MONASH
UNIVERSITY

AUSTRALIA

MISSING OBSERVATIONS AND PANEL DATA

A MONTE-CARLO ANALYSIS

László Mátyás and László Lovrics

Working Paper No. 2/91

February 1991

DEPARTMENT OF ECONOMETRICS

ISSN 1032-3813

ISBN 0 86746 980 3

MISSING OBSERVATIONS AND PANEL DATA

A MONTE-CARLO ANALYSIS

by László Mátyás* and László Lovrics**

* Monash University, Australia and Budapest University of Economics

** Budapest University of Economics

Working Paper No. 2/91

February 1991

DEPARTMENT OF ECONOMETRICS, FACULTY OF ECONOMICS COMMERCE & MANAGEMENT

MONASH UNIVERSITY, CLAYTON, VICTORIA 3168, AUSTRALIA.

Missing Observations and Panel Data a Monte–Carlo Analysis

by László Mátyás* and László Lovrics**

* Monash University, Australia and Budapest University of Economics

** Budapest University of Economics

Abstract: By means of Monte–Carlo experiments the loss of efficiency of the main error components' models estimators is analysed if from a panel data base those individuals are dropped for which the observations are not complete. An empirical risk function has been estimated. This can help to measure the risk of the use of complete sub–panels instead of the original but incomplete ones.

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

I. Introduction

Over the last two decades the use of panel data has attracted a lot of attention in econometrics (*e.g., Balestra–Nerlove* [1966], *Wallace–Hussain* [1969], *Hsiao* [1986]). 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 (*Wansbeek–Kapteyn* [1989], *Baltagi* [1985], *Biorn* [1981]), but these are in general quite complex. Another possibility is 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.)

It has been shown (*Verbeek–Nijman* [1990]) that if we have unbalanced or incomplete panels and we use the usual estimators of panel models based on a balanced and complete sub–panel, these are (asymptotically) unbiased and consistent (except the OLS) under quite general and reasonable conditions in the case when the observations are missing at random (so there is no selectivity bias present). But an important practical problem has been ignored in the literature. Which is better: to use the (more complex) unbalanced methods or to use the traditional balanced

methods based on a balanced sub–panel of the original data base? Or, in a more general form: What should we do when we have "holes" in the panel and we can suppose that these missing observations are missing at random? Do we apply more complex tools (estimators) or we can just drop from the panel those individuals for which there are missing observation(s)?

Because the consistency of the main usual estimators is guaranteed whether we use unbalanced panels or balanced sub–panels, the problem to analyse is the increase of risk (or the loss of efficiency) of the estimation caused by the use of balanced sub–panels. If this loss is not important it is not worth it to use the more complex unbalanced methods. We can just drop those individuals for which the observations are not complete. On the contrary, if the loss of efficiency is important (so the risk of the estimation increases) the more complex methods developed for unbalanced panels should be used.

In this paper we analyse the behaviour of the error components model through a Monte–Carlo experiment—because it is the most commonly used—and we focus our attention on the risk of two estimators: the within and the feasible GLS (FGLS) (*Hsiao* [1986]).

II. The design of the Experiment

The basic framework of our experiments is defined as follows:

- The analysed model has an error components specification with both individual and time effects:

$$y = X\beta + u\,,$$

or

$$y_{it} = X'_{it}\beta + u_{it}\,,$$

where u_{it} can be decomposed as $u_{it} = \mu_i + \lambda_t + v_{it}$ (or with matrix notations $u = \mu \otimes l_T + \widetilde{I}\lambda + v$), μ_i is the random variable of individual effects, μ is the random vector of individual effects ($N \times 1$), λ_t is the random variable of time effects, λ is the random vector of time effects ($T \times 1$). v_{it} is the usual error term, and v is the vector of the error terms, N is the number of observed individuals and T is the length of the observed time series.

- The risk of the estimators is represented by the expected value of a quadratic loss function $((\widehat{\beta} - \beta)'(\widehat{\beta} - \beta))$.

