MONASH
UNIVERSITY

AUSTRALIA

SIMULTANEOUS ERROR COMPONENTS MODELS WHEN PANEL DATA ARE INCOMPLETE

László Mátyás and László Lovrics

Working Paper No. 11/91

November 1991

DEPARTMENT OF ECONOMETRICS

ISSN 1032-3813

ISBN 0 86746 958 7

SIMULTANEOUS ERROR COMPONENTS MODELS WHEN PANEL DATA ARE INCOMPLETE

László Mátyás and László Lovrics

Working Paper No. 11/91

November 1991

DEPARTMENT OF ECONOMETRICS, FACULTY OF ECONOMICS COMMERCE & MANAGEMENT

MONASH UNIVERSITY, CLAYTON, VICTORIA 3168, AUSTRALIA.

Simultaneous Error Components Models When Panel Data Are Incomplete

László Mátyás* and László Lovrics**

* Monash University, Australia and Budapest University of Economics

** Budapest University of Economics

Abstract: The purpose of this paper is to investigate the loss of efficiency of the simultaneous error components model's estimators in the case of incomplete panels. The static and the dynamic cases were analysed, when from a panel data base those individuals are dropped for which the observations are not complete.

Key words: Panel data, simultaneous error components model, missing observations, unbalanced panels, Monte–Carlo experiment, risk function.

Introduction

The error components approach is the classical way in econometric modelling to pool time series and cross section data together. In recent years —besides some special problems related to the single equation model— the attention was focused on the simultaneous case (*Baltagi* [1981], *Prucha* [1985], *Balestra–Varadharajan-Krishnakumar* [1987], *Krishnakumar* [1988], *Lovrics–Mátyás* [1990], *Baltagi* [1981], *Baltagi–Li* [1991]). All the main studies analysed the complete panel data case, that is their basic assumption was that no one observation is missing. However in practice panel data bases may be incomplete. This means that either the individuals present in the data base are not observed during the same period (unbalanced panels) or there may be "holes" in the observation panel. When we want to estimate an econometric model with these kinds of incomplete panels we have two possibilities. We can use appropriate (unbalanced) estimation methods (for the single equation model see for example *Mátyás–Lovrics* [1991], *Wansbeek–Kapteyn* [1989], *Baltagi* [1981], *Biorn* [1981]). Unfortunately for the simultaneous case such methods have not been developed. Another possibility is to drop from the panel those individuals for which the observations are not complete and carry out the estimation on a balanced and complete sub–panel of the original one. (Obviously panels like rotating ones cannot be treated in this way.) For the moment this is the unique way to proceed in the simultaneous case.

It has been shown (*Verbeek–Nijman* [1990], *Heckman* [1979]) that if unbalanced or incomplete panels are available and the usual estimators of panel models based on a balanced and complete sub–panel are used, these are consistent under quite general and reasonable conditions when the observations are missing at random (so there is no selectivity bias present).

From a practical point of view it is very important to analyse the increase of risk of an estimator due to the simultaneous effect of the small sample bias and the loss of efficiency caused by the use of balanced sub–panels.

In this paper we analyse the behaviour of the simultaneous error components model's limited information estimators in the static and "semi asymptotic" case through a Monte–Carlo experiment.

1. The framework

The basic framework of our experiment was defined as follows:

- The first step was the specification of the model.
- The risk of the estimators was represented by a quadratic loss function.
- For a given model, given parameter values and given sample size, we generated one hundred samples (sometimes much more), and for the given sample we performed all the analysed estimations.

In the experiment, first we chose a model with a given parameter structure, then we generated the samples and performed the estimations for the entire sample NT (N is the number of observed individuals, T is the length of the time series). In the second round we supposed that the missing observations are affecting only the one individual which was dropped from the sample. The specific individual was chosen randomly from a uniform distribution. So in this case the sample size was $(N-1)T$. Then we performed again one hundred Monte–Carlo experiments. Next we supposed that more and more individuals have to be dropped from the sample (each time one more than previously) up to 75% of the original sample size. For each sample size one hundred Monte–Carlo experiments were performed. In the next step we modified the (complete) sample size NT (in the entire experiment $5 \leq N < 50$ and $5 \leq T < 30$) and then we came back to the first step. For all the possible scenarios defined by the above mentioned parameters these steps were carried out. We wanted to answer the following two important questions:

- Does the model specification affect the risk of the estimation?
- In which form does the risk of the estimation depend on the size of the balanced sub–sample and on the proportion of the original (unbalanced) sample and the balanced sub–sample size?

Our analysis was quite difficult. The degrees of freedom of the problem is very important so many more samples had to be generated and many more regressions had to be estimated than is usual for a simple Monte–Carlo study.

