M O N A S H

U N I V E R S I T Y

AUSTRALIA

LINEAR MODELS FOR PANEL DATA

László Mátyás and P. Sevestre

Working Paper No. 6/93

June 1993

DEPARTMENT OF ECONOMETRICS

ISSN 1032-3813

ISBN 0 7326 0374 9

LINEAR MODELS FOR PANEL DATA

László Mátyás and P. Sevestre

Working Paper No. 6/93

June 1993

DEPARTMENT OF ECONOMETRICS, FACULTY OF ECONOMICS COMMERCE & MANAGEMENT

MONASH UNIVERSITY, CLAYTON, VICTORIA 3168, AUSTRALIA.

Preliminary version
Please do not quote

LINEAR MODELS FOR PANEL DATA

L. MATYAS[1] and P. SEVESTRE[2]

1 Monash University, Australia and Budapest University of Economics

2 Université d'Evry and ERUDITE - Université de Paris XII.

One of the main changes in applied econometrics during the last twenty years has been the increasing use of panel data (i.e. data which consist of repeated observations on firms, households, sectors, regions, countries,...etc). The reasons for this development are diverse. The evolution of economic theory, of statistical and econometric methods as well as the increasing means of data collection and computing facilities have favoured this evolution.

Indeed, developments in economic theory are more and more based upon microeconomic models relying on individual agents' behaviours and their possible heterogeneity. Panel data thus offer a better framework to match both the level of analysis of these models and the data available.

Moreover, since the pionneering works by Kuh [1959], Mundlak [1961], Hoch [1962], Balestra and Nerlove [1966], Wallace and Hussain [1965] and Maddala [1971], specific econometric methods dealing with such data have progressed considerably. They allow advantage to be taken of the particular characteristics of the data in a more efficient way. In particular, as Mundlak [1961] emphasized, these data allow the evaluation of the relative influence of factors affecting a variable, "free of individual heterogeneity". This possibility of taking into account the individual unobserved heterogeneity while estimating the coefficients of a regression model is probably the most notable advantage offered by panel data to applied economists.

At the same time, availability of such panel data sets has considerably increased and it is now common to have longitudinal data about firms and/or households which are obtained from repeated surveys, from administrative sources and from various publications.[3]

The aim of this paper is to emphasize the benefits, for applied economists, in using panel data for studying firm and consumer behaviours and to present the most usual models and methods used for dealing with them in a linear context. The plan of the paper is as follows: section 1 is devoted to a general discussion of the advantages and difficulties associated with the use of panel data in applied econometrics. In section 2, we give some illustrations of the superiority of panel data over aggregate time-series in getting reliable answers to questions about firms and households behaviour. Then, in section 3, we present the usual linear models for panel data. Sections 4 and 5 are devoted to the presentation of the most common estimation methods and tests in linear models with unobserved heterogeneity. In section 6, specific problems due to incomplete panels, selection, attrition and false panels are dealt with.

---

3 For example, firms which are quoted on a stock exchange generally have to make their accounts public. These accounts are then collected by various institutions and can be obtained for applied studies.

1 - Advantages and difficulties of using panel data in applied econometric studies

Panel data present important advantages over the usual cross sections and/or time-series:

- They have a double dimension, individual and temporal. This is a very important advantage. This double dimension allows the dynamics of behaviour and their heterogeneity to be taken into account simultaneously, something which is not possible with time-series nor with cross sections. Using the former amounts assuming, at least implicitly, that behaviour is homogeneous[4], whereas the latter does not allow one to estimate dynamic models. Moreover, let us again recall that this double dimension enables the influence of unobserved characteristics of individuals on their behaviour to be taken into account, as long as it is assumed to be stable over time.

- Another feature of panel data is that, generally, the dataset is large. It is not exceptional to have samples with several hundreds, and even several thousands, of observations. Indeed, if there are ten years of observations on two hundred individuals, this amounts to two thousand observations. Most often, these samples contain observations about a large number of individuals but cover only a short period. One of the main advantages of this large number of observations is that it leads to estimators whose properties can generally be analysed using asymptotic theory given the large number of individuals ("N-asymptotics" or semi-asymptotics). This means that if the necessary assumptions are satisfied, using a consistent estimator leads to estimates that can be assumed to be very close to the true value of the unknown parameters. In general, estimates based on panel data have high precision.

- A third interesting characteristic of these data sets is their important variability. The large differences observed between households, firms, sectors, regions and countries give rise to a very large "between-individuals variability". This between-individuals variability is most often much more important than its temporal counterpart, the within-individuals variability. This characteristics is rather obvious. Differences that can be observed on production, investment, capital stock,etc... between a firm which employs 100,000 workers and that of one employing 20 workers are much more important than the change over time of the value of these variables for the same firms. An intuitive way of interpreting the advantages associated with this large variability is to think about it in terms of information: panel data contain much information. This allows the specification of rather sophisticated models which could not be estimated with single time-series or cross sections. Moreover, these data have a larger discriminating power between alternative specifications. The well-known econometric problem that alternative specifications fit the data equally well is very rarely met when using panel data.

- A last but not minor advantage of these data sets is that most often, they consist of statistical observations at a micro level. This feature constitutes a double advantage. First, as it has previously been mentioned, many theoretical models are based on microeconomic beha-

---

4 Although it is possible to specify aggregate models in which heterogeneity is taken into account via regimes and/or the introduction of variables representing the distribution of a given variable across the total population.

viour; these datasets match better with the level of analysis than aggregate time-series. Second, these data sets allow the avoidance of some of the difficulties associated with aggregation (see Gorman [1953], Theil [1954], Kuh [1959], Edwards and Orcutt [1969], Blundell [1988], Körösi and Matyas (1991) among many others).

Although advantages of using panel data are important, their use can also present some difficulties:

- First of all, heterogeneity of individuals is generally important and it may be difficult to formalize it, but if it is ignored when present, it results in an ill specified model and biased estimates.

- Another difficulty is the non negligible frequency of missing observations (non responses and/or outliers) and the problem of panel attrition, which may result in the loss of representativity of the data.

The problem of outliers is a difficult one. In fact, it is not enough that, for an individual, an observation is far away from other observations on the same individual to consider it as an outlier. For example, the facts that the production of a firm is multiplied by four or five, or that the income of an household is doubled are not necessarily aberrant. They can be explained by a merger (for the firm) and a marriage (for the household). It is necessary to look at apparent outliers with caution. First of all, they should not be automatically taken out of the sample. In order to check whether an observation is an outlier, one can look at other related variables in order to see whether the change of their values is similar. If the production of a firm increases by 300% but its capital stock and its employment do not vary similarly, one may question the reliability of this observation.

If the outlier appears to be really an aberration, one can either eliminate it or try to make a correction. In the latter case, it is necessary that the available information is sufficient to ensure that the corrected figures are reliable. In the former case, it is important to stress that one should not eliminate all the observations on individuals which present such aberrant observations. The reason is that this would lead to a sample in which only those individuals which have a "quiet history" are included and could result in a selection bias problem. If the outlier is not a real one (for example, it results from a merger), then it is highly recommended not to discard it. On the contrary, once such particular observations have been identified it is important to see whether the events which are at the origin of these particular variations have an influence on the phenomenon under study.

Concerning the non responses (either voluntary or involuntary) which are rather numerous in panel data sets, the same kind of observations as those made about outliers apply. In particular, it is preferable not to eliminate all individuals for which some data are lacking. This could lead to an important loss in the number of observations and to eventual selection bias.

In order to conclude about the difficulties associated with the use of panel data, it is important to stress that a few years ago, it was common to work with panel data sets which were complete in the sense that individuals which were not observed during the whole sample period were excluded from the dataset. The development of methods to deal with incomplete panels and the fact that this could lead to selection bias have discouraged this practice. The usual practice now is to work with the maximum number of observations, even though some individuals are not present in the sample during all the period under consideration (i.e. to work with unbalanced panels).

Nevertheless, the usual presentation of estimation techniques for panel data rests on the assumption of completeness of the sample. We follow this "custom" in most of the paper but we will present the adaptations of the usual methods for incomplete panel data sets as well.

2- Assessing the superiority of panel data for the study of microeconomic units behaviour: some illustrative examples

It is not difficult to find in the literature sentences which claim the superiority of panel data over aggregate time-series for the study of economic behaviour. For example, in his survey about consumer behaviour, Blundell [1988] says"... therefore, the most persuasive level of analysis must be at the individual consumer or household level... The clear attraction of individual level data is that they avoid aggregation bias". Another example can be found in Hamermesh [1992]: "Obtaining large panels of annual data on firms is a useful step forward, as they allow us to circumvent any potential difficulties caused by heterogeneity of firms behaviour in non linear models of long run employment determination". It would be easy to find other quotations arguing in the same way.[5]

Historically, the first empirical studies based on the use of panel data sets were concerned with firms behaviour. More specifically, the question was essentially to evaluate the contribution of the production factors, given the heterogeneity that characterizes firms (Mundlak [1961], Hoch [1962]). Later on, other aspects of firms' behaviour (investment, labour demand, dividends distribution) were empirically studied. Consumers' behaviour, in relation to labour supply and consumption has also been the subject of many applied econometric studies using panel data. In this section, we do not survey all these studies. We focus on how panel data allow us to take into account both behavioural heterogeneity and dynamics.

As mentioned above, the estimation of production functions using firms' panel data has a rather long history in applied econometrics. Indeed, pioneering works in this area are the papers by Mundlak [1961] and by Hoch [1962]. In his paper, Mundlak provides estimates of Cobb-Douglas production function for 66 farms in Israël, observed over the period 1954-1958. One of his most striking results is that excluding the specific farm effects from the regression results in biased elasticities of production with respect to the various inputs: "In both sets of comparison the firm effect turns out to be highly significant. The implication is that the usual regression which is computed by not allowing for the firm effect is likely to be subject to a bias." Hoch's

---

5 For other defences of the usefulness of panel data, one can refer to Hsiao [1985,1986], Dormont [1989], Klevmarken [1989] or Baltagi and Raj [1992].

paper is devoted to the same subject. He estimates a Cobb-Douglas production function with individual and time fixed effects: the former "...will reflect differences in technical efficiency between firms;...The time constant...will reflect change in technical efficiency over time and differences in weather between years." This last interpretation of time dummies is due to the fact that Hoch considers 63 Minnesota farms whose output and inputs were observed over the years 1946 through 1961. Moreover, in his discussion about the farm effects, he considers the possibility that they represent, as Mundlak emphasized, the "entrepreneurial capacity". His estimation results show that both the farm and time specific effects are significantly different from zero.

Since then, many other papers have been devoted to the estimation of production functions[6]. In particular, the impact of technological progress or that of R&D on production have been considered in many studies (Ringstad [1971], Mairesse [1978], Griliches [1979,1980], Mansfield [1980], Schankerman [1981], Griliches and Mairesse [1983, 1984], Cunéo and Mairesse [1984], Griliches [1986], Jaffé [1986], Odagiri and Iwata [1986], Cette and Szpiro [1988], Fecher and Perelman [1989], Lichtenberg and Siegel [1989], Griliches and Mairesse [1990], Hall and Mairesse [1992]...[7]), as well as the influence of the skill of workers on productivity (Mairesse and Sassenou [1989], Sevestre [1990],Crépon and Mairesse [1993]). Whatever aspect they emphasize, these studies constitute good illustrations of the advantages and difficulties associated with the use of panel data in applied econometrics. The first issue is how firms' heterogeneity is taken into account in these studies. Most often, a Cobb-Douglas production function such as

$$(2.1)\quad \log Q_{nt} = a_0 + a_1 t + a_2 \log K_{nt} + \sum_j a_{3j} \log L_{jnt} + a_4 \log KR\&D_{nt} + u_n + w_{nt}$$

is estimated,[8] where K is a measure of the physical capital stock, L_j, j=1,...,J are different labour inputs (classified according to their skill level or to their direct or indirect relation with production) and $KR\&D$ is a measure of the stock of "scientific knowledge" of the firm. Unobserved heterogeneity is taken into account via the individual specific effects u_n.

A rather general observation that can be made about the econometric results reported in these studies is that, most often, estimates of the parameters obtained using the individual variability of the data (i.e. associated with the Between or the OLS regressions) are significantly different from those obtained when restricted to their time variability (i.e. using the Within, or the first (or longer)-differences regressions). Since all the resulting estimators are theoretically consistent (when N tends to ∞), whenever the model is correctly specified,[9] this has led researchers to look for possible explanations of these discrepancies.

---

6 Note the increasing literature on the estimation of production function frontiers which do not fit strictly in the framework that we consider here (on this topic, one can refer in particular to the survey by Schmidt and Sickles [1984] and to recent papers by Cornwell, Schmidt and Sickles [1990] and Kumbhakar [1990]).

7 Most of these studies are surveyed in Mohnen [1992] and Mairesse and Sassenou [1992].

8 A less restrictive specification has been considered by Sevestre [1990] and Crépon and Mairesse [1993] where a translog production function is also considered. Because of the squared and cross terms, this allows some firm variability in the elasticities of production with respect to the inputs.