- For a given model (for example, the model with two exogenous variables: $y = a_1x_1 + a_2x_2 + u$), given parameter values, and given sample size, we generated one hundred samples, and for the given sample we performed both within and the FGLS estimations.
- We analysed linear models up to five exogenous variables with β parameter values 0.5, 1, 2, 5, and 10, and with variance components $Var(\mu_i)$, $Var(\lambda_t)$ and $Var(v_{it})$ values 0.5, 1, 2, and 5. (Several scenarios are possible from the combination of these parameters.)
- In the data generation process the exogenous variables were generated by a stationary autoregresive process.

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. 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 $10 < N < 50$ and $5 < T < 20$) and then we came back to the first step. We performed these steps for all the possible scenarios defined by the above mentioned parameters. We wanted to answer the following 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?
- Is it possible to define a unique empirical function which defines the risk of the estimation as function of the size of the balanced sub–panel and eventualy more factors?

Our analysis was quite difficult. The degree 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. In total we performed about 500,000 regressions which would be impossible without the new very fast PCs. This is a possible explanation of why nobody tried to answer these simple but relevant questions up to now.

III. Monte–Carlo Results

The results of the Monte–Carlo experiments were very satisfactory.

- We found that the risk of the estimation does not depend essentially on the parameters β of the model. Although there were some differences depending on these parameters, this variability always remained within a well defined interval.
- It seems that from the point of view of the experiment there is no important difference between the within and the FGLS estimators. For large samples ($NT > 300$) the within had a bit lower risk while for smaller samples the FGLS performed a bit better. This is easy to understand, because the loss of information due to orthogonal projection of the within estimator is very significant when the sample size is not large enough.
- Our initial supposition was that the risk of an estimator depends on the actual sample size, the relative sample size (the proportion of the actual sample size where the individuals with missing observations were dropped and the original sample size) and the variance $Var(v_{it})$ of the real error term. It was surprising that the relative sample size practically never had a significative effect on the risk of the estimation.
- We found that the values of N and T separately did not effect the risk while $N > 20$ and $T > 5$. Then only the whole NT actual sample size had importance.
- For very short time series or few observed individuals the risk of both estimators increased explosively even if NT remained large enough.

Based on the Monte–Carlo experiment it is possible to define an *empirical risk function* which gives us information about the risk of an estimator in function of the actual sample size and the $Var(v_{it})$ error variance. This empirical risk function is:

$$RF = \frac{1}{a + b\,ass + c\,var(v_{it})},$$

where RF is the risk of the estimation represented by the quadratic loss, a, b and c are parameters (estimated from the Monte–Carlo experiments) and ass is the actual sample size. We found that the value of the parameter a is within the interval [3, 7] and the value of b is within the interval [0.025, 0.04]. The result of the estimation of the overall risk function (based on more than 200,000 different samples and the above listed different scenarios) is summed up in the following tables.

Risk function, within estimator

variable	coefficient	t–statistic	R^2
const. (a)	7.790	46.45	
ass (b)	0.033	66.90	
$Var(v_{it})$ (c)	-3.580	-32.06	
			0.93

Risk function, FGLS estimator

variable	coefficient	t–statistic	R^2
const. (a)	6.91	40.07	
ass (b)	0.031	60.45	
$Var(v_{it})$ (c)	-3.08	-26.84	
			0.91

Using the above empirical risk function we are able to tell for a given case what is the loss of efficiency if we use a balanced sub–panel while ommiting those individuals for which there are missing observations. The necessary variance of the v_{it} error term can be estimated from the within residual. (The graphs of the overall risk function can be found in the appendix.)

IV. Conclusions

We have seen that the risk of the usual estimators of the error components models increases in function of the actual sample size and the residual variance when we drop from the panel those individuals for which observations are missing. Using the results of the empirical risk function we can decide in a specific analysis whether it is worth it to use the more complex unbalanced methods for the estimation of an error components model, or the use of an balanced sub–panel may be satisfactory. However in general we can say that if the remaining actual sample is large enough ($NT > 250$), the estimations based on a complete sub–panel are nearly as good as if the estimations were based on the entire panel (that is the loss of efficiency is negligible).