1.1 The general model

Let us consider the following general simultaneous model:

$$Y\Gamma + X\beta + U = 0$$

where

Y $=[y_1, \ldots, y_M]$ is the $(NT \times M)$ matrix of the endogenous variables;

X $=[X_1, \ldots, X_K]$ is the $(NT \times K)$ matrix of the exogenous variables;

u $=[u_1, \ldots, u_M]$ is the $(NT \times M)$ matrix of the residuals;

Γ $=[\gamma_1^*, \ldots, \gamma_M^*]$ is the $(M \times M)$ parameter matrix related to the endogenous variables;

β $=[\beta_1 \ldots, \beta_M]$ is the $(K \times M)$ parameter matrix related to the predetermined variables;

N is the number of individuals observed and

T is the length of the time–series.

Taking the usual assumptions, the typical structural equation, say the j–th one, can be written as

$$y_j = Y_j\alpha_j + X_j\beta_j + u_j = Z_j\gamma_j + u_j \tag{1}$$

where

$$Z_j = [Y_j, X_j] \text{ and}$$
$$\gamma_j' = [\alpha_j'\beta_j'] .$$

The error component structure is given by the decomposition of the residual:

$$u_j = (I_N \otimes L_T)\mu_j + (L_N \otimes I_T)\lambda_j + v_j \tag{2}$$

where I_N and I_T are identity matrices of order N and T respectively, L_N and L_T are unit vectors of order N and T, respectively.
$\mu_j' = (\mu_{1j}, \ldots, \mu_{Nj})$ is the vector of individual effects,
$\lambda_j' = (\lambda_{1j}, \ldots, \lambda_{Tj})$ is the vector of time effects,
$v_j' = (v_{11j}, \ldots, v_{NTj})$ is the vector of "pure" residual effects,
and they are independent two by two with the following properties:

$$E(\mu_j) = 0 , \quad E(\lambda_j) = 0 , \quad E(v_j) = 0 \quad j = 1, \ldots, M$$

$$E(\mu_j\mu_j') = \sigma_{\mu jj}^2 I_N , \quad E(\lambda_j\lambda_l') = \sigma_{\lambda jj}^2 I_T , \quad E(v_j v_j') = \sigma_{vjj}^2 I_{NT} .$$

The covariance matrices of the residuals are

$$E(u_j u_j') = \Sigma_{jj} = \sigma_{\mu jj}^2 (I_N \otimes J_T) + \sigma_{\lambda jj}^2 (J_N \otimes I_T) + \sigma_{vjj}^2 I_{NT} .$$

where J_N and J_T are unit matrices of order N and T, respectively.

1.2 The limited information estimators

Let us define the following operators:
$M_1 = B_n = (I_N \otimes \frac{J_T}{T}) - \frac{J_{NT}}{NT}$ (the "between individuals" operator),
$M_2 = B_t = (\frac{J_N}{N} \otimes I_T) - \frac{J_{NT}}{NT}$ (the "between time" operator),
$M_3 = \frac{J_{NT}}{NT}$
$M_4 = W^* = I_{NT} - (I_N \otimes \frac{J_T}{T}) - (\frac{J_N}{N} \otimes I_T) + M_3$, (the "within" operator).
Then the Within estimator of model (1) is

$$\hat{\gamma}_{jw} = [Z_j' W^* X (X' W^* X)^{-1} X' W^* Z_j]^{-1} \times$$
$$\times Z_j' W^* X (X' W^* X)^{-1} X' W^* y_j .$$

The generalized two–steps least squares estimator of model (1) (G2SLS) is

$$\hat{\gamma}_{j,G2SLS} = [Z_j' \Sigma_{jj}^{-1} X (X' \Sigma_{jj}^{-1} X)^{-1} X' \Sigma_{jj}^{-1} Z_j]^{-1} \times$$
$$[Z_j' \Sigma_{jj}^{-1} X (X' \Sigma_{jj}^{-1} X)^{-1} X' \Sigma_{jj}^{-1} y_j] . \tag{3}$$

It is clear that for the G2SLS estimator to be operational we need the estimated values of the variance components. Taking into consideration the spectral decomposition of Σ_{jj}:

$$\Sigma_{jj} = (\sigma^2_{vjj} + T\sigma^2_{\mu jj} + N\sigma^2_{\lambda jj})M_3 + \\ + (\sigma^2_{vjj} + T\sigma^2_{\mu jj})B_n + (\sigma^2_{vjj} + N\sigma^2_{\lambda jj})B_t + \sigma^2_{vjj}W^*$$

we easily can estimate the variance components:

$$\widehat{\sigma}^2_{i,jj} = \frac{1}{m_i}\widehat{u}'_j M_i \widehat{u}_j \quad i = 1,2,4$$

where m_i-s are the ranks of the M_i operators ($m_1 = (N-1)$, $m_2 = (T-1)$, $m_4 = (N-1)(T-1)$), and

$$\widehat{\sigma}^2_{3,jj} = \widehat{\sigma}^2_{1,jj} + \widehat{\sigma}^2_{2,jj} - \widehat{\sigma}^2_{4,jj}$$

and finally we get

$$\widehat{\Sigma}_{jj} = \sum_{i=1}^{4} \widehat{\sigma}^2_{i,jj} M_i \,.$$

In the following sections we will focus our attention —besides the OLS estimator— on

1. the Within estimator,
2. the theoretical G2SLS estimator (when the elements of Σ_{jj} are exactly known),
3. the feasible G2SLS estimator (when the elements of Σ_{jj} are estimated from the Within residual) and
4. the OLS based feasible G2SLS estimator (when the elements of Σ_{jj} are biasedly estimated from the OLS residual).

