656</td>
</tr>
<tr>
<td>E</td>
<td>625</td>
</tr>
<tr>
<td>F</td>
<td>629</td>
</tr>
<tr>
<td>G</td>
<td>621</td>
</tr>
<tr>
<td>H</td>
<td>621</td>
</tr>
<tr>
<td>I</td>
<td>634</td>
</tr>
<tr>
<td>J</td>
<td>634</td>
</tr>
<tr>
<td>K</td>
<td>631</td>
</tr>
<tr>
<td>L</td>
<td>600</td>
</tr>
<tr>
<td>M</td>
<td>600</td>
</tr>
<tr>
<td>N</td>
<td>603</td>
</tr>
<tr>
<td>O</td>
<td>608</td>
</tr>
<tr>
<td>P</td>
<td>604</td>
</tr>
<tr>
<td>Q</td>
<td>581</td>
</tr>
<tr>
<td>R</td>
<td>577</td>
</tr>
<tr>
<td>S</td>
<td>576</td>
</tr>
<tr>
<td>T</td>
<td>577</td>
</tr>
<tr>
<td>U</td>
<td>550</td>
</tr>
<tr>
<td>V</td>
<td>550</td>
</tr>
<tr>
<td>W</td>
<td>538</td>
</tr>
<tr>
<td>X</td>
<td>527</td>
</tr>
<tr>
<td>Y</td>
<td>442</td>
</tr>
<tr>
<td>Z</td>
<td>415</td>
</tr>
</tbody>
</table>

$^a$ Measured in tens of milliseconds on an ICL 2966. All programs are written in ALGOL 68. The entries are averages over the 10 companies.

$^b$ Grid search method; $p$ is increased in steps of 0.1 from -0.95 to 0.95 and for each value of $p$ 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.

$^c$ Direct maximization of the likelihood. We used the EO4 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 $p$ on the basis of one observation. Thus we do not report results for any of these four methods for patterns Y and Z.
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:

\[ y_t = \beta_1 + \beta_2 x_t + \varepsilon_t, \quad \varepsilon_t = \rho \varepsilon_{t-1} + u_t, \quad u_t \sim NID(0,0.0036). \]

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

\[ x_t = \exp(0.04t) + w_t, \quad w_t \sim NID(0,0.0009). \]

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

\[ x_t \sim NID(0,0.0625). \]

We consider three values of \( \rho \): 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 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.

Some results for \( N=20 \) are given in table 2 for \( \rho \), and in table 3 for \( \beta_2 \). To save space we only present some selected patterns, and only means and RMSE's. Each number presented is based on 100 replications. The main impression from table 2 is that the different estimators for \( \rho \) 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 \( \rho \) 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...
<table>
<thead>
<tr>
<th>Pattern</th>
<th>Trending</th>
<th>Non-trending</th>
<th>Number of observations missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ML</td>
<td>COCO</td>
<td>PWCO</td>
</tr>
<tr>
<td>complete</td>
<td>49</td>
<td>38</td>
<td>46</td>
</tr>
<tr>
<td>C</td>
<td>48</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td>G</td>
<td>47</td>
<td>41</td>
<td>43</td>
</tr>
<tr>
<td>J</td>
<td>46</td>
<td>43</td>
<td>45</td>
</tr>
<tr>
<td>p=0.8</td>
<td>M</td>
<td>45</td>
<td>43</td>
</tr>
<tr>
<td>P</td>
<td>45</td>
<td>46</td>
<td>41</td>
</tr>
<tr>
<td>Q</td>
<td>44</td>
<td>45</td>
<td>39</td>
</tr>
<tr>
<td>T</td>
<td>42</td>
<td>47</td>
<td>38</td>
</tr>
<tr>
<td>V</td>
<td>40</td>
<td>50</td>
<td>36</td>
</tr>
<tr>
<td>W</td>
<td>43</td>
<td>47</td>
<td>38</td>
</tr>
<tr>
<td>complete</td>
<td>36</td>
<td>33</td>
<td>34</td>
</tr>
<tr>
<td>C</td>
<td>34</td>
<td>34</td>
<td>32</td>
</tr>
<tr>
<td>G</td>
<td>33</td>
<td>37</td>
<td>30</td>
</tr>
<tr>
<td>J</td>
<td>32</td>
<td>37</td>
<td>31</td>
</tr>
<tr>
<td>p=0.6</td>
<td>M</td>
<td>32</td>
<td>38</td>
</tr>
<tr>
<td>P</td>
<td>30</td>
<td>42</td>
<td>30</td>
</tr>
<tr>
<td>Q</td>
<td>29</td>
<td>43</td>
<td>27</td>
</tr>
<tr>
<td>T</td>
<td>28</td>
<td>42</td>
<td>26</td>
</tr>
<tr>
<td>V</td>
<td>26</td>
<td>43</td>
<td>25</td>
</tr>
<tr>
<td>W</td>
<td>29</td>
<td>43</td>
<td>27</td>
</tr>
<tr>
<td>complete</td>
<td>-76</td>
<td>12</td>
<td>-75</td>
</tr>
<tr>
<td>C</td>
<td>-76</td>
<td>12</td>
<td>-75</td>
</tr>
<tr>
<td>J</td>
<td>-76</td>
<td>13</td>
<td>-72</td>
</tr>
<tr>
<td>p=-0.8</td>
<td>M</td>
<td>-77</td>
<td>12</td>
</tr>
<tr>
<td>P</td>
<td>-75</td>
<td>17</td>
<td>-65</td>
</tr>
<tr>
<td>Q</td>
<td>-76</td>
<td>14</td>
<td>-72</td>
</tr>
<tr>
<td>T</td>
<td>-76</td>
<td>13</td>
<td>-75</td>
</tr>
<tr>
<td>W</td>
<td>-76</td>
<td>15</td>
<td>-71</td>
</tr>
</tbody>
</table>
Table 3. RMSE's of estimators for $\beta_2$ (×1000), N=20^a)