References

Balestra, P — Nerlove, M. [1966]: Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas; *Econometrica*, Vol. 34, pp. 585–612.

Baltagi, B. H. [1985]: Pooling Cross–Sections with Unequal Time–Series Lengths; *Economic Letters*, Vol. 18, pp. 133–136.

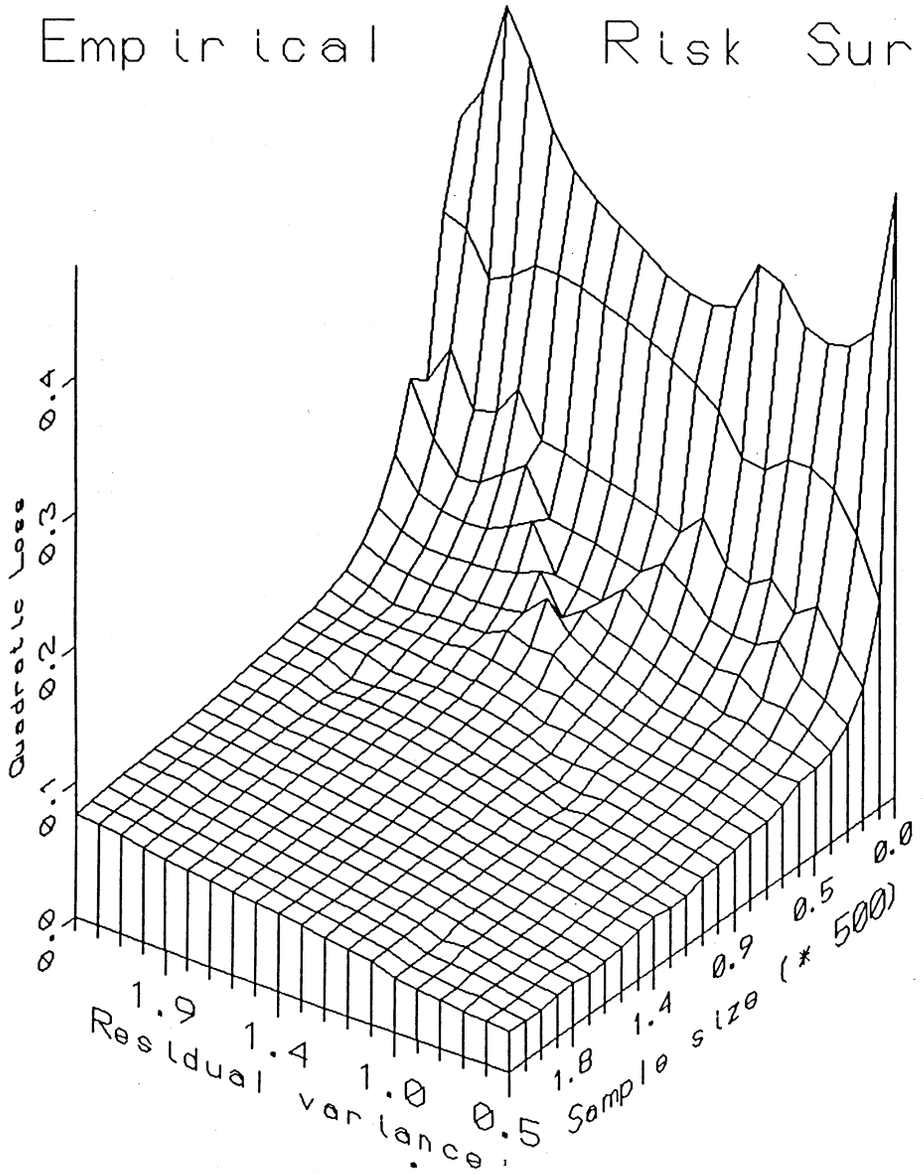
Biorn, E. [1981]: Estimating Economic Relations from Incomplete Cross Section and Time Series Data; *Journal of Econometrics*, Vol. 16, pp. 221–236.