It is easy to derive other 2SLS like estimators by choosing different instrumental variables (*Baltagi* [1981], *Baltagi–Li* [1991], *Ahn–Schmidt* [1990], *Breusch–Mizon–Schmidt* [1989]), but their analysis is out of our scope, mainly because the extra instruments are redundant when the estimation is performed by equation (see *Baltagi–Li* [1991]).

2. Monte–Carlo results for the static case

2.1 The model

The base of the Monte–Carlo study was the following simple model:

$$\begin{aligned} y^{(1)}_{it} &= y^{(2)}_{it}\alpha_1 + X^{(1)}_{it}\alpha_2 + X^{(2)}_{it}\alpha_3 + u^{(1)}_{it} \\ y^{(2)}_{it} &= y^{(1)}_{it}\beta_1 + X^{(3)}_{it}\beta_2 + X^{(4)}_{it}\beta_3 + u^{(2)}_{it} \end{aligned} \tag{4}$$

Each equation in the system is identified.

The data generating process was based on the generation of the variables $X^{(j)}$ ($j = 1,2,3,4$) and the reduced form of model (4).

First of all we have generated the variables $X^{(j)}$ $(j = 1, \ldots, 4)$ with the following process:

$$\begin{cases} X_{i0}^{(j)} = \varepsilon_{i0}^{(j)}/(1-\gamma_j) & \text{assuming that the process is stationary} \\ X_{it}^{(j)} = X_{it-1}^{(j)}\gamma_j + \varepsilon_{it}^{(j)} & (i = 1 \ldots N)\ (j = 1,2,3,4) \end{cases}$$

where

$$\varepsilon_{it}^{(j)} \sim \mathcal{N}(0, \sigma_{\varepsilon_j})$$

In the explosive model case ($\gamma_j > 1$) the initial values were the simple ε noise variables.

Using the above variables and the reduced form for the first equation we generated the $y^{(1)}$ variable. The variable $y^{(2)}$ is produced by using the second equation of the structural form.

In order to control the data generating process first we generated the $y^{(1)}$ variable again from the structural form and it was — obviously — the same as obtained from the reduced form. Second, we matched the error variables from the structural form with those generated with the random variable generator and used in the reduced form.

2.2 Numerical results

The first important conclusion of the analysis is that the loss (that is the risk) of the estimators in focus depends only on the actual sample size. The relative sample size (the proportion of the actual sample size to the original one) doesn't have any effect on the risk of these estimators. Even in the case when the sample truncation was relatively important (e.g., 50% or over) this had no effect on the estimators if the remaining sample was large enough.

As we would expect the parameterization of the model doesn't have any effect on the behaviour the estimators.

For small sample sizes (always actual) the OLS estimator has much smaller risk than the other estimators. Here sample size $N \times T = 100$ seems to be the critical value. In extreme cases, that is when the remaining number of observed individuals is very small (under 10–15) or the time series are very short (less than 5 periods), even for samples over 100 the OLS has performed at least as well as the four other estimators. For all other sample sizes the loss of the OLS was much larger than that of the competing estimators.

The real surprise of the experiment was the good performance of the Within estimator. In samples over 100 it has the same behaviour as the three other G2SLS estimators. In fact it was impossible to find any difference in the behaviour of these four estimators: they seem to have the same small sample properties.

To be more specific let us focus our attention on the α_1 and β_1 parameters, that is the parameter of the endogenous variables of the first and the second equation. These have a central role because they are the source of the small sample bias and this is always larger than the bias of the other parameters. It turned out that in samples under 50 the bias of the OLS estimator for these parameters is never larger than 100% and in average is about 30–60%, while the bias of the other estimators can be ten times larger. For samples over 50 the bias of the OLS is about 30–50% and the same magnitude remains regardless the sample size. The bias of the other estimators decreases quite rapidly from 50–70% for samples between 50 and 100, to less than 3–4% for samples over 600–700 observations.(Graph 1 and 2 show the bias of the G2SLS(OLS) estimator for the α_1 parameter in the case when all the parameters of the model are equal to 0.5.)

When we modified the initial simulation model (4), it turned out that the larger the model (the larger the number of endogenous variables in one equation), the larger the minimum sample size necessary to get a "good" estimation result. For an equation with three endogenous and five exogenous variables this minimun sample size seems to be between 150 and 250. (However we have to admit here the limits of our study: it is practically impossible to analyse properly the effects of the size of a model on the behaviour of its estimators. The degrees of freedom of the problem is too large to fit in this framework.)