9 See section 4.

Two main possible sources of biases are considered. First, particular attention is paid in most of these studies to the correlation between the specific firm effects and the production factors. Indeed, if one assumes, following Mundlak [1961], that these effects account for the quality of the management, it is hardly believable that they are uncorrelated with the quantities of the different inputs used by the firm. In this case, among the usual estimators, only the Within estimator, as well as the estimators of the model written in differences are consistent. Unfortunately, the resulting estimates are often not really plausible, which means that there exists other sources of bias than the correlation between the effects and the regressors. Then, the second invoked cause of bias is the existence of measurement errors on the production factors quantities. These can originate in the ignorance of degrees of utilisation, in errors in the construction of the capital stock series (for both the physical capital and the R&D capital) or in real measurement errors. It has been shown (see Griliches and Hausman [1986]) that the Within estimator, as well as the "difference estimators" are very sensitive to measurement errors, while the OLS and Between estimators are more robust against these errors. That is, the OLS and Between estimators biases come essentially from the correlation of the specific effects with the regressors, while those of the Within and difference estimators result from the existence of measurement errors.

Another possible explanation of the discrepancy between these estimators has been considered by Mairesse [1988] and Griliches and Mairesse [1988]. They consider the following model:

$$(2.2)\quad \log(Q/L)_{nt} = a_{0i} + a_{2i}\log(K/L)_{nt} + w_{nt}$$

where Q/L is a measure of the output per employee and K/L is the capital per employee and where various assumptions are made about the coefficients a_{0i} and a_{1i}. They can be all equal, different and fixed or different and random with means a_0 and a_1 and variances σ^2_{a0} and σ^2_{a1}.

They show that the dispersion of individual estimates of the production elasticity with respect to capital (a_1) is high and come from three sources: the sampling variance, a true heterogeneity in the parameters and possible misspecifications of the model (in particular, it can be the case that all the a_{0i} and a_{1i}'s do not have the same mean). Given that the Within and Between estimators can be interpreted in terms of weighted means of individual estimators, the existence of such heterogeneity could also help in explaining the unfortunately usual discrepancy between the Within and Between estimates of production function parameters. Nevertheless, it seems more likely that the main causes of these biases are the first two previously mentioned. The problem is then to use estimation methods which can deal with both correlated specific effects and measurement errors. These are either instrumental variables methods or the Π matrix approach proposed by Chamberlain [1982,1984].

Another domain of the firms' behaviour in which panel data studies are getting more and more numerous is the demand for inputs, i.e. labour demand, investment in physical assets and investment in R&D. Before presenting some issues relative to labour demand, let us first consider some panel data studies concerned with investment in physical capital. A particular interest has been granted in the last few years to the Q model (see Abel [1980]). According to this model,

the level of investment depends on the ratio of the increase in the value of the firm generated by this investment, to the price of investing one supplementary unit of capital; this ratio being known as the Tobin's marginal q:[10]

$$(2.3)\quad \left(\frac{I}{K}\right)_{nt} = c + \frac{1}{b}\left((q_{nt}-1)\,\frac{p_t^I}{p_t}\right)$$

where p_t^I is the price of capital, and p_t that of output.

If the q ratio increases, then it is profitable to the firm to invest. One important feature of this model is that, although its formulation is not explicitly dynamic, it does rely on the assumption that capital adjustments are costly and make the investment decisions of all future periods interdependent. Firms' expectations are also taken into account in q. Unfortunately, this model is not estimable because we do not have observations on q. Nevertheless, under some supplementary assumptions (constant returns to scale, efficiency of the stock market), it has been shown (see Hayashi [1982]), that the marginal q equals the average or Tobin's q defined as the ratio of the stock market's valuation of the firm's capital to the replacement cost of this capital. The estimable model is then

$$(2.4)\quad \left(\frac{I}{K}\right)_{nt} = c + \frac{1}{b}\left(\left(\frac{V_{nt}}{(1-\delta)p_t^I K_{n,t-1}}-1\right)\frac{p_t^I}{p_t}\right)$$

which is estimated as

$$(2.5)\quad \left(\frac{I}{K}\right)_{nt} = c + \frac{1}{b}\,Q_{nt}+\varepsilon_{nt}$$

where the disturbances ε_{nt} enter the model in a somewhat ad-hoc way. Many empirical studies based on aggregate or semi-aggregate time-series have been devoted to the empirical evaluation of this model (see Abel [1980], Oulton [1981], Summers [1981], Blanchard and Wyplosz [1983], Abel and Blanchard [1986], Chan-Lee [1986], Chan-Lee and Torres [1987], ...). The main conclusion that can be drawn from these studies is that, though q has some explanatory power of investment variations, it leaves an important part of these variations unexplained. The inclusion of supplementary explanatory variables such as lagged values of q, profits or variations in demand appear to significantly improve the results. Among the questions that have arisen from these deceptive results, one concerns the aggregation problem. Is it possible that, due to the heterogeneity of firms, this model could be rejected when using aggregate data (i.e. when ignoring this heterogeneity), whereas using panel data (i.e. taking this heterogeneity into account) could lead to more satisfactory results? A number of studies using panel data to test this model are devoted to this question (see Salinger and Summers [1983], Chappell and Cheng [1982], Schaller [1990], Hayashi and Inoue [1991], Blundell, Bond, Devereux and Schiantarelli [1992] among others).

---

10 Details about the derivation of this model can be found in Abel [1980], Hayashi [1982], or Blundell et al. [1992] among others.

Heterogeneity of firms can be taken into account in various ways. First, the computation of the q variable for each firm can allow for differences in effective tax rates, for the asymmetric treatment of profits and losses by the tax system. Moreover, in order to account for unobserved heterogeneity, the disturbances of the model are generally assumed to contain specific effects. The main difficulty which arises when estimating this model is that the q variable cannot be considered to be uncorrelated with the disturbances.[11] Then, one has to use instrumental variables methods such as those presented in section 5 below. This is done in Blundell et al. [1992] who estimate a model like (2.5) with a panel of 532 British firms observed over the period 1971-1986. Their results lead to the rejection of the model because the idiosyncratic component of the disturbances exhibits some serial correlation. Then, they rewrite this model as autoregressive along the lines suggested by Durbin [1960], and estimate it with the appropriate methods (see section 5.3). This last model is not rejected and they obtain an estimate of b which equals 0.01 (and an estimate of 0.24 for the serial correlation coefficient). Nevertheless, these estimates lead to a very slow adjustment process of capital. Moreover, cash-flow and output variables, when added to the model, prove to have a significant influence on investment. This result still holds when more heterogeneity is taken into account by considering different sub-samples of firms according to the existence or not of a liquidity constraint measured by the capacity to pay dividends (see Fazzari et al. [1988] and Bond and Meghir [1990] though their model is different from model (2.5)), or by the relationships with banks (see Hoshi et al. [1991]).[12] These results show that the basic Q model does not seem to be very relevant for explaining investment variations. This means that the rejection of this model, found in studies using aggregate time-series data, cannot be explained by aggregation issues. This is the theoretical model which has to be reconsidered (see for example the extension of this model proposed by Malinvaud [1987]).

Labour demand is another aspect of the firms' behaviour where panel data prove to be useful for getting a better understanding of this behaviour. Indeed, since the pioneering work of Oi [1962], the quasi-fixity of labour is undisputed. The question is then to evaluate, as precisely as possible, the costs that firms face when they want to adjust their employment to its desired level and the consequences of these costs in terms of the employment dynamics. It is then easy to understand that using aggregate time-series data is not likely to provide relevant results. Indeed, as it has been emphasized by Hamermesh [1990], the use of aggregate data is disqualified as soon as there exists some heterogeneity across firms. Indeed, it is probably the case that in some firms, employment increases while it decreases in some others, due for example to different business conditions in different sectors. Then, at the macro level, observed variations of employment will underestimate the true employment movements. This will lead to an overestimation of employment stickiness, i.e. to an overestimation of adjustment costs. Working with panel data at the firm level will, on the contrary, allow a correct evaluation of these costs. Nevertheless, the analysis can be pursued and one can imagine that firms employ different categories of workers, whose employment levels do not necessarily vary in the same direction. The same problem as above arises, so it seems preferable to work with firms' data which allow

---

11 Contemporaneous shocks will obviously affect the value of the firm and the heterogeneity in tax rates can hardly be considered to be uncorrelated with the unobserved heterogeneity.

12 See also Devereux and Schiantarelli [1990].

a disaggregation of employment in different categories.[13] Using a panel data set concerning 580 French firms observed over the period 1975-1983, Bresson et al. [1992a] have estimated the following model for three categories of workers (engineers and technicians, skilled workers, unskilled workers):

$$(2.6)\ \log L_{j,t} = \mu_j \log L_{j,t-1} + \alpha_{0j} \log Q_t + \alpha_{1j} \log Q_{t-1} + \sum_{i \neq j} \gamma_{0i} \log\left(\frac{w_j}{w_i}\right)_t + \sum_{i \neq j} \gamma_{1i} \log\left(\frac{w_j}{w_i}\right)_{t-1}$$

$$+ \beta_{0j} \log\left(\frac{w}{c}\right)_t + \beta_{1j} \log\left(\frac{w}{c}\right)_{t-1} + \gamma_j D'_t + \delta_j + \omega_{j,t} \ , \ j = 1,2,3$$

Since the model is autoregressive, they have used a generalized method of moments estimator similar to that proposed by Arellano and Bond [1991].[14] Their results show that the dynamics of engineers and technicians employment and that of skilled workers is indeed very different from that of unskilled workers; the former exhibiting a rather fast adjustment whereas the latter have a slower speed of adjustment. Moreover, the impact of a change in expected production has a stronger and faster influence on the unskilled workers level of employment. Another result being worthy of notice is that, as was expected, the implied adjustment delay for total employment was longer when the sample was not restricted to firms in which employment variations were of the same sign than when it was. Last, the estimation of the delay was about a third of the one obtained for a similar period and model estimated with French aggregate time-series (see Maurel [1990]).

The other aspect of employment dynamics to which several panel data studies have been devoted in the last few years is the form of adjustment costs. Until recently, it was assumed that these costs were quadratic, which, let us recall, implies both convexity and a complete symmetry of hiring and firing costs, a feature that has been known to be unrealistic for a long time (Oi [1962]). Several studies have considered the case of fixed costs (Hamermesh [1989,1990]) or asymmetries with quadratic or non-quadratic adjustment costs (Bresson, Kramarz and Sevestre [1992b], Jaramillo, Schiantarelli and Sembenelli [1992], Lockwood and Manning [1992]). All these studies confirm what had been suspected for a long time, but without the possibility to really test for it, i.e. that labour adjustment costs are not quadratic. The non linearity of the models obtained under these assumptions reinforces the advantage of panel data over aggregate time-series since consistent aggregation of such non linear models is known to be impossible (see Hamermesh [1992]).

Another domain where this question of aggregation over economic agents is known to be of importance is that of consumer demand.[15] As was stressed by Blundell [1988] for example, consumption elasticities with respect to prices and income, are likely to vary across households,

---

13 One could still continue and consider that many firms own several plants in which employment does not necessarily evolve in the same direction. It would then be better to work with data at the plant level rather than at the firm level. Unfortunately, while data about employment often exist at this level of disaggregation, it is rarely the case for variables which determine employment.

14 See section 5.3 below.

15 e.g. see Gorman [1953], Mullbauer [1975,1976], Lau [1982], Jorgenson et al. [1982] and Stoker [1984].

depending for example on the family size, on tenure, on employment status,etc... This implies that, as soon as the household population structure with respect to these characteristics changes, these elasticities vary at the macroeconomic level over time. Moreover, even if one ignores this interdependance, using aggregate time-series makes it difficult to identify pure price and income elasticities on the one hand, and the effects of the households characteristics (i.e. of their distribution in the population at the macro level) on the other hand, because their respective variations are probably correlated. It seems preferable to study household consumption behaviour using microeconomic data as the influence of households characteristics, both observed and unobserved, can more readily be taken into account. Typical models are of the form:

$$(2.7)\quad S_{it}^g = \beta^g \; LogY_{it} + \sum_j \gamma_j^g \; Logp_{jt}^g + \sum_k \delta_k^g \; D_{kit}^g + \alpha_i^g + \varepsilon_{it}^g \quad i = 1,\ldots,N;\; t = 1,\ldots,T$$

where the superscript g means that the equation relates to the commodity (group) g, S is the share of expenditure for this commodity, Y is the disposable income, p_j, $j = 1,\ldots,J$ are relative prices and D_k are dummies which account for various socio-economic characteristics of the households. In this model, the constant term is supposed to vary across households. These household specific terms are generally assumed to be correlated with the regressors and then are considered as fixed constants to be estimated. This is possible without any restriction if true panel data are available. If not, i.e. if the available data are only repeated cross-sections, it is generally assumed that these terms vary with the measured households characteristics. In some cases, other coefficients are also allowed to vary across individuals, depending on their characteristics. For example, Blundell et al. [1989a] have estimated a model close to (2.7)[16] for 6 commodity groups, using 15 pooled cross-sections covering more than 63000 households over the period 1970-1984. Their results show that the intercepts of the equations for each commodity group strongly depend on household characteristics, as does the income elasticity.