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Trending</th>
<th>Non-trending</th>
<th>Number of observations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ML COCO COPW COMA PWCO PWPW PWMA ML2</td>
<td>ML COCO COPW COMA PCO PWPW PWMA ML2</td>
<td>missing</td>
</tr>
<tr>
<td>complete</td>
<td>108 136 109 109 147 109 109 109</td>
<td>37 38 37 38 37 38 38 0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>108 133 109 110 142 109 111 109</td>
<td>39 40 40 39 40 40 40 1</td>
<td>1</td>
</tr>
<tr>
<td>$\rho=0.8$ M</td>
<td>109 134 110 111 146 110 112 110</td>
<td>40 41 41 40 41 41 41 3</td>
<td>3</td>
</tr>
<tr>
<td>W</td>
<td>110 187 114 113 196 115 112 112</td>
<td>61 72 62 69 61 69 62 5</td>
<td>5</td>
</tr>
<tr>
<td>complete</td>
<td>77 93 77 77 97 77 77 77</td>
<td>42 43 43 43 43 43 43 0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>77 91 77 77 94 77 77 77</td>
<td>44 45 44 44 44 44 44 1</td>
<td>1</td>
</tr>
<tr>
<td>$\rho=0.6$ M</td>
<td>77 91 77 78 95 77 78 77</td>
<td>44 45 45 45 45 45 45 3</td>
<td>3</td>
</tr>
<tr>
<td>W</td>
<td>78 116 79 80 118 79 81 78</td>
<td>66 78 66 78 66 66 68 5</td>
<td>5</td>
</tr>
<tr>
<td>complete</td>
<td>25 25 25 25 25 25 25 25</td>
<td>67 74 76 76 73 74 74 75</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>25 25 25 25 25 25 25 25</td>
<td>75 79 81 81 78 79 80 81</td>
<td>1</td>
</tr>
<tr>
<td>$\rho=-0.8$ M</td>
<td>25 25 25 25 25 25 25 25</td>
<td>77 78 84 81 77 82 79 83</td>
<td>3</td>
</tr>
<tr>
<td>W</td>
<td>29 29 28 29 29 29 30 28</td>
<td>76 84 76 82 81 76 84 75</td>
<td>5</td>
</tr>
</tbody>
</table>

^a) Since all estimators are unbiased, the RMSE's are also standard errors. Given that the entries of the table are based on 100 replications we can approximate their standard errors by the square root of $2\times(s.e.)^4/100$, if s.e. is the entry we are concerned with. This equals $\sqrt{2}/10$ times $(s.e.)^2=0.14\times(s.e.)^2$. For example, for the north-west entry 108, the associated standard error (×1000) = $100\times0.14\times(0.108)^2=1.6$; for the south-west entry 29 it is 0.1.
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 $p$ and trending $x_t$. The case of a negative $p$ 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 therefore 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 $p$ is now generally 0.10 or less.