3. Monte–Carlo results for the dynamic case

3.1 The model

The basis of the analysis here was the following simple model:

$$\begin{aligned} y_{it}^{(1)} &= y_{it}^{(2)}\alpha_1 + y_{it-1}^{(1)}\alpha^* + X_{it}^{(1)}\alpha_2 + X_{it}^{(2)}\alpha_3 + u_{it}^{(1)} \\ y_{it}^{(2)} &= y_{it}^{(1)}\beta_1 + X_{it}^{(3)}\beta_2 + X_{it}^{(4)}\beta_3 + u_{it}^{(2)} \end{aligned} \tag{5}$$

The generation of the endogenous and exogenous variables was like the static case. The only point of difference in the data generating process was the initial value of the lagged dependent variable. We chose the most simple solution. We generated supplementary observations for the residuals and the exogenous variables and, using these in the reduced form, we obtained the initial values of the endogenous variables.

3.2 Numerical results

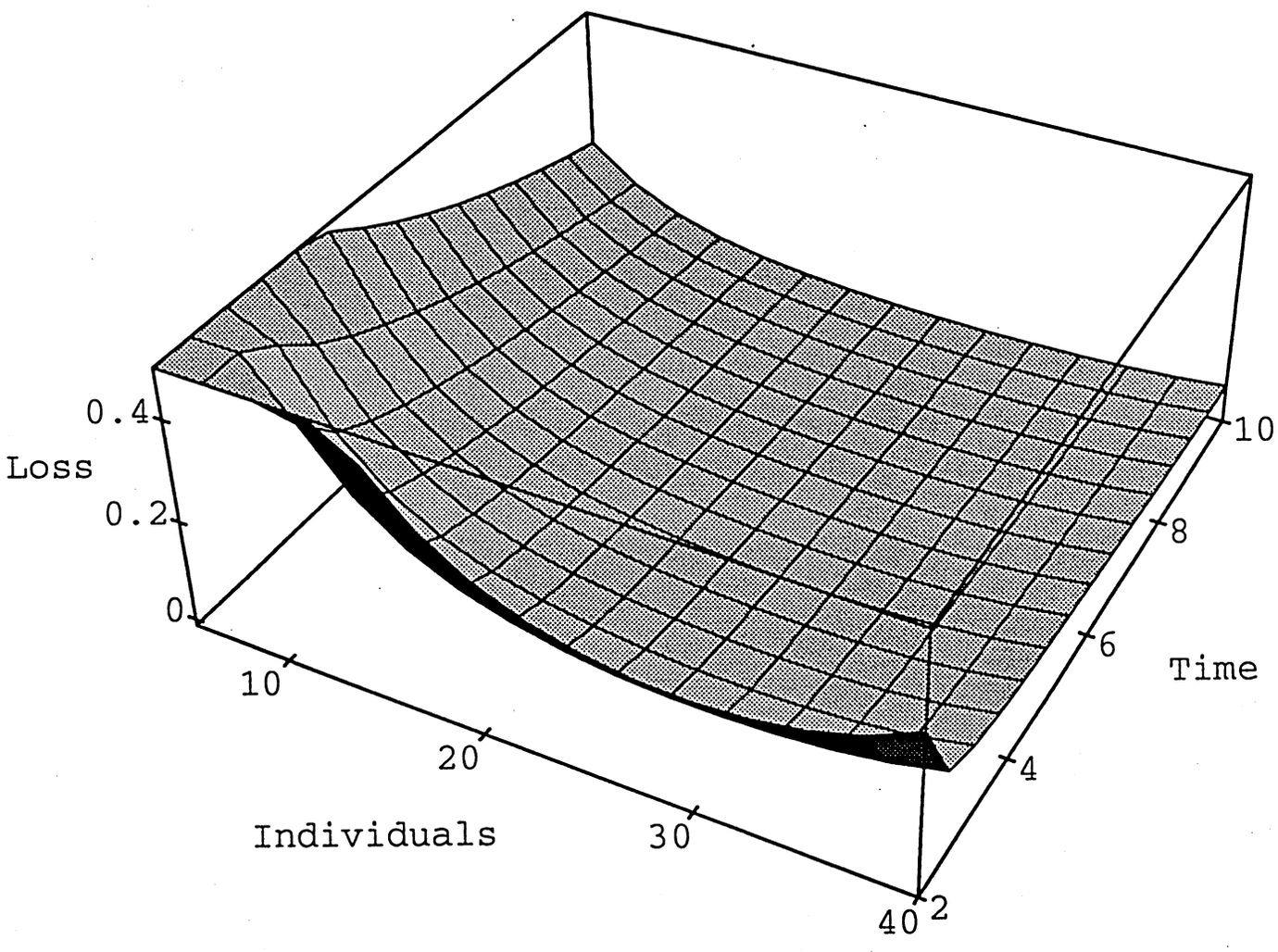
In the explosive (non stationary) case (that is when $\alpha^* > 1$) the results are similar to those obtained for the single equation models, that is all the main estimators (including the OLS) are consistent. However if α^* is close to 1 and/or N or T are small (even if the sample size $N \times T$ is large) large biases may occur. In graph 3 it can be seen clearly that both N and T should be large to have a fairly small loss. It seems that N should be larger than 30 and T should be larger than 8 or 9. This latter condition seems to be quite strong, because in practice we rarely have such long time series.

