ON ECONOMETRIC MODELING OF INCOMPLETE DATA*

Theo E. NIJMAN and Franz C. PALM, Amsterdam

ABSTRACT

In econometric analysis of time series, it is usually assumed that the data are available for the time periods which are considered appropriate for the model. We discuss results on identification and estimation of dynamic models when values of the endogenous variable are regularly missing. The available data are assumed to be sampled at regular intervals of length k and can be linear combinations of the realizations of the variable over a finite number of periods.

1. INTRODUCTION

Econometric modeling of time series has made substantial progress in recent years. The time series approach to modeling consists in deriving the specification of the model at least in part from the information in the data. This approach is useful when sufficient sample information is available on a set of variables for which the process has been approximately stable through time.

Quite often in empirical economics, data are not complete or they have been completed by including values constructed from related series for the unobserved variables. For instance, many variables in the quarterly national accounts for many countries are constructed rather than observed. If

---

*) The results presented in this invited paper have been obtained in a joint research project of the two authors. The presentation at the 9th Symposium on Operations Research in Osnabrück, August 27-29, 1984, was made by the second author.

the constructed series are subsequently used as if they were observations on the process to be modeled, statistical inference can be heavily flawed. To assess the properties of the methods used in constructing values for the unobserved variables, assumptions on the model relating missing observations and data are needed. Moreover, if this model is completely specified, its parameters can be consistently and efficiently estimated, provided they are identified, and optimal 'forecasts' of the missing observations can be generated. Thereby, a pure data-based approach will probably be inadequate. A structural approach to modeling using economic theory and other a priori subject matter considerations may have to be adopted because the incomplete sample is insufficiently informative on the structure of the model. For instance, it is possible that some parameters of interest in the model are not identified or that the likelihood is rather flat in a large neighborhood of the true parameter values.

2. AN OUTLINE OF THE PROBLEM OF REGULARLY MISSING OBSERVATIONS

To formalize the problem of incomplete data with regular nonstochastic sampling, consider the joint density function of the vector of observations y and missing values m, which is assumed to depend on the parameters of interest θ

$$D(y,m \mid \theta). \tag{2.1}$$

The density function (2.1), which completely specifies the model, is conditional on initial values in dynamic models and on a vector of exogenous variables x in causal models and models with leading indicators. For the simplicity of exposition, the conditioning variables are not included. The data generation process (DGP) is obtained by model reduction through marginalization of (2.1) with respect to m

(2.2) $$D(y \mid \theta) = \int D(y,m \mid \theta)dm.$$

The parameters θ are identified iff for all $\theta_1, \theta_2 \in \Theta$, with Θ being the parameter space, $D(y|\theta_1) = D(y|\theta_2)$ implies that $\theta_1 = \theta_2$. As we will see below, θ is often not identified. We can reparametrize the DGP by introducing a vector of identified parameters θ^* in the DGP denoted as $D(y|\theta^*)$ where the elements of θ^* are functions of θ, $\theta^* = f(\theta)$. The parameters of interest θ are identified if f is injective.

Specifying (2.2) indirectly via (2.1) can be useful for generating overidentifying restrictions on θ in (2.2), which then can be tested. It can be needed for the interpretation of the parameters of the DGP. Moreover, if one is interested in forecasting the missing observations, the conditional density of m given y is of importance, which can be readily obtained when (2.1) is specified

(2.3) $$D(m \mid y, \theta) = D(y,m \mid \theta) \; / \; D(y|\theta).$$

To illustrate these points, consider a time series y_t which is annually observed and for which a quarterly ARMA(1,1)-model is appropriate

(2.4) $$y_t = \phi \, y_{t-1} + \varepsilon_t - \omega \, \varepsilon_{t-1},$$

with ε_t being i.i.d. $N(0,\sigma^2)$, $|\phi| < 1$, $|\omega| \leq 1$, $\phi \neq \omega$. For the ease of notation, we use T_k to denote the index set $T_k = \{k,2k,\ldots,T\}$, where $\frac{T}{k}$ is the sample size, assumed to be an integer. The DGP corresponding to (2.4)

(2.5) $$y_t = \phi^4 y_{t-4} + (1 + \phi L + \phi^2 L^2 + \phi^3 L^3)(1 - \omega L)\varepsilon_t,$$

for $t \in T_4$ can be reparametrized as an ARMA (1,1)-model for yearly data

(2.6) $$y_t = \phi^* y_{t-4} + (1 - \omega^* L^4) v_t \quad ,$$

with $\phi^* = \phi^4$, v_t being i.i.d. $N(0, \sigma_v^2)$ for $t \in T_4$, and L being the lag operator.
The parameters σ_v^2 and ω^* satisfy the relationships
$\sigma_v^2 (1 + \omega^{*2}) = \sigma^2[1 + (\phi - \omega)^2 + (\phi^2 - \omega\phi)^2 + (\phi^3 - \omega\phi^2)^2 + \phi^6\omega^2]$;
$\omega^*\sigma_v^2 = \phi^3 \omega \sigma^2$; $|\omega^*| \leq 1$.