Table 3 makes it clear that for trended data and positive $p$ 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 $p=0.8$ or $p=0.6$, ML, COPW, PWPW, PWMA, ML2 have a similar performance. For $p=-0.8$ and trending data, the data are stretched very thinly after the autoregressive transformation so that all estimators of $B_2$ 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 $\beta_2$ 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 estimator for $\rho$; 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

Table 4. RMSE's of selected estimators of $\beta_2$ ($\times1000$) $\rho=0.8$, $N=20$

<table>
<thead>
<tr>
<th>pattern</th>
<th>trending</th>
<th>non-trending</th>
<th>number of observ. missing</th>
<th>number of gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ML</td>
<td>PWCO</td>
<td>PWPW</td>
<td>PWMA</td>
</tr>
<tr>
<td>complete</td>
<td>108</td>
<td>147</td>
<td>109</td>
<td>109</td>
</tr>
<tr>
<td>A</td>
<td>110</td>
<td>154</td>
<td>111</td>
<td>109</td>
</tr>
<tr>
<td>B</td>
<td>109</td>
<td>151</td>
<td>109</td>
<td>110</td>
</tr>
<tr>
<td>C</td>
<td>108</td>
<td>142</td>
<td>109</td>
<td>111</td>
</tr>
<tr>
<td>D</td>
<td>108</td>
<td>148</td>
<td>110</td>
<td>116</td>
</tr>
<tr>
<td>E</td>
<td>109</td>
<td>176</td>
<td>110</td>
<td>110</td>
</tr>
<tr>
<td>F</td>
<td>108</td>
<td>151</td>
<td>108</td>
<td>110</td>
</tr>
<tr>
<td>G</td>
<td>109</td>
<td>143</td>
<td>109</td>
<td>113</td>
</tr>
<tr>
<td>H</td>
<td>108</td>
<td>150</td>
<td>108</td>
<td>122</td>
</tr>
<tr>
<td>I</td>
<td>110</td>
<td>150</td>
<td>109</td>
<td>110</td>
</tr>
<tr>
<td>J</td>
<td>110</td>
<td>150</td>
<td>111</td>
<td>110</td>
</tr>
<tr>
<td>K</td>
<td>111</td>
<td>160</td>
<td>112</td>
<td>116</td>
</tr>
<tr>
<td>L</td>
<td>108</td>
<td>184</td>
<td>110</td>
<td>107</td>
</tr>
<tr>
<td>M</td>
<td>109</td>
<td>146</td>
<td>110</td>
<td>112</td>
</tr>
<tr>
<td>N</td>
<td>108</td>
<td>147</td>
<td>109</td>
<td>129</td>
</tr>
<tr>
<td>O</td>
<td>111</td>
<td>150</td>
<td>115</td>
<td>114</td>
</tr>
<tr>
<td>P</td>
<td>111</td>
<td>157</td>
<td>111</td>
<td>118</td>
</tr>
<tr>
<td>Q</td>
<td>110</td>
<td>196</td>
<td>114</td>
<td>112</td>
</tr>
<tr>
<td>R</td>
<td>109</td>
<td>150</td>
<td>108</td>
<td>129</td>
</tr>
<tr>
<td>S</td>
<td>107</td>
<td>180</td>
<td>108</td>
<td>109</td>
</tr>
<tr>
<td>T</td>
<td>109</td>
<td>147</td>
<td>109</td>
<td>112</td>
</tr>
<tr>
<td>U</td>
<td>106</td>
<td>200</td>
<td>108</td>
<td>110</td>
</tr>
<tr>
<td>V</td>
<td>109</td>
<td>148</td>
<td>111</td>
<td>114</td>
</tr>
<tr>
<td>W</td>
<td>110</td>
<td>196</td>
<td>115</td>
<td>112</td>
</tr>
<tr>
<td>X</td>
<td>110</td>
<td>153</td>
<td>110</td>
<td>113</td>
</tr>
</tbody>
</table>
according to $x_t = \exp(at)$. Then transformation (22) carries $x_t$ over into 