In the stationary case the OLS estimator has the usual behaviour: it is not consistent. However its loss does not exceed in general 15–20% if $T > 4$ and $N > 10$ (see graph 4). The other analysed estimators have properties similar to each other: they are very sensitive to the size of N and/or T. The bias of the estimators may be quite important even if the sample size is very large (about one thousand), but T is not at least (near to) 10. In Graph 5 the cyclical large losses are due to small T values. Moreover the loss of these estimators is smaller than the loss of the OLS estimator only in the case when the number of the observed individuals is at least 40 and the time series is quite long ($T > 20$). This means that in practice the OLS estimator is to be preferred nearly always.

Conclusion

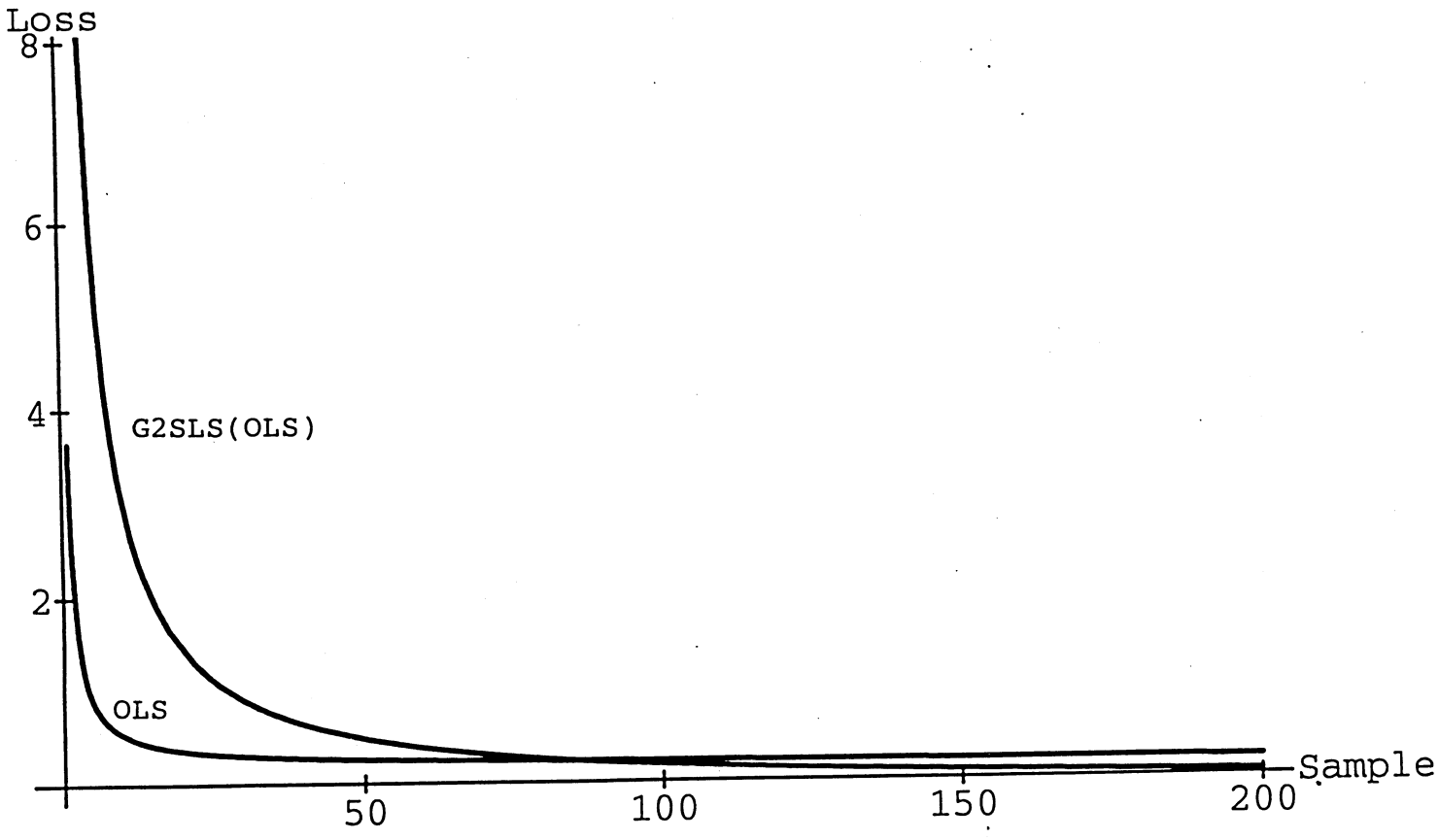
This analysis of the OLS, Within, and Generalized 2SLS estimators shows that from incomplete panel data those individuals for which the time series are not complete can be dropped without major loss of efficiency, as far as the remaining number of individuals is larger than 25 and the sample size is larger than 100. In the case of dynamic models not only the number of individuals should exceed this limit, but the length of the time series should be at least 20 as well, to get acceptable results by the use of the G2SLS estimators. For the static case when N is small and for the dynamic case when $T < 20$ and the sample size does not exceed 1000, OLS in general has a smaller bias than the G2SLS estimator.