Despite the evident advantage of panel data for studying consumption behaviour of households, some difficulties arise, which do not exist, or are at least less problematic, when working with aggregate time-series. Among these difficulties, we can mention errors in the measurement of income (e.g. see Altonji and Siow [1987] and Aasness, Biorn and Skjerpen [1988]), zero expenditures (indeed, since survey interviews generally cover, for each household, a rather short period, many of them report, for some commodities, zero expenditures; but this does not mean that they do not consume this good; e.g. see Deaton and Irish [1984], Keen [1986], Blundell and Meghir [1987], Robin [1991], Meghir and Robin [1992]). Another important aspect of consumption behaviour which has been studied using panel data is that of consumption dynamics (see Hall and Mishkin [1982], Shapiro [1983], and for various extensions Bernanke [1984], Hayashi [1985], Mariger [1987], Altonji and Siow [1987], Zeldes [1989] and Lam [1991] as well as the survey in Robin [1992]).

Before moving to the presentation of specific econometric methods to estimate and test these models, let us just mention that there are other aspects of firms' and consumers' behaviour which have been explored using panel data. R&D investment has been considered by Mairesse and Siu [1984], Hall and Hayashi [1988] and Hall [1990] among others; studies of the payment

---

16 The only difference is that they include a supplementary regressor, defined as the log of income squared.

of dividends by firms have been made by Chowdhury and Miles [1989], Kim and Maddala [1992] and Malécot [1992]. The determinants of direct foreign investment were recently explored by Balestra and Negassi [1992] and by Mathieu [1993]. Labour supply was considered in a large number of studies using panel or pseudo-panel data (e.g. see Heckman and MaCurdy [1980, 1982], MaCurdy [1981, 1983], Browning, Deaton and Irish [1985], Altonji [1986], Lilja [1986], Ham [1986], Jakubson [1988], Blundell, Browning and Meghir (1989), Bover [1991], and, for a survey, Laisney, Pohlmeier and Staat [1992]).

3 - Linear models with unobserved heterogeneity: an introduction

As pointed out above, one of the main advantages of panel data is that they allow us to take into account the heterogeneity of individuals (and, possibly, that of time periods). There are basically four ways to formalize unobserved heterogeneity: the fixed effects model, the error components model (the most commonly used) and their generalizations, the varying coefficients models and the random coefficients models.

3.1 The fixed effects and the varying coefficients models[17]

The fixed effects model is a relevant specification when the sample cannot be considered as a (random) drawing from a population. This is the case, for example, if the data are relative to geographical regions, economic sectors, countries, etc. Another situation where this specification appears to be adequate is when one suspects the presence of correlation between the unobserved heterogeneity factor and the explanatory variables of the model. An example is the number of years at school in an earnings function. It is indeed difficult to consider this variable as uncorrelated with the unobserved ability of the person.

In these cases, it is common to represent the unobserved heterogeneity by individual coefficients associated with dummy variables:[18]

$$(3.1)\quad y_{nt} = a_0 + a_n + \sum_{k=1}^{K} b_k x_{knt} + w_{nt}, \qquad n = 1,\ldots,N;\ \ t = 1,\ldots,T.$$

17 For the sake of simplicity, we will restrict ourselves to the case of individual heterogeneity. The case of time periods heterogeneity can be analysed in a similar way (see Balestra [1992]).

18 In fact, this model is not identified because the sum of the dummy variables related to the individuals equals the constant variable. It is then necessary to impose a constraint on the parameters. It is generally assumed that the sum of the fixed effects coefficients is zero.

where a_0 is the constant term, $a_n, n = 1,\ldots,N$ are the coefficients accounting for individual specific effects, $x_{knt}, k = 1,\ldots,K$ are the explanatory variables and $b_k, k = 1,\ldots,K$ their coefficients.

Assuming that the error term is a white noise, this model leads to:

$$E y_{nt} = a_0 + a_n + \sum_{k=1}^{K} b_k x_{knt}$$

and,

$$E y_{nt} y_{n't'} = \delta_{nn'}\ \delta_{tt'}\ \sigma_w^2$$

The individual heterogeneity then appears in the first order moments of the endogenous variable. This model is a standard regression model and as long as the usual assumptions about the disturbances are satisfied, the OLS is a BLU estimator.[19]

In this model, heterogeneity is taken into account via an "individualization" of the constant term. The coefficients of the real explanatory variables are the same for all individuals. This can be considered as too restrictive and one can think of a model where all coefficients vary across individuals:

$$(3.2)\quad y_{nt} = a_0 + a_n + \sum_{k=1}^{K} b_{kn} x_{knt} + w_{nt}, \qquad n = 1,\ldots,N;\ \ t = 1,\ldots,T.$$

with

$$b_{kn} = b_k + a_{kn}$$

Then,

$$E y_{nt} = a_0 + a_n + \sum_{k=1}^{K} (b_k + a_{kn}) x_{knt}$$

and,

$$E y_{nt} y_{n't'} = \delta_{nn'}\ \delta_{tt'}\ \sigma_w^2$$

Again, under standard assumptions, the OLS estimator of this model is BLUE.

This model, as the fixed effects model, is particularly relevant when one wishes to measure explicitly the differences in the behaviour of various individuals. But it must be noticed that it is not possible to make out-of-sample predictions with such models since there is no possibility of having an out-of-sample measure for the individual coefficients. Another drawback related to this model is that if the number of individuals is large, the number of coefficients to estimate can be quite unrealistic. For example, with a sample of 200 individuals and 5 explanatory variables, one gets 1000 coefficients to estimate! In this case, it can be a better choice to consider a random representation of the individual specificities.

---

19 The OLS on this model is often called Least Squares with Dummy Variables (LSDV), Within or covariance estimator.

3.2 The error components and the random coefficients models

In these models, the individual (and time) specific effects are incorporated into the model by using random variables.[20]

The error components model can be written as:

$$(3.3)\quad y_{nt} = a_0 + \sum_{k=1}^{K} b_k x_{knt} + u_n + w_{nt}, \qquad n = 1,\ldots,N; \quad t = 1,\ldots,T.$$

where a_0, x_{knt}, and b_k are defined as above and where:

$$Ew_{nt} = 0, \qquad Ew_{nt}w_{n't'} = \delta_{nn'}\ \delta_{tt'}\ \sigma_w^2$$

and,

$$Eu_n = 0, \qquad Eu_n u_{n'} = \delta_{nn'}\ \sigma_u^2, \qquad Eu_n w_{n't'} = 0.$$

Then,

$$Ey_{nt} = a_0 + \sum_{k=1}^{K} b_k x_{knt}$$

and,

$$Ey_{nt}y_{n't'} = \delta_{nn'}\ \delta_{tt'}\ \sigma_w^2 + \delta_{nn'}\ \sigma_u^2.$$

Unlike for the previous models, the error components model assumes that the individual heterogeneity can be formalized at the variance level. The presence of a random specific effect introduces a particular form of serial correlation in the disturbances of an individual. Then, OLS is not the best way to estimate the model. It is then necessary to use Feasible-GLS (or ML) methods to get asymptotically efficient estimates of the parameters.

One way to interpret the error components model is to consider it as a special case of the random coefficients model, where only the constant term is individualized and the specific effects are random.

Let us now consider the general form of the random coefficients model. It can be written as:

$$(3.4)\quad y_{nt} = a_0 + u_n + \sum_{k=1}^{K} b_{kn} x_{knt} + w_{nt}, \qquad n = 1,\ldots,N; \quad t = 1,\ldots,T.$$

with

$$b_{kn} = b_k + \mu_{kn}$$

---

20 Nevertheless, it is generally assumed that the best representation of time specific effects is obtained with time dummies, i.e. these effects are fixed rather than random. It is indeed difficult to think of these effects as a random drawing from a population.

and,

$$Ew_{nt}=0, \quad Ew_{nt}w_{n't'}=\delta_{nn'}\ \delta_{tt'}\ \sigma_w^2 \quad E\mu_{kn}=0, \quad E\mu_{kn}\mu_{k'n'}=\delta_{kk'}\ \delta_{nn'}\ \sigma_{\mu_k}^2 \quad [21]$$

$$E\mu_{kn}u_{n'} = E\mu_{kn}w_{n't}=0$$

This model can be rewritten as:

$$(3.5)\quad y_{nt}=a_0+\sum_{k=1}^{K} b_k x_{knt}+u_n+\sum_{k=1}^{K}\mu_{kn}x_{knt}+w_{nt}, \qquad n=1,\ldots,N;\ \ t=1,\ldots,T.$$

Then,

$$Ey_{nt}=a_0+\sum_{k=1}^{K} b_k x_{knt}$$

and,

$$Ey_{nt}y_{n't'}=\delta_{nn'}\ \delta_{tt'}\ \sigma_w^2+\delta_{nn'}\ \sigma_u^2+\sum_{k=1}^{K} x_{knt}^2\ \delta_{kk'}\ \delta_{nn'}\ \sigma_{\mu_k}^2.$$

The main problem related to the estimation of this model is that its disturbances are serially correlated and heteroscedastic. The OLS again is not a BLU estimator and it is preferable to use the Feasible-GLS.

Since the estimation problems related to static panel data models are mainly due to the presence of random specific effects in terms of heteroscedasticity and/or serial correlation and of the possible correlation of these effects with the regressors, we next concentrate on these two problems.

4 - Estimation and testing when specific effects are not correlated with the regressors

4.1 Estimation of a single equation

Let us consider the error components model:

$$(4.1)\quad y_{nt} = \sum_{k=1}^{K} b_k\ x_{knt}+\varepsilon_{nt}, \quad n=1,\ldots,N;\ \ t=1,\ldots,T.$$

with

$$\varepsilon_{nt}=u_n+w_{nt}$$

where

---

21 This assumption of absence of correlation between the coefficients is often relaxed.

$$Eu_n = 0, \quad Eu_n u_{n'} = \delta_{nn'} \ \sigma_u^2$$

$$Ew_{nt} = 0, \quad Ew_{nt} w_{n't'} = \delta_{nn'} \ \delta_{tt'} \ \sigma_w^2$$

$$Eu_n w_{n't} = 0, \ \forall n, n', t$$

As it has been stressed above, in this model, the disturbances are serially correlated over time for each individual, but there is no correlation across individuals. Moreover, the structure of serial correlation is particular in the sense that it is constant over time. This is a noticeable difference with the autocorrelation of disturbances in usual time-series models which is generally assumed to diminish over time.

Rewriting the model in matrix form:

$$(4.1') \quad \underset{(NT,1)}{y} = \underset{(NT,K)}{X} \ \underset{(K,1)}{b} + \underset{(NT,1)}{\varepsilon}$$

with

- $E\varepsilon = 0$
- $V\varepsilon = E\varepsilon \ \varepsilon' = \Omega$

where Ω is not scalar since the disturbances are serially correlated. It can be written as:

$$\Omega = \sigma_w^2 \left[I_N \otimes (I_T - J_T) + \frac{\sigma_w^2 + T\sigma_u^2}{\sigma_w^2} \ I_N \otimes \frac{J_T}{T} \right]$$

where I_N is the (N,N) identity matrix, I_T is the (T,T) identity matrix and J_T is a (T,T) matrix of ones.

Given the particular structure of this matrix, it can be shown[22] that the GLS estimator amounts to the OLS one applied to the following transformed model:

$$(4.2) \quad y_{nt} + (\sqrt{\theta^2} - 1)y_{n.} = \left[x_{nt} + (\sqrt{\theta^2} - 1)x_{n.} \right] b + \varepsilon_{nt} + (\sqrt{\theta^2} - 1)\varepsilon_{n.}$$

with, $\Theta^2 = \sigma_w^2 \ / \ (\sigma_w^2 + T \ \sigma_u^2)$.

Unfortunately, the GLS estimator is not a feasible estimator since the variances σ_u^2 and σ_w^2 are unknown. The common practice is to estimate them using the residual variances associated to the Within and Between regressions; the Within regression being defined as the OLS estimator applied to the model written in terms of differences from individual means:

$$(4.3) \quad y_{nt} - y_{n.} = (X_{nt} - X_{n.})b + \varepsilon_{nt} - \varepsilon_{n.}$$

whereas the Between regression is simply the OLS estimator applied to the model written in terms of individual means[23]:

$$(4.4) \quad y_{n.} = X_{n.} b + \varepsilon_{n.}$$

---

22 See Maddala [1971], or Matyas [1992].

23 The individual means of the variables are generated by applying the projector $I_N \otimes J_T/T$ to the initial vector of observations, whereas differences from individual means come from the application of $I_N \otimes (I_T - J_T/T)$ to these initial observations.

It can be shown that, as long as the degrees of freedom associated to each of these regressions are correctly calculated (i.e. $N(T-1) - K_w$ for the Within regression and $N-K_b$ for the Between one, where K_w and K_b are the number of explanatory variables in each of these regressions), then, the estimated residual variance of the Within regression is a consistent estimate of σ_w^2, whereas that of the Between one is a consistent estimation of $\sigma_w^2/T + \sigma_u^2$. [24]

Then, the Feasible-GLS estimator can be obtained by applying OLS to the model (4.2) where Θ is replaced by its consistent estimate obtained as explained above. It is consistent and asymptotically efficient either when only N tends to infinity or when N and T tend to infinity.