The vector θ^* is chosen to be $\theta^* = (\phi^*, \omega^*, \sigma_v^2)$. The model (2.4) is only locally identified if $\phi \neq 0$. Notice that changing the sign of ϕ and ω simultaneously does not change the value of the likelihood function. If for instance the sign of ϕ is a priori known, global identification is achieved and a minimum mean square error predictor for the missing value in period $T+1$ say, can be computed as:

(2.7) $$E(y_{T+1} \mid y_t, t \in T_4) = \frac{(\phi - \omega)}{1 - \omega L} E(y_T \mid y_t, t \in T_4) \quad ,$$

with $E(y_{t'} \mid y_t, t \in T_4) = y_{t'}$ if $t' \in T_4$ and being a smoothed value (using e.g. the fixed interval smoother) for $t' \notin T_4$.

When a strictly exogenous variable x_t, assumed to be observed for $t \in T_1$, is included in (2.4), the model

(2.8) $$y_t = \phi y_{t-1} + \beta x_t + \varepsilon_t - \omega \varepsilon_{t-1} \quad , \quad \beta \neq 0 \quad ,$$

is identified (for $k = 4$) provided $\phi \neq 0$ and at least two of the variables x_{t-i}, $i = 0,\ldots,3$, are not collinear such that β and the sign of ϕ can be determined from their coefficients. This can be easily checked by considering the DGP associated with (2.8)

(2.9) $$y_t = \phi^4 y_{t-4} + \beta (1 + \phi L + \phi^2 L^2 + \phi^3 L^3) x_t +$$
$$+ (1 + \phi L + \phi^2 L^2 + \phi^3 L^3)(1 - \omega L)\varepsilon_t \quad .$$

Quite obviously, when x_t is a constant, global identification is not achieved.

When a temporal aggregate $\bar{y}_t = \sum_{i=0}^{k-1} y_{t-i}$ is observed, the DGP corresponding to (2.4) for $k = 4$ is given in (2.5) with y_t and ε_t replaced by their aggregate values $\bar{y}_t$ and $\bar{\varepsilon}_t$. It can be represented as an ARMA(1,1)-model in L^4. The data are informative on the sign of the parameters ϕ and ω, which are identified even if $\phi = 0$.

In the literature on missing observations, the identification problem has received little attention compared with estimation problems. There are at least two reasons for this. First, part of the literature is on irregularly missing observations, for which the model is usually identified - at least in large samples (see e.g. Dunsmuir and Robinson (1981)). When all variables in the model are observed for a sufficient number of consecutive periods, the model is identified. This is ususally the case in static models and in some dynamic models (see e.g. Dagenais (1973), Harvey and Pierse (1984), Hsiao (1979), Palm and Nijman (1982) (hereafter denoted by PN) among many others). The requirements discussed above are not fulfilled in monthly or quarterly dynamic macroeconomic models when some variables are only observed annually.

We have analyzed univariate ARMA-models and linear dynamic regression models with moving average disturbances (ARMAX-models), where we assume that the endogenous variable is observed every kth period (skipped observations) or its aggregate value over k periods is available. Other sampling schemes also fit into our framework. Missing exogenous variables can be analyzed along the lines outlined in this section provided their process can be specified so that they can be endogenized and included in model (2.1) which is again complete.