Graph 1: Quadratic Loss of the Parameter α_1 for the G2SLS(OLS)



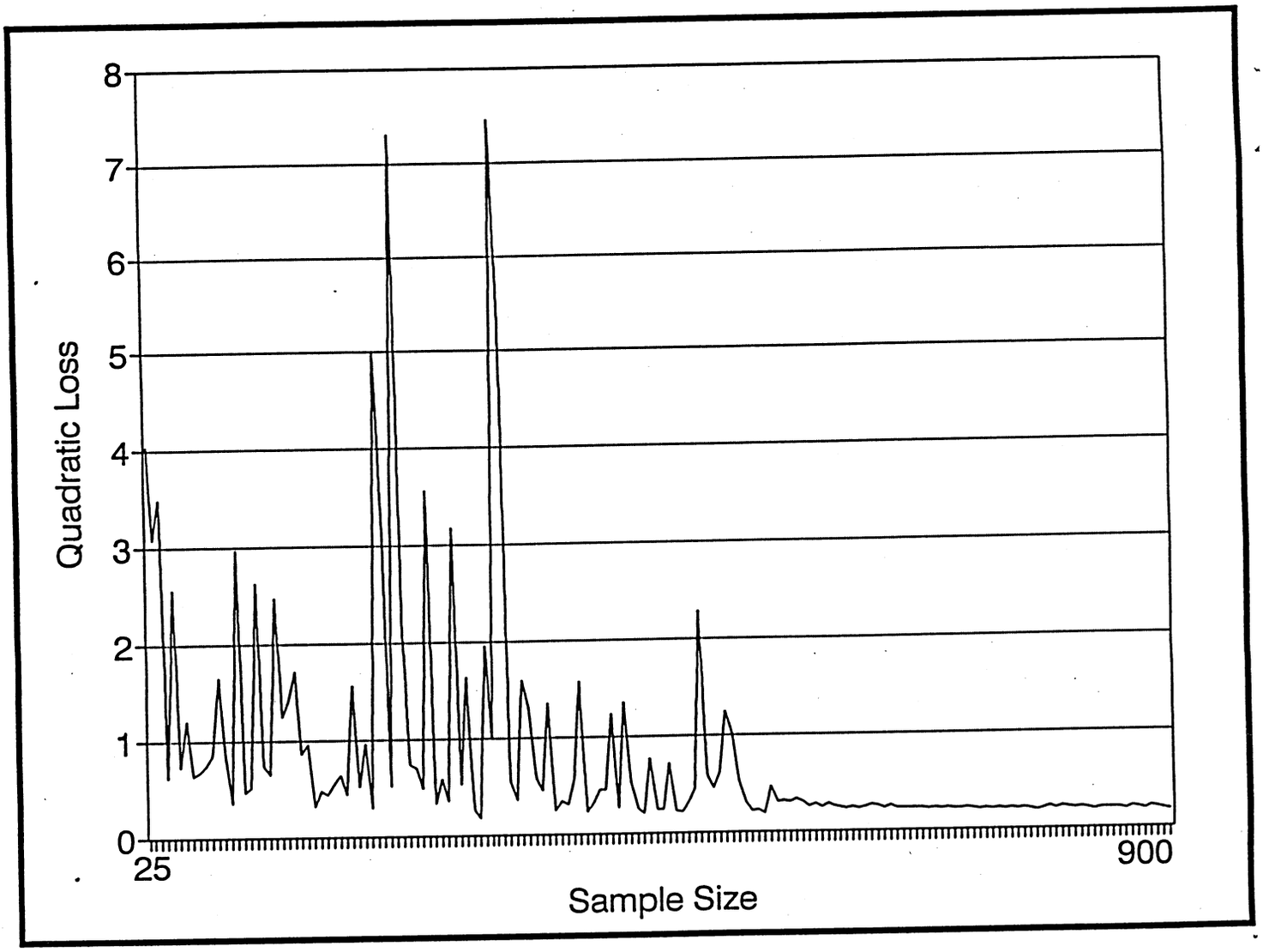
Note: The minimum sample size calculated was 25, and all the parameters are equal to 0.5.