It has been shown that all the different estimators suitable for the estimation of an error components models can be obtained as OLS on a transformed model (see Maddala [1971], Matyas [1992]). It is then useful to consider these estimators as members of a class of estimators (that one can name as the λ-class), which can be defined as OLS estimators on the following model:

$$(4.5)\quad y_{nt} + (\sqrt{\lambda}-1)y_{n.} = [X_{nt} + (\sqrt{\lambda}-1)X_{n.}]\ b + \varepsilon_{nt} + (\sqrt{\lambda}-1)\varepsilon_{n.}$$

If $\lambda = 0$, the Within estimator is obtained; $\lambda = \Theta$ corresponds to the GLS estimator and $\lambda = \hat{\Theta}$ to the Feasible-GLS estimator ; if $\lambda = 1$, the OLS estimator is obtained and $\lambda = \infty$ is related to the Between estimator.

This parametrization of the different estimators has the advantage of making clear that all of them use, with various weights, the Within and the Between variabilities of the data. Unfortunately, the good behaviour of these methods (consistency for $N \to \infty$ with T fixed,[25] etc...) relies on the assumption that there is no correlation between the disturbances and the regressors. This assumption is frequently violated, either because the estimated equation belongs to a system (some of the explanatory variables are then endogenous), or because some of these explanatory variables are subject to measurement errors. Another possibility which is often considered in panel data models is that the specific effects are correlated with these variables. This last problem will be considered later on and we now concentrate on the simultaneous equations and measurement error models.

---

24 There exists many other ways to estimate these variances (e.g. see T. Amemiya [1971]). Nevertheless, Maddala and Mount [1973] have shown that the way one estimates them does not have a significant effect on the behaviour of the second step of the Feasible-GLS estimator.

25 One generally concentrate on the consistency of estimators for $N \to \infty$ and fixed T because most panel data sets involve a large number of individuals but a limited number of periods. Looking at the asymptotic behaviour of estimators for $N \to \infty$ and fixed T is often called semi-asymptotics.

4.2 Estimation of simultaneous equations and models with measurement errors

As it is well-known, except in particular cases, the OLS estimator is not consistent for a system of equations. It is then necessary to use instrumental variables or the maximum likelihood estimation method. Since there are no particular difficulties associated with the estimation of simultaneous equations when the individual effects are fixed,[26] let us concentrate on the case of error components.

As long as one considers an error components model, it is necessary to take into account the induced serial correlation. For example, the usual Two-Stage Least Squares estimator must be adapted to this case. This entails the definition of the instrumental variables as $\Omega_{mm}^{-1}X$ instead of X, where X is the set of exogenous variables in the system and Ω_{mm} is the variance-covariance matrix of the disturbances of the m^{th} equation, which has the usual block diagonal structure (see the definition of Ω above). Since the elements of this matrix (the variances) are unknown, it is necessary, as a first step, to estimate them. This can be carried out by first estimating the m^{th} equation using the Within Two-Stage Least Squares estimator, i.e. by using W X as instruments, where $W = I_N \otimes (I_T - J_T/T)$, then computing the estimated residuals of the m^{th} equation and estimating the variance components by analysis of variance.[27] Another way to deal with this estimation problem has been proposed by Baltagi [1981]. His suggestion is to estimate the model by Two-Stage Least Squares on the Between and Within transformed models. This allows one to get consistent estimators of the variance components which can be used in order to compute Feasible-GLS on the following system:

$$(4.6)\quad \begin{pmatrix} X' W y_m \\ X' B y_m \end{pmatrix} = \begin{pmatrix} X' W Z_m \\ X' B Z_m \end{pmatrix} \alpha_m + \begin{pmatrix} X' W u_m \\ X' B u_m \end{pmatrix}$$

It can be noticed that the two estimators have the same asymptotic properties.

It is also possible to define a Generalized Three-Stage Least Squares estimator, which amounts to apply the Generalized Method of Moments to the complete system. Instruments are defined as $(I \otimes X)D^{-1}$, where D is the block-diagonal matrix containing the variance-covariance matrices Ω_{mm}, $m=1,...,M$.[16] It is also possible to use $\Sigma^{-1}(I \otimes X)$ as instruments, where Σ is the variance-covariance matrix of the whole system. In both cases, one has to estimate all the variance components, which can be made along the same lines as above. Lastly, a simultaneous equations model can also be estimated by the maximum likelihood method, but its derivation is too complex to be useful and presented here.[16]

The difficulties related to the correlation that exists between disturbances and regressors can also be found when one or several explanatory variables are subject to measurement errors. If the model does not contain specific effects and the measurement error is a white noise, none of the previously mentioned λ-class estimators is consistent. The only

---

26 In this case, one only has to apply the Within transformation to the model before using the usual simultaneous estimation methods.

27 Details can be found in Krishnakumar (1988, 1992).

least-squares-type estimator which is consistent, when only N tends to infinity, is the Between periods estimator, i.e. the estimator which uses the Between-periods variation of the observations. Unfortunately, given the limited number of periods of most panel data sets, this estimator generally has a rather poor behaviour. Fortunately, in that situation, other consistent estimators can be obtained by applying instrumental variables to one cross-section, i.e. one period, with instruments defined as the explanatory variables for another period:

$$(4.7)\quad \hat{\beta}_{ts} = \frac{\sum_{i=1}^{N} (x_{it} - x_{.t})(y_{is} - y_{.s})}{\sum_{i=1}^{N} (x_{it} - x_{.t})(x_{is} - x_{.s})}, t,s = 1,\ldots,T; t \neq s$$

These estimators have been called by Biorn [1992], "base estimators". Moreover, consistent estimators can also be obtained by combining inconsistent ones (see Biorn (1992)). For example, it is possible to combine the Between-Individual and the Within-Individual-Period estimators to get a consistent estimator:

$$(4.8)\quad \hat{\beta} = \frac{(T-1)B_{xx}\hat{\beta}_B - R_{xx}\hat{B}_R}{(T-1)B_{xx} - R_{xx}}$$

where B_{xx} and R_{xx} are respectively the Between-Individual and Within-Individual-Period variations of the observations of the explanatory variable(s).

Another way to get consistent estimators along the same lines is to combine difference estimators (see Griliches and Hausman [1986], Biorn [1992]). These difference estimators are just the OLS estimator applied to the model written in first or higher differences. For example, in the simple case when the model contains only one exogenous variable measured with error, the estimator defined as

$$(4.9)\quad \hat{\beta} = \frac{\sum_{i=1}^{N} (y_{it} - y_{is})(x_{it} - x_{is}) - \sum_{i=1}^{N} (y_{iz} - y_{iq})(x_{iz} - x_{iq})}{\sum_{i=1}^{N} (x_{it} - x_{is})^2 - \sum_{i=1}^{N} (x_{iz} - x_{iq})^2}, \quad t \neq s, \quad z \neq q, \quad (t,s) \neq (z,q)$$

is consistent when $N \to \infty$ and T is fixed. Obviously, there exists many more estimators of this kind and all these estimators can be combined in such a way that one gets more efficient ones (see Hansen [1982] or Gourieroux, Monfort and Trognon [1985], Biorn [1992]).

Things are a little bit more complicated if the measurement errors have an error components structure. In particular, the instrumental variables estimators presented above are no longer consistent since measurement errors are in that case serially correlated. The same remark applies to the Between-Period estimator. Again, it is possible to get consistent estimators by an appropriate combination of inconsistent "base" or "difference" estimators (see Biorn [1992]).

4.3 Testing for the absence of specific effects

As seen in the previous discussion, the eventual presence of specific effects in a regression model must be tested for. In the case where this presence only affects the efficiency of the λ-class estimators, the tests are based on the existence of a specific component in the variance of the disturbances. When the specific effects affect the consistency of these estimators, then Hausman's type tests can be applied.

a) Testing for the structure of disturbances when consistency of the λ-class estimators is not affected

The first procedure to test for the absence of individual effects is nothing but an analysis of variance test. Testing the absence of these effects amounts to testing $\sigma_u^2 = 0$.

This test can easily be implemented using the estimated variances of the disturbances of the Between and Within regressions. Under normality of all components of the disturbances, the random variable

$$(N(T-1)-K_w)\frac{\hat{\sigma}_w^2}{\sigma_w^2} = \frac{\hat{\varepsilon}'_W \hat{\varepsilon}_W}{\sigma_w^2}$$

has a χ^2 distribution with $(N(T-1)-K_w)$ degrees of freedom.

Since, under the same assumption, the variable

$$(N-K_b)T\frac{\frac{\hat{\sigma}_w^2}{T} + \hat{\sigma}_u^2}{\sigma_w^2 + T\sigma_u^2} = T\frac{\hat{\varepsilon}'_B \hat{\varepsilon}_B}{\sigma_w^2 + T\sigma_u^2}$$

has a χ^2 distribution with $(N-Kb)$ degrees of freedom, under H_0: $\sigma_u^2 = 0$, the test statistics defined as:

$$\frac{\sigma_w^2}{\sigma_w^2 + T\sigma_u^2} T \frac{\hat{\varepsilon}'_B \hat{\varepsilon}_B}{\hat{\varepsilon}'_W \hat{\varepsilon}_W} \rightarrow F(N-K_b, N(T-1)-K_w)$$

Then,

$$(4.10)\quad T\frac{\hat{\varepsilon}'_B \hat{\varepsilon}_B}{\hat{\varepsilon}'_W \hat{\varepsilon}_W} \rightarrow F(N-K_b, N(T-1)-K_w)$$

When N is large enough, if T times the estimated variance of the disturbances associated with the Between regression is higher than the one associated with the Within regression, one rejects the absence of specific effects. In that case, it can be concluded that there exists some unobserved heterogeneity.

Another test has been proposed by Breusch and Pagan [1980]. This test is a Lagrange multiplier test of the H_0 hypothesis that $\sigma_u^2 = 0$. Under H_0, the statistic

$$(4.11)\quad g = \frac{NT}{2(T-1)}\left[\sum_{n=1}^{N}\left(\sum_{t=1}^{T}\hat{\varepsilon}_{nt}\right)^2 - 1\right]^2$$

is asymptotically (for $N \to \infty$) distributed as a χ^2 with one degree of freedom.

It is interesting to note that this statistic requires only the estimation of the model by OLS. Various extensions of this test were proposed in the literature. In particular, Baltagi-Chang-Li [1990] proposed a test of the absence of individual effects when time effects are present. It must be noticed that in this case, it is also required that T tends to infinity for the test to work. Another noticeable extension is the one-sided test proposed by Honda (1985), taking into account that, under H_1, the variance of the specific effects cannot be negative. This test amounts to compute the following test statistics

$$(4.12)\quad \sqrt{\frac{NT}{2(N-1)}}\,\frac{T\sum_{n=1}^{N}\hat{\varepsilon}_n^2}{\sum_{n=1}^{N}\sum_{t=1}^{T}\hat{\varepsilon}_{nt}^2}$$

which is asymptotically (for $N \to \infty$) distributed as $N(0,1)$.[28]

In the case of a system, the OLS, Between and Within estimators are no longer consistent and these tests cannot be applied directly. Nevertheless, by using the previously mentioned Within and Between Two Stage Least Squares estimators, one can test, at least asymptotically, the absence of specific individual effects along the same lines. Things are more complicated in the case of measurement errors, since the presence of specific effects can affect the consistency of some estimators.

b) Testing for the structure of disturbances when there are measurement errors

It has been mentioned above that when the explanatory variables are subject to measurement errors, usual estimators are inconsistent. Moreover, some of the estimators based upon "base" or "difference" estimators are only consistent when there are no specific effects. This allows the design of tests based on the Hausman's idea. Consider that one has an estimator $\hat{b}_1$ which is consistent under both assumptions of presence or absence of specific effects (e.g. the estimator given by equation (4.9)) and another one $(\hat{b}_2)$ which is only

---

28 For a presentation of other tests and of their relative performances, see Moulton and Randolph [1989], Baltagi, Chang and Li [1992].

consistent when there are no specific effects but which is more efficient than $\hat{b}_1$ (e.g. consider a combination of several estimators such as those given by equation (4.7)). Then, one can state that the test statistic

(4.13) $g_h = (\hat{b}_2 - \hat{b}_1)'(V(\hat{b}_2 - \hat{b}_1))^{-1}(\hat{b}_2 - \hat{b}_1)$

is asymptotically (again, for $N \to \infty$) distributed as a χ^2 with k degrees of freedom where k is the number of regressors.

It must be noticed that despite its apparent simplicity, this test is not easy to implement since one has to calculate the variance of differences, which can be tedious.

Although the usual way of presenting inference and estimation in panel data always begins with the above methods, it is almost undisputable that, in practice, the assumption of non-correlation between the specific effects and the disturbances is quite frequently violated. In that case, most of the presented methods are inconsistent. Then, other estimation techniques have to be considered and testing for the absence of such correlation is therefore necessary to determine which estimator should be used.

5 - Estimation and testing when specific effects are correlated with the regressors

5.1 Estimation of static models with correlated specific effects

The assumption of uncorrelated specific effects is often criticized because it is likely not to be satisfied in many situations. For example, in the estimation of an earnings function, it is difficult to consider that there is no correlation between the unobserved ability of the individuals and their human capital as measured by their qualifications.