$$x_t \left(1 - \frac{\exp(a)}{1 - \rho^2}\right)^{\frac{1}{2}} \text{ if there is no gap between } x_t \text{ and } x_{t-1};$$

if there is a gap of one, then $x_t$ becomes 

$$x_t \left(1 - \frac{\exp(2a)}{1 - \rho^4}\right)^{\frac{1}{4}}.$$ 

The ratio of these two expressions equals 

$$\left(1 + \frac{\exp(a)}{1 + \rho^2}\right)^{\frac{1}{2}} \text{ or, for } \rho \text{ close to 1 and small } a, \text{ roughly } \sqrt{2}.$$ 

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 O, 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 $\rho$ does not lie in the interval $(-1,1)$. The standard approach taken for that event is to set $\rho$ equal to $-0.99999$ or $0.99999$. In the case where $\rho$ is equal to $0.99999$, the Cochran-Orcutt transformation turns the ones
corresponding to the constant term practically into zeros. Consequently, $\beta_1$ 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 $\beta_1$ 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 $\beta_1$, 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 $\rho$ 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 $\rho$ are heavily biased. The approximations of the standard errors of $\beta_2$ 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 $\rho$, which are used to compute $\Omega$.

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

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

For $x_t$ the U.S. GNP data are taken, as in Maeshiro (1976, 1979) and in Park and Mitchell (1981) ($t=1950, \ldots, 1969$). 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 $\rho$ (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.
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:

\[ \theta = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y \]  

(A.1)

\[ \text{tr}\left( \frac{\partial \hat{\mu}}{\partial \sigma^2} \Omega \right) = e'\left( \frac{\partial \hat{\mu}}{\partial \sigma^2} \right) e \]  

(A.2)

\[ \text{tr}\left( \frac{\partial \hat{\mu}}{\partial \rho} \Omega \right) = e'\left( \frac{\partial \hat{\mu}}{\partial \rho} \right) e \]  

(A.3)

with \( e = y - X\theta \) and \( \hat{\Omega} = \sigma^2 V \). Of course, (27) follows immediately from (A.1).

First consider (A.2). Since

\[ \frac{\partial \hat{\mu}}{\partial \sigma^2} = -\frac{4 \hat{\mu} - 1}{\sigma^2} \]  

(A.4)

(A.2) reduces to

\[ \text{tr}\left( -\frac{4 \hat{\mu} - 1}{\sigma^2} \sigma^2 V \right) = e'\left( -\frac{4 \hat{\mu} - 1}{\sigma^2} \right) e \]  

(A.5)

Using

\[ Qe = \begin{bmatrix} e_1 \\ t_2 e_1 \\ e_2 - \rho t_2 e_1 \\ \vdots \\ t_m e_1 \\ e_m - \rho t_m e_{m-1} \end{bmatrix} \]  

(A.6)
we can rewrite (A.5) as

\[ \sigma^2_e = \frac{1}{m} e' V_*^{-1} e = \frac{1}{m} e' \hat{Q} \hat{\Delta}^{-1} \hat{Q} e = \]

\[ = \frac{1}{m} \left( e_1^2 + \sum_{i=2}^{m} (1 - \hat{\rho}^i) (e_i - \hat{\rho}^i e_{i-1})^2 \right), \]  

(A.7)

which is (28).