3. A SUMMARY OF THE RESULTS

3.1 IDENTIFICATION

The results on identification in the presence of incomplete data obtained by Nijman (1984) and PN (1984a) are in line with the findings for the identification of dynamic models with other types of unobservables (see e.g. Maravall (1979) and Nowak (1983) for errors in variables models). The main conclusions are:

- AR processes and autoregressive regression models are more easily identified than models with MA-parameters.
- The presence of exogenous variables in the model facilitates the identification problem.
- Similarly, observing temporal aggregates instead of a sample with skipped data is helpful for the parameter identification.
  Heuristically, the lack of identification decreases as the number of lags increases relative to the number of unknown parameters in the lag polynomials.
- Local maxima of the likelihood function not only occur when the model is only locally identified. They can also arise in finite samples when the model is formally identified. Several starting values for the parameter estimates will have to be used in an iterative ML-estimation procedure to check that indeed the global maximum of the likelihood function of the globally identified model has been found.
- In finite samples, the likelihood function is often very flat so that the parameters cannot be determined with a reasonable degree of accuracy although they are formally identified.
- Even a slight overparametrization of the model can lead to a singular Hessian matrix of the log-likelihood function when the sample is incomplete, whereas no anomaly would occur with a complete data set. This

finding has obvious implications for a top-down modeling approach when data are missing.

To conclude, identification of the parameters in models for missing observations requires much care in applied work.

3.2 EFFICIENT ESTIMATION

When the DGP in (2.2) has been obtained and θ is identified, efficient estimation is in fact a standard problem of non-linear optimization possibly subject to overidentifying restrictions on θ^* implied by $\theta^* = f(\theta)$. It could be solved in two steps using asymptotic nonlinear least squares proposed by Gouriéroux et al. (1983) where in the first step the likelihood function is maximized with respect to θ^* and in the second step the 'asymptotic model' $\theta^* = f(\theta)$ is solved.
In some cases, explicit expressions for the gradient of the log-likelihood function are available and can be used in a scoring algorithm (see e.g. Hsiao (1979), PN (1982)).

However, several procedures have the advantage of not requiring the model reduction to get the DGP. For instance, the EM-algorithm (see e.g. Dempster et al. (1977)), applies to $D(y,m \mid \theta)$ and yields the value of θ which maximizes the likelihood function in (2.2). Kalman filtering techniques applied to the model in prediction error decomposition form (see e.g. Harvey (1981) and Harvey and Pierse (1984)) appeared to be very useful to obtain efficient estimates of our models and smoothed values for the missing observations. Standard errors for the parameter estimates can be computed by means of the procedures proposed in Watson and Engle (1983).

3.3 CONSISTENT ESTIMATION

Although efficient estimation is in principle feasible, it

requires that the DGP is completely specified. Moreover, the number of parameters to be estimated can be large causing numerical problems when fully efficient joint estimation is attempted. The almost unidentifiability of the parameters discussed in section 3.1 can lead to problems with parameter estimation.

Quite often, one is only interested in a subset of θ, say θ_1. Provided it is identified, it can be consistently estimated (under the usual regularity conditions).
One way to obtain estimates consists in substituting proxy variables for the missing data and then applying a standard estimation method. This is frequently done in applied work when constructed series (see e.g. Boot et al. (1967)) are substituted for the missing observations.

Consider for example the dynamic regression model (2.8) with $\omega = 0$ and y_t being observed every second period. If we substitute a proxy, say $\hat{y}_t$ for y_t whenever y_t is not observed (when y_t is observed, $\hat{y}_t \equiv y_t$), we get in matrix notation

$$\hat{y} = \hat{X}\delta + w, \tag{3.1}$$

with $\delta = (\phi, \beta)'$, the columns of $\hat{X}$ consisting of the lagged values of $\hat{y}_t$ and of x_t respectively, $w = \varepsilon + (X - \hat{X})\delta + \hat{y} - y$.
Consistent estimation of δ by OLS applied to (3.1) requires that $\operatorname{plim}_{T\to\infty} T^{-1}\hat{X}' w = 0$, which is satisfied for special models (see e.g. Dagenais (1973) for a model with missing exogenous variables, and PN (1984b)) or when the proxies have been chosen such that the above condition holds true. Otherwise, an instrumental variables (IV) procedure can be applied to (3.1). In any case, some structural considerations on the model are needed for selecting proxy and/or instrumental variables.

The main conclusions about proxy variables estimators are:

- OLS applied to a model in which data-based interpolations along the lines of Boot et al. (1967) are substituted for the missing values of the endogenous variable can be strongly biased (see e.g. PN (1984a)).
- The composite nature of the disturbance term w in (3.1) causes problems with the estimation of standard errors (SEs). Several methods to obtain consistent estimates and bounds for the SEs are presented in Nijman (1984) and PN (1984b).
- The accuracy of the consistent estimators can be improved by extending the conditioning set of the proxies, by selecting more appropriate IVs and by accounting for the correlation of the disturbance w (see Nijman and Palm (1984a) (hereafter denoted as NP) and PN(1984b).
- Consistent proxy variables estimators which take into account the most important restrictions on the parameters are almost as efficient as ML estimates (see NP (1984a) and PN (1984b)).
- The implementation of the consistent proxy variables estimators and the formulae for the SEs in applied work is feasible.