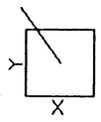
Hsiao, C. [1986]: *Analysis of Panel Data*; Cambridge University Press, Cambridge.

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.

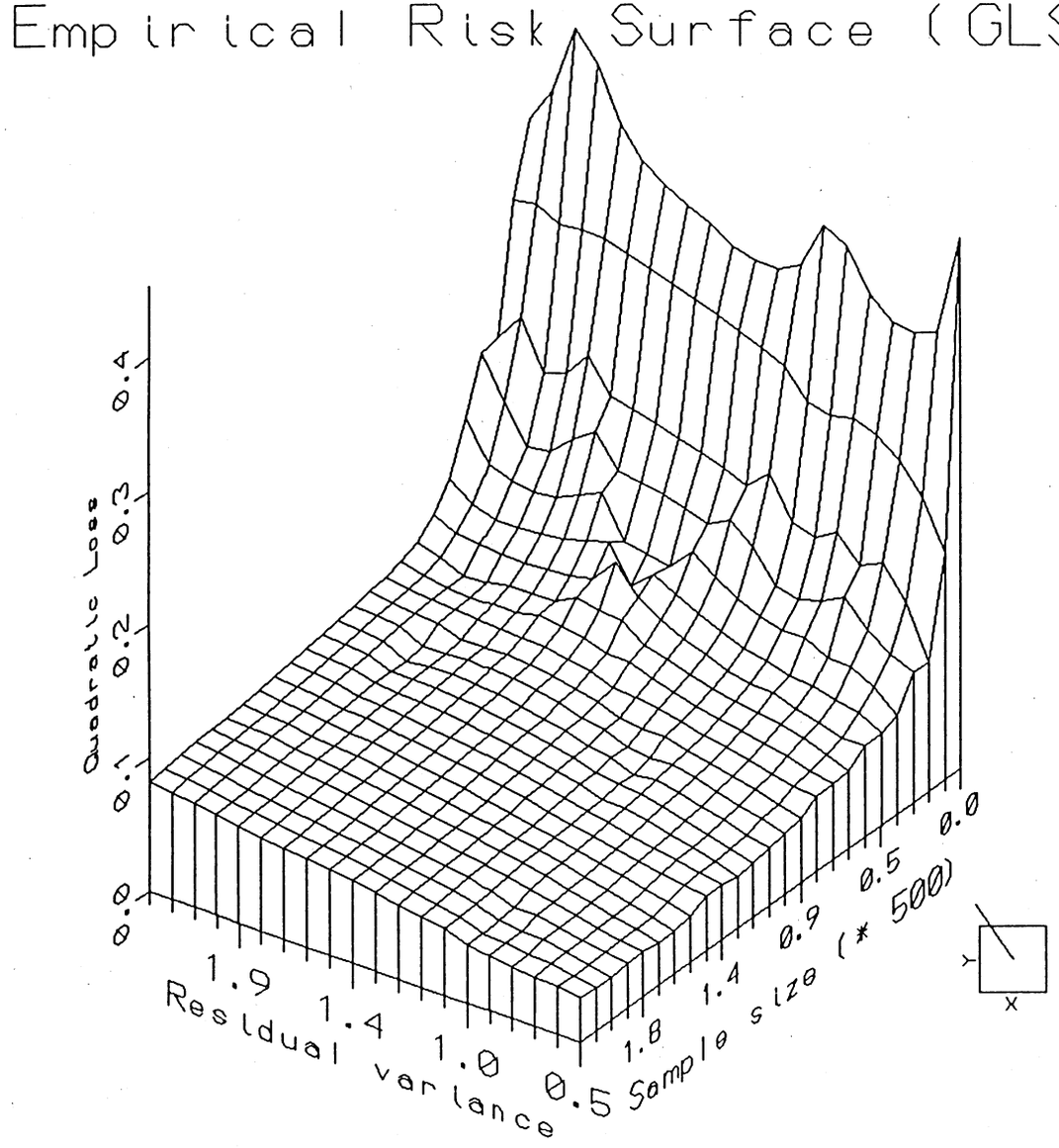
Wallace, T. D. — Hussain, A. [1969]: The Use of Error Components Models in Combining Cross Section with Time Series Data; *Econometrics*, Vol. 37, pp. 55–72.