Now consider (A.3). As

\[ \Omega^{-1}_{\rho} \frac{\partial \Omega}{\partial \rho} = \sigma^2_e \frac{\partial V_*^{-1}}{\partial \rho} \]

(A.8)

(A.3) reduces to

\[ \text{tr} \left( \frac{\partial V_*}{\partial \rho} - V_* \right) = \sigma^2_e \frac{\partial V_*^{-1}}{\partial \rho} \]  

(A.9)

In view of (19), there holds

\[ \frac{\partial V_*^{-1}}{\partial \rho} = \frac{\partial Q'}{\partial \rho} \Delta^{-1} Q + Q' \frac{\partial \Delta^{-1}}{\partial \rho} Q + Q' \Delta^{-1} \frac{\partial Q}{\partial \rho}, \]  

(A.10)

so

\[ \text{tr} \left( \frac{\partial V_*}{\partial \rho} - V_* \right) = \text{tr} \left( \frac{\partial V_*^{-1}}{\partial \rho} \right) = \]

\[ = \text{tr} \left( \frac{\partial Q'}{\partial \rho} (Q')^{-1} + \frac{\partial \Delta^{-1}}{\partial \rho} \Delta + \frac{\partial Q}{\partial \rho} Q^{-1} \right) = \]

\[ = 2 \text{tr} \left( \frac{\partial Q}{\partial \rho} Q^{-1} \right) + \text{tr} \left( \frac{\partial \Delta^{-1}}{\partial \rho} \Delta \right). \]  

(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
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):

\[
\begin{align*}
\frac{\partial e}{\partial \rho} & = 2e'Q'\Delta^{-1} \frac{\partial Q}{\partial \rho} e + e'Q' \frac{\partial \Delta^{-1}}{\partial \rho} Qe = \\
& = 2e'Q'\Delta^{-1} \frac{\partial Q}{\partial \rho} e - e'Q'\Delta^{-1} \frac{\partial \Delta^{-1}}{\partial \rho} Qe .
\end{align*}
\]

(A.13)

Since:

\[
\frac{\partial e}{\partial \rho} = - \begin{bmatrix}
0 \\
t_2 \rho^{-1} e_1 \\
\vdots \\
t_m \rho^{-1} e_{m-1}
\end{bmatrix}
\]

(A.14)

\[
\frac{\partial \Delta}{\partial \rho} = -2 \begin{bmatrix}
0 & 2t_2 \rho^{-1} \\
t_2 \rho & \ddots \\
\vdots & \ddots & \ddots \\
t_m \rho & \cdots & 2t_m \rho^{-1}
\end{bmatrix}
\]

(A.15)

(A.13) can be further written as:
\[
\frac{\partial V}{\partial \rho} e' = -2 \sum_{i=2}^{m} (1-\rho_{i-1})^{-2} \rho_{i-1}^{2} \left( e_{i} - \rho_{i} e_{i-1} \right) + \\
+ 2 \sum_{i=2}^{m} (1-\rho_{i-1})^{-2} \rho_{i-1}^{2} \left( e_{i} - \rho_{i} e_{i-1} \right)^{2} = \\
= 2 \sum_{i=2}^{m} (1-\rho_{i-1})^{-2} \rho_{i-1}^{2} \left( e_{i} - \rho_{i} e_{i-1} \right) \left( e_{i} - \rho_{i} e_{i-1} \right). \quad (A.16)
\]

Putting \( \rho = \hat{\rho} \) in this expression gives the RHS of (A.9) and hence of (A.3), apart from the factor \( \sigma_{\varepsilon}^{2} \). Combining (A.3), (A.9), (A.12) and (A.16) gives (29).
APPENDIX B. THE INFORMATION MATRIX

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

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

(B.1)

The information matrix $I$ corresponding with the likelihood function is

$$
I = \begin{bmatrix}
X_1^\prime \Omega^{-1} X_1 & 0 \\
0 & \frac{1}{2} \Psi
\end{bmatrix}
$$

(B.2)

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

$$
\psi_{22} = \text{tr}\left( \frac{\partial \Omega^{-1}}{\partial \sigma_\varepsilon^2} - \frac{\partial \Omega^{-1}}{\partial \sigma_\varepsilon^2} \right) = \text{tr}(\sigma_\varepsilon^{-4} V_*^{-1} \times \sigma_\varepsilon^{-2} V_* \times \sigma_\varepsilon^{-4} V_*^* \times \sigma_\varepsilon^{-2} V_*^*) = m \sigma_\varepsilon^{-4}.
$$

(B.3)

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

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

(B.4)

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

$$
\psi_{11} = \text{tr}\left( \frac{\partial \Omega^{-1}}{\partial \rho} - \frac{\partial \Omega^{-1}}{\partial \rho} \right) = \text{tr}\left( -\frac{\partial \Omega^{-1}}{\partial \rho} V_* \times \frac{\partial \Omega^{-1}}{\partial \rho} V_* \right).
$$