  NP (1984a) and Nijman (1984) use proxy variables estimators to analyze the aggregate quarterly demand for labor in the private sector in the Netherlands (1967-1981). Univariate and bivariate time series processes and simple causal models are used to generate the proxies. The specification of these schemes is checked empirically and confronted with those implied by the structural model for the labor market. These schemes are direct extensions of interpolation procedures proposed in the literature (see e.g. Fernandez (1981) and Litterman (1983)). Subsequently, the proxies are used to estimate the structural model and to obtain standard errors.

4. FINAL REMARKS

More details on the results presented in this note can be found in the references cited above which can be obtained from the authors. The conclusions reached for missing observations are expected to hold for models with other types of unobservables as well. The results on consistent estimation when data are incomplete also apply to models with rational expectations (see PN (1984b) for more details). Also, as pointed out above the identification problem with incomplete data has common features with that of errors in variables models.
To end, it should be obvious that much care and reliable a priori information are required for the econometric analysis of incomplete samples.

REFERENCES

(1) J.C.G.BOOT, W. FEIBES, and J.H.C. LISMAN (1967): Further methods of derivation of quarterly figures from annual data. Applied Statistics, 16, 65-75.

(2) M.G. DAGENAIS (1973): The use of incomplete observations in multiple regression analysis, a generalized least squares approach. Journal of Econometrics, 1, 317-328.

(3) A.P. DEMPSTER, N.M. LAIRD, and D.B. RUBIN (1977): Maximum likelihood from incomplete data via the EM-algorithm. Journal of the Royal Statistical Society, B, 39, 1-38.

(4) W. DUNSMUIR, P.M. ROBINSON (1981): Estimation of time series models in the presence of missing data. J.A.S.A., 76, 560-568.

(5) R.B. FERNANDEZ (1981): A methodological note on the estimation of time series. The Review of Economics and Statistics, 63, 471-476.

(6) CH. GOURIEROUX, A. MONFORT, and A. TROGNON (1983): Nonlinear asymptotic least squares. Paper presented at the ESEM in Pisa.

(7) A.C. HARVEY (1981): Time Series Models, Oxford, Philip Allan.

(8) A.C. HARVEY, R.G. PIERSE (1984): Estimating missing observations in economic time series. J.A.S.A., 79, 125-131.

(9) CH. HSIAO (1979): Linear regression using both temporally aggregated and temporally disaggregated data. Journal of Econometrics, 10, 243-252.

(10) R.B. LITTERMAN (1983): A random walk, Markov model for the distribution of time series. Journal of Business and Economic Statistics, 1, 169-173.

(11) A. MARAVALL (1979): Identification in dynamic shock-error models. Berlin, Springer-Verlag.

(12) T.E. NIJMAN (1984): Missing Observations in Dynamic Macroeconomic Modelling. Free University, Amsterdam, forthcoming doctoral dissertation.

(13) T.E. NIJMAN, F.C. PALM (1984a): Consistent estimation of a regression model with incompletely observed exogenous variable. Free University, Amsterdam, mimeographed.

(14) T.E. NIJMAN, F.C. PALM (1984b): Missing observations in a quarterly model for the aggregate labor market in the Netherlands. Free University, Amsterdam, Research-memorandum.

(15) E. NOWAK (1983): Identification of the dynamic shock-error model with autocorrelated errors.Journal of Econometrics, 23, 211-222.

(16) F.C. PALM, T.E. NIJMAN (1982): Linear regression using both temporally aggregated and temporally disaggregated data. Journal of Econometrics, 19, 333-343.

(17) F.C. PALM, T.E. NIJMAN (1984a): Missing observations in the dynamic regression model. Free University, Amsterdam, mimeographed, forthcoming in Econometrica.

(18) F.C. PALM, T.E. NIJMAN (1984b): Consistent estimation using proxy-variables in models with unobserved variables. Free University, Amsterdam, Researchmemorandum.

(19) M.W. WATSON, R.F. ENGLE (1983): Alternative algorithms for the estimation of dynamic factors, MIMIC and varying coefficient regression models. Journal of Econometrics, 23, 385-400.

Theo E. NIJMAN and Franz C. PALM, Faculteit der Economische Wetenschappen, Vrije Universiteit, Postbus 7161, 1007 MC Amsterdam, Nederland.