The consequence of such correlation on the properties of the OLS, GLS, Feasible-GLS and Between estimators is that they become inconsistent (when only N tends to infinity; when T also tends to infinity, GLS and Feasible-GLS are consistent). Among the estimators which have been considered above, only the Within remains consistent. This is rather intuitive since this estimator is nothing but OLS on a transformed model from which individual effects have been washed away. Moreover, under the assumption that the correlation between these effects and the regressors does not depend on time, Mundlak [1978] has shown that this estimator is BLUE, but the necessary assumption for this can be considered rather restrictive (see Chamberlain [1982,1984]). Another drawback related to this method is that is does not allow the estimation of coefficients associated with variables constant over time since they disappear with the Within transformation. This is why Hausman and Taylor [1981], Amemiya and MaCurdy [1986] and Breusch, Mizon and Schmidt [1989] have proposed estimating these models with correlated effects by the method of instrumental variables. Consider that the model to be estimated can be written as:

(5.1) $y_{nt} = X_{nt}b + Z_n c + \varepsilon_{nt}$

It is possible to estimate the parameters b by using the Within estimator. It is easily shown, using the Frisch-Waugh theorem, that estimating c then amounts to estimate the model:

(5.2) $y_{n.} - X_{n.}\hat{b} = Z_n c + \bar{\varepsilon}_{n.}$

But it must be kept in mind that it has been assumed that there is a correlation between the individuals effect and the regressors, so the OLS on this transformed model does not lead to a consistent estimator. Hausman and Taylor have proposed the assumption that, among the X and Z variables, a number K_1 of X variables (X_1) exist which are uncorrelated with the disturbances, as well as a number P_1 of Z variables (Z_1) sharing the same property. Then assuming that K_1 is greater than the number of Z variables which are correlated with the effects (to ensure identifiability), it is possible to use instrumental variables with X_1 and Z_1 as instruments for estimating model (5.2). This in fact amounts to estimating the model (5.1) with the $G_{HT} = [\, W, X_1, Z_1 \,]$ set of instrumental variables. In order to achieve efficiency, the model (5.1) should be first transformed by $\Omega^{-1/2}$, so that its disturbances have a scalar covariance matrix. Breusch, Mizon and Schmidt [1989] showed that the G_{HT} set of instruments is equivalent to $[\, WX_1, WX_2, BX_1, Z_1 \,]$; that is, the instrumental variables are X_1 variables expressed both in deviations from individual means and as individual means, X_2 variables expressed in terms of deviations from individual means and Z_1 variables used as they are.

Amemiya and MaCurdy [1986] have proposed another set of instrumental variables defined as $G_{AM} = [\, WX_1, WX_2, X_1^*, Z_1 \,]$ where

$$X_1^* = \begin{pmatrix} X_{11} & X_{12} & \cdot & \cdot & \cdot & X_{1T} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ X_{N1} & X_{N2} & \cdot & \cdot & \cdot & X_{NT} \end{pmatrix} \otimes e_T$$

In contrast to the Hausman-Taylor estimator, the instruments for the endogenous variables are not the individual means of the X_1 variables but the observations vectors corresponding to each period. It must be noticed that for this estimator to be consistent, the X_1 variables must be uncorrelated with the specific effects at each period, whereas for Hausman-Taylor estimator, this absence of correlation must be satisfied only for the means over the time periods.

Breusch, Mizon and Schmidt [1989] have proposed yet another extension of the set of instruments. They suggest the use of $G_{BMS} = [\, WX_1, WX_2, BX_1, (WX_1)^*, (WX_2)^*, Z_1 \,]$, i.e. deviations from means of the X_1 and X_2 variables, as well as the vectors of these deviations for each time period (i.e. the X_1^* matrix above, where observations are in terms of deviations from means), the time means of the X_1 variables and the Z_1 variables.

5.2 Testing for correlation between the regressors and the specific effects in static models

The existence of a correlation between the regressors and the specific effects leads to the non-nullity of the expectation of these specific effects, conditional on the exogenous variables: $E(u_n/X_{nt}) \neq 0$. Mundlak [1978] proposed approximating this conditional expectation by:

(5.3) $E(u_n/X_{nt}) = \sum_{\tau} X_{n\tau}\Psi_{\tau}$

Then, it is possible to write

(5.4) $u_n = \sum_{\tau} X_{n\tau}\Psi_{\tau} + \Phi_n$

where Φ is uncorrelated with X and is distributed as a normal variable with mean 0 and variance σ^2.

Considering that the correlation between the specific effects and the X variables can be considered in terms of individual means, Mundlak proposed simplifying (5.3) and to write:

(5.5) $E(u_n/X_{n\cdot}) = X_{n\cdot}.\Psi + \Phi_{n\cdot}$.

This assumes that all the Ψ coefficients are identical, whatever τ.

Then, the model can be rewritten as:

(5.6) $y_{nt} = X_{nt}b + X_{n\cdot}.\Psi + \Phi_{n\cdot} + w_{nt}$

where the specific effects Φ are now uncorrelated with the variables X. This model is a "standard" error components model, except that it contains more explanatory variables than the initial one. Their coefficients are directly linked to the existence of a correlation beween the specific effects and the original explanatory variables. Then, testing for such a correlation amounts only to testing the nullity of the Ψ parameters in model (5.6). This is just a classical F test.[29]

Hausman [1978] proposed another test for the absence of correlation between the regressors and the effects. His test relies on the fact that, if the specific effects are correlated with the explanatory variables, the Within estimator is consistent while the Feasible-GLS is not, but under the null hypothesis of no correlation, both estimators are consistent, and the Feasible-GLS is asymptotically efficient. Then, comparing estimates obtained by these two

---

29 It must be noticed that the model under consideration remains an error components model. The estimations must be carried out by Feasible-GLS in order to ensure that the test statistic has a Fisher distribution.

methods allows an inference about the correlation of the effects: if their difference is "small", one can expect that there is no correlation, and vice-versa, if it is "large". More precisely, Hausman [1978] shows that the test statistic

$$(5.7)\quad g_h = (\hat{b}_w - \hat{b}_{mcqg})'(V\hat{b}_w - V\hat{b}_{mcqg})^{-1}(\hat{b}_w - \hat{b}_{mcqg})$$

is asymptotically distributed as a χ^2 with K degrees of freedom, for $N \to \infty$. Hausman and Taylor [1981] and Baltagi [1989] showed that this test can also be based on the difference between the GLS and Within estimators, between the GLS and Between estimators, between the Between and Within estimators or between the GLS and OLS estimators, with the corresponding variances. This test is also identical to the one proposed by Mundlak [1978].

5.3 Dynamic models

The problem with the estimation of dynamic (or autoregressive) error components models such as:[30]

$$(5.8)\quad y_{nt} = \alpha y_{n,t-1} + \sum_k \beta_k x_{knt} + u_n + w_{nt} \quad n = 1,\ldots,N;\ t = 1,\ldots,T.$$

is the correlation between the lagged endogenous variable and the individual effects. This problem is well-known in classical econometrics as the non-consistency of the least-squares method for dynamic models with autocorrelated errors.

It can be shown that none of the usual estimators of error components models is consistent when only N tends to infinity[31] (see Sevestre and Trognon [1983, 1985, 1992], Nickell [1982]).

It is therefore necessary to find other estimation methods. Obviously, instrumental variables methods are good candidates. Several such estimators have been proposed, based on different instruments sets and/or different ways of (re)writing the model.

In their seminal paper, Balestra and Nerlove [1966] proposed estimating model (5.8) by using lagged values of the X variables as instruments. This estimator is consistent as long as the X variables are exogenous, i.e. do not exhibit any correlation with the individual effects and the non specific disturbances. As stressed above, the assumption of non-correlation between the effects and the regressors is frequently questioned. If not satisfied, the Balestra-Nerlove estimator is inconsistent.[32]

---

30 An extension of this model has been proposed by Holtz-Eakin, Newey and Rosen [1988]. They suggest specifying the disturbances of the model as:
$u_n\lambda_t + w_{nt}$ where the λ_t's are time specific effects.

31 The only exception is when it is assumed that the initial observations y_{n0} are uncorrelated with the individual effects. In that case, GLS lead to consistent estimations of the parameters.

32 Except if it can be assumed that at least one of the X variables does not suffer from such correlation.

In order to avoid this problem, Anderson and Hsiao [1982] proposed rewriting the model in first differences:

$$(5.9)\quad y_{nt} - y_{n,t-1} = a(y_{n,t-1} - y_{n,t-2}) + \sum_k b_k(x_{knt} - x_{kn,t-1}) + w_{nt} - w_{n,t-1}.$$

Even if the primary cause of the inconsistency of the least squares method disappears with this transformation, it induces an MA type autocorrelation in the disturbances of the transformed model. The least squares estimator then does not lead to consistent estimates. This is why Anderson and Hsiao proposed estimating this model using instrumental variables methods. They proposed using as instruments the first differences of the X's of the model as well as either $y_{n,t-2}$ or ($y_{n,t-2}$ - $y_{n,t-3}$). Given the assumption of non-autocorrelation of the w_{nt})'s, these variables are valid instruments.[33] Various simulation studies (Arellano and Bond [1990], Sevestre and Trognon [1990]) showed that these estimators are not very efficient. This is why Arellano and Bond [1990] proposed estimating model (5.9) by a GMM type estimator. Their idea is to use all the possible orthogonality conditions that exist between the disturbances of model (5.9) and the possible instruments. They proposed using all the first differences ($y_{n,t-i}$ - $y_{n,t-i-1}$) which are not correlated with (w_{nt} - $w_{n,t-1}$), i.e. with T=4 periods of observations, one can use y_{n0} for the first period of estimation (t=2), y_{n0} and y_{n1} for t=3, y_{n0}, y_{n1} and y_{n2} for t=4. Moreover, in order to improve the efficiency of the estimators, Arellano and Bond proposed taking into account the autocorrelated structure of the disturbances in the first-order differenced model. These are MA(1) if the w_{nt}'s are i.i.d. or may follow higher order MA processes if the w_{nt}'s are autocorrelated.

Another way to deal with the problem of correlation between the individual effects and the explanatory variables is to write the model in differences from the individual means:

$$(5.10)\quad y_{nt} - y_{n.} = \alpha(y_{n,t-1} - y_{n.,-1}) + \sum_k \beta_k(x_{knt} - x_{kn.}) + w_{nt} - w_{n.}, \quad n = 1,\ldots,N;\ t = 1,\ldots,T.$$

Again, the OLS on this model is not consistent because $w_{nt} - w_{n.}$ are correlated with the lagged endogenous variable. But one can use lagged values of the differences $X_{nt} - X_{n.}$ as instruments. Nevertheless, this estimator requires strict exogeneity of the variables X.

One major problem common to all these instrumental variables estimators is that they have poor efficiency. This is an important problem for the applied econometrician since a lack of efficiency can lead to unreliable estimates. The most obvious way to improve efficiency is to take into account the structure of the serial correlation. This was proposed by Arellano and Bond (1990). A generalised version of the Balestra-Nerlove estimator can also be suggested. It amounts to apply instrumental variables on the following transformed model:

$$(5.11)\quad y_{nt} + (\sqrt{\theta}-1)y_{n.} = (y_{n,t-1} + (\sqrt{\theta}-1)y_{n,.-1})\alpha + (X_{nt} + (\sqrt{\theta}-1)X_{n.})\beta + \varepsilon_{nt} + (\sqrt{\theta}-1)\varepsilon_{n.}$$

---

33 If the w_{nt}'s are autocorrelated according to an MA(q) process, it is then possible to use $y_{n,t-q-2}$ or $y_{n,t-q-2} - y_{n,t-q-3}$ as instruments.

Another way to estimate the model with a two-step method is based on the fact that the Within and OLS estimators respectively underestimate and overestimate the true value of the coefficient of the lagged endogenous variable. Since these two estimators belong to the previously analysed λ-class of estimators, it can be shown that there exist a value for λ which ensures the consistency of this estimator. This value is given by (See Sevestre and Trognon [1983]):

$$\lambda^* = (K(1-\rho))/\left(\frac{1-\alpha^T}{1-\alpha}\frac{Ey_{n0}u_n}{\sigma^2} + K(1-\rho+T\rho)\right)$$

where:

$$K = \frac{T-1-T\alpha+\alpha^T}{T(1-\alpha)^2}$$

and $\rho = \sigma_u^2/(\sigma_u^2+\sigma_w^2)$

This estimator is the OLS applied to the model:[34]

$$(5.12)\quad y_{nt} + (\sqrt{\lambda^*}-1)y_{n.} = \left(X_{nt} + (\sqrt{\lambda^*}-1)X_{n.}\right)\beta + \varepsilon_{nt} + (\sqrt{\lambda^*}-1)\varepsilon_{n.}$$

It must be noticed that, as for the previous two-step estimator, one has to estimate the parameters related to θ and λ. This can be done using the residuals obtained using one of the instrumental variables estimators presented above.

But the best way to get a consistent and efficient estimator is to use the Maximum Likelihood method. An easy way to get the maximum likelihood estimators of all the parameters of the model was suggested by Blundell and Smith [1990]. In order to write the likelihood function, one can first recall that the problem associated with the estimation of an autoregressive error components model is mainly that of the correlation of the individual effects with the initial observations. Under the assumption of normality of these specific effects, they propose decomposing them as:

$$(5.13)\quad u_n = \psi u_{n0} + \nu_n$$

where u_{n0} is the disturbance entering the definition of the initial observations:

$$(5.14)\quad y_{n0} = \phi z_n + \beta x_{n0} + u_{n0}$$

and where ν_n is now uncorrelated with the initial observations. Then, the log-likelihood can be written as:

$$LogL = -\frac{NT}{2}\log 2\pi - \frac{N}{2}\log det\Omega - \frac{N}{2}\log\sigma_u^2 - \frac{1}{2}\sum_i \varepsilon_i'\Omega^{-1}\varepsilon_i - \frac{1}{2\sigma_u^2}\sum_i u_{i0}^2$$

---

34 It can be noticed that as long as the initial observations are uncorrelated with the specific effects, one gets $Ey_{n0}u_n = 0$, and the λ^* estimator is exactly identical to the GLS estimator.

with $\varepsilon_n' = (y_{n1} - \alpha y_{n0} - \beta x_{n1} - \gamma z_n - \psi u_{n0}, \ldots, y_{nT} - \alpha y_{nT-1} - \beta x_{nT} - \gamma z_n - \psi u_{n0})$

and $u_{n0} = y_{n0} - \phi z_n - \beta x_{n0}$

and Ω is the usual variance-covariance matrix of error components models.