(B.5)
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 $\text{tr}(P)=\text{tr}(P')$ and $\text{tr}(AB)=\text{tr}(BA)$, one easily obtains

$$
\psi_{11} = 2 \text{tr}\left(\frac{\partial Q}{\partial \rho} Q^{-1} - \frac{\partial Q}{\partial \rho} Q^{-1} \Delta^{-1}\right) - 4 \text{tr}\left(\frac{\partial Q}{\partial \rho} Q^{-1} \frac{\partial A}{\partial \rho} \Delta^{-1}\right) + 
+ 2 \text{tr}\left(\frac{\partial Q}{\partial \rho} V^* \frac{\partial Q'}{\partial \rho} \Delta^{-1}\right) + \text{tr}\left(\frac{\partial A}{\partial \rho} \Delta^{-1} \frac{\partial A}{\partial \rho} \Delta^{-1}\right). 
$$

(B.6)

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:

$$
\frac{\partial Q'}{\partial \rho} = -\left(0_m, t_2^\rho e_1, \ldots, t_2^\rho e_{m-1}\right),
$$

(B.7)

and so:

$$
\left[ \frac{\partial Q}{\partial \rho} V^* \frac{\partial Q'}{\partial \rho} \right]_{ii} = t_1^\rho t_i^{2t_i-2} e_i^* V e_i = t_1^\rho t_i^{2t_i-2},
$$

(B.8)

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

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

(B.9)

The fourth term equals

$$
4 \sum_{i=2}^{m} (1-\rho_i)^{-2} t_i^{4t_i-2}. 
$$

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

$$
\psi_{11} = 2 \sum_{i=2}^{m} (1-\rho_i)^{-2} \frac{2t_i}{\rho_i} \{ (1-\rho_i)^{2t_i} + 2\rho_i \} =
2 \sum_{i=2}^{m} (1-\rho_i)^{-2} (1+\rho_i)^{2t_i} \frac{2t_i}{\rho_i}.
$$

Together, (B.3), (B.4) and (B.11) give the elements of $\Psi$, the lower right
part of the information matrix.
Notes

1) Due to a referee.

2) 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 those of COCO and PWPW.

3) A full set of tables with simulation results is available on request.

4) 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.
References


Honohan, P. and C. McCarthy, 1982, On the use of Durbin-Watson type statistics where there are missing observations. The Statistician 31, pp. 149-152.


IN 1983 REEDS VERSCHENEN

126 H.H. Tigelaar
Identification of noisy linear systems with multiple arma inputs.

127 J.P.C. Kleijnen
Statistical Analysis of Steady-State Simulations: Survey of Recent Progress.

128 A.J. de Zeeuw
Two notes on Nash and Information.

129 H.L. Theuns en A.M.L. Passier-Grootjans
Toeristische ontwikkeling - voorwaarden en systematiek; een selectief literatuuroverzicht.

130 J. Plasmans en V. Somers
A Maximum Likelihood Estimation Method of a Three Market Disequilibrium Model.

131 R. van Montfort, R. Schippers, R. Heuts
Johnson $S_U$-transformations for parameter estimation in arma-models when data are non-gaussian.

132 J. Glombowski en M. Krüger
On the Role of Distribution in Different Theories of Cyclical Growth.

133 J.W.A. Vingerhoets en H.J.A. Coppens
Internationale Grondstoffenovereenkomsten. Effecten, kosten en oligopolisten.

134 W.J. Oomens
The economic interpretation of the advertising effect of Lydia Pinkham.

135 J.P.C. Kleijnen
Regression analysis: assumptions, alternatives, applications.

136 J.P.C. Kleijnen
On the interpretation of variables.

137 G. van der Laan en A.J.J. Talman
Simplicial approximation of solutions to the nonlinear complementarity problem with lower and upper bounds.
IN 1984 REEDS VERSCHENEN

138 G.J. Cuypers, J.P.C. Kleijnen en J.W.M. van Rooyen
Testing the Mean of an Asymmetric Population:
Four Procedures Evaluated