Graph 2: Quadratic Loss of the Parameter α_1 for the OLS and G2SLS(OLS)



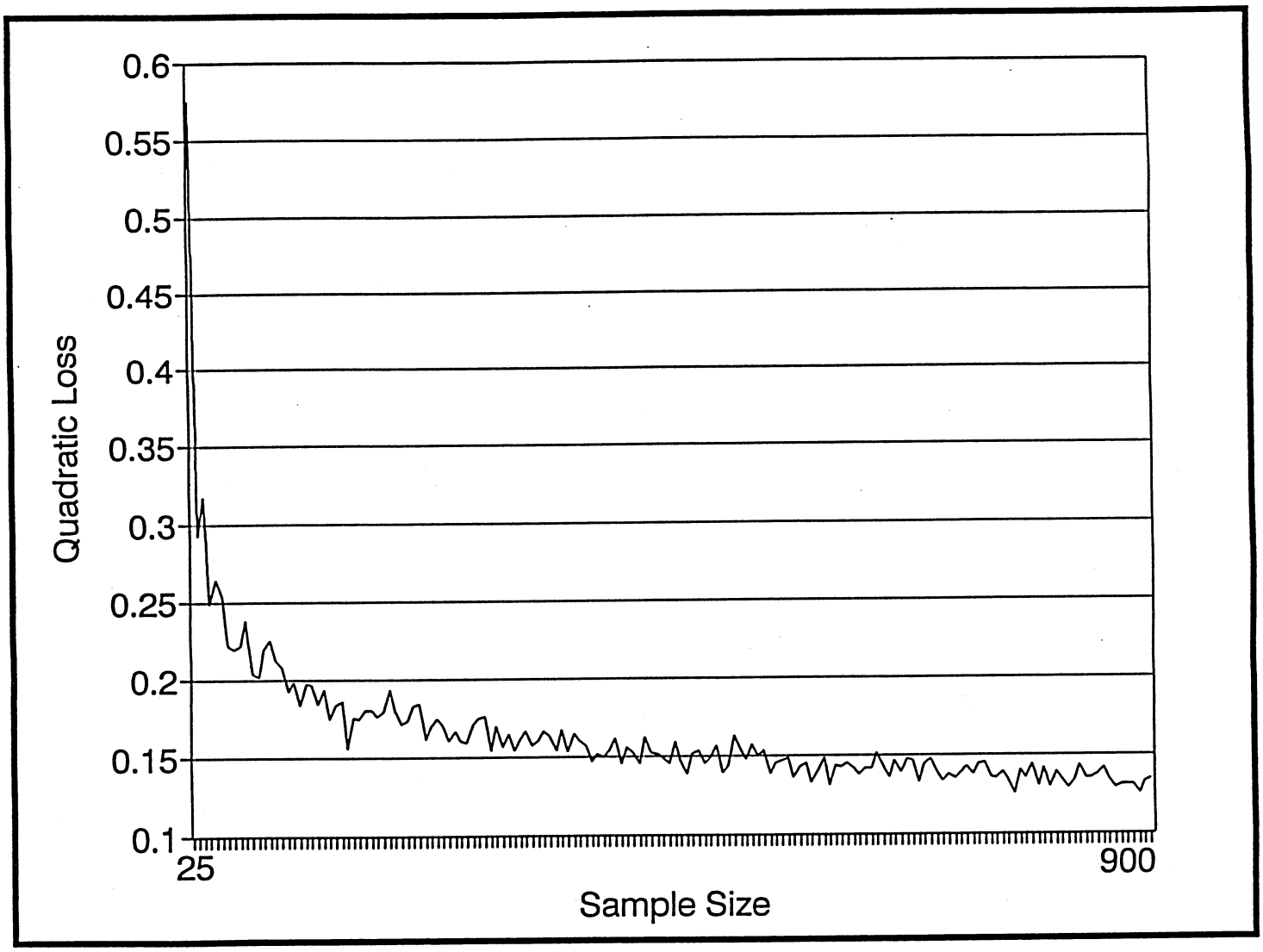
Note: The minimum sample size calculated was 25 and all the parameters are equal to 0.5.

Graph 3: Quadratic Loss of the OLS for the First Equation ($\alpha^* > 1$) of the Dynamic Model