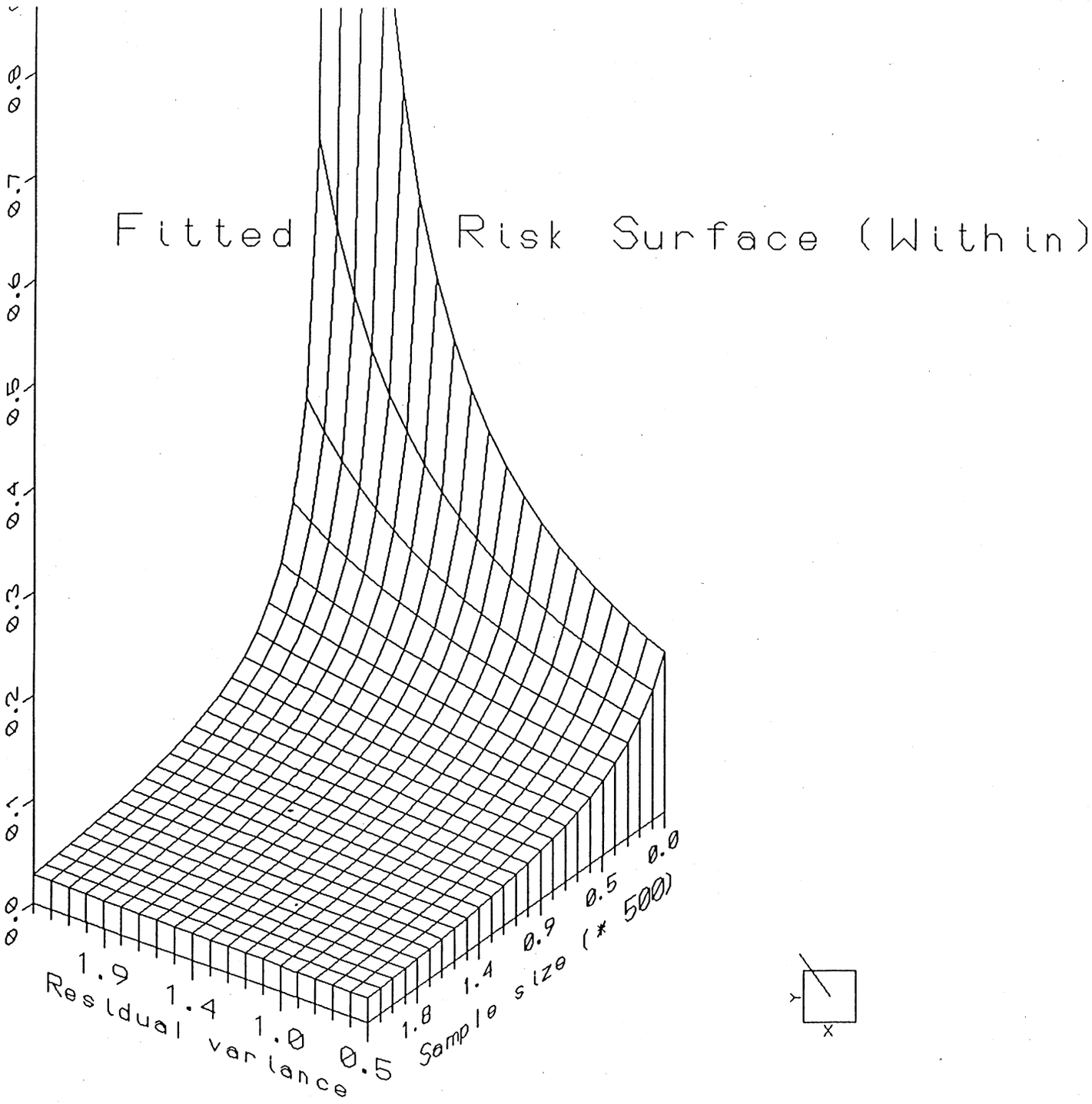
Wansbeek, T. J. — Kapteyn, A. [1989]: Estimation of the Error Components Model with Incomplete Panels; *Journal of Econometrics*, Vol. 41, pp. 341–361.