Estimators of the parameters cannot be got directly by maximizing this log-likelihood with respect to the unknown parameters. One has to use an iterative procedure. Fortunately, there exists an alternative way to get estimators which are asymptotically equivalent to the maximum-likelihood ones (see Sevestre and Trognon [1990, 1992]). These estimators are obtained from the following multi-step procedure:

1) Estimate equation (5.14) by OLS and get the estimated residuals $\hat{u}_{n0}$.

2) Estimate the following model by instrumental variables:

$$(5.15) \quad y_{nt} = \alpha y_{n,t-1} + \sum_k \beta_k x_{knt} + \psi_{\hat{u}_{n0}} + \nu_n + w_{nt} \quad n = 1,\ldots,N;\ t = 1,\ldots,T.$$

and get estimates of the variances of ν_n and w_{nt}.

3) Estimate the previous model by Feasible-GLS and iterate. The resulting estimator is asymptotically equivalent to the maximum likelihood estimator. It is then consistent and asymptotically efficient. Moreover, according to a Monte Carlo simulation, it overrides all other estimators in terms of small sample bias and efficiency (as long as the model is correctly specified).

6 - Estimation problems associated with unbalanced panel data and "false" panel data

It has been assumed so far that the sample available for the estimation of the model is balanced, i.e. that all individuals are observed over all the time periods without any missing observation, i.e. all T observations for each of the N individuals in the sample are recorded. Unfortunately, this is very rare in practice and frequently we have to deal with samples where all individuals are not observed for the whole period covered. For a long time, practitioners used to restrict themselves to "balanced" sub-panels, i.e. to those individuals in the panel which were observed over the complete period. This has two main disadvantages. First, it led to a rather large loss of observations and resulted in much loss of efficiency (see for example Matyas and Lovrics [1991], or Baltagi and Li [1990] who show that the GLS on a balanced sample can be much less efficient than the OLS on an unbalanced sample). Second, it can result in a selection bias problem if the observations are not missing at random , i.e. the endogenous variable of the model is correlated with the process explaining the pattern of absence of the observations.

Fortunately, in the recent years, extensions of the usual estimators have been developed to deal with the case of incomplete panels (see Nijman and Verbeek [1992]). Let us first consider the case where missing observations are missing at random.

6.1 Estimation with unbalanced panel data without selection bias[35]

Let us assume that the model to estimate is a fixed effects model:

$$(6.1)\quad y_{nt} = X_{nt}b + f_n + w_{nt} \quad n = 1,\ldots,N;\quad t = 1,\ldots,T.$$

As shown earlier, the OLS on this model leads to the well-known Within or LSDV estimator. While this estimator is BLU in the case of a balanced panel data set, it is not if the sample is unbalanced. In this case, the Within transformation leads to heteroscedastic disturbances:

$$Var(w_{nt} - w_{n.}) = \sigma_w^2 + \frac{1}{T_n}\sigma_w^2 - \frac{2}{T_n}\sigma_w^2$$
$$= \sigma_w^2\left(\frac{T_n - 1}{T_n}\right)$$

It is then necessary to transform the model to get the OLS BLUE. This transformation is easy since it only amounts to multiplying all observations by $\sqrt{T_n/(T_n - 1)}$. Now, applying the OLS to the transformed model

$$(6.2)\quad \tilde{y}_{nt} - \tilde{y}_{n.} = (\tilde{X}_{nt} - \tilde{X}_{n.})b + \tilde{w}_{nt} - \tilde{w}_{n.}$$

where the tilda sign means that the variables have been transformed as indicated above, allows one to get the BLU estimator of b.

Things are more complicated if we want to estimate an error components model. In this case, BLU estimators can be obtained by applying the OLS to the following transformed model (Baltagi [1985]):

$$(6.3)\quad y_{nt} + (\sqrt{\theta_n} - 1)y_{n.} = (X_{nt} + (\sqrt{\theta_n} - 1)X_{n.})b + \varepsilon_{nt} + (\sqrt{\theta_n} - 1)\varepsilon_{n.}$$

with

$$\theta_n = \frac{\sigma_w^2}{(\sigma_w^2 + T_n\sigma_u^2)}$$

The problem is getting consistent estimates of the variances σ_u^2 and σ_w^2. When the sample is balanced, these estimates can be obtained by applying the OLS to the Between and Within transformed models, i.e. on the models written in terms of individual means and in differences from these individual means. Things get more complicated when the sample

---

35 We restrict ourselves to the most usual estimation methods of the error components model. Almost all estimation methods can be extended to the case of incomplete panels rather easily.

is unbalanced. It is still possible to get consistent estimates of σ_w^2 by estimating model (6.2) (the Within model) by OLS. But through the estimation of the Between model, it is not possible to get a consistent estimate of $\sigma_w^2 + T_n\sigma_u^2$. Indeed, in this case, we estimate

$$(6.4) \quad y_{n.} = x_{n.}b + \varepsilon_{n.}$$

with,

$$E\varepsilon_{n.} = 0, \quad V\varepsilon_{n.} = \sigma_u^2 + \frac{1}{T_n}\sigma_w^2$$

Then, the problem of heteroscedasticity emerges as is the case with the Within model. Unfortunately, here, the transformation relies on the knowledge of the unknown parameters.

Several solutions have been proposed in the literature. One is to use the analysis of variance with unequal cells method (Baltagi [1985]). Nijman and Verbeek [1992] consider the possibility of using the Maximum Likelihood method but come to the conclusion that this leads to rather complex calculus, so they propose adapting the usual formulae to estimate σ_u^2:

$$\hat{\sigma}_u^2 = \frac{1}{N}\sum_n \left((\bar{y}_n - \bar{X}_n'\hat{b}_B)^2 - \frac{1}{T_n}\hat{\sigma}_w^2 \right)$$

where $\hat{b}_B$ is the Between estimator. It is then easy to compute $\theta_n = \hat{\sigma}_w^2/(\hat{\sigma}_w^2 + T_n\hat{\sigma}_u^2)$ and the transformed data as in (6.3). The OLS estimator on the transformed model leads to consistent and asymptotically efficient estimates.

Unfortunately, the extensions of these methods to models where there are both individual and time specific effects appear much more complicated to implement (see Wansbeek and Kapteyn [1989).

6.2 Testing and adjusting for selection bias

As outlined above, making the sample balanced, i.e. deleting those individuals which are not observed over the complete period, leads, at the least, to efficiency loss. But it can also have worse consequences: it can imply selection bias(es) if the process explaining how an individual is absent from the sample is not independent of the endogenous variable. It is then important to test for this possible selection bias. Nijman and Verbeek [1992] show that adapting the usual Lagrange multiplier test to panel data leads to a complex calculus since numerical integration over at least two dimensions is required. They propose using a kind of Hausman test based on the differences of estimates obtained using the balanced and unbalanced panels. For example, if we consider the estimation of an error components model, it is possible to compute the following statistic

$$(6.5) \quad Q = (\hat{b}_{GLS,B} - \hat{b}_{GLS,U})'(V_{GLS,B} - V_{GLS,U})^{-}(\hat{b}_{GLS,B} - \hat{b}_{GLS,U})$$

where the symbol "-" means the generalized inverse of the matrix. Under the null hypothesis, this statistic is asymptotically distributed as a χ^2 with K degrees of freedom where K is the

number of coefficients in b. This test helps to get an idea whether making the sample balanced leads to selection bias.

But unfortunately, it may well happen that even an unbalanced sample leads to estimators subject to selection bias. In this case, the detection of the bias cannot rely on the comparison of estimates obtained from the whole sample with those obtained from the balanced sub-sample. If one is able to write a model describing the process of selection, it is possible to use the generalized Heckman [1979] procedure. In effect, it adds to the model a supplementary term which adjusts estimates for selection bias. Then we can test for this bias by testing the significance of the coefficient associated with the correction term. When one uses panel data and wants to take individual effects into account, this is rather burdensome. Nijman and Verbeek [1992] propose instead some variable addition tests which are more easy to implement. These consist of testing the significance of some variables included as supplementary regressors in the model which was estimated using the unbalanced panel. These variables are T_n, the number of observations of each individual n, c_n, a dummy variable which equals 1 if individual n is observed over the total period, but 0 otherwise, and a variable $r_{n,t-1}$, indicating whether the n^{th} individual is observed in the preceding period.

The estimated model including these supplementary variables can be considered as an approximation of the model including the true correction term(s). Estimating a model when the sample is subject to selection bias by Heckman's two-step method or by Maximum Likelihood is indeed very complicated and difficult to implement in practice (more details can be found in Nijman and Verbeek [1992]).

6.3 Estimation with "false" panel data (repeated cross-sections)

Another difficulty that practitioners often meet is that panel data sets are not always available. In many countries, there are repeated surveys about, for instance, consumers, but in these surveys, different individuals are involved in each time period. In fact, these are repeated cross-sections rather than true panel data. The question is then to see whether they can be used (with all their advantages) as true panel data sets.

Deaton [1985] was the first to address this question. He considered the case of a fixed effects model[36] such as:

$$(6.6)\quad y_{nt} = a_n + \sum_{k=1}^{K} b_k x_{knt} + w_{nt}, \qquad t = 1,\ldots,T.$$

Obviously, since each individual is observed only once, it is not possible to estimate this model. Deaton [1985] then proposed constructing cohorts, i.e. groups of individuals

---

36 If the specific effects were assumed to be random and uncorrelated with the regressors, OLS applied to the pooled sample would result in a consistent, though not efficient, estimator.

sharing some characteristics such as age, residential area, profession,etc... Denoting the empirical means of the variables within a cohort by $\bar{y}_{ct}, \bar{X}_{ct}$ where c is the index for cohorts, the problem is to get consistent estimators for the model

$$(6.7) \quad \bar{y}_{ct} = \bar{X}_{ct}'b + \bar{a}_{ct} + \bar{w}_{ct}, c = 1,\ldots,C; \quad t = 1,\ldots,T$$

where $\bar{a}_{ct}$ depends on time since the mean is computed over individuals which are different at each time period. The problem then is that we have CT observations and $CT+K$ parameters to estimate (the $\bar{a}_{ct}$'s and b). This is obviously impossible and one has to consider the assumption $\bar{a}_{ct} = \bar{a}_c$ as reasonable, as soon as the number of individuals in each cohort is large enough. Then, a natural way to estimate the b coefficients is to use the Within estimator.

Unfortunately, it has been shown (Verbeek and Nijman [1992]) that even for rather large cohort sizes, this estimator suffers from small sample biases which can be quite large. So, another approach is required, where the size of the cohorts is not required to be very large. Deaton [1985] proposed considering model (6.7) as a model with measurement errors. The cohorts means $\bar{y}_{ct}, \bar{X}_{ct}$ are then assumed to be error-ridden measures of the true population means y^*_{ct}, X^*_{ct}. Assuming that

$$\begin{pmatrix} \bar{y}_{ct} \\ \bar{X}_{ct} \end{pmatrix} - \begin{pmatrix} y^*_{ct} \\ X^*_{ct} \end{pmatrix} \; i.i.d. \; \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{yy} & \sigma_{yx} \\ \sigma_{yx}' & \Sigma_{xx} \end{pmatrix} \right)$$

the following estimator

$$(6.8) \quad \hat{b}_D = \left(\frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{X}_{ct} - \bar{X}_c)(\bar{X}_{ct} - \bar{X}_c)' - \frac{T-1}{T} \hat{\Sigma} \right)^{-1} \left(\frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{X}_{ct} - \bar{X}_c)(\bar{y}_{ct} - \bar{y}_c) - \frac{T-1}{T} \hat{\sigma}_{xy} \right)$$

is consistent, where Σ can be estimated by[37]

$$\hat{\Sigma} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{C} \sum_{c=1}^{C} \frac{1}{N_c^2} \sum_{n \in cohort c} (X_{nt} - \bar{X}_{ct})(X_{nt} - \bar{X}_{ct})'$$

and σ_{xy} can be estimated in the same way.

If the number of individuals in each cohort (N_c) tends to infinity, this estimator is asymptotically equivalent to the previous Within estimator, since the measurement errors tend to zero. If this number is finite, then, this estimator is consistent for C, T (or both) going to infinity, as long as the matrix to be inverted in (6.8) is non-singular.

---

37 This estimator is valid as long as Σ does not depend on c and t.

7 - Conclusion.

In this paper, our aim was firstly to emphasize the benefits, for applied economists, in using panel data for studying firms' and consumers' behaviours and then to provide an overview of linear models estimation methods allowing one to take advantage of the particular structure of these data.