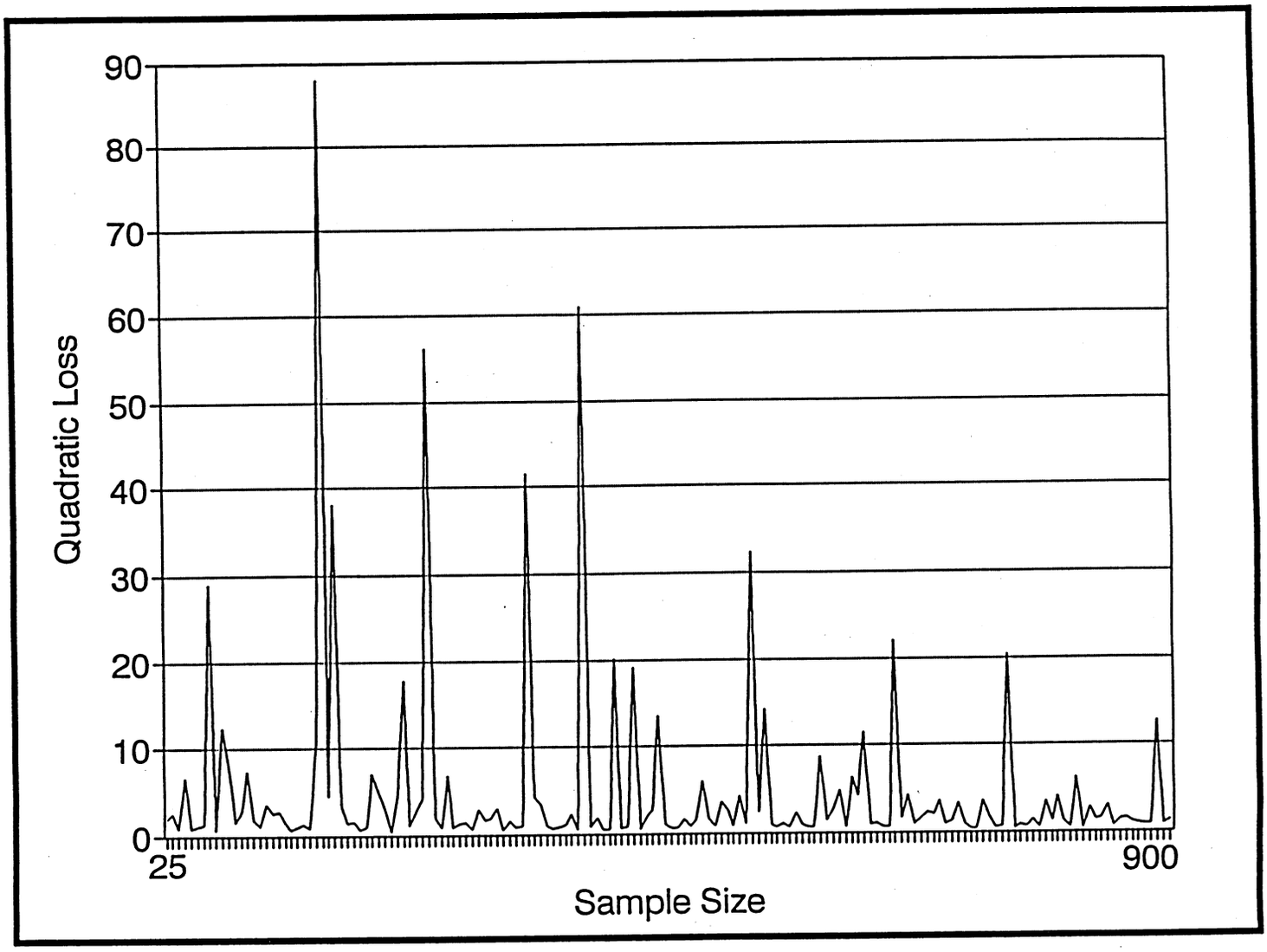
Note: The minimum sample size calculated was 25 and all the parameters are equal to 0.5 except $\alpha^* = 1.1$.

Graph 4: Quadratic Loss of the OLS for the First Equation of the Dynamic Model



Note: The minimum sample size calculated was 25 and all the parameters are equal to 0.5.

Graph 5: Quadratic Loss of the G2SLS(OLS) for the First Equation of the Dynamic Model



Note: The minimum sample size calculated was 25 and all the parameters are equal to 0.5.

References

Ahn, S. C. – Schmidt, P. [1990]: Efficient Estimation of Models for Dynamic Panel Data; *Paper presented at the Third Conference on Panel Data*, Paris.

Balestra, P. – Varadharajan-Krishnakumar, J. [1987]: Full Information Estimations of a System of Simultaneous Equations with Error Component Structure; *Econometric Theory*, **3**, pp 223–246.

Baltagi, B.H. [1981]: Simultaneous Equations with Error Components; *Journal of Econometrics*, **17**, pp 189–200.

Baltagi, B. H. – Li, Q. [1991]: A Note on the Estimation of Simultaneous Equations with Error Components; *Texas A & M University*, Department of Economics.

Biorn, E. [1981]: Estimating Economic Relations from Incomplete Cross Section and Time Series Data; *Journal of Econometrics*, **16**, pp. 221–236.

Breusch, T. S. – Mizon, G. E. – Schmidt, P. [1989]: Efficient Estimation Using Panel Data; *Econometrica*, **57**, pp. 695–700.

Heckman, J. [1979]: Sample Selection Bias as a Specification Error; *Econometrica*, **47**, pp. 153–161.

Krishnakumar, J. [1988]: *Estimation of Simultaneous Equation Models with Error Components Structure*; Springer – Verlag, 312.

Lovrics, L. – Mátyás, L. [1990]: Small Sample Properties of Simultaneous Error Components Models; *Economics Letters*, **32**, pp. 25–34.

Mátyás, L. – Lovrics, L. [1991]: Missing Observations and Panel Data; *Economics Letters*, **37**, pp 39–44.

Verbeek, M. — Nijman, T. [1990]: Testing for Selectivity Bias in Panel Data Models; *Paper Presented in Barcelona at the 6th World Congress of the Econometric Society*, Barcelona.

Prucha, I. R. [1985]: Maximum Likelihood and Instrumental Variable Estimation in Simultaneous Equation System with Error Components; *International Economic Review*, **26**, pp 491–506.

Wansbeek, T. J. — Kapteyn, A. [1989]: Estimation of the Error Components Model with Incomplete Panels; *Journal of Econometrics*, **41**, pp. 341–361.