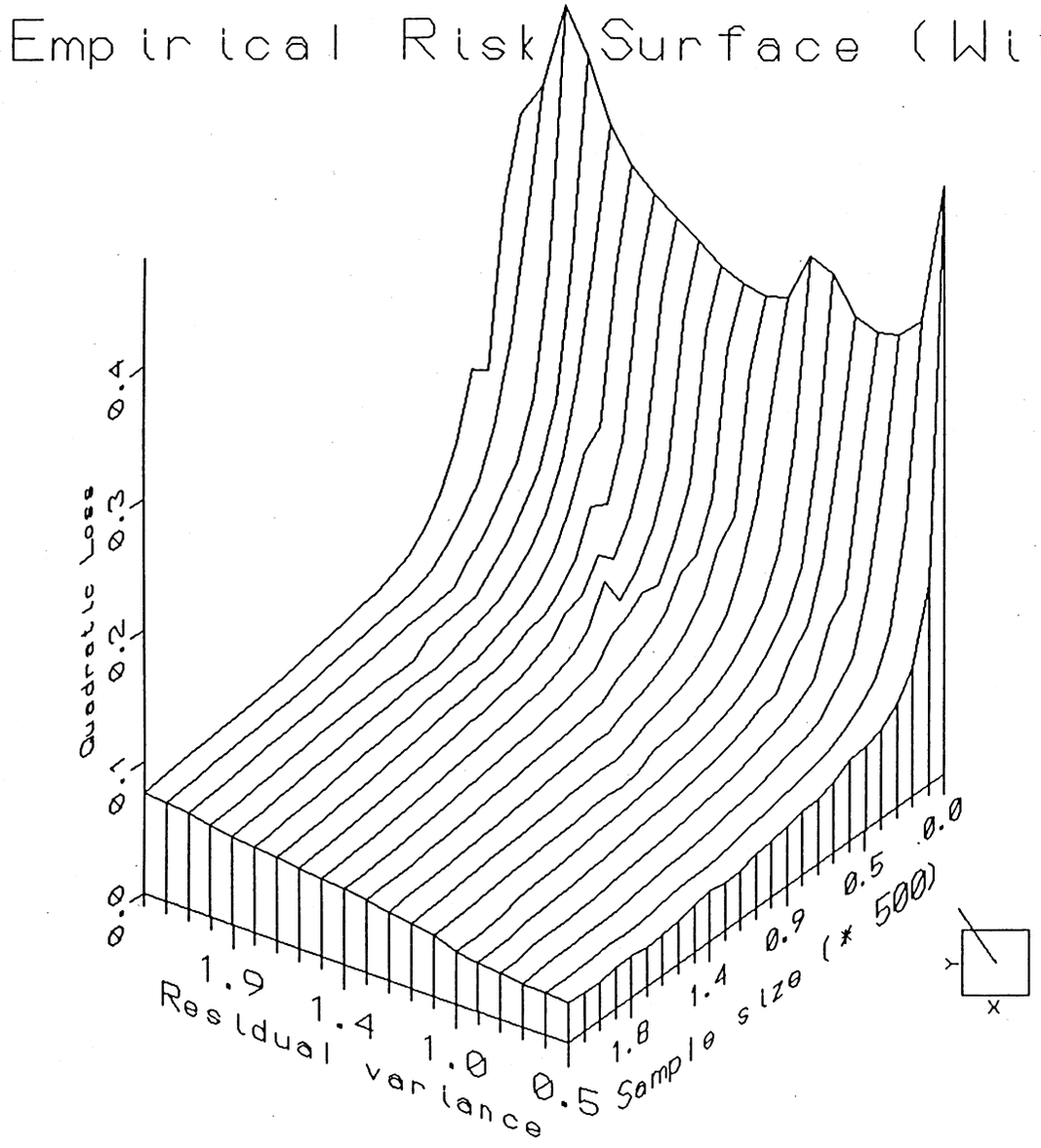




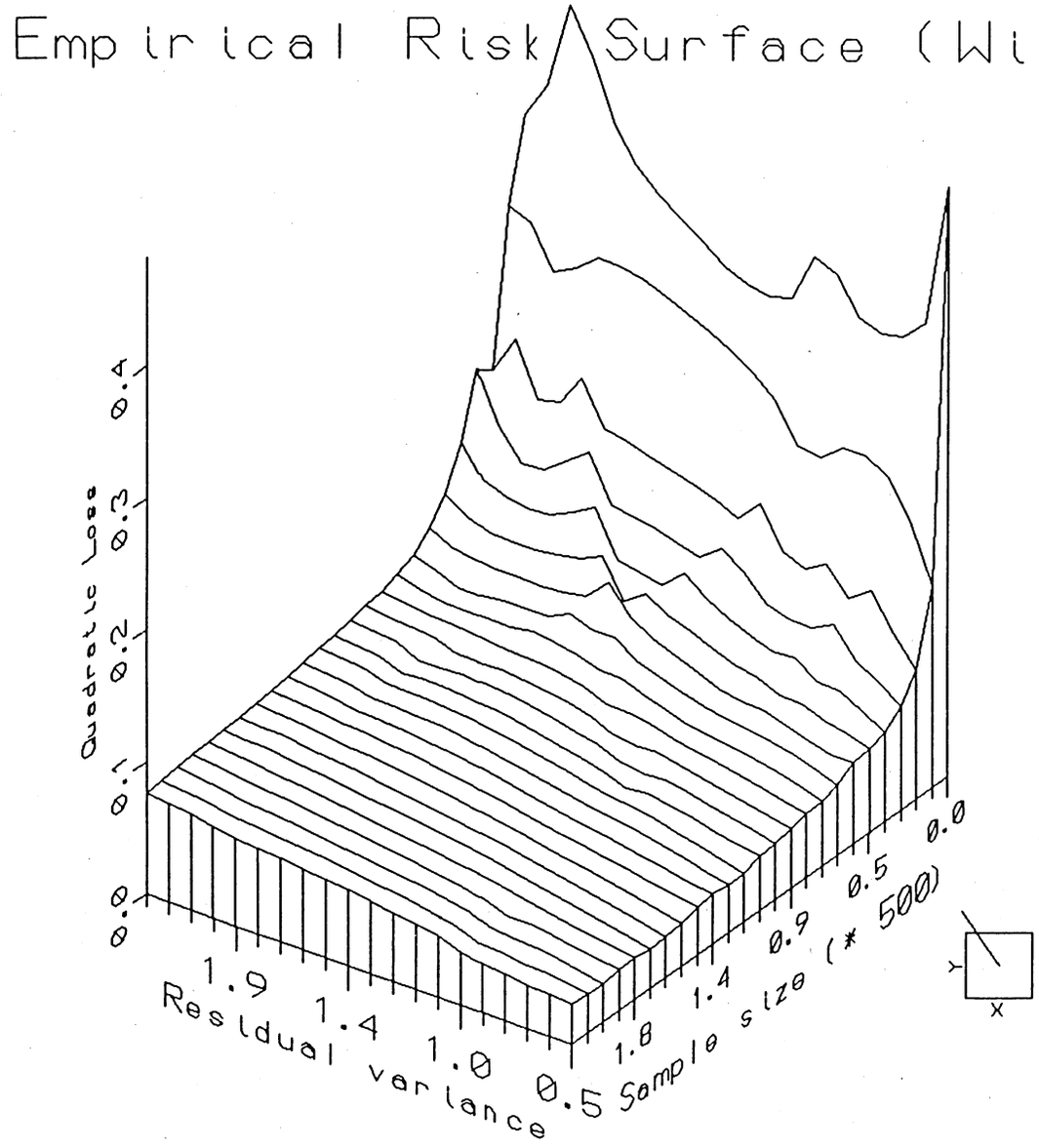
(It can be seen that practically there is no difference between the within and the FGLS empirical risk surfaces)