Panel data can and probably will change dramatically the way we look at and think about applied and theoretical economics. Rather sophisticated microeconomic models can be estimated with such data and, in the long run, its use may lead towards a better understanding of the link between micro behaviour and macro relations.

The data is available, the methods are more and more relevant, so one can think that the use of panel data in applied econometrics is likely to spread more and more.

References.

AASNESS, J., BIORN E. and T. SKJERPEN (1988) "Engel Functions, Panel Data and Latent Variables", Mimeo, University of Oslo.

ABEL, A.B. (1980) "Empirical Investment Equations: An Integrative Framework", in On The State of Macroeconomics, K. Brunner and A. Metzler eds., Carnegie-Rochester Conference Series, 12.

ABEL, A.B. and O. BLANCHARD (1986) "The Present Value of Profits and Cyclical Movements in Investment", Econometrica, 54, pp. 249-273.

ALTONJI, J.G. (1986) "Intertemporal Substitution in Labour Supply: Evidence from Micro-Data", Journal of Political Economy, 94, pp. 176-215.

ALTONJI, J.G. and A. SIOW (1987) "Testing the response of consumption to income change with (noisy) panel data", Quarterly Journal of Economics, 102, pp. 293-328.

AMEMIYA, T. (1971) "The Estimation of Variances in a Variance-Components Model", International Economic Review, 12, pp. 1-13.

AMEMIYA, T. and T.E. MACURDY (1986) " Instrumental Variable Estimation of an Error Components Model", Econometrica, 54, pp. 869-880.

ANDERSON, T.W. and C. HSIAO (1982) "Formulation and Estimation of Dynamic Models Using Panel Data", Journal of Econometrics, 18, pp.578-606.

ARELLANO, M. and S. BOND (1991) "Some Tests of Specification for Panel Data: Monte-Carlo Evidence and an Application to Employment Equations", Review of Economic Studies, 58, pp. 277-297.

BALESTRA, P. and M. NERLOVE (1966) "Pooling Cross-Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas", Econometrica, 34, pp. 585-612.

BALESTRA, P. (1992) "Fixed Effects Modeld and Fixed Coefficients Models", in The Econometrics of Panel Data. Handbook of Theory and Applications, L. Matyas and P. Sevestre eds., Kluwer Academic Publishers.

BALESTRA, P. and S. NEGASSI (1992) "A Random Coefficient Simultaneous Equation System with an Application to Direct Foreign Investment by French Firms", Empirical Economics, 17, pp. 205-220.

BALTAGI, B. (1981) "Pooling: An Experimental Study of Alternative Testing and Estimation Procedures in a Two-Way Error Components Model", Journal of Econometrics, 17, pp. 21-49.

BALTAGI, B. (1985) "Pooling Cross-Sections with Unequal Time Series Lengths", Economics Letters, 18, pp. 133-136.

BALTAGI, B. (1989) "The Equivalence of the Boothe-MacKinnon and the Hausman Specification Tests in the Context of Panel Data", Econometric Theory, 5, p.454.

BALTAGI, B. and Q. LI (1990) "A Comparison of Variance Components Estimators Using Balanced Versus Unbalanced Data", Econometric Theory, 6, pp. 283-285.

BALTAGI, B., CHANG Y. and Q. LI (1992) "Monte Carlo Results on Several New and Existing Tests for the Error Components Model", Journal of Econometrics, 54, pp. 95-120.

BALTAGI, B. and B. RAJ (1992) "A Survey of Recent Theoretical Developments in the Econometrics of Panel Data", Empirical Economics, 17, pp. 85-109.

BIORN, E. (1992) "Panel Data with Measurement Errors", in The Econometrics of Panel Data. Handbook of Theory and Applications, L. Matyas and P. Sevestre eds., Kluwer Academic Publishers.

BLANCHARD, O. and C. WYPLOSZ (1983) "An Empirical Structural Model of Aggregate Demand", Journal of Monetary Economics, N°7, pp. 1-28.

BLUNDELL, R. (1988) "Consumer Behaviour: Theory and Empirical Evidence - A survey", The Economic Journal, 98, pp. 16-65.

BLUNDELL, R. and C. MEGHIR (1987) "Bivariate Alternatives to the Tobit Model", Journal of Econometrics, 34, pp. 179-200.

BLUNDELL, R., BROWNING M. and C. MEGHIR (1989) "A Microeconometric Model of Intertemporal Substitution and Consumer Demand", University College London, Discussion Paper N°89-11.

BLUNDELL, R., PASHARDES P. and G. WEBER (1989) "What Do We Learn About Consumer Demand Patterns From Micro-Data?", University College London, Discussion Paper N°89-18.

BLUNDELL, R. and R. SMITH (1991) "Conditions Initiales et Estimation Efficace dans les Modèles Dynamiques sur Données de Panel", Annales d'Economie et de Statistique, N°20-21, pp. 109-124.

BLUNDELL, R., BOND S., M. DEVEREUX and F. SCHIANTARELLI (1992) "Investment and Tobin's Q: Evidence from Company Panel Data", Journal of Econometrics, 51, pp.233-257.

BOND, S. and C. MEGHIR (1990) "Dynamic Investment Models and the Firm's Financial Policy", Institute for Fiscal Studies Discussion Paper N° W90/17.

BOVER, O. (1991) "Relaxing Intertemporal Separability: A Rational Habits Model of Labour Supply Estimated From Panel Data", Journal of Labour Economics, 9, pp. 85-100.

BRESSON, G., KRAMARZ F. and P. SEVESTRE (1992a) "Heterogeneous Labour and the Dynamics of Aggregate Labour Demand: Some Estimations Using Panel Data", Empirical Economics, 17, N°1, pp. 153-168.

BRESSON, G., KRAMARZ F. and P. SEVESTRE (1992b) "Labour Demand with Heterogeneous Workers and Non Linear Asymmetric Adjustment Costs, INSEE Discussion Paper.

BREUSCH, T. S. and A.R. PAGAN (1979) "A Simple Test for Heteroscedasticity and Random Coefficient Variation", Econometrica, 47, pp. 1287-1294.

BREUSCH, T.S., MIZON G.E. and P. SCHMIDT (1989) "Efficient Estimation Using Panel Data", Econometrica, 57, pp. 695-700.

BROWNING, M., DEATON A. and M. IRISH (1985) "A Profitable Approach to Labour Supply and Commodity Demands over the Life-Cycle", Econometrica, 53, pp. 503-543.

CHAMBERLAIN, G. (1982) "Multivariate Regression Models for Panel Data", Journal of Econometrics, 18, pp. 5-46.

CHAMBERLAIN, G. (1984) "Panel Data", in Handbook of Econometrics, Z. Griliches and M.D. Intriligator eds., North-Holland.

CHAN-LEE, J. (1986) "Pure Profit Rates and Tobin's q in Nine OECD Countries", OECD Economic Studies, N°7.

CHAN-LEE, J. and R. TORRES (1987) "q de Tobin et Taux d'Accumulation en France", Annales d'Economie et de Statistique, N°5, pp. 37-48.

CHAPPELL, H. and D. CHENG (1982) "Expectations, Tobin's q and Investment: A Note", Journal of Finance, 37, pp. 231-236.

CHOWDHURY, G. and D. MILES (1989) "Modelling Companies Debt and Dividend Decisions With Company Accounts Data", Applied Economics, 21, pp. 1483-1508.

CORNWELL, C., SCHMIDT P. and R. SICKLES (1990) "Production Frontiers with Cross-Sectional and Time-Series Variation in Efficiency Levels", Journal of Econometrics, 46, pp. 185-200.

CREPON, B. and J. MAIRESSE (1993) "Productivité, Recherche-Développement, et Qualifications", Mimeo, INSEE, Paris.

CUNEO, P. and J. MAIRESSE (1984) "Productivity and R&D at the Firm Level in French Manufacturing" in R&D, Patents and Productivity, Z. Griliches ed., University of Chicago Press, pp. 375-392.

DEATON, A. (1985) "Panel Data from Time Series of Cross-Sections", Journal of Econometrics, 30, pp. 109-126.

DEATON, A., and M. IRISH (1984) "Statistical Models for Zero Expenditures in Household Budgets", Journal of Public Economics, 23, pp. 59-80.

DEVEREUX, M. and F. SCHIANTARELLI (1990) "Investment, Financial Factors and Cash-Flow: Evidence From UK Panel Data", in Asymmetric Information, Corporate Finance and Investment", R. Hubbard ed., University of Chicago Press.

DORMONT, B. (1989) "Petite Apologie des Données de Panel", Economie et Prévision, N°87, pp. 19-32.

DURBIN, J. (1960) "Estimation of Parameters in Time-Series Regression Models", Journal of the Royal Statistical Society, Series B, 22, pp. 139-153.

EDWARDS, J.B. and G.H. ORCUTT (1969) "Should Aggregation Prior to Estimation Be the Rule?", Review of Economics and Statistics, 51, pp. 409-420.

FAZZARI, S., HUBBARD R. and B. PETERSEN (1988) "Financing Constraints and Corporate Investment", Brookings Papers on Economic Activity, N°1, pp. 141-195.

FECHER, F. and S. PERELMAN [1989] "Productivité, Progrès Technique et Efficacité: Une Etude Comparative de 14 Secteurs Industriels Belges", Annales d'Economie et de Statistique, N°13, pp. 93-118.

GORMAN, W. (1953) "Community Preference Fields", Econometrica, 21, pp. 63-80.

GOURIEROUX, C., MONFORT A. and A. TROGNON (1985) "Moindres Carrés Asymptotiques", Annales de l'INSEE, N°58, pp. 91-122.

GRILICHES, Z. (1979) "Issues in Assessing the Contribution of R&D to Productivity Growth", Bell Journal of Economics, 10, N°1, pp. 92-116.

GRILICHES, Z. (1980) "Returns to Research and Development Expenditures in the Private Sector", in New Developments in Productivity Measurement and Analysis, J. Kendrick et B. Vaccara eds., University of Chicago Press, pp. 419-461.

GRILICHES, Z. and J. MAIRESSE (1983) "Comparing Productivity Growth: An Exploration of French and US Industrial and Firm Data", European Economic Review, 21, pp. 89-119.

GRILICHES, Z. and J. MAIRESSE (1984) "Productivity and R&D at the Firm Level", in R&D, Patents and Productivity, Z. Griliches ed., University of Chicago Press, pp. 339-374.

GRILICHES, Z. (1986) "Productivity, R&D and Basic Research at the Firm Level in the 1970's", American Economic Review, 76, pp. 141-154.

GRILICHES, Z. and J. HAUSMAN (1986) "Errors in Variables in Panel Data", Journal of Econometrics, 31, pp. 93-118.

GRILICHES, Z. and J. MAIRESSE (1990) "R&D and Productivity Growth: Comparing Japanese and US Manufacturing Firms", in Productivity Growth in Japan and the United States, C. Hulten ed., University of Chicago Press, pp. 317-348.

HALL, B. and F. HAYASHI (1988) "Research and Development as an Investment", NBER Working Paper N°2973.

HALL, B. (1990) "Research and Development Investment at the Firm Level: Does the Source of Financing Matter?", Mimeo, University of California at Berkeley.

HALL, B. and J. MAIRESSE (1992) "Exploring the Relationship between R&D and Productivity Growth in French Manufacturing Firms", NBER Working Paper N° 3485.

HALL, R. and F. MISHKIN (1982) "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households", Econometrica, 50, pp. 461-481.

HAM, J.C. (1986) "Testing Whether Unemployment Represents Intertemporal Labour Suply Behaviour", Review of Economic Studies, 53, pp. 559-578.

HAMERMESH, D. (1989) "Labour Demand and the Structure of Adjustment Costs", American Economic Review, 79, pp. 674-689.

HAMERMESH, D. (1990) "Aggregate Employment Dynamics and Lumpy Adjustment Costs", NBER Working Paper n°3229.

HAMERMESH, D. (1992) "Spatial and Temporal Aggregation and the Dynamics of Labour Demand", paper presented at the Symposium on Labour Demand and Equilibrium Wage Formation, Amsterdam, January.

HANSEN, L.P. (1982) "Large Sample Properties of Generalized Method of Moments Estimators", Econometrica, 50, pp. 1029-1054.

HAUSMAN, J. (1978) "Specification Tests in Econometrics", Econometrica, 46, pp. 1251-1271.

HAUSMAN, J. and W.E. TAYLOR (1981) "Panel Data and Unobservable Individual Effects, Econometrica, 49, pp. 1377-1398.

HAYASHI, F. (1982) "Tobin's Marginal q and Average q: A Neo-Classical Interpretation", Econometrica, 50, pp. 213-224.

HAYASHI, F. (1985) "The Permanent Income Hypothesis and Consumption Durability: Analysis Based on Japanese Panel Data", Quarterly Journal of Economics, 100, pp. 1083-1113.

HAYASHI, F. and T. INOUE (1991) "The Relation Between Firm Growth and q With Multiple Capital Goods: Theory and Evidence From Japanese Panel Data", Econometrica, 59, pp. 731-754.

HECKMAN, J. (1979) "Sample Selection Bias as a Specification Error", Econometrica, 47, pp. 153-161.

HECKMAN, J. and T. MACURDY (1980) "A Life-Cycle Model of Female Labor Supply", Review of Economic Studies, 47, pp. 47-74.