Empirical Risk Surface (GLS)
Quadratic Loss
0.0
0.1
0.2
0.3
0.4
1.9
1.4
1.0
0.5
Residual variance
1.8
1.4
0.9
0.5
0.0
Sample size (* 500)
Y
X

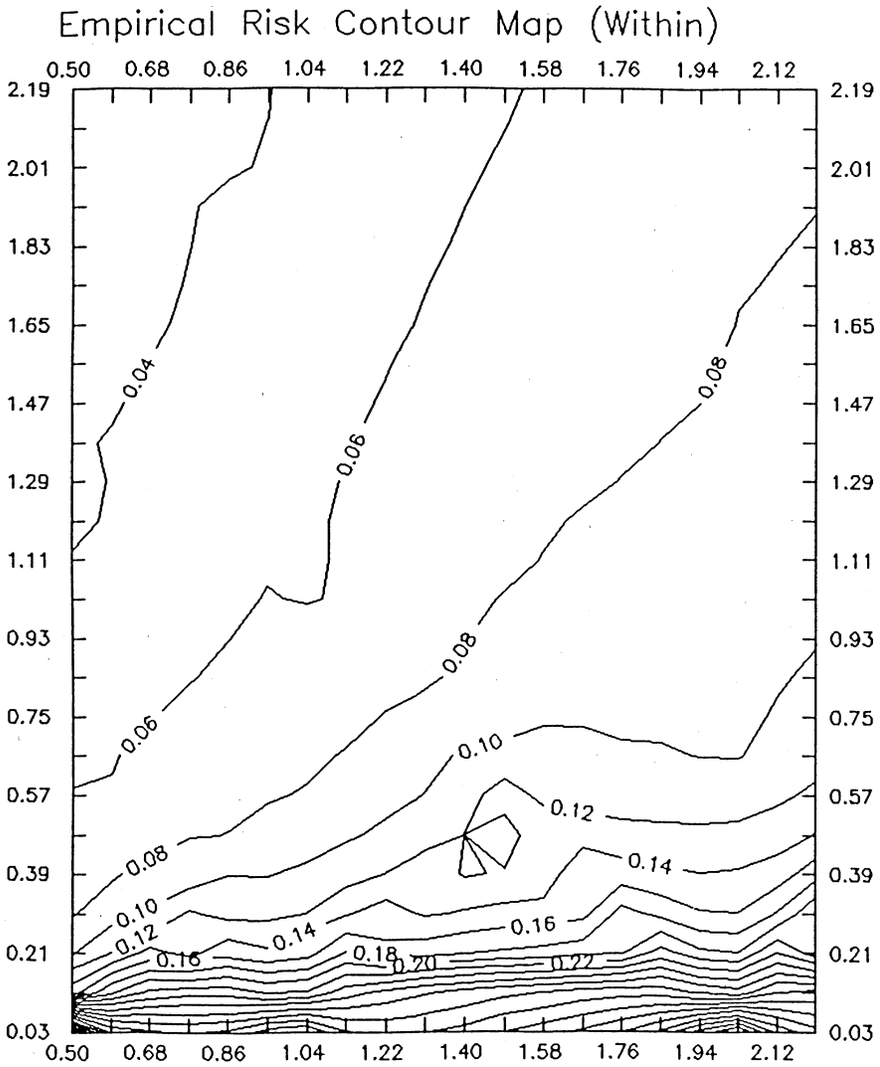

Fitted Risk Surface (Within)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
1.9
1.4
1.0
0.5
Residual variance
1.8
1.4
0.9
0.5
0.0
Sample size (* 500)
Y
X



(One convex line is the risk function for a given residual variance.)



(One line shows that for a given sample size how does the risk change.)


Empirical Risk Contour Map (Within)
0.50 0.68 0.86 1.04 1.22 1.40 1.58 1.76 1.94 2.12
2.19 2.01 1.83 1.65 1.47 1.29 1.11 0.93 0.75 0.57 0.39 0.21 0.03
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22