HECKMAN, J. and T. MACURDY (1982) "Corrigendum on A Life-Cycle Model of Female Labor Supply", Review of Economic Studies, 49, pp. 659-660.

HOCH, I. (1962) "Estimation of production function parameters combining time series and cross-section data", Econometrica, 30, pp. 34-53.

HOLTZ-EAKIN, D., NEWEY W. and H.S. ROSEN (1988) "Estimating Vector Autoregressions with Panel Data", Econometrica, 56, pp. 1371-1395.

HOSHI, T., KASHYAP A. and D. SCHARFSTEIN (1991) "Corporate Structure, Liquidity and Investment: Evidence from Japanese Industrial Groups", Quarterly Journal of Economics, 106, pp.

HONDA, Y. (1985) "Testing the Error Components Model with Non-Normal Disturbances", Review of Economic Studies, 52, pp. 681-690.

HSIAO, C. (1985) "Benefits and Limitations of Panel Data", Econometric Reviews, 4, pp. 121-174.

HSIAO, C. (1986) Analysis of Panel Data, Econometric Society Monographs, Cambridge University Press.

JAFFE, A. (1986) "Technological Opportunity and Spillovers of R&D", American Economic Review, 76, pp. 984-1001.

JAKUBSON, G. (1988) "The Sensitivity of Labor Supply Parameter Estimates to Unobserved Individual Effects: Fixed- and Random-Effects Estimates in a Nonlinear Model Using Panel Data", Journal of Labour Economics, 6, pp. 302-329.

JARAMILLO, F., SCHIANTARELLI F. and A. SEMBENELLI (1992) "Are Adjustment Costs for Labour Asymmetric? An Econometric Test on Panel Data for Italy", paper presented at the Symposium on Labour Demand and Equilibrium Wage Formation, Amsterdam, January.

JORGENSON, D.W., LAU L.J. and T. STOKER (1982) "The Transcendental Logarithmic Model of Aggregate Consumer Behaviour", in Advances in Econometrics, R. Bassman and G. Rhodes eds., JAI Press.

KEEN, M.J. (1986) "Zero Expenditures and the Estimation of Engel Curves", Journal of Applied Econometrics, 1, pp. 277-286.

KIM, B. and G.S. MADDALA (1992) "Estimation and Specification Analysis of Models of Dividend Behaviour Based on Censored Panel Data", Empirical Economics, 17, pp. 111-124.

KLEVMARKEN, N.A. (1989) "Panel Studies: What Can We Learn From Them? Introduction", European Economic Review, 33, pp. 523-529.

KÖRÖSI, G. and L. MATYAS (1991) "Cointegration and Aggregation", Monash University, Department of Econometrics, Working Paper 16/91.

KRISHNAKUMAR, J. (1992) "Simultaneous Equations", in The Econometrics of Panel Data. Handbook of Theory and Applications, L. Matyas and P. Sevestre eds., Kluwer Academic Publishers.

KUH, E. (1959) "The Validity of Cross-Sectionally Estimated Behavior Equations in Time Series Applications", Econometrica, 34, pp. 335-348.

KUMBHAKAR, S.C. (1990) "Production frontiers, Panel Data, and Time-Variant Technical Inefficiency", Journal of Econometrics, 46, pp. 201-211.

LAISNEY, F., POHLMEIER W. and M. STAAT (1992) "Estimation of Labour Supply Functions Using Panel Data: A Survey", in The Econometrics of Panel Data, Handbook of Theory and Applications, L. Matyas and P. Sevestre eds., Kluwer Academic Publishers.

LAM, P.S. (1991) "Permanent Income, Liquidity and Adjustments of Automobile Stocks: Evidence from Panel Data", Quarterly Journal of Economics, 106, pp. 203-230.

LAU, L.J. (1982) "A Note on the Fundamental Theorem of Exact Aggregation", Economic Letters, 9, pp. 119-126.

LICHTENBERG, F. and D. SIEGEL (1989) "The Impact of R&D Investment on Productivity: New Evidence Using Linked R&D - LED Data", NBER Working Paper N°2901.

LILJA, R. (1986) "Econometric Analyses of Family Labour Supply over the Life-Cycle Using US Panel Data", The Helsinki School of Economics, Helsinki.

LOCKWOOD, B. A. and A. MANNING (1992) "The Importance of Linear Hiring and Firing Costs: Some Evidence from UK Manufacturing", paper presented at the Symposium on Labour Demand and Equilibrium Wage Formation, Amsterdam, January.

MACURDY, T. (1981) "An Empirical Model of Labor Supply in a Life-Cycle Setting", Journal of Political Economy, 89, pp. 1059-1085.

MACURDY, T. (1983) "A Simple Scheme for Estimating an Intertemporal Model of Labor Supply and Consumption in the Presence of Taxes and Uncertainty", International Economic Review, 24, pp. 265-290.

MADDALA, G.S. (1971) "The Use of Variance Components Models in Pooling Cross-Section and Time-Series Data", Econometrica, 39, pp. 341-358.

MADDALA, G.S. and T.D. MOUNT (1973) "A Comparative Study of Alternative Estimators for Variance Components Models Used in Econometric Applications", Journal of the American Statistical Association, 68, pp. 324-328.

MAIRESSE, J. (1978) "New Estimates of Embodied and Disembodied Technical Progress", Annales de l'INSEE, N°30-31, pp. 681-720.

MAIRESSE, J. and A. SIU (1984) "An Extended Accelerator Model of R&D and Physical Investment", in R&D, Patents and Productivity, Z. Griliches ed., University of Chicago Press.

MAIRESSE, J. (1988) "Les Lois de la Production Ne Sont Plus Ce qu'Elles Etaient: Une Introduction à l'Econométrie des Panels", Revue Economique, 39, pp. 225-271.

MAIRESSE, J. and Z. GRILICHES (1988) "Hétérogénéité et Panels: Y-a-t-il des Fonctions de Production stables?", in Essais en l'Honneur de Edmond Malinvaud, Economica, pp. 1010-1054.

MAIRESSE, J. and M. SASSENOU (1989) "Les Facteurs Qualitatifs de la Productivité: Un Essai d'Evaluation", Economie et Prévision, N°91, pp. 35-42.

MAIRESSE, J. (1990) "Time-Series and Cross-Sectional Estimates on Panel Data: Why Are They Different and Why Should They Be Equal?", in Panel Data and Labor Market Studies, J. Hartog, G. Ridder and J. Theeuwes eds., North-Holland, pp. 81-95.

MAIRESSE, J. and M. SASSENOU (1992) "Recherche-Développement et Productivité: Un Panorama des Etudes Econométriques sur Données d'Entreprises", in Recherche et Technologie, J. De Bandt et D. Foray eds., Economica.

MALECOT, J.F. (1992) "Modelling Companies'Dividend Policy Using Account Panel Data", in The Econometrics of Panel Data, Handbook of Theory and Applications, L. Matyas and P. Sevestre eds., Kluwer Academic Publishers.

MALINVAUD, E. (1987) "Capital Productif, Incertitudes et Profitabilité", Annales d'Economie et de Statistique, N°5, pp. 1-36.

MANSFIELD, E. (1980) "Basic Research and Productivity Increase in Manufacturing", American Economic Review, 70, pp. 863-873.

MARIGER, R.P. (1987) "A Life-Cycle Consumption Model with Liquidity Constraints: Theory and Empirical Results", Econometrica, 55, pp. 533-537.

MATHIEU, C. (1993) "L'Implantation des Entreprises Françaises aux Etats-Unis: Une Analyse Empirique", ERUDITE Working Paper, Université de Paris XII.

MATYAS, L. (1992) "Error Components Models", in The Econometrics of Panel Data. Handbook of Theory and Applications, L. Matyas and P. Sevestre eds., Kluwer Academic Publishers.

MATYAS, L. and L. LOVRICS (1991) "Missing Observations and Panel Data", Economics Letters, 37, pp. 39-44.

MAUREL, F. (1990) "Dynamique de l'Emploi et Tendance de la Productivité dans les Années Quatre-Vingt", Economie et Statistique, N° 237-238, pp. 151-162.

MEGHIR, C. and J.M. ROBIN (1992) "Frequency of Purchase and the Estimation of Demand Systems", Journal of Econometrics, 53, pp. 53-85.

MOHNEN, P. (1992) "The Relationship Between R&D and Productivity Growth in Canada and Other Major Industrialized Countries", Minister of Supply and Services, Canada.

MULLBAUER, J. (1975) "Aggregation, Income Distribution and COnsumer Demand", Review of Economic Studies, 42, pp. 523-543.

MULLBAUER, J. (1976) "Community Preferences and the Representative Consumer", Econometrica, 44, pp. 979-1000.

MUNDLAK, Y. (1961) "Empirical Production Function Free of Management Bias", Journal of Farm Economics, 43, N°1, pp. 44-56.

MUNDLAK, Y. (1978) "On the Pooling of Time Series and Cross Section Data", Econometrica, 46, pp. 69-85.

NICKELL, S. (1981) "Biases in Dynamic Models with Fixed Effects", Econometrica, 49, pp. 1417-1426.

ODAGIRI, H. and H. IWATA (1986) "The Impact of R&D on Productivity Increase in Japanese Manufacturing Companies", Research Policy, 15, pp. 13-19.

OI, W.Y. (1962) "Labour as a Quasi-Fixed Factor", Journal of Political Economy, LXX, N°6, pp. 538-555.

OULTON, N. (1981) "Aggregate Investment and Tobin's q: The Evidence from Britain", Oxford Economic Papers, 33, pp.

RINGSTAD, V. (1971) "Estimating Production Functions and Technical Change from Micro-Data", Central Bureau of Statistics of Norway, Oslo.

ROBIN, J.M. (1991) "Short-Run Fluctuations of Households'Purchases", INSEE-CREST Discussion Paper.

ROBIN, J.M. (1992) "Consumption Dynamics and Panel Data: A Survey" in The Econometrics of Panel Data, Handbook of Theory and Applications, L. Matyas and P. Sevestre eds., Kluwer Academic Publishers.

SALINGER, M. and L. SUMMERS (1983) "Tax Reform and Corporate Investment: A Microeconometric Simulation Study", in Behavioural Simulation Methods in Tax Policy Analysis, M. Feldstein ed., University of Chicago Press.

SCHALLER, H. (1990) "A Re-Examination of the Q theory of Investment Using United-States Firm Data", Journal of Applied Econometrics, 5, pp.

SCHANKERMAN, M. (1981) "The Effects of Double-Counting and Expanding on the Measured Returns to R&D", Review of Economics and Statistics, 63, pp. 454-458.

SHAPIRO, M.D. (1983) "The Permanent Income Hypothesis and the Interest Rate: Some Evidence from Panel Data", Economics Letters, 14, pp. 93-100.

SCHMIDT, P. and R. SICKLES [1984] "Production Frontiers and Panel Data", Journal of Business and Economic Statistics, 2, pp. 367-374.

SEVESTRE, P. and A. TROGNON (1983) "Propriétés de Grands Echantillons d'une Classe d'Estimateurs des Modèles Autorégressifs à Erreurs Composees", Annales de l'INSEE ,N° 50, pp. 25-49.

SEVESTRE, P. and A.TROGNON (1985) "A Note on Autoregressive Error Components Models", Journal of Econometrics, 28, pp.231-245.

SEVESTRE, P. "Qualification de la Main d'Oeuvre et Productivité du Travail", Economie et Statistique, N°237-238, pp. 109-120.

SEVESTRE, P. and A. TROGNON (1990) " Consistent Estimation Methods for Dynamic Error Components Models: Small and Large Sample Properties", mimeo.

SEVESTRE, and A. TROGNON (1992) "Linear Dynamic Models", in The Econometrics of Panel Data. Handbook of Theory and Applications, L. Matyas and P. Sevestre eds., Kluwer Academic Publishers.

STOKER, T. (1984) "Completeness, Distribution Restrictions and the Form of Aggregate Functions" Econometrica, 52, pp. 887-908.

SUMMERS, L. (1981) "Taxation and Corporate Investment: A q-Theory Approach", Brookings Papers on Economic Activity, N°1, pp. 67-127

SZPIRO, D. and G. CETTE (1988) "Productivité et Progrès Technique dans l'Industrie sur la Période 1972-1984", Cahiers Economiques et Monétaires N°28, Banque de France.

THEIL, H. (1954) Linear Aggregation of Economic Relations, Noth-Holland.

VERBEEK, M. and T. NIJMAN (1992) "Incomplete Panels and Selection Bias", in The Econometrics of Panel Data. Handbook of Theory and Applications, L. Matyas and P. Sevestre eds., Kluwer Academic Publishers.

VERBEEK, M. and T. NIJMAN (1992) "Can Cohort Data Be Treated as Genuine Panel Data?", Empirical Economics, 17, pp. 9-24.

WALLACE, T.D. and A. HUSSAIN (1969) "The Use of Error Components Models in Combining Time-Series with Cross-Section Data", Econometrica, 37, pp. 55-72.

WANSBEEK, T.J. and A. KAPTEYN (1989) "Estimation of the Error Components Model with Incomplete Panels", Journal of Econometrics, 41, pp. 341-361.

ZELDES, S.P. (1989) "Consumption and Liquidity Constraints: An Empirical Investigation", Journal of Political Economy, 97, pp. 305-346.